NOTTINGHAM TRENT UNIVERSITY

# Transcription Start Site selection within a single cluster and G quadruplex structures: a novel mechanism regulating gene expression

**Arif Anwer Surani**

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy (PhD)

**May 2021**

## Copyright statement:

*ARIF SURANI*
(N0764615)

# Abstract

The 5’ untranslated region (UTR) in a messenger RNA (mRNA) can greatly influence translation. Depending on where the transcription starts, the 5’ UTR will contain (or not) regulatory elements that can modulate mRNA translation. Although studies have demonstrated the effect of differential transcription start sites (TSSs) usage on translation efficiency, these are mainly restricted to TSSs in different clusters many nucleotides apart. However, it is currently unknown if there is any relevance to the TSS variation within a single cluster. Here, we present our findings on single cluster TSS-mediated regulation of protein expression, using mainly AGAP2 as an example.

Using 5’ RLM-RACE, we have identified different TSSs usage for *AGAP2* mRNA in chronic myeloid leukaemia (CML) and prostate cancer (PC) cell lines, giving rise to populations of transcripts with variable lengths of 5’ UTR. The population of longer 5’ UTR were relatively higher in CML cell lines ($P < 0.05$), and those extra nucleotides contained the consensus sequence for a G-quadruplex (G4). The G4 formation was verified by CD spectroscopy. Additionally, we developed an immunoprecipitation method termed ‘GRIP’ [G4 RNA Immunoprecipitation] and demonstrated the existence of these RNA secondary structures in the living cells.

To study the impact of the longer 5’ UTR and the G4 on translation efficiency, we cloned three 5’ UTR isoforms (shorter, longer and mutated-longer version) into a bicistronic plasmid and reported a significant decrease in luciferase activity by the G4 in the longer 5’ UTR ($P < 0.001$). This result coincides with the discrepancy noted in AGAP2 mRNA and protein levels in these cell lines. Furthermore, polysome fractionation studies also confirmed that mRNA with longer 5’ UTR associated less prominently with polyribosomes ($P < 0.001$).

Our bioinformatics pipeline has identified 4,920 transcripts in the FANTOM database that contained putative G4 sequences between the major and upstream TSSs within the same TSS cluster. By integrating the NCI-60 microarray and SWATH-MS database, we curated a list of genes that displayed discrepancies in RNA and protein expression with a significantly higher level of G4 forming TSS; and validated our findings in another gene target (*HK1*). This highlights that the TSS-G4 mediated mechanism is not only limited to *AGAP2* expression regulation but is also implicated in controlling the expression of other genes.

# Acknowledgement

I would like to acknowledge and thank my director of studies, Dr Cristina Montiel-Duarte for giving me this amazing learning opportunity to become an independent scientist, and for providing me with balanced supervision, with supervised and independent learning opportunities. I am truly inspired by her critical and scientific acumen. I would also like to thank my co-supervisor Dr Christos Polytarchou and my project independent assessor Dr Stéphanie McArdle for their support, encouragement, and critical insights. I would also like to acknowledge the assistance of Dr Keith Spriggs in designing the bioinformatics pipeline. A special thanks to the John van Geest Cancer Research Centre for providing me with the facility to carry out my research and the NTU VC scholarship for funding my research.

A special thanks to my colleagues in JvG and IBRC for their constant support, motivation, and for the great times of my life. Also, my loving family for their unconditional love, trust, and support in every aspect of my life.

# Publications and conference proceedings

## Publications:

SURANI, A. A., Spriggs, K., UFER, C., POLYTARCHOU, C., & MONTIEL-DUARTE, C. 2020. Transcription Start Site selection and G quadruplex structures: novel players regulating the translatability of mRNA. Molecular Cell (*under submission*)

SURANI, A. A., COLOMBO, S., BARLOW, G., FOULDS, G., & MONTIEL-DUARTE, C. 2020. Optimising cell synchronisation using nocodazole or double thymidine block. Springer Protocols (*In press*).

DOUSH, Y., SURANI, A. A., NAVARRO-CORCUERA, A., MCARDLE, S., BILLETT, E. E. & MONTIEL-DUARTE, C. 2019. SP1 and RARα regulate AGAP2 expression in cancer. Scientific Reports, 9, 390.

## Conference Attendance:

### Oral communications:

- Oral presentation, Transcription in Health and Disease conference (Nov 2019) organised by the Biochemical Society, St. Paul's Centre, London, UK.
- Oral presentation, Science and Technology Annual Research Conference (May 2019), Nottingham Trent University, UK.

### Poster Presentations:

- Poster presentation, RNA2020 (June 2020) online conference organised by RNA Society.
- Poster presentation, Translation UK conference (Jul 2019) organised by the Biochemical Society, Strathclyde Business School, Glasgow, UK.
- Flash poster presentation, Bioscience Research day, University of Nottingham (Sep 2019).
- Poster presentation, Science and Technology Annual Research Conference (May 2019), Nottingham Trent University, UK (**2nd Best Poster award**).
- Poster presentation, Translation UK conference (Jul 2018), University of Manchester, UK.
- Poster presentation, Postgraduate Poster Competition organised by the Royal Society of Biology, De Montford University, Leicester, UK (**Best Poster award**).

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

**5’ RACE:** 5’ Rapid Amplification of cDNA Ends

**AFE:** Alternative First Exon

**ATI:** Alternative Transcription Initiation

**AUAP:** Abridged Universal Amplification Primer

**BTZ:** Bortezomib

**CAGE:** Cap Analysis Gene Expression

**CD:** Circular Dichroism

**CIP:** Calf Intestinal Phosphatase

**CML:** Chronic Myeloid Leukaemia

**eIF:** eukaryotic Initiation Factors

**FANTOM:** Functional Annotation of the Mammalian Genome

**FLUC:** Firefly luciferase

**G4:** G quadruplex

**GO:** Gene Ontology

**GRIP:** G4 RNA Immunoprecipitation

**GSP:** Gene-Specific Primer

**Indels:** Insertion–deletion mutations

**Inr:** Initiator

**IRES:** Internal Ribosomal Entry Sites

**lncRNA:** long non-coding RNA

**miRNA:** micro-RNA

**mRNA:** Messenger RNA

**mRNP:** messenger ribonucleoprotein complex

**ncRNA:** non-coding RNA

**NGS:** Next Generation Sequencing

**NTC:** No Template Control

**ORF:** Open Reading Frame

**PABP:** Poly(A)-binding protein

**PC:** Prostate Cancer

**PCA:** Principal Component Analysis

**PCR:** Polymerase Chain Reaction

**PIC:** Pre-initiation complex

**qRT-PCR:** quantitative reverse transcription PCR

**RBP:** RNA Binding Protein

**RIN:** RNA Integrity Number

**RLUC:** Renilla luciferase

**RNAP:** RNA polymerase

**RNP:** Ribonucleoprotein

**RT:** Reverse Transcriptase

**SEASTAR:** Systematic Evaluation of Alternative STArt site in RNA

**SNP:** Single Nucleotide Polymorphisms

**TAP:** Tobacco Acid Pyrophosphatase

**TF:** Transcription Factors

**TPM:** Tags Per Million

**TSS:** Transcription Start Site

**uAUG:** upstream start codon

**Ub:** Ubiquitin

**uORF:** upstream Open Reading Frame

**UPP:** Ubiquitin Proteasome Pathway

**UTR:** untranslated region

# Chapter 1:
# General Introduction

Our group has been studying the role and the regulation of *AGAP2* (ArfGAP with GTPase domain, ankyrin repeat and PH domain 2) gene in different diseases and cancers. Our previous work has highlighted that the levels of AGAP2 mRNA and protein are not always correlated. This thesis attempt to explain a novel mechanism responsible for regulating *AGAP2* gene expression. The introduction chapter will cover the basics of eukaryotic gene expression and summarise up to date knowledge of all possible mechanisms involved in regulating gene expression.

## 1.1 Eukaryotic gene expression

Gene expression is a process by which the genetic code in the DNA is used to direct the synthesis of proteins or non-coding RNAs, shaping the cellular phenotype. In order for a protein-coding gene to express, the DNA code is first transcribed to a messenger RNA (mRNA) intermediate which is subsequently translated to protein. The process of gene expression is complex and comprise of different stages including regulatory processes that convert the DNA signal into the protein product. The genetic code (DNA) is first converted to a pre-mRNA which undergoes processing that include splicing, capping, and polyadenylation to name a few. The mature mRNA is subsequently exported from the nucleus into the cytoplasm where it is converted to protein in a process called translation, using the mRNA code to link the constituent amino acids to form a protein. The process of gene expression is non-linear and is regulated at different stages during the process and include transcriptional regulation (section 1.2.1), post-transcriptional regulation (section 1.2.2), and translational regulation and post-translational processing (section 1.2.3). The process of gene expression and the detail of the regulatory mechanisms taking place are described in detail below and also overviewed in Figure 1.1.

The eukaryotic gene expression is a complex and highly regulated process and is far more complicated compared to prokaryotic organisms. In eukaryotic organisms, the process of transcription and translation are separated spatially and temporally due to the presence of a nuclear membrane. Moreover, the regulation may occur at different levels during the gene expression process. The introduction section presented here will primarily focus on the mechanism and regulation of gene expression found in eukaryotic cells.

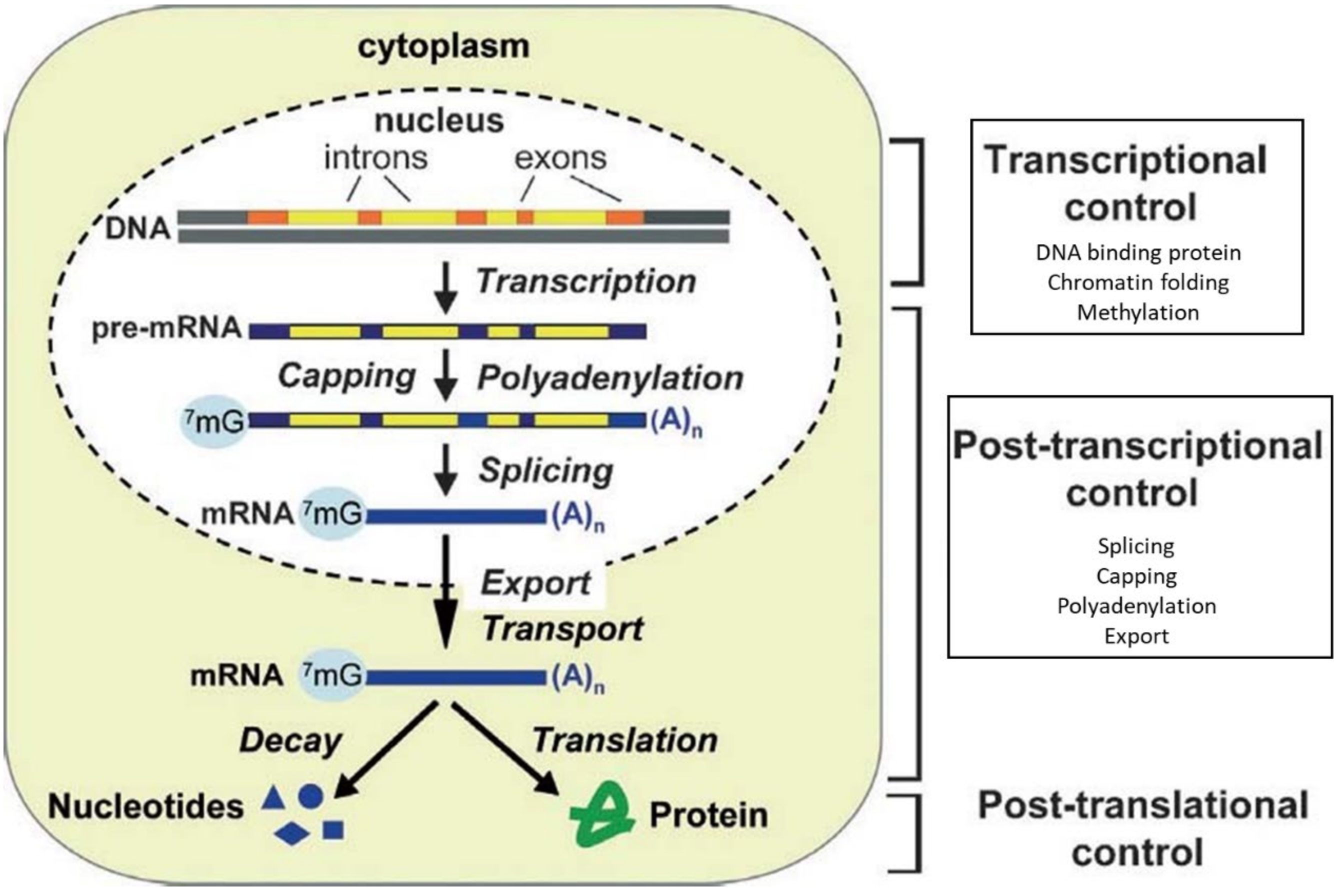


**Figure 1.1: Overview of eukaryotic gene expression.** Steps involved in the expression of protein-coding genes. The genetic code in the DNA (base sequence) is first transcribed into pre-mRNA which undergo extensive processing before export to the cytoplasm and is used to direct protein synthesis. The regulation taking place at different levels of gene expression process is also depicted. Image adapted from Halbeisen *et al.* (2007) (CC BY-NC).

### 1.1.1 Mechanism of Transcription:

The process of gene expression starts with transcription which is a multistep process that involves the unwinding of the double-stranded DNA and conversion of the DNA sequence, in either of the two strands of the genomic DNA, into mRNA transcript. The transcription of all the protein-coding genes (mRNA) and non-coding genes (most snRNA and miRNA) are mediated by RNA polymerase II (RNA Pol II) (Reviewed in Miglani, 2014). The RNA Pol II along with accompanying protein complexes assembles on the DNA sequence at the 5’ end of a gene, called the promoter, where the transcription begins. The transcription typically initiates at a defined position referred to as the transcription start site (TSS) which is embedded within the core promoter region. The phases of transcription can be divided into preinitiation, initiation, elongation, and termination.

#### 1.1.1.1 Preinitiation:

Preinitiation is the first phase of transcription and involves the formation of the preinitiation complex (PIC). The PIC consists of general and specific transcription factors (TFs), coactivators, and RNA pol II enzyme and is formed at the gene promoter region that serves as a binding platform for the assembly of these components (Hampsey, 1998). Different eukaryotic promoter sequences are known and contain certain consensus patterns such as TATA, CAAT, and GC boxes. The TATA box has the consensus sequence of TATAWAAR [W = (A/T), R = (A/G)], where the 5’ T is located at -30 or -31 bp upstream of the TSS (Reviewed in Danino *et al.,* 2015). The TATA box is bound by the general transcription factor TFIID and in some cases requires TFIIA to facilitate TFIID binding with the TATA box (Butler and Kadonaga, 2002). The CAAT box is usually found between -80 and -60 nucleotides upstream of the TSS and has the consensus GGNCAATCT [N = (A/T/G/C)], which is bound by the transcription factor CTF (CAAT box transcription factor) (Miglani, 2014). The GC box has the consensus GGGCGG which is often present in multiple copies and is recognised and bound by the SP1 transcription factor (Miglani, 2014). The core promoter is a short (+/- 50 bp) region around the TSS that contain the cis-acting elements including the short sequence elements (TATA box) and an Initiator (Inr) element that encompasses the TSS and mainly bound by TAF1 and TAF2 subunits of TFIID (Kaufmann and Smale, 1994, Smale, 2001, Smale and Baltimore, 1989).

The assembly of the PIC begins when the activator binds to one of the promoter sequences and recruit coactivators and complexes that mediate ATP dependant chromatin relaxation. The TFIID which contain the TATA-binding protein (TBP) subunit binds to the TATA box, serving as a nucleation point for PIC assembly (Buratowski *et al.,* 1989). In absence of the TATA box, other subunits of TFIID associate with the promoter region and facilitate the formation of PIC (Ranish *et al.,* 1999). The binding of TFIID is followed by the recruitment of general transcription factors including TFIIA, TFIIB, TFIIE, TFIIF, and TFIIH. The TFIIB subsequently binds to basal recognition elements which are located immediately upstream and downstream of the TATA box (Deng and Roberts, 2007). The TATA–TBP–TFIIB complex provides support for the binding of the TFIIF–Pol II complex. The TFIIF escorts the RNA pol II to the complex and recruits TFIIE and TFIIH which complete PIC assembly (Reviewed in Ghosh and Van Duyne, 1996) (*Figure 1.2*). The TFIIE and TFIIH are involved in unwinding the promoter region, transcription initiation and promoter clearance (Reviewed in Thomas and Chiang, 2006).

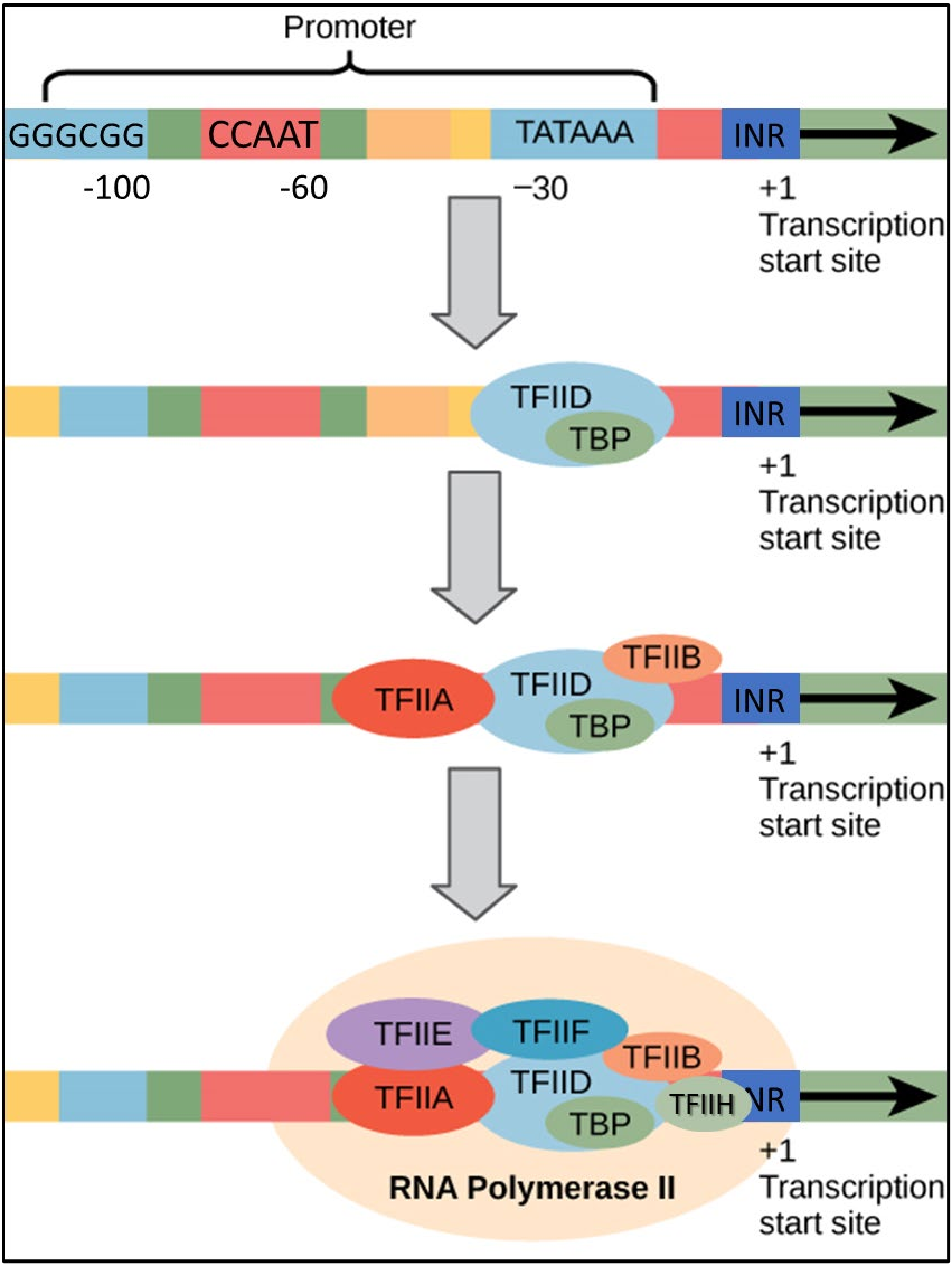


**Figure 1.2: Schematic summary of transcription initiation in eukaryotes.** The eukaryotic transcription is initiated by binding of TBP subunit of TFIID to the TATA box. It is followed by the binding of general transcription factors (TFIIB, Pol II-TFIIF complex, TFIIE and TFIIH) in a stepwise fashion. The RNA pol II promoter elements (TATA, CAT, GC box) are shown with their location relative to TSS (+1). INR: initiator element; RNA pol II: RNA polymerase II; TF: transcription factor; TBP: TATA-binding protein. Image adapted and modified from Avissar *et al.* (2018) (CC BY 4.0).

### 1.1.1.2 Initiation, elongation, and termination:

After the formation of PIC, the DNA strands at the TSS are separated to generate an open promoter complex. The RNA transcription is initiated beginning with two initiating NTPs dictated by the DNA base pairing and formation of the first phosphodiester bond leading to the initial transcribing complexes (ITC). The ITC proceed into the gene through a DNA scrunching mechanism (unwinding DNA and pulling strands into itself) before reaching a critical length (Cheung *et al.,* 2011). After transcribing about 20-30 nucleotides downstream of the TSS, the RNA pol II pauses and disconnects from the promoter elements with partial disassembly of PIC in a process called the promoter escape or clearance (Reviewed in Gupta *et al.,* 2016). It is followed by the formation of transcription elongation complex on the RNA pol II which commence productive elongation facilitated by a wide variety of elongation factors (Reviewed in Sims *et al.,* 2004). The termination usually occurs downstream of the 3' end of the gene, releasing the nascent RNA and disassembly of elongation machinery. The transcription termination factor 2 (TTF2) is a well-characterized factor associated with RNA pol II that disassemble the elongating complex using ATP hydrolysis (Jiang *et al.,* 2004). The termination of transcription is functionally connected to cleavage and polyadenylation.

### 1.1.1.3 Co-transcriptional events:

During the transcription, the protein complexes (including the capping enzyme complex, spliceosome, and 3' end processing machinery) approach the primary transcript to facilitate the mRNA maturation process. The pre-mRNA undergoes two important processing steps including modifying both ends of mRNA and selective elimination of the intronic fragments within the coding sequence. The modification at the 5' end includes the addition of a 7-methylguanosine (m7G) cap structure to the first nucleotide at the 5' termini of the elongating mRNA. The 5' cap structure plays an important role in RNA stability, resistance from exonucleases, nuclear export, and translation initiation (Shimotohno *et al.,* 1977). The 3' end of the pre-mRNA is modified by the addition of 40–200 adenine residues to generate a poly-A tail. The tailing reaction is facilitated by poly-A polymerase and catalysed by cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CSTF) that guide AAUAAA dependent poly(A) tail addition (Colgan and Manley, 1997). The poly-A tail acts as a stabiliser of intact mRNA and also impact translation and localisation (Reviewed in Yuan *et al.,* 2021). The pre-mRNA also undergoes splicing to remove any introns. The RNA splicing is mediated by a multi-subunit complex called spliceosome that detects the consensus region at 5' donor and 3' acceptor end, releasing the introns (Reviewed in Matera and Wang, 2014). The alternative splicing produces multiple mRNA isoforms from a single gene and can regulate protein composition by changing the coding sequences between the isoforms, generating proteomic diversity (Reviewed in Stamm *et al.,* 2005). The mature mRNA is then exported from the nucleus by forming messenger ribonucleoprotein (mRNP) complexes that are translocated through nuclear pore embedded in the nuclear envelope.

### 1.1.2 Mechanism of Translation:

The translation is a process by which the information contained in mRNA is used to direct the synthesis of polypeptides. The genetic code in the mRNA consists of triplets of adjacent ribonucleotides called codon that are recognised by transfer RNAs (tRNAs) which incorporates a specific amino acid depending on the codon seqeunce. The ribosomes are the central multi-subunit complex for translation which facilitate the interpretation of the codons in the mRNA and form peptide bonds between amino acids. The process of translation generally entails three stages: initiation, elongation, and termination.

#### 1.1.2.1 Translation Initiation:

As reviewed in Haimov *et al.* (2015), Hinnebusch and Lorsch (2012), and Sonenberg and Hinnebusch (2009), the initiation of translation begins with the formation of 43S PIC which is formed by the binding of the ternary complex (TC) to 40S ribosomal subunit. The TC is composed of initiator methionyl-tRNA (Met-tRNA$_i$) and the GTP-bound form of eukaryotic initiation factor 2 (eIF2). The multiprotein eIF3 complex, together with eIF1, eIF1A and eIF5, promotes the binding of the 40S ribosomal subunit to the mRNA. This complex is recruited to the mRNA 5' cap by poly(A)-binding protein (PABP), eIF3, eIF4B, eIF4H, and heterotrimeric factor eIF4F. The eIF4F complex is composed of eIF4E (cap-binding protein), eIF4G (scaffold protein), and eIF4A (RNA helicase). The recruitment of 43S PIC to 5' mRNA cap is followed by a scanning process that involves the unwinding of secondary structures in 5' termini of mRNA region by the eIF4A, whose helicase activity is enhanced by eIF4B and eIF4H. Subsequently, the initiation complex scans the 5' untranslated region (UTR) of mRNA in the 5' to 3' direction until it encounters the start codon. The recognition of the start codon triggers the hydrolysis of the GTP bound to eIF2, displacement of the initiation factors and concomitant joining of 60S ribosomal subunit to complete the assembly of the 80S translating ribosome (*Figure 1.3*). This process is catalysed by eIF5B that triggers the GTP hydrolysis and release of eIFs. The PABP at the 3' end interacts with the eIF4E and eIF4G at the 5' mRNA cap to form a closed-loop configuration that enhances translation (Tomek and Wollenhaupt, 2012). Although most of the mRNAs are translated using the canonical cap-dependant scanning mechanism, alternative non-canonical translation initiations are also been reported demonstrating scanning-free translation initiation and cap-independent translation through internal ribosome entry sites (IRES) (Haimov *et al.,* 2015, Komar *et al.,* 2012).

#### 1.1.2.2 Translation Elongation:

The translation initiation culminates with the formation of 80S ribosome with the Met-tRNA$_i$ base-paired to AUG in the P (peptidyl) site of the ribosome and the second codon in the A (aminoacyl) site ready to accept the aminoacyl-tRNA carrying the next amino acid guided by the codon context. It is followed by the formation of a peptide bond between the first two amino acids in the P site catalysed by the peptidyl transferase. The deacylated tRNA is accommodated at the E (exit) site of the ribosome before being released and recycled. The elongation continues with the translocation of the ribosome to the next codon and delivery and shuttling of the tRNA through A, P, and E sites. The elongation, in contrast to translation initiation, requires a minimal set of elongation factors that are well-conserved between prokaryote and eukaryotes (Reviewed in Dever *et al.,* 2018). Certain antibiotics and compounds including cycloheximide have shown to inhibit translation elongation (Schneider-Poetsch *et al.,* 2010). Cycloheximide-mediated elongation inhibition has been used by different techniques to evaluate ribosome density and their distribution profiles (Chassé *et al.,* 2016, Duncan and Mata, 2017).

### 1.1.2.3 Translation termination:

The translation terminates when the ribosome reaches the stop codon (UAG, UAA, and UGA). The process of termination includes the recognition of a stop codon in the A ribosomal site and hydrolysis of peptidyl-tRNA in the P site, releasing the nascent polypeptide chain. The process is mediated by the release factors (RF), eRF1 and eRF3, which disassemble the complex, followed by the recycling of its constituents to participate in the next rounds of translation (Reviewed in Jackson *et al.,* 2012).

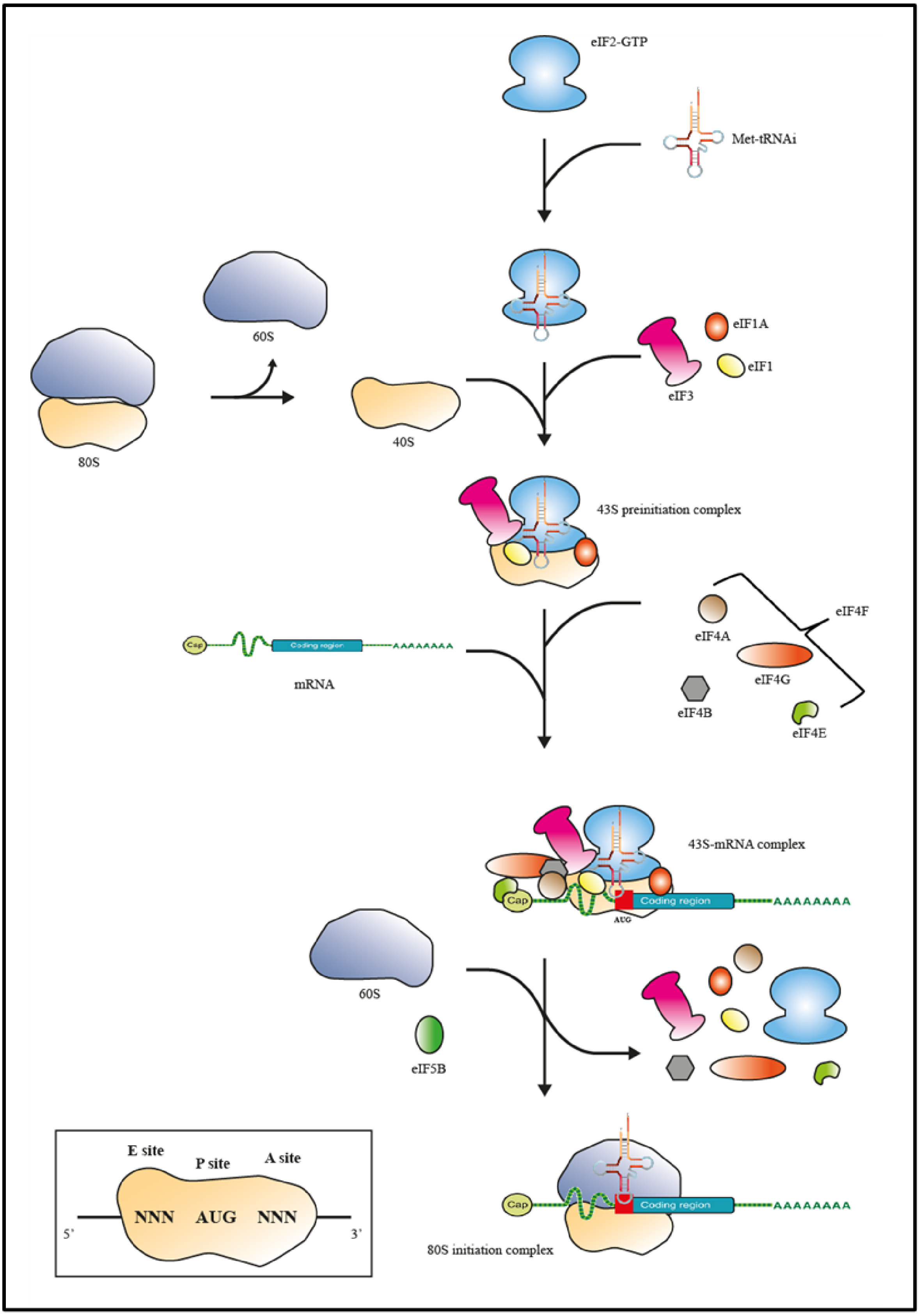


**Figure 1.3: Overview of eukaryotic translation initiation.** The eukaryotic translation begins with the formation of the 43S preinitiation complex (ternary complex [eIF2/GTP/Met-tRNAi), 40S ribosomal subunit, and other eIFs]. The complex is recruited to the mRNA 5' cap structure and associate with eIF4F forming initiation complex that scan the mRNA until it encounters the start codon. The 60S ribosomal subunit then joins the mRNA bound 40S ribosomal subunit, facilitated by eIF5B, to form 80S translating ribosome with dissociation and recycling of eIFs. The three binding sites for tRNA: A (aminoacyl), P (peptidyl), and E (exit) site are shown in the 40S ribosomal subunit.

## 1.2 Regulation of gene expression

The regulation of gene expression is a highly sophisticated, multi-step, and non-linear process that controls the output of a gene. It is an intricate mechanism fundamental to all biological processes and involves dynamic coordination between multiple events, shaping the gene's functional product. A regulated gene expression is essential for the specialised cell states in a multicellular organism that contains the same genetic material but displays differing phenotypes. The gene regulation occurs at multiple levels including transcription, post-transcriptional, translation, and post-translational processing.

### 1.2.1 Transcriptional regulation:

Transcription is the first level of gene expression control and is likely the most regulated and evolutionary conserved step (Schena, 1989). The regulation of transcription occurs primarily at two distinct interconnected levels involving the interaction between specific DNA motifs and DNA binding protein (transcription factors) and three-dimensional folding of the chromatin and its regulators.

The transcription factors (TFs) are proteins that recognise and bind to the specific DNA sequence in the promoter region and regulate gene expression. These TFs plays an important role in patterning cell types and orchestrating specialised cell programs (Lee and Young, 2013). A variety of TFs have been classified that regulate the initiation and elongation steps of the transcription and could be broadly grouped into general and specific TFs (Reviewed in Lee and Young, 2013). Most of the eukaryotic TFs also associate with other protein complexes called cofactors that do not display DNA-binding properties and either stimulate (coactivators) or suppress (corepressors) gene expression through TF binding. These cofactors include the mediator complex, P300, and general TFs that regulate transcription via different mechanisms (Conaway and Conaway, 2011, Juven-Gershon and Kadonaga, 2010, Malik and Roeder, 2010). The DNA sequence context is equally important as it guides the binding of TF and serves as a gateway to transcription. These cis-acting sequence elements (promoters, enhancers, silencers and insulators) are usually present in the proximity of the gene transcription initiation site and facilitate binding of TFs and assembly of the transcription machinery. Conversely, these binding motifs could be also present in the distal region away from the gene and cooperatively bind TFs and recruit cofactors to mediate long-range interaction with the promoter, modulating its activity (Krivega and Dean, 2012).

The chromatin structure significantly impacts all the aspects of transcription and is tightly regulated by a variety of mechanisms including chromatin remodelling, modification of histone, nucleosome dynamics, and chromosomal interactions (Reviewed in Li *et al.,* 2007). The accessibility of the DNA for transcription is dependent on its nucleosome packaging. The histone tails in the nucleosome complex are subjected to a wide range of modifications including lysine (K) acetylation in histone 3 (H3K27) and di/tri-methylation (me) of H3K4 that are associated with transcriptional activation (euchromatin state) (Bonn *et al.,* 2012). On the other end, the tri-methylation on histone 3 (H3K27me3, H3K9me3) suppresses transcription (heterochromatin state) (Nakayama *et al.,* 2001, Simon and Kingston, 2009). The regulation also occurs by protein complexes that alter the chromosome architecture and consequently lead to transient unwrapping of the DNA from histone octamer and nucleosome sliding (Smith and

Peterson, 2005). These remodelling complexes including ATP-dependent complexes (SWI/SNF, CHD1, ISWI protein family) and histone chaperone proteins (FACT), cooperatively facilitate chromosome remodelling along with modifications in the histone tail. Additionally, long-range chromosome interaction mediated by chromosome structure proteins (e.g. cohesion) also shapes gene expression (Xu *et al.,* 2016).

DNA methylation also plays an important role in controlling gene expression by inhibiting the binding of TFs and recruiting proteins that silence gene expression. The majority of DNA methylation occurs on cytosines in the CpG context and around 70% of promoters are associated with a CpG island (Saxonov *et al.,* 2006). The methylation of DNA contributes to the spatiotemporal regulation of gene expression with dense promoter methylation associated with transcriptional repression (Reviewed in Moore *et al.,* 2013). However, in some cases, promoter hypermethylation also enhances the transcriptional activity (Smith *et al.,* 2020).

### 1.2.2 Post-transcriptional regulation:

The post-transcriptional control of gene expression is mediated by the regulatory elements in the transcribed mRNA. It plays an important role in diverse cellular processes such as development (Kuersten and Goodwin, 2003), metabolism (Kim and Kyung Lee, 2012), and cell division (Hengst and Reed, 1996). These post-transcriptional regulatory mechanisms create a mismatch between mRNA and protein levels, setting the protein level independently from their mRNA concentration. The technological advances in the recent decades have characterised multiple post-transcriptional regulatory elements that regulate gene expression, some of these processes also occur co-transcriptionally and determine the nature and translational potential of the mRNA (Merkhofer *et al.,* 2014). The major post-transcriptional regulatory mechanisms include alternative splicing and polyadenylation, mRNA decay, and regulation mediated by elements encoded in the mRNA 5' and 3' UTR. The section presented here further details the regulatory features embedded in the mRNA UTRs. These elements variably impact the mRNA translational potential and are group according to the 5' and 3' region of the mRNA. The different regulatory elements in the 5' and 3' UTR are overviewed in *Figure 1.4*.

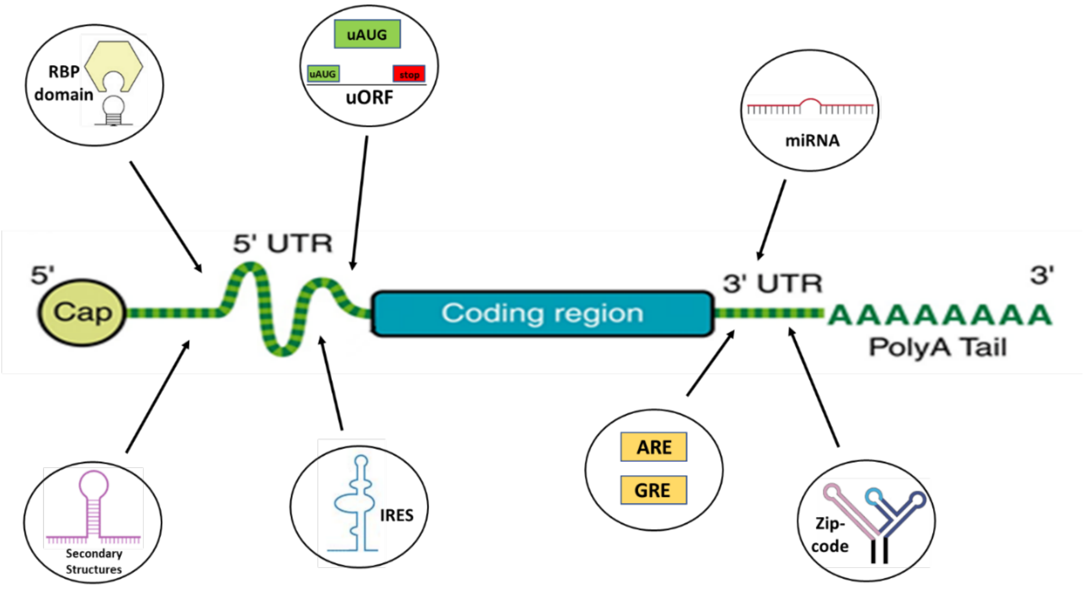


**Figure 1.4: Post-transcriptional regulatory elements in the mRNA UTR.** The general structure of eukaryotic mRNA, highlighting some post-transcriptional regulatory features in the 5′ and 3′ UTR of the mRNA that affects gene expression. ARE: AU-rich element; GRE: GU-rich element; IRES: internal ribosome entry site; miRNA: microRNA; PTC: premature termination codon; RBP: RNA binding protein; uAUG: upstream start codon; uORF: upstream open reading frame.

### 1.2.2.1 5′ UTR regulatory elements:

The 5′ UTR is the area between the mRNA cap structure and the start codon and serves as a platform for different regulatory features that impact translation initiation and scanning. The median length of 5′ UTR is longer in humans (218 nucleotides) compared to other higher eukaryotes (Leppek *et al.,* 2018). The length of 5′ UTR tends to be longer for transcripts that presumably have a regulatory role such as mRNA encoding for TFs, protooncogenes and growth factors (Davuluri *et al.,* 2000). The prominent regulatory elements in the 5′ UTR include RNA binding proteins (RBP) domain, upstream open reading frame (uORF), upstream start codon (uAUG), RNA secondary structures, and internal ribosome entry sites (IRES).

The RBP recognise specific motifs in the 5′ UTR and interact with the translational apparatus to modulate gene expression. These RBP also recognise binding motifs in the 3′ UTR to regulate mRNA stability and localisation. The classical example of regulation by RBP is the iron response elements (IREs) in the 5′ UTR that control the ferritin mRNA translation according to the intracellular iron level (Reviewed in Goss and Theil, 2011). In response to low intracellular RNA levels, the iron regulatory proteins (IRP1 and IRP2) bind to the specific stem-loop structures called IRE (Iron Response Element) that are located in the 5′ UTR of mRNAs including ferritin light and heavy chain and mitochondrial aconitase, thereby preventing their translation. The RBP, unlike their DNA counterpart (TFs), binds to secondary structures in addition to the primary sequence. The RBPs have shown to be involved in various aspects of RNA metabolism including RNA maturation, biogenesis, localisation, and turnover (Reviewed in Hentze *et al.,* 2018). A study by Keene and Tenenbaum (2002) has proposed a model of post-transcriptional gene expression in which RBP regulate a group of functionally related genes, contributing to the specification of cellular state and identity. An interesting example is the regulation of P21 expression by two antagonising RBPs (CUGBP1 and calreticulin) which establishes the final level of p21 mRNA and determine if the cell would proliferate or undergo senescence (Iakova *et al.,* 2004). Likewise, Kumar *et al.* (2021) have demonstrated that RBP (La and HuR) cooperatively modulate translation repression of *PDCD4* mRNA. In addition to the cytoplasmic post-transcriptional regulation, RBPs also assist in splicing the pre-mRNA and 3′ end processing, as part of a heterogeneous nuclear ribonucleoproteins complex (Xu *et al.,* 2001).

The uAUG and uORF are the major regulatory elements in the 5′ UTR. Studies have shown that approximately 50% of the 5′ UTR isoforms contain single or multiple uORFs (Davuluri *et al.,* 2000, Ingolia *et al.,* 2011). The uORF is defined by a start codon that is out-of-frame and upstream of the main coding sequence with an in-frame stop codon. In contrast, the uAUG is upstream, out-of-frame start codon without a downstream stop codon. Both uORF and uAUG primarily function as a translation repressor by limiting the ribosome access to the downstream AUG of the principal ORF (Matsui *et al.,* 2007). The translation of a wide variety of mRNAs is tightly controlled by the presence of uORF in the mRNA 5′ UTR (Reviewed in Chatterjee and Pal, 2009). The amplitude of repression mediated by these elements depends on the sequence context of the uAUG (Morris and Geballe, 2000). In addition to acting as a decoy, the short peptide translated from uORF may act in a cis-fashion and stall the ribosome at the uORF, reducing initiation from the downstream principal ORF (Oyama *et al.,* 2004). The stalled ribosome at the uORF can also undergo reinitiation and access the downstream main AUG to begin translation. The mechanism of delayed reinitiation has been noted for different genes under various conditions including stress response, amino acid starvation, and apoptosis (Beznosková *et al.,* 2015, Dever *et al.,* 1992, Proud, 2005, Szamecz *et al.,* 2008).

High ordered structures, including G quadruplexes, are also prevalent in the 5′ UTR and frequently noted in genes with regulatory functions (Reviewed in Araujo *et al.,* 2012). Davuluri *et al.* (2000) performed a regression tree analysis and noted that a large majority of transcripts (>90%) that were highly regulated also contained stable secondary structures in their 5′ UTR. These structures when present near the 5′ cap rendering the 5′ cap less accessible and blocking the formation of the PIC (Pickering and Willis, 2005). These structures could also impede the scanning process and studies have shown that a 5′ UTR structure with the minimum free energy of -50 kcal/mol is sufficient to impact the 43S ribosome scanning process (Kozak, 1989, Pelletier and Sonenberg, 1985). Other than inhibiting translation, the secondary structures (stem-loops) also associate with RBP to modulate gene expression (Fraser *et al.,* 2008, Goss and Theil, 2011). These secondary structures in the 5′ UTR are unwinded by cellular RNA helicases including eIF4A, RNA helicase A, DDX3, and DHX29, to name a few (Bourgeois *et al.,* 2016). The RNA helicases play an important role in gene expression and their aberrant expression has been linked to different pathologies including cancers (Robert and Pelletier, 2013).

The 5′ UTR could fold into an extended and multidomain structure that serves as an internal site for ribosomal recruitment, mediating cap-independent translation (Reviewed in Kozak, 2001). These IRES elements, discovered initially in viruses, are also reported in mammalian mRNA with more than 10% of mRNAs described in the literature found to contain an IRES element (Spriggs *et al.,* 2008, Weingarten-Gabbay *et al.,* 2016). The IRESs are implicated in facilitating translation initiation during stress conditions when the cap-mediated translation is compromised (Spriggs *et al.,* 2009). The IRES containing mRNAs have been shown to encode regulatory proteins such as TFs, growth factors, protooncogene, and homeodomain proteins (Lacerda *et al.,* 2017). The IRES trans-acting factors (ITAFs) have been shown to remodel cellular IRESs and act as a chaperone to induce structural changes in the mRNA following cellular stress to activate IRES-mediated translation. This mechanism of regulation has been reported for several mRNAs including *MYC* (Cobbold *et al.,* 2010), *APAF1* (Mitchell *et al.,* 2003), and *BAG1* (Pickering *et al.,* 2004).

### 1.2.2.2 3′ UTR regulatory elements:

The 3′ UTR plays an important role in mRNA localisation, stability, and translation. The 3′ UTR also act as a scaffold to facilitate protein-protein interactions (Berkovits and Mayr, 2015). Studies have shown that more than 50% of the genes generate 3′ UTR isoforms using alternative cleavage and polyadenylation (Reviewed in Mayr, 2016). The length of the 3′ UTR has dramatically increased during evolution with a median length of 1200 bp noted in humans

(Lianoglou *et al.,* 2013). The longer 3’ UTRs contain diverse regulatory elements including the binding sites for miRNAs and RBPs.

miRNAs are class of short non-coding RNA that are on average 22bp in length. The mature miRNAs are incorporated into the multimeric protein–RNA complex called RNA-induced silencing complex (RISC) that regulate gene expression. In most cases, the miRNA interacts with the 3’ UTR of the target mRNA and induces mRNA degradation and translation repression (Huntzinger and Izaurralde, 2011, Ipsaro and Joshua-Tor, 2015). The 3’ UTR contains a seed region that binds with partial complementarity to the miRNA. Friedman *et al.* (2009) reported that more than 60% of human protein-coding genes could be potentially regulated post-transcriptionally by miRNA. The miRNA suppresses the mRNA expression by recruiting RISC (RNA-induced silencing complex) that degrade the target transcript (Kawamata and Tomari, 2010). miRNA binding sites have been also identified in the 5’ UTR and the coding region (Xu *et al.,* 2014). The miRNAs are generally repressive; however, translational activation has been also reported in some cases (O'Brien *et al.,* 2018).

In addition to miRNA binding sites, the 3’ UTR also contain AU-rich and GU-rich elements that negatively affect gene expression (Barreau *et al.,* 2005, Vlasova *et al.,* 2008). These elements bind to RBPs and have shown to destabilize mRNAs, repress translation, and has been also noted in some cases to increase protein synthesis (Kontoyiannis *et al.,* 1999, Lindstein *et al.,* 1989). Studies have identified various RBPs such as AUF1, KSRP, and TTP that promote the decay of AU-rich containing mRNAs (Liao *et al.,* 2007, Stoecklin *et al.,* 2003). The 3’ UTR also encompasses mRNA localisation signal (zip codes) which are short repetitive sequences or stem-loop structures that bind to ribonucleoprotein (RNP) complexes to assist their export to specific cytoplasmic locations (Jansen, 2001).

### 1.2.3 Translation regulation and post-translational processing:

Translational control plays an important role in regulating gene expression and defining the proteome. The translational regulation can occur at different steps during the translation process, but the majority of regulation takes place at the initiation phase (Reviewed in Lackner and Bähler, 2008). The process of translation could be globally regulated by modulating one or more components of the core translation machinery. Such global changes in the profile and activation states of key eIFs have been frequently noted in cancers (Bjornsti and Houghton, 2004). One such example has been reported by Boussemart *et al.* (2014) who showed an increased expression of eIF4F, leading to therapy resistance and metastasis in BRAF-mutated tumours. However, changes in the levels of certain general translation initiation factors do not always affect the global translation pattern. In the case of eIF4E, only the translation of a specific subset of eIF4E-sensitive mRNAs is affected: those that contain long, highly structured 5’ UTR (Boussemart *et al.,* 2014, Roux and Topisirovic, 2018). The specific translational control is induced by the cis and trans-acting factors in the selected mRNA. Some of the cis and trans-acting elements in the 5’ and 3’ UTR have been described above (*Figure 1.4*). Moreover, the codon usage bias has been exhibited to locally exert their effect on the gene’s expression level. A positive correlation has been noted between the expression level of a gene and the degree of its codon bias (Plotkin and Kudla, 2011). The differential codon usage has been also shown to influence the accuracy and efficiency of protein synthesis (Tuller *et al.,* 2010, Zhou *et al.,* 2009).

Long noncoding RNA (lncRNA) have been shown to play a role in protein translation and are also implicated in regulating mRNA stability and transcription (Reviewed in Song *et al.,* 2021). LncRNAs have been shown to interact with RBPs to modulate translation of specific mRNA, for example, lncRNA *AFAP1-AS1* has been observed to associate with *AUF1* to promote translation of ERBB2 that encode for HER-2 protein (Han *et al.,* 2020). The lncRNA also interact with the component of translational apparatus to modify their phosphorylation status (Xu *et al.,* 2020). Moreover, the lncRNA acts as a molecular sponge or decoy for miRNA to regulate protein expression (Sun *et al.,* 2016, Xu *et al.,* 2019).

Post-translational modifications (PTMs) have been shown to increase the diversity of the proteome by covalent addition, removal or folding of functional groups in the protein that drastically change their functional properties (Higgins and Hames, 1999). Currently, more than 3 X $10^6$ experimentally verified PTMs of more than 69 types have been defined for greater than 1 X $10^6$ proteins in the PTMcode2 database (Minguez *et al.,* 2012, Minguez *et al.,* 2014). Some of these modifications include phosphorylation, methylation, acetylation, ubiquitination, and glycosylation. The covalent attachment of the small ubiquitin protein (76 amino acid) targets the substrate protein for degradation by the 26S proteasome complex (Hershko and Ciechanover, 1998). The protein regulation by ubiquitin-mediated degradation contributes to the vital cellular processes including transcription, cell-cycle regulation, signal-transduction, and antigen presentation (Reviewed in Zheng and Shabek, 2017). The combination of multiple PTMs on a protein surface constitutes a PTM code that is recognised by different effectors, stimulating or inhibiting downstream events (Lothrop *et al.,* 2013). Moreover, different PTM modifications interact with each other and the crosstalk between PTMs are essential for optimal gene expression, DNA damage response, and chromatin organisation (Badeaux and Shi, 2013, Parkes and Niranjan, 2019, Venne *et al.,* 2014).

Post-translational modification has been also shown to influence chromatin folding and modulate DNA accessibility. A wide variety of PTM modifications of histone tail has been described in the literature and include acetylation, ubiquitination, sumoylation, phosphorylation, and methylation (Reviewed in Tolsma and Hansen, 2019). These modifications stimulate structural and dynamic changes in the nucleosome, impacting the accessibility of DNA sequences near the nucleosome region and thus affect the DNA dependent processes such as transcription, replication, and DNA repair (Radman-Livaja and Rando, 2010). The PTM code of the histone shapes the chromatin state at a given locus. Certain PTM signatures at histone tails such as acetylation (H3K9ac, H3K27ac, H4K16ac, H3K14ac, H3K18ac), phosphorylation (H3S10ph), and methylation (H3K4me1, H3K4me2, H3K4me3, H3K79me1, H3K36me3) are associated with relaxed or open chromatin conformation (euchromatin states). On the other hand, certain histone PTM signatures including methylation at certain regions (H3K9me3, H3K9me2, H3K27me3, H4K20me3) and ubiquitination (H2Aub) are repressive and associated with compact or closed chromatin conformation (heterochromatin states) (Reviewed in Alaskhar Alhamwe *et al.,* 2018).

## 1.3 Relevance of Transcription Start Site selection

The TSS is described as a position at the 5' end of a gene sequence from where the transcription begins. It defines the 5' gene boundary and the beginning of the 5' UTR. Identifying the TSS is crucial for the characterisation of putative promoter motifs and provides important information about the gene expression pattern and regulation. The transcription of protein-coding genes is initiated by RNA Pol II in the core promoter region that is +/-50 nucleotides with respect to the TSS (Hampsey, 1998). The core promoter directs the initiation of transcription at the TSS.

The examination of the pattern of transcription initiation identified two different modes of initiation namely focused and dispersed (Lenhard *et al.,* 2012, Kadonaga, 2012). In the focus mode of initiation, the transcription begins at a single predominant TSSs or within a small region of few nucleotides. In contrast, the dispersed initiation demonstrates multiple weak initiation sites spread over a broad region of 50-100 nucleotides within the core promoter. Carninci *et al.* (2006) have noted that the TATA boxes are overly represented in the promoters that show the focus mode of initiation. A majority of TATA boxes (>70%) are situated within a tightly defined region with a preferred position of -31 or -30 relative to the TSS (Carninci *et al.,* 2006). Structural studies also confirmed this finding and reported a distance of 30 nucleotides from the TATA box to the active centre of RNA Pol II (Smale and Kadonaga, 2003). The precise selection of the TSSs could be predicted in the case of promoters with TATA boxes, with the TATA box being the anchor of PIC and guiding the polymerase to the predicted transcription starting site. However, the factors that decide the TSSs in the dispersed mode of initiation are not yet known.

The understanding of the exact position of the TSS is critical for two reasons. Firstly, mapping the TSS identifies the regulatory regions in the core promoter that immediately flank the TSS. Secondly, the TSSs determine the nature of the 5' UTR and the regulatory features present in the alternative TSS isoforms. Alternative transcription initiation producing multiple TSS isoforms has been frequently reported in the literature and create diversity in the human transcriptome (Davuluri *et al.,* 2008, Landry *et al.,* 2003).

### 1.3.1 Alternative Transcription Initiation (ATI) and gene expression:

In the mammalian genome, greater than 50% of the genes have alternative promoters (Kimura *et al.,* 2006). In humans, about 30% to 50% of the genes use alternative promoters, with on average 4 TSSs per gene (Forrest *et al.,* 2014, Davuluri *et al.,* 2008). Alternative transcription initiation and termination contribute more to genomic diversity than alternative splicing (Reyes and Huber, 2018). The selective use of alternative transcript isoforms has been shown to play a significant role in regulating gene expression during development, cell differentiation, and cell specialisation (Ding *et al.,* 2007, Pozner *et al.,* 2007, Vu and Hoffman, 1994). The transcription initiation from alternative promoters could follow two patterns: initiation from alternative promoters that changes the ORF (*Figure 1.5A*) or initiation from a nearby promoter that does not change the ORF but yield divergent 5' UTR (*Figure 1.5B*).

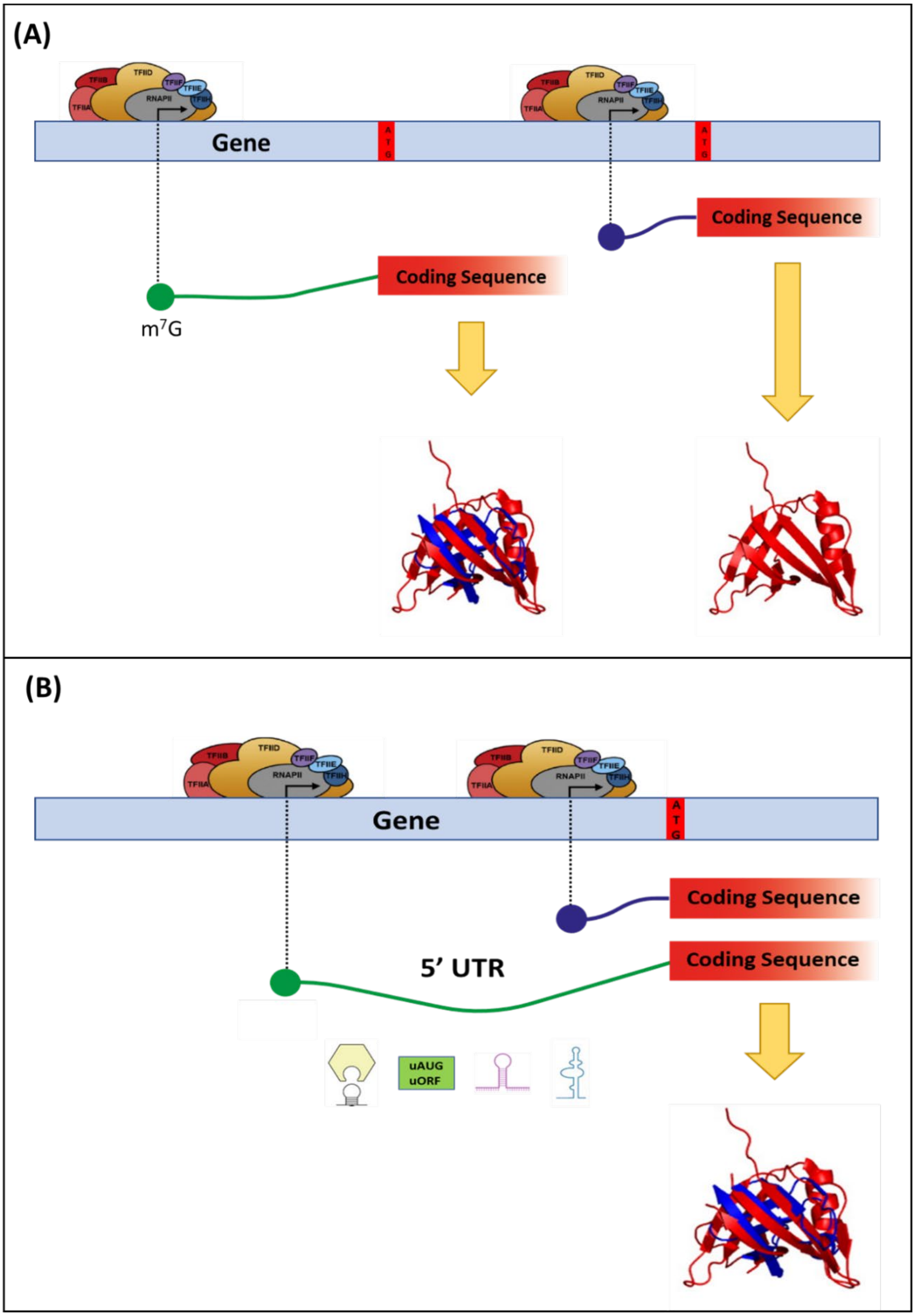


**Figure 1.5: Transcription initiation patterns from alternative promoters. (A)** Initiation from multiple promoters that change the ORF leading to the production of distinct proteins. The example shows alternative transcription initiation from a distal and intronic promoter. The protein isoform generated from the intronic promoter lacks the N terminal coding sequence (in blue). **(B)** Transcription initiation from alternative promoters that do not change the ORF but changes the nature of the 5' UTR. Both the TSS isoforms encode for the identical protein but differ in their translational potential due to the presence of the regulatory elements in the longer 5' UTR isoform. The structure of the protein is shown for illustrative purpose and to highlight that missing N terminal (in blue). The regulatory elements in the longer 5' UTR from left to right: binding domain for RNA binding protein, upstream open reading frame/start codon (AUG), secondary structures, and internal ribosome entry sites (IRES).

The ATI from the distal promoters could change the ORF and generate the mRNA isoform that encodes novel proteins. For example, the transcription initiation of *RUNX1* (runt-related transcription factor 1), an important regulator of early haematopoiesis, is mediated by two functionally distinct promoters that differ in their coding region, producing protein isoforms that differ in their biological functions (Pozner *et al.,* 2007). Likewise, *LEF1* (Lymphoid Enhancer Binding Factor 1), which regulate the transcription of Wnt/β-catenin target genes, is transcribed using two alternative promoters. The earlier promoter produces a full-length functional protein, whereas the intronic promoter produces a non-functional protein, which lacks the β-catenin binding domain and suppresses Wnt-mediated regulation of target genes (Arce *et al.,* 2006). Other related examples include *HOMER1* (Goossens *et al.,* 2007), *GNAS* (Weinstein *et al.,* 2007), and *CDKN2A* (Quelle *et al.,* 1995). The use of alternative promoters could also generate protein isoforms with truncated peptide sequence at the N-terminal, for example, the *CTNNA3* (catenin-cadherin associated protein, α3) in which the smaller isoform is deficient in the β-catenin binding domain and is unable to restore cell-cell adhesion (Goossens *et al.,* 2007).

The ATI also produce mRNA isoforms that encode for identical protein but differing in their 5’ UTR region and affects the mRNA translational potential. The human *BBOX1* gene, which produces a key enzyme for fatty acid metabolism, is transcribed from three distinct promoters generating heterogeneous tissue-specific 5’ UTRs (Rigault *et al.,* 2006). Similarly, *SHOX* (short stature homeobox), involved in bone growth and development, uses two alternative promoters, one of them producing an alternative longer 5’ UTR containing seven uAUGs that suppress mRNA translation (Blaschke *et al.,* 2003). Interestingly, the distinct 5’ UTR of the *XIAP* mRNA (X-chromosome linked inhibitor of apoptosis) contains an IRES. The shorter *XIAP* 5’ UTR maintains basal expression under normal growth conditions. On the other hand, the longer IRES-containing isoform is implicated in cap-independent translation during stress (Riley *et al.,* 2010). The *FGF1* (fibroblast growth factor 1) mRNA is another example of alternative IRES encoding 5’ UTR that regulate mRNA expression during cellular stress (Martineau *et al.,* 2004). Additionally, the alternative 5’ UTR could also contain secondary structures and uORF that strongly inhibit translation. Altered expression of the longer 5’ UTR of tumour suppressor *BRCA1* (Breast Cancer gene 1) mRNA could decrease the expression and contribute to the sporadic ovarian cancers and breast cancers (Sobczak and Krzyzosiak, 2002).

### 1.3.2 Transcription start site selection within a single promoter region:

Previous studies examining ATI have evaluated TSSs derived from multiple promoter regions. These TSSs are usually separated by a significant genomic space and encode distinct 5’ UTR regions that harbour regulatory elements. However, the work of Carninci *et al.* (2006), Forrest *et al.* (2014), Karlsson *et al.* (2017), Kawaji *et al.* (2006), and Suzuki *et al.* (2001) have identified the presence of multiple TSSs within a given core promoter region. Forrest *et al.* (2014), as part of the FANTOM consortium, have mapped TSSs and their usage in a large subset of human and mouse primary cells, cell lines and tissues. The authors documented that a large set of promoters displayed several closely spaced TSSs with independent cell type specific expression profiles. Kawaji *et al.* (2016) have also demonstrated the regional and positional bias in TSS distribution, highlighting that the TSSs are ‘tissue-specifically’ utilised. These findings suggested a potential regulatory role of multiple TSS within a promoter region and their tissue-specific TSSs distribution patterns. These multiple TSSs, which are separated by just a few nucleotides, could impact the translational potential of the mRNA isoform.

The landmark study by Rojas-Duran and Gilbert (2012) has shown that the transcript isoforms derived from alternative starting sites that are separated by only 50-200 bp showed large differences in translational activity. The authors have demonstrated that a seemingly minor change in transcription initiation could have a major impact on mRNA translational potential. Likewise, a study by Wang *et al.* (2016) has reported isoform-specific translational differences by combining polysome profiling with high throughput 5′ mRNA end sequencing. Although most of the transcript isoforms analysed by the study were presumably derived from alternative promoters, some of the TSSs analysed were separated by less than 200 bp and showed a significant difference in polysome association (isoform-divergent translation). It signifies that the minor changes in TSS selection within a single promoter, like ATI from multiple promoters, could incorporate regulatory elements which significantly impact translational potential.

The studies highlighted above point to the functional role of multiple TSSs within a single promoter region. The selection of alternative TSSs within a single promoter region could change the 5′ UTR by few nucleotides and incorporate regulatory elements that affect mRNA translatability. Despite their existence, the consequences of differential TSS selection are not yet known. The dominant starting sites (major TSS) have usually been the focus of studies that analysed ATI-mediated regulation (Arribere and Gilbert, 2013, Dieudonné *et al.,* 2015, Li *et al.,* 2019). A recent study by Xu *et al.* (2019) underestimates the role of multiple TSSs and have argued that there is only one optimal TSS per gene and alternative TSSs within the promoter region are the products of molecular errors. However, single-cell analysis of TSSs by Karlsson *et al.* (2017) have shown that the multiple TSSs in the promoter region are coregulated and are not stochastically expressed. Moreover, the study by Wang *et al.* (2016) have also showed differential isoform-specific polysome association. The study by Wang *et al.* (2016) together with the findings of Kawaji *et al.* (2016) and Rojas-Duran and Gilbert (2012) highlight the potential regulatory role of multiple TSSs in the single promoter region and the need to further study their functional significance.

There is a gap in the literature regarding the functional consequences of ATI within a single promoter. The alternate TSSs within a core promoter could also produce divergent 5′ UTR isoforms, differing by only a few nucleotides, without changing the ORF and would yield identical protein products. Further studies are required to understand the impact of TSS selection within the promotor region and would generate novel insights.

### 1.3.3 Mapping transcription initiation sites:

Given the significance of TSS selection in the gene regulation process, several attempts have been made to map the TSSs at individual or genome-wide levels. The main methods used in the literature are cap analysis of gene expression (CAGE) (Takahashi *et al.,* 2012), oligo-capping (Hashimoto *et al.,* 2004), robust analysis of 5′-transcript ends (5′-RATE) (Gowda *et al.,* 2006), rapid amplification of 5′ cDNA ends (Frohman *et al.,* 1988), and their variations.

CAGE is a commonly used method for high-throughput profiling of the TSSs and is based on the specific chemical oxidation of the 2',3'-diol structure at the 5′ ends on the cap nucleotide. It is followed by biotinylation that enables selective capture of the capped message using streptavidin immunoprecipitation and purification of full-length

RNA-cDNA hybrid. Subsequently, single-stranded DNA is released and ligated to linkers containing recognition sites for MmeI endonuclease that produces 20 bp CAGE tags. The tags are amplified, concatenated, cloned and sequenced (Kodzius *et al.,* 2006). In the oligo-capping approach, the 5' mRNA cap structure is removed and replaced with oligoribonucleotides to specifically amplify 5' end sequences of mRNAs (Maruyama and Sugano, 1994). The ligation reaction is made cap-specific by removing the 5' phosphate group from the uncapped mRNA using alkaline phosphatase. In the 5' RACE technique, the 5' end of the selected mRNA is amplified using a gene-specific oligonucleotide primer (Frohman *et al.,* 1988). The technique has been subsequently improved by incorporating an RNA-ligase mediated oligo-capping step to specifically amplify selected full-length mRNA (Maruyama and Sugano, 1994, Volloch *et al.,* 1994) (section 2.2.5, Figure 2.3).

The high-throughput mapping of TSSs has facilitated the development of TSS databases that characterise the dynamically changing TSS landscape in a diverse range of samples. The three notable TSS databases include the Functional Annotation of Mammalian Genomes 5 (FANTOM5) project (Forrest *et al.,* 2014), DataBase of Transcription Start Sites (DBTSS) (Suzuki *et al.,* 2002), and the Encyclopedia of DNA Elements (ENCODE) project (Dunham *et al.,* 2012). The TSSs in the FANTOM database are identified using CAGE technology adapted for single-molecule sequencing with a median depth of 4 million mapped tags per sample to precisely detect transcription initiation activities (Kanamori-Katayama *et al.,* 2011). The FANTOM edition 5 contains the TSS data across 975 human samples and 399 mouse samples. The human samples included 573 primary cell lines, 152 human post-mortem tissues and 250 different cancer cell lines covering distinct cancer subtypes (Lizio *et al.,* 2015). The DBTSS is based on the oligo-capping approach adapted to massively parallel NGS (Tsuchihara *et al.,* 2009). The recent version of the DBTSS contains the TSS profile from 20 tissues and 7 cell cultures (Yamashita *et al.,* 2012). The DBTSS database has been recently expanded to include the genomic and epigenome variation dataset of the Japanese population (Suzuki *et al.,* 2018). The ENCODE project has profiled the TSS of 36 cell lines using CAGE sequencing (Batut *et al.,* 2013, Dunham *et al.,* 2012). The FANTOM database is clearly the largest available collection of TSS profiles on a single platform.

The TSS distribution profiles of the samples in the FANTOM database was produced using modified CAGE protocol optimised for single-molecule sequencing (Kanamori-Katayama *et al.,* 2011). In the protocol used by the authors, the RNA was first reverse transcribed, and the mRNA cap was biotinylated and captured on magnetic streptavidin beads. Subsequently, the RNA/cDNA hybrid molecules were washed to remove unbound molecules and the single-stranded cDNA was released following treatment with RNase H and RNase I. The released cDNA was poly-A tailed, blocked, and loaded onto the HeliScope flow cell channel for high-precision sequencing (*Figure 1.6*). The sequencing data was analysed, and library sizes were adjusted by relative log expression. The CAGE peaks were annotated to determine promoter region, gene associations, ontology, co-expression, and motif analysis. To identify the CAGE peaks across the genome, the authors employed decomposition-based peak identification, where the CAGE tags were first clustered based on proximity and the tags wider than 49 bp were decomposed into non-overlapping subregions using independent component analysis (Forrest *et al.,* 2014).

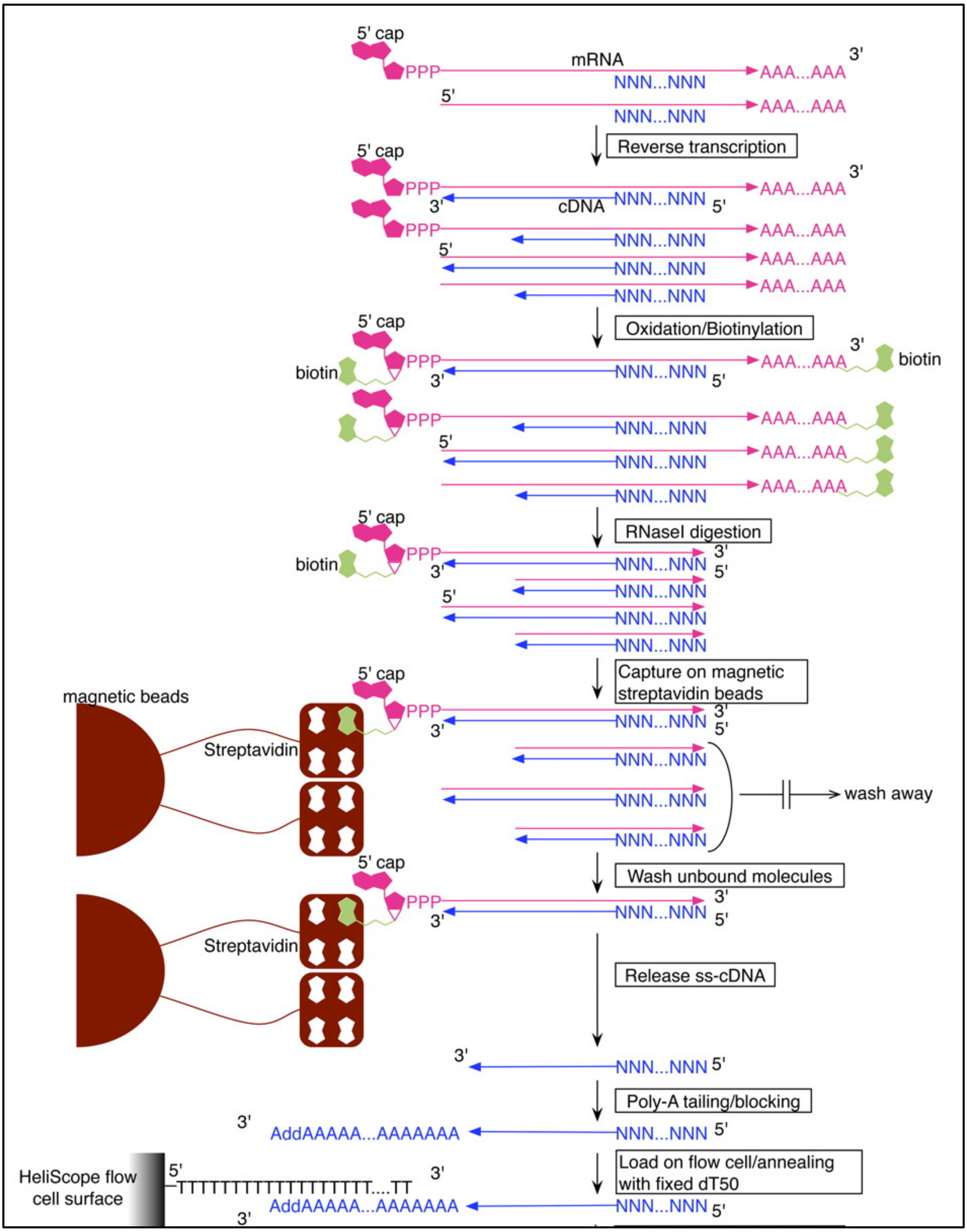


**Figure 1.6: Schematics of HeliScopeCAGE for high-throughput determination of TSSs in the FANTOM database.** The RNA is reverse transcribed using random primers and the mRNA cap is oxidised with sodium peroxide ($Na_2O_2$) and biotinylated with biotin hydrazine. It is followed by the digestion of single-stranded RNA, that are not reverse transcribed, using RNase I. The biotinylated RNA/cDNA hybrid molecules are captured using magnetic streptavidin beads and unbound molecules are washed away. Next, the single-stranded cDNA is released using RNase H and RNase I followed by heat treatment. The release cDNA is polyA tailed using terminal deoxynucleotidyl transferase and dATP and then blocked using biotin- dideoxy ATP. The blocked tailed cDNA is loaded on the HeliScope flow cell channel and anneals with the dT50 surface and sequenced. Figure adapted from Kanamori-Katayama *et al.* (2011) (CC BY-NC 4.0).

## 1.4 G-quadruplex structure and gene regulation

G-quadruplexes (G4) are non-canonical, four-stranded, secondary structures that are formed by sequences rich in guanine nucleic acid (Smith and Feigon, 1992). These guanine-rich sequences spontaneously fold into a cyclical planar arrangement held together by Hoogsteen base pairing (N1–N6 and N2–N7) to form a G-quartet (*Figure 1.7*). Three or more G-quartet stack onto one another, separated by ~3.3 Å in vertical stacking, and form a stable right-handed helical G4 structure (Forman *et al.,* 2000). These structures are stabilised by monovalent cations (K+ > NH4+> Na+ > Li+) which interact with the negatively charged carbonyl oxygen atom placed at the centre of the G4 (Burge *et al.,* 2006, Hardin *et al.,* 1992). Studies have found that a classical consensus sequence capable of forming G4 can be described as $G_3$+$N_{1–7}$$G_3$+$N_{1–7}$$G_3$+$N_{1–7}$$G_3$+ where N corresponds to any nucleotide (A, G, T, C, or U) (Huppert and Balasubramanian, 2005, Todd *et al.,* 2005). However, recent studies have shown the classical consensus definition is flexible and imperfect G quadruplex with longer loop and interrupted G runs also exists (Puig Lombardi and Londoño-Vallejo, 2019, Varizhuk *et al.,* 2017).

The capacity of guanosine monophosphate to self-aggregate into a four-stranded helical structure has been known since the early 19th century (Bang, 1910). Then, 50 years later, the work of Gellert *et al.* (1962) using fibre diffraction revealed the aggregation of guanosine monophosphate into a gelatinous substance in the aqueous solution. Subsequent biophysical studies using conserved DNA sequence of telomeric repeats or immunoglobulin switch regions have shown the formation of stable G4 structures *in vitro* (Sen and Gilbert, 1988, Sundquist and Klug, 1989). Since then, numerous sequences have been identified in DNA and RNA that folds into a stable G4. The G4 can adopt a wide variety of topological conformations depending on inter or intramolecular folding of G rich strands. The intermolecular folding can arise from different combinations of strand orientation resulting in a parallel or anti-parallel G4 structure (Burge *et al.,* 2006, Banco and Ferré-D'Amaré, 2021).

The G4 structures are found in both DNA and RNA. The fundamental difference between the DNA and RNA G4 is the presence of ribose sugar and uracil instead of thymine residue in the RNA G4. The RNA G4 are thermodynamically more stable and less hydrated compared to their DNA G4 counterpart (Joachimi *et al.,* 2009). The presence of the 2'OH group in the ribose sugar enhances the stability of RNA G4 due to the increase in intramolecular interactions. It also restricts the RNA G4 topology to a parallel conformation, where all the strands of G4 are oriented in a similar direction, via stearic constraints on the glycosidic torsion angle (Tang and Shafer, 2006).

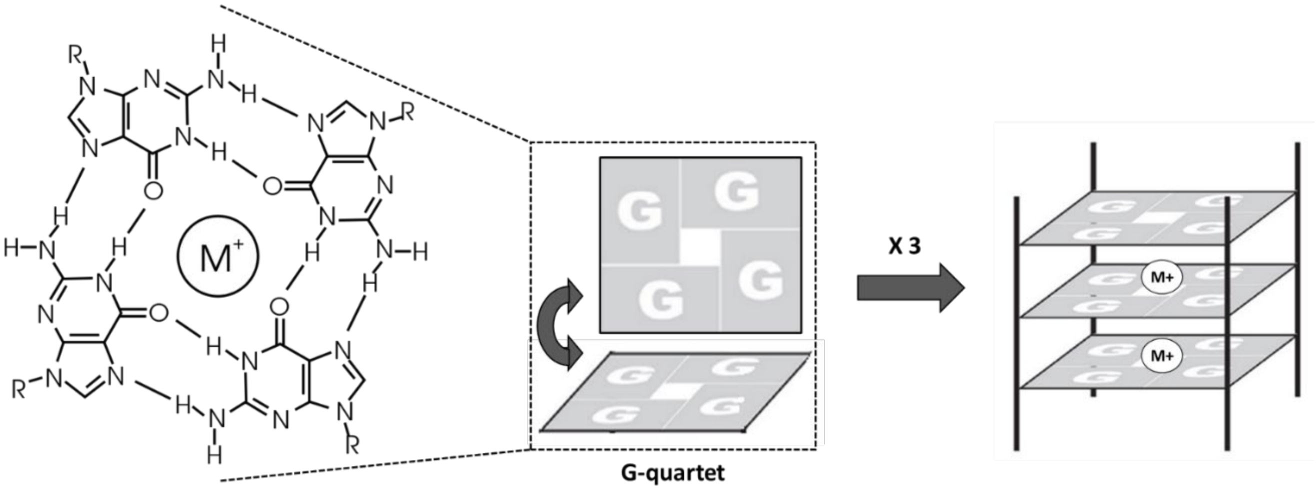


**Figure 1.7: Structure of G quadruplex.** Chemical structure of G-quartet with a metal ion (M+) coordinated to carbonyl oxygen atoms. G-quartets are stabilized by Hoogsteen hydrogen bonding (N1–N6 and N2–N7). Three or more planar G-quartets stacks on top of one another, forming four-stranded helical structures.

## 1.4.1 Functional consequence of DNA G4:

A seminal study by Huppert and Balasubramanian (2005) demonstrated more than 3.7 x $10^5$ G4 forming sequences in the human genome. Recently, a high-throughput study has identified a further 4.5 X $10^5$ previously uncharacterized putative G4-forming sequences with long loop and bulges (Chambers *et al.,* 2015). The Genome-wide bioinformatics analysis has shown that G4s are non-randomly distributed, correlating with important genomic regions including promoters of proto-oncogenes, mutational hotspots, and immunological switch regions (Huppert and Balasubramanian, 2005, Simonsson, 2001). The DNA G4 have been also reported in other genomic regions including ribosomal DNA (Drygin *et al.,* 2009), mitochondrial DNA (Wanrooij *et al.,* 2010), the region of replication initiation (Besnard *et al.,* 2012), and retrotransposon elements (Lexa *et al.,* 2014).

The earliest biologically relevant G4 sequences were observed in the telomeric repeat region (Sen and Gilbert, 1988). The G4 formation by the consensus sequence (GGGGGAGCTGGGGAAGGTGGG) in the telomeric repeat region was shown to inhibit telomerase activity (Sen and Gilbert, 1988). It establishes a mechanistic link between G4 formation and telomere maintenance. As reviewed in Fouquerel *et al.* (2016), a G4 stabilising ligand could interfere with telomere repair and inhibit the growth of cancer cells. The work of Vannier *et al.* (2012) supported this notion and demonstrated that the mouse regulator of telomere elongation helicase 1 (*RTEL1*) unwinds the telomeric G4 to ensure the stability of the telomere. The impaired regulation of the G4 structures has been also shown to promote genomic instability. The G4 stabilisation has been shown to hamper Pif1 helicase activity that is responsible for DNA resection and homologous recombination (Jimeno *et al.,* 2018). Moreover, the G4 structures function as a potential sensor of oxidative damage to the DNA by reactive oxygen species and mediate transcriptional activation (Fleming *et al.,* 2017). Furthermore, the G4 motifs have been proposed to have a biological role in DNA replication. These structures have been shown to stall the progression of the DNA replication fork in absence of G4-unwinding helicases (Lopes *et al.,* 2011).

The G4s have been reported in the promoter region of many important genes including c-*MYC, VEGF*, *BCL2*, *KRAS*, c-*KIT*, and *TERT* (Reviewed in Yang, 2019). G4 structures in the promoter region have been shown to have a regulatory role. In the case of c-*MYC* and *KRAS*, stabilisation of G4 using a ligand resulted in reduced mRNA transcription (Cogoi and Xodo, 2006, Siddiqui-Jain *et al.,* 2002). The link between G4 and transcription has been also exhibited by Nguyen *et al.* (2014), showing that impaired function of G4-unwinding helicases (BLM and WRN) alters the expression profiles of genes containing G4 motifs in the promoter region (Johnson *et al.,* 2010, Nguyen *et al.,* 2014). A study by Mao *et al.* (2018) have provided evidence that the G4 structures in the promoter region decrease the local methylation at the CpG island by inhibiting DNA methyltransferase 1 enzymatic activity, contributing to elevated gene expression.

### 1.4.2 Functional consequence of RNA G4:

RNA G4 sequencing has revealed the widespread formation of G4 in the human transcriptome. The rG4-seq pioneered by Kwok *et al.* (2016a) has indicated more than 3000 putative RNA G4, mapped to more than 2000 genes, that increased to greater than 11000 presumed RNA G4 motifs upon treatment with G4 stabilising ligand (pyridostatin). The RNA G4 has been shown to be enriched in the mRNA 5' and 3' UTR (Beaudoin and Perreault, 2013, Bugaut and Balasubramanian, 2012), R loop (RNA:DNA hybrid) (Xiao *et al.,* 2013), non-coding RNA (Rouleau *et al.,* 2018), ribosomal RNA (Mestre-Fos *et al.,* 2019), IRES (Morris *et al.,* 2010), and RNA introns (Weldon *et al.,* 2018, Marcel *et al.,* 2011).

The RNA G4, like DNA G4, plays an important role in regulating gene expression. These structures have been shown to modulate transcriptional and co-transcriptional regulation. The RNA G4 consensus sequence in the nascent RNA could form complementary base pairing with the non-template DNA strand, resulting in the RNA:DNA hybrid-containing G4s in the R loop (Skourti-Stathaki and Proudfoot, 2014). Recently, Ribeiro de Almeida *et al.* (2018) has demonstrated the unique mechanism of post-transcriptional formation of hybrid G4 in the R loop of mouse immunoglobulin heavy chain (IHC) that promote class switching in the IHC locus. Intronic RNA G4s have been shown to regulate the alternative splicing of pre-mRNA. In a study by Weldon *et al.* (2018), the authors have reported a restrictive set of G4 ligands that interact with the G4 forming sequence of the Bcl-X pre-mRNA and shift the splicing from the dominant Bcl-XL (anti-apoptotic isoform) to the Bcl-XS (pro-apoptotic isoform). Likewise, the G4 motifs in the *TP53* intron region 3 regulate the splicing of intron 2, producing distinct p53 isoforms (Marcel *et al.,* 2011).

RNA G4s in the 3' UTR have been shown to play an important role in mRNA localisation. A study by Subramanian *et al.* (2011) has demonstrated that RNA G4 serve as a neurite localisation signal and deletion of the G4 sequence from *αCaMKII* and *PSD-95* resulted in the loss of mRNA translocation to dendrites. Ishiguro *et al.* (2016) has reported a TAR DNA-binding protein (TDP-43) capable of binding G4-containing mRNAs and facilitate their intracellular transport to distal neurite for local translation. Moreover, the presence of G4 in the 3' UTR of *FADS2* was shown to impede the binding of miRNA mir331-3p, regulating RNA interference (Rouleau *et al.,* 2017). Although the RNA G4 forming sequence are depleted in the coding region, their presence in the coding region is generally associated with translational suppression (Reviewed in Kharel *et al.,* 2020). The translation inhibition is mediated by obstructing the elongating ribosomes that cause ribosome stalling and dissociation (Murat *et al.,* 2014). The RNA G4 are also indicated in different lncRNAs (Reviewed in Jayaraj *et al.,* 2012). In the lncRNA *TERRA* (telomeric repeat-associated RNA), its presence has been demonstrated to contribute to the telomere homeostasis by interacting with telomeric protein TRF2 (Biffi *et al.,* 2012).

#### 1.4.2.1 Regulatory role of 5' UTR G4:

Huppert *et al.* (2008) revealed about 4,141 potential G4 motifs in the 5' UTR region of human mRNA. The overrepresentation of G4 in the mRNA 5' UTR points towards the important regulatory role of RNA G4. A variety of studies, using a cell-free system and cell-based reporter assay, have highlighted that the 5' UTR G4 are generally associated with a reduction in translational efficiency (*Table 1*). However, in some instances, a contrasting role for G4 has been noted. A study by Bonnal *et al.* (2003) showed that the formation of G4 motifs in the *FGF2* (fibroblast growth factor 2) IRES promoted translation. Likewise, Agarwala *et al.* (2013) noted the activating role of RNA G4 in the translation of *TGFβ2* (transforming growth factor β2). In addition, the stable G4 structure in the 5' UTR could also promote the formation of 80S translating ribosome on upstream AUG, reducing the translation from downstream principal ORF (Murat *et al.,* 2018). Recently, the 5' UTR G4 has been also indicated to facilitate distal localisation of the mRNA for local translation, contributing to distinctive regional proteomes (Maltby *et al.,* 2020).

The 5' UTR G4s could also regulate the translational output of the alternative TSS isoforms. The differential TSS selection could incorporate these secondary structures in the gene isoforms and control their translational output. However, the significance of alternatively transcribed G4 in the context of TSS selection within the core promoter is not yet defined.

**Table 1.1**: **Translational impact of 5' UTR G4 in cell-free and cell-based reporter assay.**

| Gene | Protein Product | RNA G4 Sequence (5' to 3') | Change in protein expression* | Reference |
|---|---|---|---|---|
| *NRAS* | GTPase NRAS | GGGAGGGGCGGGUCUGGG | ~ 70% decrease | (Kumari *et al.,* 2007) |
| *ZIC1* | Zinc finger protein 1 | GGGUGGGGGGGCGGGGGAGGCCGGGG | ~ 80% decrease | (Arora *et al.,* 2008) |
| *MT3-MMP* | Matrix metallopeptidase 16 | GAGGGAGGGAGGGAGAGGGA | ~ 55% decrease | (Morris and Basu, 2009) |
| *ERS1* | Estrogen receptor α | GGGUAGGGGCAAAGGGGCTGGGG | ~ 85% decrease | (Balkwill *et al.,* 2009) |
| *EBAG9* | Estrogen receptor binding site associated antigen 9 | GGAGCCUCCGCCGGGCGGGCGGGGAGGGGGAGGGGGCAGGUUUUGA | ~ 45% decrease | (Beaudoin and Perreault, 2010) |
| *FZD2* | Frizzled family receptor | GGGGAAGAAGCGCAGUCUCCGGUGGGGG | ~ 60% decrease | (Beaudoin and Perreault, 2010) |

| | | | | |
|---|---|---|---|---|
| | | CGGGGGCGGGGGGGGCGCCA<br>AGGAGCCGGG | | |
| *NCAM2* | Neural cell adhesion molecule 2 | GGAGGAGCGGCGGGGCUGCG<br>GGCGGCUGG<br>GGCACCGCGGGAGCG-<br>GCGGCGGCGG | ~ 35% decrease | (Beaudoin and Perreault, 2010) |
| *BCl-2* | Apoptosis regulator Bcl-2 | GGGGGCCGUGGGGUGGGAGC<br>UGGGG | ~ 50% decrease | (Shahid *et al.,* 2010) |
| *TRF2* | Telomeric repeat binding factor 2 | CGGGAGGGCGGGGAGGGC | ~ 65% decrease | (Gomez *et al.,* 2010) |
| *THRA* | Thyroid hormone receptor α | GGGUGCUGUGCCCUAGGGCC<br>UGGGUGGCAG<br>GGGGUGGGUGGCCUGUGGG | ~ 35% decrease | (Beaudoin and Perreault, 2010) |

*Percent change in protein expression of reporter gene induced by the G4 consensus sequence relative to the mutated or deleted G4 sequence.

## 1.5 Tools for detecting G4

The significance of G4 structures and their role in regulating cell processes in the context of health and disease has necessitated the generation of tools to predict and demonstrate their formation in the genome and transcriptome. A variety of biophysical techniques have been developed to validate the G4 formation by specific sequences including nuclear magnetic resonance (NMR) (Mathad and Yang, 2011), circular dichroism (CD) spectroscopy (Del Villar-Guerra *et al.,* 2018), x-ray crystallography (Campbell and Parkinson, 2007), thermal stability analysis (UV melting) (Rachwal and Fox, 2007), and fluorescence resonance energy transfer (FRET) (Mergny *et al.,* 2001). However, these techniques are not suitable for high-throughput detection of G4 motifs on a genomic level as they can interrogate only a single stretch of sequence for G4 identification in a single run.

### 1.5.1 Computational approaches for G4 prediction:

To scan the G4 motif on a larger scale, numerous predictive algorithms have been developed in the last decades. These algorithms were guided by biophysical and biochemical techniques that defined the rules and criteria of G4 folding. *Table 2* presents the list of different computational, open-source, prediction tools to detect G4 forming sequence.

The earlier approaches were based on an expression matching algorithm that searches for sequences strictly following the classical G4 motif sequences (GX-N1-7-GX-N1-7-GX-N1-7-GX). The recent approaches allow for greater flexibility and imperfection tolerance (longer loops, mismatches, incomplete G-runs, bulges) while searching for G4 forming sequence. The computational approaches to identify putative G4 sequences can be categorised into four groups, namely, classical expression matching, sliding windows, scoring-based, and machine learning approaches. In the sliding windows approach, the density of G-runs is calculated in a given sequence window to generate a score. Other scoring-based approaches use a flexible motif definition to produce a score. The score computed indicates the propensity of the input sequence to form G4. More recently, machine learning approaches have been developed that are motif independent and do not rely on predefined motif definitions.

**Table 1.2: Selected computational prediction tool to detect G4 motif in DNA/RNA.**

| Name | Salient features | Language | Reference |
|---|---|---|---|
| **Classical expression matching** | | | |
| Quadparser | Folding Rule: $G_3+N_{1-7}G_3+N_{1-7}G_3+N_{1-7}G_3+$ | C++,<br>Python | (Huppert and Balasubramanian, 2005) |
| ImGQfinder | Facilitate detection of imperfect G4 that contain interrupted or truncated G-runs. | Web | (Varizhuk *et al.,* 2017) |
| **Sliding window approach** | | | |

| | | | |
|---|---|---|---|
| G4P calculator | Compute G4 forming potential based on G-run density in a sequence.<br>Criteria used: G-run length = ≥ 3; number of G-runs per window = ≥4; window length = 100 bp; and sliding interval length = 20 bp. | C Sharp | (Eddy and Maizels, 2006) |
| G4Hunter | Accounts for the G-richness and G-skewness in the input sequence and output a score that indicates the propensity to form G4.<br>Each position in a sequence is given a score between −4 and 4 (0 for A and T, negative for C, and positive for G (G=1, GG=2, GGG=3, GGGG=4). | Python/R | (Bedrat *et al.,* 2016) |
| **Scoring based approaches** | | | |
| QGRS Mapper | Motif definition: $G_xN_{y1}G_xN_{y2}G_xN_{y3}G_x$ where X ≥ 2, default maximum length of 30 bp, default minimum G size =2. The length of G4, loop length, and loop criteria could be set by the user. | We, Perl, Java | (Kikin *et al.,* 2006) |
| pqsfinder | The algorithm has three logical steps: identification of all possible G-run quartets, assignment of a score, and overlap resolution. | Web, R | (Hon *et al.,* 2017) |
| **Machine learning** | | | |
| G4RNA screener | Artificial neural network-based algorithm that evaluates the similarity of the input sequence to known RNA G4 motif and output a score. | Python | (Garant *et al.,* 2017) |
| Quadron | Modelling based on tree-based gradient boosting machines that predict the intramolecular G4 formation. | R | (Sahakyan *et al.,* 2017) |
| **Complementary tool** | | | |
| Vienna RNA (RNA fold) | Estimate RNA G4 folding energy and examine the competition between RNA G4 and other secondary structures by comparing minimum free energy. | Web | (Lorenz *et al.,* 2013) |

### 1.5.2 Biochemical and molecular approaches to detect G4 structures:

Different molecular and biochemical approaches, in addition to the biophysical techniques described above, have been used to examine the formation of G4 structures *in vitro* and *in cellula*. Attempts have been made to interrogate the G4 folding by the G4 consensus sequences in the DNA and RNA for the selected targets and on a genome-wide scale.

For DNA G4, biochemical techniques including electrophoretic mobility shift assay (EMSA), dimethylsulfate (DMS) footprinting, and DNA polymerase stop assay have been used to demonstrate the *in vitro* formation of G4 structure (Reviewed in Sun and Hurley, 2010). The first line of evidence for the formation of DNA G4 *in vivo* was presented by Schaffitzel *et al.* (2001) using a G4 structure-specific antibody that stained the G4 in the telomeres of ciliates. Subsequently, G4 formation has been demonstrated in mammalian cells using different antibodies (BG4, 1H6) combined with complementary approaches such as the use of stabilising ligand and depletion of the G4 helicase FANCJ (Biffi *et al.,* 2013, Henderson *et al.,* 2013). Besides antibody and stabilising ligands, small molecules and probes with fluorescent properties have been developed and employed for real-time monitoring of G4 in the living cells after treatment with a G4 stabilising ligand (pyridostatin) (Shivalingam *et al.,* 2015, Zhang *et al.,* 2018). The G4 mapping was also adapted to next-generation sequencing (NGS) to identify G4-specific polymerase stalling sites in the purified human genomic DNA under different stabilising conditions (Chambers *et al.,* 2015). Moreover, the G4 landscape *in celluo* has been also determined using BG4-mediated chromatin immunoprecipitation followed by NGS sequencing (ChIP-seq) (Hänsel-Hertsch *et al.,* 2018).

The G4 formation in the RNA has been reported *in vitro* using different biochemical approaches including in-line probing (Beaudoin *et al.,* 2013), structural analysis using selective 2'-hydroxyl acylation or dimethyl sulphate coupled with lithium ion-based primer extension (SHALiPE/ DMSLiPE) (Kwok *et al.,* 2016b), and RNase T1 footprinting assay in which guanine in the G4 is protected against the RNase T1 digestion (Morris and Basu, 2009). The Dominguez group has developed a new strategy to confirm the formation of G4 by using a 7-deazaguanine analogue that contains carbon, instead of nitrogen, at position 7 and is no longer able to form Hoogsteen base pairing to yield a stable G4 (Figure *1.8A*). The authors performed RNA footprinting of long 7-deazaguanine-substituted RNAs (FOLDeR) to confirm the formation of RNA G4 in the entire functional pre-mRNA and identified differences in the native and deazaguanine-substituted analogue by footprinting assay and RNase H cleavage patterns (Weldon *et al.,* 2016, Weldon *et al.,* 2017). In addition, Kwok and Balasubramanian (2015) have also described a novel approach wherein RNA G4-induced reverse transcriptase stalling is coupled to ligation-mediated PCR to positionally map RNA G4 in selected mRNA transcripts (Figure *1.8B*). This method has been also adapted to high-throughput RNA sequencing (rG4-seq) to generate an *in vitro* transcriptome-wide map of canonical and noncanonical RNA G4 structures (Kwok *et al.,* 2016a).

The formation of RNA G4 inside the cell has been suggested by different studies. The work of Guo and Bartel (2016), using complementary structural probing approaches (DMS and SHAPE), has demonstrated more than 10,000 RNA G4 putative sites. However, the authors noted that most of the RNA G4 at these sites are globally unfolded in the cells. Biffi *et al.* (2014) have visualised RNA G4 in the cytoplasm that could be selectively stabilised using RNA G4 specific ligand (carboxypyridostatin) (Figure *1.8C*). Chen *et al.* (2018) have developed an RNA-G4 specific fluorescent probe (QUMA-1) that could be used for real-time visualization of RNA G4 in living cells. Likewise, Laguerre *et al.* (2015) have designed a Naphtho-template-assisted synthetic G-quartet (N-TASQ) fluorescent probe for the detection of RNA G4 motifs in the live cells. The TASQ ligands (BioTASQ) have been also recently used by the same group for transcriptome-wide detection of RNA G4 in human cells (Yang *et al.,* 2018). The authors have developed BioTASQ-mediated G4 RNA immunoprecipitation and sequencing protocol (G4RP-seq), also comprising of a crosslinking step, to capture the transient RNA G4 landscape of the human transcriptome (Figure *1.8D*). Recently, work by Lat *et al.* (2020) has evidenced self-biotinylation of DNA/RNA G4 structures, facilitated by the peroxidase activity of hemin complexed to them, that could facilitate tagging of G4-containing RNA and DNA in the living cells. Even though numerous studies highlighted above have supported the formation of RNA G4 inside the cell, there is still a lack of in-depth understanding about their folding inside the cells.

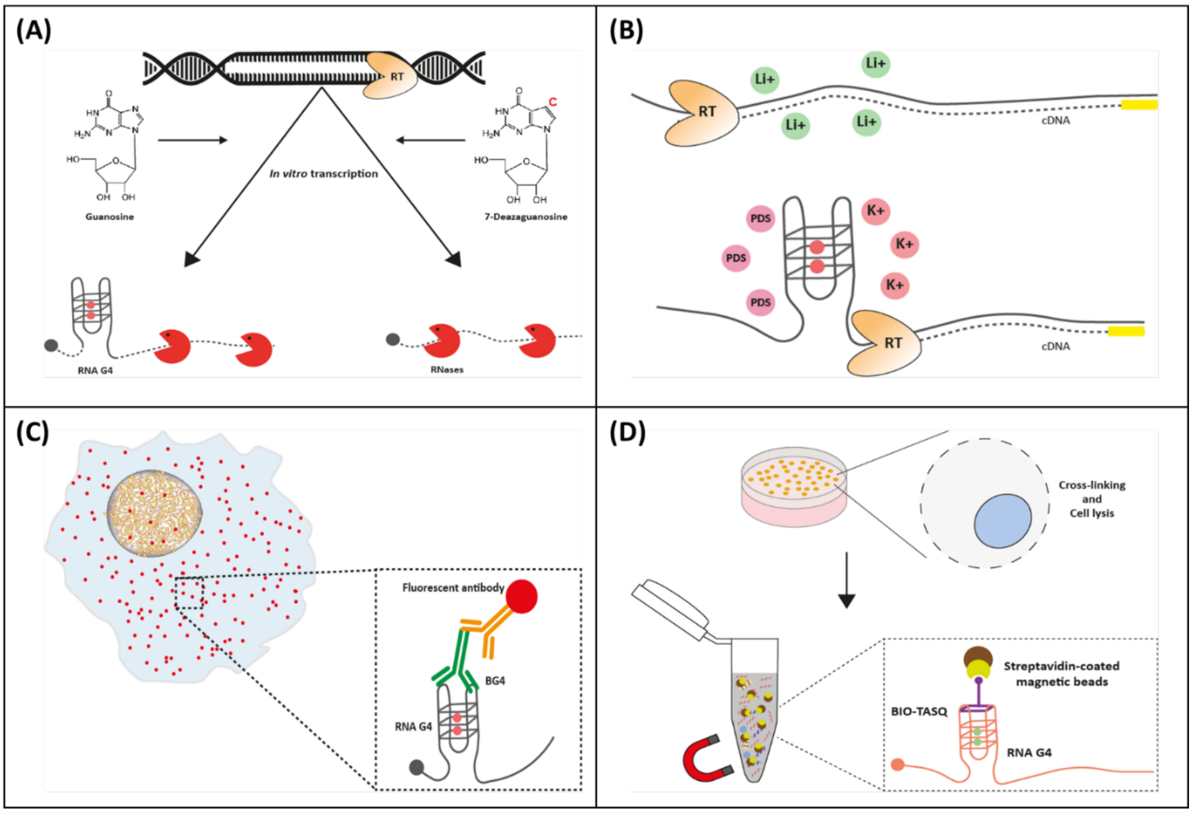


**Figure 1.8: Selected techniques to map and detect RNA G4 motif. (A)** Footprinting of long 7-deazaguanine-substituted RNAs (FOLDeR). The long functional pre-mRNA is transcribed *in vitro* using either guanosine or 7-deazaguanosine analogue. The 7-deazaguanosine has N7 substituted to Carbon [C] and is unable to form Hoogsteen base pairing to yield a stable G4. The G4 structured regions are subsequently mapped using ribonucleases (RNases). **(B)** G4-induced reverse transcriptase stalling (rG4-seqRNA). Two consecutive sequencing runs are performed, under stabilising ($K^+$ and pyridostatin-PDS) or non-stabilising ($Li^+$) conditions to detect G4 dependent polymerase stalling. **(C)** RNA G4 visualisation in the cytoplasm using structure-specific BG4 antibody in combination with fluorophore tagged secondary antibodies. **(D)** G4-RNA-specific precipitation (G4RP) using affinity capture by small-molecule ligand (BioTASQ). The cells were crosslinked and lysed and the RNA G4 in the lysate was captured using template-assisted synthetic G-quartets coupled to biotin. The BioTASQ bound to RNA G4 was pull-downed using streptavidin-coated magnetic beads and sequenced to determine transiently folded G4-RNAs.

## 1.6 Discrepancy in RNA and protein levels

The complex interplay between a variety of regulatory elements that act at different levels of gene expression results in discordant mRNA and protein expression profiles. As anticipated, a variety of studies have demonstrated a weak correlation between mRNA and protein levels (Reviewed in Vogel and Marcotte, 2012). Different studies examining the correlation between mRNA and protein abundance in both prokaryotic and eukaryotic systems have reported a correlation of about 0.40, implying that mRNA concentration could only explain 40% of the differences in the protein levels (Maier *et al.,* 2009, de Sousa Abreu *et al.,* 2009). Despite such a weak correlation, mRNA has been widely used as a proxy of their protein product. The undue reliance on the RNA measurement to predict gene expression could be explained by the relative effortless acquisition of transcriptomic profiles compared to the obtention of the proteomic data that is technically more challenging. The mRNA expression alone provides an incomplete picture and the higher mRNA levels might not correlate to the increase in protein concentration due to the sophisticated post-transcription, translation, and post-translational regulatory processes.

### 1.6.1 AGAP2 as a case study:

*AGAP2* (Arf GAP with GTP-binding protein-like domain, ankyrin repeat and PH domain 2, where Arf is ADP ribosylation factor and GAP is GTPase activating protein) is a member of the centaurin gamma 1 GTPase superfamily. It is transcribed from the CENTG1 gene locus located on chromosome 12 in reverse orientation (O'Leary *et al.,* 2016). *AGAP2* belongs to the Arf GAP protein family and act as a GTPase switches for Arfs and regulate Arf-mediated signalling (Navarro-Corcuera *et al.,* 2020). *AGAP2* was initially identified by Ye *et al.* (2000) and further characterised by Xia *et al.* (2003). Previously, it was referred as GGAP2 (GTP-binding and GTPase-activating protein 2) because of its bifunctional GTP-binding and GTPase-activating activities. It was also showed that unlike other Arf GAPs, AGAP2 can intrinsically activate its GTPase activity either via intermolecular or intramolecular interactions and the Arf GAP domain is essential for GTPase activation (Xia *et al.,* 2003).

Alternative splicing results in two isoforms namely isoform 1 (PIKE-L) and isoform 2 (PIKE-A or centaurin gamma 1) (Kahn *et al.,* 2008). These two isoforms share most of their DNA sequences except for the sequences in the first exon and intron. The start of isoform 1 (PIKE-L) is located downstream to the first exon in isoform 2 (PIKE-A) and lacks the initial sequences present in the N-terminus of isoform 2 (*Figure 1.9*). The expression of isoform 1 has been noted to be brain-specific (Ye and Snyder, 2004), possibly due to the methylation of the CpG islands in the promoter region corresponding to isoform 1, but not isoform 2, that results in a tissue-specific expression profile (*Figure 1.9*).

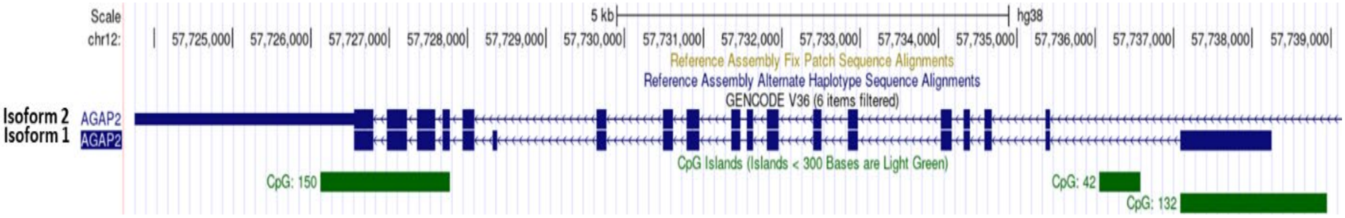


**Figure 1.9: Isoforms for AGAP2.** Alternative splicing produces two isoforms for AGAP2. Isoform 1 (PIKE-L) and isoform 2 (PIKE-A). The first exon for isoform 2 is located upstream of the first exon for isoform 1. The promoter region corresponding to isoform 1 is enriched for CpG island. Figure is downloaded from UCSC genome bowser (http://genome.ucsc.edu)

The isoform 2 (hereafter referred to as *AGAP2*) is ubiquitously expressed with increased expression noted in the brain followed by spleen, thymus and peripheral blood leukocyte (Elkahloun *et al.,* 1997, Nagase *et al.,* 1996). AGAP2 is a potent activator of Akt, a major downstream effector of the PI3K-Akt signalling pathway, and stabilises AKT in active conformation (Ahn *et al.,* 2004b). AGAP2 also interact with insulin receptor tyrosine kinase (IRTK), suppressing AMP-activated protein kinase (AMPK) phosphorylation. This association enhances hepatic IRTK and plays a key role in mediating insulin signal transduction and regulating hepatic glucose homeostasis (Chan *et al.,* 2011). The physiological interactions of AGAP2 are summarised in *Table 3*.

**Table 1.3: AGAP2 interacting proteins and their functions.**

| Interacting Protein | Functions | Reference |
|---|---|---|
| Akt | Enhancing Akt kinase activity | (Ahn *et al.,* 2004b) |
| Insulin receptor | Inhibiting AMPK phosphorylation and enhance hepatic IRTK activity | (Chan *et al.,* 2011) |
| STAT5A | Enhancing STAT5A phosphorylation and mediate prolactin induced mammary gland development | (Chan *et al.,* 2010) |
| AP1 | Regulating the intracellular distribution of AP1 and affects AP1/RAB4 endosomal compartment | (Nie *et al.,* 2005) |
| Fyn | Preventing PIKE-A degradation | (Tang *et al.,* 2007) |
| FAK | Increasing the activity of FAK and results in dissolution of the focal adhesions | (Zhu *et al.,* 2009) |
| β2-AR | Signalling and recycling of β2-ARs | (Wu *et al.,* 2013) |
| Arf1 | Increase GTPase activity of Arf1 | (Nie *et al.,* 2005) |
| AMPK | Regulating AMPK activity | (Chan *et al.,* 2011) |
| UNC5B | Inhibits UNC5B-induced apoptosis | (He *et al.,* 2011) |
| TGFβ1 | Regulate TGFβ1-receptor 2 (TGFR2) trafficking to the membrane | (Navarro-Corcuera *et al.,* 2019) |

AMPK: AMP-activated protein kinase; AP1: Clathrin adaptor protein activator protein 1; β2-AR: 2-adrenoreceptor; FAK: Focal adhesion kinase; RAB4: Ras-related protein 4A; STAT5A: Signal transducer and activator of transcription 5A; TGFβ1: Transforming growth factor beta-1; UNC5B: Unc-5 Netrin Receptor B.

The diverse interactome of AGAP2 has been implicated in cell survival, apoptosis, cell motility, and lipid metabolism (Tse *et al.,* 2013, Ahn *et al.,* 2004a). *AGAP2* is also classified as a protooncogene with increased expression noted in several cancers including brain, breast, ovarian, stomach, lung, kidney, bladder, prostate, uterine, thyroid, testicular, and skin cancer (Cai *et al.,* 2009, Ahn *et al.,* 2004a). Additionally, our group has also recently demonstrated the role of *AGAP2* in hepatic fibrosis (Navarro-Corcuera *et al.,* 2019).

### 1.6.2 Regulation of AGAP2 expression:

*AGAP2* expression is modulated by different mechanisms. Liu *et al.* (2007) have linked *AGAP2* overexpression with the amplification of the *CDK4* amplicon, as the gene is located on chromosome 12 adjacent to the *CDK4* gene. The co-expression of *AGAP2* and *CDK4* has been implicated in glioblastoma progression (Qi *et al.,* 2017). *AGAP2* can be also overexpressed without an alteration in the gene copy number, indicating other possible mechanisms. Recently, our group has characterised the promoter and transcriptional activity of *AGAP2* (Doush *et al.,* 2019). Using the reporter assay, we have identified that the -246/+36 in the *AGAP2* DNA sequence contain the minimal *AGAP2* promoter region. It drives *AGAP2* expression in prostate cancer and chronic myeloid leukaemia cell lines and binds to Specificity protein 1 (SP1) transcription factor, which is required for AGAP2 expression in these cell lines. Furthermore, we have also reported that the -475/-246 fragment in the *AGAP2* DNA sequence contains a DR5 binding site with a functional retinoic acid response element (RARE), inducing AGAP2 promoter activity and expression on treatment with ATRA (all-trans retinoic acid).

Post-translational modification of AGAP2 has been also shown to regulate AGAP2 expression and activity and has been implicated in multiple signalling pathways and disease processes (Reviewed in Navarro-Corcuera *et al.,* 2020). Several serine and tyrosine residues in different domains of AGAP2 protein could be targeted for phosphorylation by multiple kinases and affect AGAP2 activity. Phosphorylation of AGAP2 at Ser-279 by Cyclin-dependent kinase 5 (Cdk5) increase its GTPase activity and has been also shown to further activate AKT kinase activity, leading to increase cell migration and invasion in human glioblastoma (Liu *et al.,* 2008). Likewise, AGAP2 phosphorylation on Ser-351 and Ser-377 by AMP-activated protein kinase (AMPK) stimulate AGAP2 interaction with anchor protein (14-3-3β) and promote its translocation to the nucleus (Zhang *et al.,* 2019). The phosphorylation of different domains of AGAP2 and their impact on signalling pathways, cellular processes, and disease pathogenesis are depicted in *Figure 1.10*.

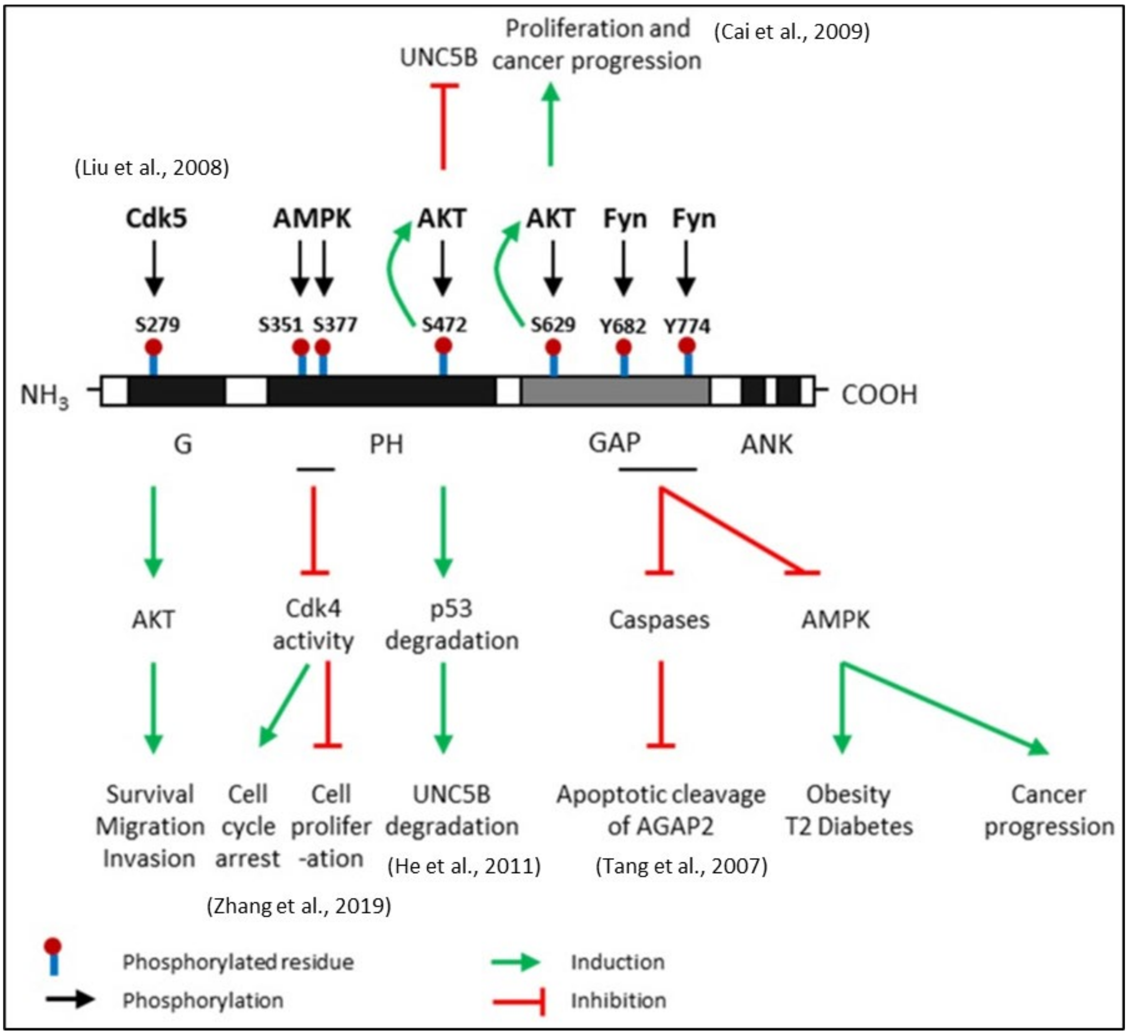


**Figure 1.10: Post-translational modifications of AGAP2 and their impact:** Effect of phosphorylation of serine and tyrosine residues in different domains of AGAP2 and their impact on AGAP2 protein activity. The consequences of phosphorylation on different signalling pathways and disease processes are highlighted by the green (induction) and red (inhibition) arrows. The different domains of AGAP2 depicted include GTPase domain (G domain), pleckstrin homology (PH) domain, GTPase-activating proteins (GAP domain), and ankyrin (ANK) repeats. AMPK: AMP-activated protein kinase; Cdk5: Cyclin-dependent kinase 5; UNC5B: Uncoordinated-5 netrin receptor B. Figure adapted and modified from a previously published paper by our group Navarro-Corcuera *et al.* (2020) (CC BY).

### 1.6.3 Correlation between *AGAP2* mRNA and protein levels:

Preliminary work in our lab has revealed a mismatch between *AGAP2* mRNA and protein levels in a subset of Chronic Myeloid Leukaemia (CML) and Prostate cancer (PC) cell lines (section 3.1.1). The observed discrepancy in the mRNA and protein level has captured our attention and persuaded us to explore the gene regulatory mechanisms responsible for mediating such inconsistency. Our preliminary experiments have also displayed differential usage of *AGAP2* TSSs in PC and CML cell lines with a G4 consensus sequence between the two TSSs (section 3.1.1). Provided the relevance of 5' UTR G4 in regulating gene expression, we aimed to explore the contribution of 5' UTR G4 forming sequence, that is incorporated by the selection of alternate TSS, in regulating *AGAP2* expression in PC and CML cell lines. It would constitute an alternative mechanism of gene regulation that would involve regulation of the gene output by the selection of TSS-isoforms harbouring elements that impact mRNA translation output.

## 1.7 Aims of the thesis

The general aim of this research is to examine the discrepancy in *AGAP2* mRNA and protein levels found in PC vs. CML cell lines. We planned to achieve this aim through the following:

- The characterisation of the TSSs within the *AGAP2* core promoter region of PC and CML cell lines.
- The study of *AGAP2* mRNA 5’ UTR and its influence on protein expression, with an interest in the potential role of G4.

# Chapter 2:

# Methods and Materials

## 2.1 Materials

The Key resource table below details all the reagents and resources used in the current study:

**Table 2.1: Key Resource Table.**

| **Reagent or Resource** | **Source** | **Identifier** |
|---|---|---|
| **Chemicals and Reagents** | | |
| Potassium chloride | Sigma-Aldrich | Cat#P9333;<br>CAS:7447-40-7 |
| HEPES | Sigma-Aldrich | Cat#H3375;<br>CAS:7365-45-9 |
| NP-40 | Sigma-Aldrich | Cat#I3021;<br>CAS:9002-93-1 |
| Digitonin | Abcam | Cat#ab141501;<br>CAS:11024-24-1 |
| Absolute Ethanol for molecular biology | Fischer Scientific | Cat#10644795 |
| 2-Propanol for molecular biology | Sigma-Aldrich | Cat#278475 |
| TWEEN 20 | Sigma-Aldrich | Cat#P1379 |
| Nuclease free water | Promega | Cat#P1193 |
| MG132, proteasome inhibitor | Sigma-Aldrich | Cat#474790;<br>CAS:133407-82-6 |
| Bortezomib, proteasome inhibitor | Santa Cruz | Cat#sc-217785;<br>CAS:179324-69-7 |
| Sodium Deoxycholate | Sigma Aldrich | Cat#30970-25G;<br>CAS: 302-95-4 |
| 30% Hydrogen peroxide solution | Sigma Aldrich | Cat#H1009;<br>CAS: 7722-84-1 |
| Cyclohexamide | Santa Cruz | Cat#sc-3508;<br>CAS:66-81-9 |
| DNase I (RNase-free) | ThermoFisher | Cat#AM2222 |
| Complete EDTA-free protease inhibitor cocktail | Roche | Cat#5056489001 |
| SureBeads Protein G | Biorad | Cat#161-4023 |
| Ampicillin sodium salt | Sigma-Aldrich | Cat#A9518 |
| Glycogen | ThermoFisher | Cat#AM9510 |

| | | |
|---|---|---|
| Deoxyribonucleotide triphosphate: dATP<br>dGTP<br>dCTP<br>dTTP | Promega | Cat#U1205<br>Cat#U1215<br>Cat#U1225<br>Cat#U1235 |
| NheI restriction endonuclease | Promega | Cat#R6501 |
| XhoI restriction endonuclease | Promega | Cat#R6161 |
| Taq DNA polymerase | Promega | Cat#M7841 |
| LB Broth powder | Sigma-Aldrich | Cat#L3522 |
| Alkaline calf intestinal phosphatase | Promega | Cat#M1821 |
| T4 DNA Ligase | Promega | Cat#M1801 |
| TRIzol Reagent | ThermoFisher | Cat#15596026 |
| RNasin Ribonuclease Inhibitor | Promega | Cat#N2511 |
| LiCl Precipitation Solution | ThermoFisher | Cat#AM9480<br>CAS:7447-41-8 |
| Chloroform | Sigma-Aldrich | Cat#C2432<br>CAS:67-66-3 |
| Precision Plus Protein™ Dual Colour Standards | Biorad | Cat#1610374 |
| 30% Acrylamide | Severn Biotech | Cat#20-2100-10<br>CAS:79-06-1 |
| 1kb DNA Ladder | Promega | Cat#G5711 |
| 100bp DNA Ladder | Promega | Cat#G2101 |
| DNA Gel Loading Dye (6X) | ThermoFisher | Cat# R0611 |
| Precision Plus Protein™ Dual Color Standards | Biorad | Cat#1610374 |
| Nitrocellulose membrane | GE Healthcare | Cat#10600006 |
| SYBR Safe DNA Gel Stain | ThermoFisher | Cat#S33102 |
| RPMI 1640 cell culture Media | Gibco | Cat#52400025 |
| Dulbecco's Modified Eagle Medium with GlutaMAX | Gibco | Cat#10566016 |
| Dulbecco's Modified Eagle Medium/Nutrient Mixture F-12 | Gibco | Cat#11320033 |
| Iscove's Modified Dulbecco's Medium | Gibco | Cat#12440053 |
| Opti-MEM Reduced-Serum Medium | Gibco | Cat#31985062 |
| Fetal Bovine Serum | Biosera | Cat#FB1090/500 |
| Human Recombinant Insulin | Sigma Aldrich | Cat#91077C |
| Phosphate Buffered Saline, pH 7.4, without Calcium and Magnesium | Lonza | Cat#BE17-516F |

| | | |
|---|---|---|
| Dimethyl Sulfoxide (DMSO) | Sigma Aldrich | Cat#D2650<br>CAS:67-68-5 |
| 2-mercaptoethanol | Sigma Aldrich | Cat#M6250<br>CAS:60-24-2 |
| L-Glutamine (200 mM) | ThermoFisher | Cat#25030149 |
| Trypsin-EDTA (0.25%) | ThermoFisher | Cat#25200056 |
| Trypan blue | Sigma Aldrich | Cat#T8154 |
| Flasks T25, T75, T125 | Sarstedt | Cat#83.3910<br>Cat#83.3911<br>Cat#83.3912 |
| Serological pipettes 5, 10, 25 mL | Sarstedt | Cat#86.1253.025<br>Cat#86.1254.001<br>Cat#86.1685.001 |
| 96 well, 24 well, and 6 well plates | Sarstedt | Cat#82.1582.001<br>Cat#83.3922.005<br>Cat#83.3920.005 |
| 2 mL Cryovials | Sarstedt | Cat#72.379 |
| Mr. Frosty™ Freezing Container | ThermoFisher | Cat#5100-0001 |
| Amersham Protran 0.2 µm nitrocellulose membrane | GE Healthcare | Cat#GE10600001 |
| 4–20% Mini-PROTEAN Precast Protein Gels | Biorad | Cat#4561094 |
| **Critical Commercial Assays** | | |
| Dual-Luciferase Reporter Assay System | Promega | Cat#E1910 |
| ReliaPrep RNA Miniprep Systems | Promega | Cat#Z6011 |
| NucleoSpin Plasmid Columns | Fischer Scientific | Cat#11932392 |
| 5' RACE System for Rapid Amplification of cDNA Ends, version 2.0 | ThermoFisher | Cat#18374058 |
| GeneRace Kit with SuperScript III RT and TOPO TA Cloning for 5’ RLM-RACE | ThermoFisher | Cat#L150201 |
| Amaxa Cell Line Nucleofector Kit V | Lonza | Cat#VCA-1003 |
| Pierce™ BCA Protein Assay Kit | ThermoFisher | Cat#23227 |
| M-MLV Reverse Transcriptase | Promega | Cat#M1701 |
| TOPO TA Cloning Kit | ThermoFisher | Cat#K4575J10 |
| GoTaq® qPCR SYBR master mix | Promega | Cat#A6001 |
| mMESSAGE mMACHINE T7 Transcription Kit | ThermoFisher | Cat#AM1344 |
| ECL Western Blotting Substrate | Promega | Cat#W1001 |
| Flexi Rabbit Reticulocyte Lysate System | Promega | Cat#L4540 |

|  |  |  |
|---|---|---|
| Wizard SV Gel and PCR Clean-Up System | Promega | Cat#A9281 |
| jetPRIME DNA/siRNA transfection reagent | Polyplus | Cat#114-01 |
| DNeasy Blood & Tissue Kit | Qiagen | Cat#69504 |
| EZ-PCR Mycoplasma Test Kit | Biological Industries | Cat#20-700-20 |
| **Equipment** |  |  |
| Countess 3 Automated Cell Counter | Fischer Scientific | Cat#15397802 |
| Tube Rotator | Grant Bio | Cat#PTR-35 |
| Bio-Rad Mini-Protean II system | Biorad | Cat#1653308 |
| Rotor-Gene Q 2plex Platform | QIAGEN | Cat#9001550 |
| Bio-Rad T100 thermocycler | Biorad | Cat#186-1096 |
| BMGLabtech CLARIOSTAR plate reader | BMG labtech | Cat#430-501S-F |
| Nanodrop 8000 spectrophotometer | ThermoFisher | Cat#ND-8000-GL |
| Bioanalyzer System | Agilent | CAT#G2939BA |
| ImageQuantTM LAS 4000 | GE healthcare |  |
| Class II (laminar flow) biological safety cabinet | Walker Safety Cabinets |  |
| Syngene™ G:BOX Chemi XX9 | Syngene | Cat#DRXX6/1100 |
| Soniprep 150 plus (MSE) | Measuring and Scientific Equipment (MSE) | Cat#MSS150 |
| Nucleofector 2b Device | Lonza | Cat#AAB-1001 |
| iMark Microplate Absorbance Reader | Biorad | Cat#1681135 |
| Refrigerated Centrifuge | Eppendorf | Cat#5418-R |
| Mini Dry Bath | AccuBlock | Cat#D0100 |
| **Bacterial Strains** |  |  |
| DH5α | Thermo-Fisher | Cat#18265017 |
| One Shot TOP10 Chemically Competent E. coli | Thermo-Fisher | Cat#C404010 |
| **Recombinant DNA** |  |  |
| pcDNA3 RLUC POLIRES FLUC | Addgene,<br>(Poulin *et al.,* 1998) | Cat#45642;<br>RRID:Addgene_45642 |
| **Software and Algorithms** |  |  |
| GraphPad Prism 8 | GraphPad Software, Inc. | https://www.graphpad.com/scientific-software/prism/ |
| Image Studio™ Lite | Li-COR | https://www.licor.com/bio/image-studio-lite/ |

| | | |
|---|---|---|
| MetaCore Pathway Analysis | Clarivate Analytics | https://portal.genego.com/ |
| BaseSpace Sequence Hub | Illumina | https://basespace.illumina.com/ |
| The Integrative Genomics Viewer (IGV) | (Thorvaldsdóttir *et al.,* 2013) | http://software.broadinstitute.org/software/igv/ |
| BEDTOOLS v2.28 | (Quinlan and Hall, 2010). | https://bedtools.readthedocs.io/en/latest/index.html |
| Rstudio | Rstudio team | https://www.rstudio.com/ |
| DESeq2 | (Love *et al.,* 2014) | http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html. |
| pqsfinder | (Hon *et al.*, 2017) | http://bioconductor.org/packages/release/bioc/html/pqsfinder.html |
| Python programming language | Version 3.6.8 | https://www.python.org/ |
| Multiple sequence alignment | (Corpet, 1988) | http://multalin.toulouse.inra.fr/multalin/ |
| InteractiVenn for Venn diagram | (Heberle *et al.*, 2015) | http://www.interactivenn.net/ |

## 2.2 Methods

### 2.2.1 Cell Culture:

The following cell lines were used in the study. The source, identifier, and growth conditions are detailed below in *Table 2.2*. (**Note:** No antibiotics were added to the growth medium).

**Table 2.2: List of cell lines, their identifiers and culture conditions used in the study.**

| Cell line | Source/identifier | Growth medium |
|---|---|---|
| KU812<br>(Human chronic myelogenous leukaemia) | ATCC<br>(RRID:CVCL_0379)<br>Kindly Provided by Dr Felipe Prosper | RPMI supplemented with 2 mM L-Glutamine and 10% FBS |
| TCCS<br>(Human myelogenous leukaemia) | (Van *et al.*, 2005)<br>Kindly Provided by Dr Felipe Prosper | RPMI supplemented with 2 mM L-Glutamine and 10% FBS |
| KCL-22<br>(Human myelogenous leukaemia) | ATCC<br>Kindly Provided by the John van Geest Cancer Research Centre | RPMI supplemented with 2 mM L-Glutamine and 10% FBS |
| DU145<br>(Human prostate cancer) | ATCC<br>(RRID:CVCL_0105)<br>Kindly Provided by the John van Geest Cancer Research Centre | DMEM GlutaMAX supplemented with 10% FBS |
| PC3<br>(Human prostate adenocarcinoma) | ATCC<br>(RRID:CVCL_0035)<br>Kindly Provided by the John van Geest Cancer Research Centre | DMEM/F12 containing 2 mM L-Glutamine and 10% FBS |
| LNCaP<br>(Human prostate cancer) | ATCC<br>Kindly Provided by the John van Geest Cancer Research Centre | RPMI supplemented with 2 mM L-Glutamine and 10% FBS |

| | | |
|---|---|---|
| HepG2<br>(Hepatocellular carcinoma) | ATCC<br>(RRID:CVCL_0027)<br>Kindly Provided by Dr Maria Hatziapostolou | DMEM GlutaMAX supplemented with 10% FBS |
| HuH7<br>(Human liver cancer) | JCRB<br>(RRID:CVCL_0336) | DMEM GlutaMAX supplemented with 10% FBS |
| MCF-7<br>(Human breast adenocarcinoma) | ATCC<br>(RRID:CVCL_0031)<br>Kindly Provided by the John van Geest Cancer Research Centre | DMEM GlutaMAX supplemented with 10% FBS and 0.01 mg/mL human recombinant insulin |
| PA-1<br>(Human ovary teratocarcinoma) | ATCC<br>(RRID:CVCL_0479)<br>Kindly Provided by the John van Geest Cancer Research Centre | DMEM GlutaMAX supplemented with 10% FBS |
| SK-OV-3<br>(Human ovary adenocarcinoma) | ATCC<br>(RRID:CVCL_0532)<br>Kindly Provided by the John van Geest Cancer Research Centre | McCoy's 5a medium supplemented with 2 mM L-Glutamine and 10% FBS |
| U-2 OS<br>(Human osteosarcoma) | ATCC<br>(RRID:CVCL_0042) | DMEM GlutaMAX supplemented with 10% FBS |
| RAJI<br>(Human Burkitt's lymphoma) | ATCC<br>(RRID:CVCL_0511) | RPMI supplemented with 2 mM L-Glutamine and 10% FBS |
| KG1<br>(Human acute myelogenous leukaemia) | ATCC<br>(RRID:CVCL_0374)<br>Kindly Provided by the John van Geest Cancer Research Centre | Iscove's Modified Dulbecco's Medium, 2mM Glutamine, and 20% FBS |
| Kasumi-1<br>(Human acute myeloblastic leukaemia) | ATCC(RRID:CVCL_0589)<br>Kindly Provided by the John van Geest Cancer Research Centre | RPMI supplemented with 2 mM L-Glutamine and 10% FBS |

#### 2.2.1.1 Maintenance and passaging of human cell lines:

The cells were maintained in the growth medium as outlined in *Table 2.2*. Upon reaching 80-90% confluence, cells were passaged. In the case of adherent cell lines, the old medium was removed followed by washing the culture with phosphate buffered saline (Lonza) to remove traces of serum and dead cells. The cells were incubated with Trypsin-EDTA (1 mL per 50 $cm^2$ of surface area) and incubated in a 5% $CO_2$ incubator at 37°C for 5 minutes. After incubation, when the cells became rounded and detached, 5 volume of fresh complete medium was added to inactivate Trypsin-EDTA. The cells were then centrifuged at 300 x g for 5 minutes at room temperature and resuspended in a complete growth medium and split 1:5 into a new cell culture flask. In the case of suspension cell lines, the cells in media were directly collected for centrifugation without washing and trypsinisation steps. The cells were centrifuged at 200 x g for 5 minutes. The cell pellet was resuspended in a fresh medium and seeded at a density of 1 x $10^6$ cells/mL in a new cell culture flask.

#### 2.2.1.2 Cell counting:

The cells were counted using the automated cell counter (Fischer Scientific). After passaging the cells, as described previously, the cells were resuspended in 3 mL of fresh medium. A volume of 10 µL of cell suspension is mixed with 10 µL of Trypan blue and pipetted into a disposable counting chamber slide (ThermoFisher) and loaded onto the cell counter. The live cell count/mL was used to determine the volume of cell suspension needed to achieve the required cell density in the growth medium.

#### 2.2.1.3 Cryopreservation and thawing:

The cells were kept in a culture for a maximum of 15 passages to prevent phenotypic and global gene expression changes associated with long-term serial passaging (Mouriaux *et al.,* 2016). The frozen stocks of low passage cells were prepared by collecting and counting 1 x $10^6$ cells per mL, as described above, and resuspending the pellet in a complete growth medium (*Table 2.2*) with 5% DMSO (Sigma Aldrich). The cells were transferred into 2 mL cryovials, appropriately labelled, and cooled at a steady rate of -1°C/minute in a freezing container (ThermoFisher) at -80°C for 24 hours. The cells were then transferred to liquid nitrogen storage.

To revive the cells from the frozen stock, the cryovials were thawed at 37°C for 1-2 minutes and added to the prewarmed medium. The cells were centrifuged to remove DMSO and reconstituted in a fresh complete medium (*Table 2.2*) and transferred to a T25 flask. After 24 hours, the cells were examined under the microscope to monitor cell health and were checked for mycoplasma contamination (see below 2.1.4).

#### 2.2.1.4 Mycoplasma testing:

All cell lines used in the study were regularly checked for mycoplasma contamination, particularly after thawing, using the EZ-PCR mycoplasma test kit (Biological Industries) by following the manufacturer's protocol. Briefly, after reviving and reaching 70-80% confluence, 1 mL of supernatant from cell culture medium was centrifuged at 250 x g for 1 minute to pellet cell debris and then centrifuged at a higher speed (16,000 x g for 10 minutes) to sediment mycoplasma. The pellet (not easily visible) was resuspended in the supplied buffer solution and heated at 95°C for

3 minutes. The sample was then amplified by PCR using the provided primers to detect the 16S mycoplasma specific rRNA region. The reaction was carried out alongside the positive template control to ensure that the PCR is working correctly. The PCR products were resolved on a 2% agarose gel [2g (w/v) agarose in TAE buffer (See below 2.2.3)] and visualised using G:BOX Chemi XX9 gel imaging system (Syngene). A representative gel showing mycoplasma testing for the main cell lines used in the study is depicted below (*Figure 2.1*). As shown in *the* figure, TCCS culture (lane 5) was contaminated with mycoplasma. The cell culture flasks were appropriately discarded, and a new frozen stock of TCCS was thawed and rechecked for mycoplasma contamination.

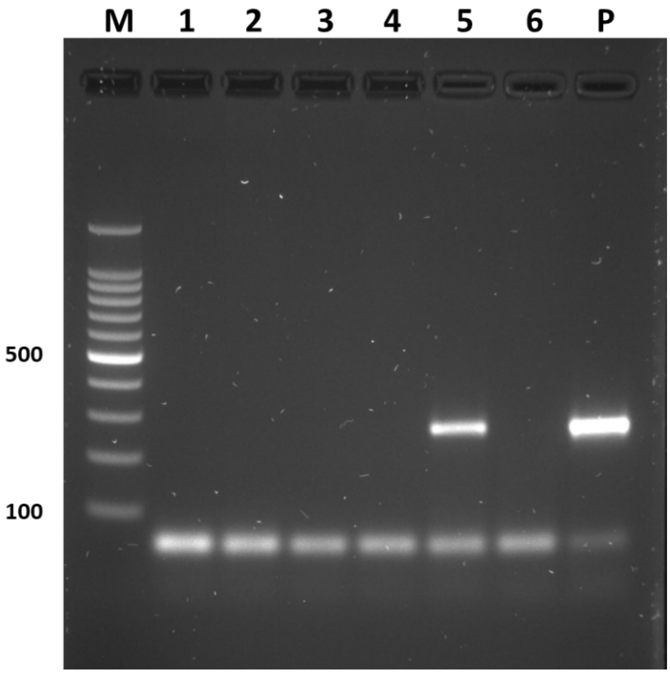


**Figure 2.1: Mycoplasma testing.** Agarose gel electrophoresis of PCR products was to detect mycoplasma contamination. Supernatant media was taken from the cell culture after reaching 70-80% confluence and centrifuged at high speed to sediment mycoplasma and amplified by PCR using supplied primers. A band at 270 bp is consistent with mycoplasma contamination. Lane M: 100 bp DNA ladder; Lane 1-6: DU145, PC3, LNCaP, KU812, TCCS (mycoplasma positive), KCL-22; Lane P: positive template control.

### 2.2.2 Gene expression analysis:

#### 2.2.2.1 RNA extraction:

The RNA was extracted using the ReliaPrep RNA Miniprep Systems (Promega) according to the manufacturer's instruction. Briefly, the cell pellet was lysed in the supplied BL buffer with 1-thioglycerol added. The lysate was then passed through a 20-gauge needle (3-4 times) to shear the genomic DNA. 100% isopropanol (35 µL per 100 µL of the BL buffer) was added to the lysate and the mix was carefully transferred to a ReliaPrep minicolumn and centrifuged at 14,000 x g for 1 minute. The sample flowthrough was discarded, and the columns were washed with the provided RNA wash solution. The on-column DNase treatment was then performed for 15 minutes with the DNase I enzyme provided with the kit. The columns were subsequently washed with column wash solution and RNA wash solution to remove DNase I and transferred to a fresh 1.5 mL collection tube. Finally, 20 µL of nuclease-free water was added to the column and centrifuged at 14,000 x g for 1 minute to elute the RNA. The RNA concentration was quantified using Nanodrop 8000 spectrophotometer (ThermoFisher) and samples were either used immediately for cDNA synthesis or stored at -80°C until further use.

#### 2.2.2.2 cDNA synthesis:

The RNA was reverse transcribed using M-MLV Reverse Transcriptase (Promega). 2 µg of RNA was added to a tube containing 1 µg of random primers (Promega) and nuclease-free water up to a volume of 15 µL. The mixture was heated at 65°C for 5 minutes to melt the secondary structures in the RNA. After incubation, the mixture was immediately cooled on ice for 2 minutes to prevent the reforming of secondary structures. The following components were then added to the mixture:

**Table 2.3: Component compositions for cDNA synthesis from RNA templates using M-MLV Reverse Transcriptase.**

| Reagents | Final Concentration | Volume (µL) |
|---|---|---|
| MMLV 5X Reaction Buffer | 1X | 5 |
| dATP (10 mM) | 500 µM | 1.25 |
| dGTP (10 mM) | 500 µM | 1.25 |
| dCTP (10 mM) | 500 µM | 1.25 |
| dTTP (10 mM) | 500 µM | 1.25 |
| Recombinant RNasin Ribonuclease Inhibitor | 25 units | 0.625 |
| M-MLV Reverse Transcriptase | 200 units | 1 |
| Nuclease free water to a final volume | - | 25 µL |

The complete reaction mix was incubated in a thermocycler (Biorad) using the following settings: Priming at 25°C for 10 minutes, reverse transcription at 37°C for 60 minutes, and reverse transcriptase inactivation at 85°C for 5 minutes. The newly synthesized cDNA was stored at -20°C until required.

#### 2.2.2.3 Quantitative reverse transcription PCR (qRT-PCR)

The quantitative reverse transcription PCR (qRT-PCR) was performed using SYBR-green fluorescence in a Rotor-Gene Q real-time PCR cycler (Qiagen). The list of primers and their concentration and annealing temperature are presented in *Table 2.4*. The PCR reaction was carried out in a total volume of 10 µL containing GoTaq qPCR SYBR master mix (Promega), forward and reverse primers, and cDNA (1-100 ng). The thermal cycler reaction conditions were as follow: initial denaturation for 5 minutes at 95°C; followed by 40 cycles, consisting of 30 seconds denaturation at 95$^{o}$C, 30 seconds primer annealing at described temperature (*Table 2.4*), and 30 seconds template extension at 72$^{o}$C. The melt-curves analysis was performed for each run to verify the specificity of amplicons. All reactions were performed in triplicate and the expression levels of the target transcripts were normalised using a housekeeping gene (*HPRT*). Data was analysed using the comparative Ct method ($2^{-\Delta\Delta Ct}$) (Livak and Schmittgen, 2001).

**Table 2.4: Primer sequences and related information for quantitative reverse transcription PCR (qRT-PCR)**

| Primer name | NCBI accession number | Sequence (5'→3') | Primer concentration (nM) | Annealing temperature |
|---|---|---|---|---|
| Outer long AGAP2 Forward | XM_005268626.2 | GACAGACGGAAGGGCGG | 500 | 60°C |
| Outer long AGAP2 Reverse | | ACAACGAACTGCCTCTGGGC | 500 | 60°C |
| Inner long AGAP2 Forward | XM_005268626.2 | GCAGGGGCGGGGAGTTCT | 100 | 63°C |
| Inner long AGAP2 Reverse | | CTTGCCAGGCTAACAACCAC | 100 | 63°C |
| Outer short AGAP2 Forward | NM_014770.4 | TCTGAGGTTTGGGGGCTGTA | 500 | 60°C |
| Outer short AGAP2 Reverse | | CAGGCGCAGTTCAGGAATGG | 500 | 60°C |
| AGAP2 Forward | NM_014770.4 | CCAGAGGTGGTTGTTAGCCTG | 500 | 65°C |
| AGAP2 Reverse | | GCGGCTCAAAGTCCATTCCT | 500 | 65°C |
| Long HK1 Forward | NM_033496.3 | AGGTTGCATGAGGGGTTGG | 250 | 60°C |
| Long HK1 Reverse | | TTTTGAGCCAGGACTCCAGC | 250 | 60°C |
| Short HK1 Forward | NM_033496.3 | TACCACAACCTGACACTGGG | 250 | 60°C |
| Short HK1 Reverse | | CACCTCGACAGGGCAAACTC | 250 | 60°C |
| *Renilla* Luciferase Forward | - | ATAACTGGTCCGCAGTGGTG | 300 | 63°C |
| *Renilla* Luciferase Reverse | | TAAGAAGAGGCCGCGTTACC | 300 | 63°C |
| NRAS Forward | NM_002524.5 | CAGAGGCAGTGGAGCTTGA | 500 | 65°C |
| NRAS Reverse | | GCTTTTCCCAACACCACCT | 100 | 65°C |
| MM16 Forward | NM_005941.5 | GCTCGTCCATCCATTGAAGC | 500 | 60°C |
| MM16 Reverse | | TGCACGAAATCCAACCGTCT | 100 | 60°C |
| PON2 Forward | NM_001018161.1<br>NM_000305.3 | GACTCCACAGCTTTGCACC | 500 | 60°C |
| PON2 Reverse | | GGCCAAATCAAACCCACGAC | 250 | 60°C |
| CKDN2A Forward | NM_001195132.2 | CCCTTTGGTTATCGCAAGCTG | 500 | 60°C |
| CKDN2A Reverse | | CCCTGTAGGACCTTCGGTGA | 250 | 60°C |
| HPRT Forward | NM_000194.3 | ATGCTGAGGATTTGGAAAGG | 500 | 60°C |
| HPRT Reverse | | AATCCAGCAGGTCAGCAAAG | 500 | 60°C |
| TBP Forward | NM_001172085.2<br>NM_003194.5 | TTCGGAGAGTTCTGGGATTG | 500 | 65°C |
| TBP Reverse | | GGATTATATTCGGCGTTTCG | 500 | 65°C |
| AGAP2 Genomic Forward | - | TTAGGATTGCACCTCGGACC | 500 | 60°C |

| | | | | |
|---|---|---|---|---|
| AGAP2 Genomic (-425) Forward: | NC_000012.12 | GTGTAGAGAGGGCAATGGGTAC | 500 | 60°C |
| AGAP2 Genomic (-218) Reverse: | | CAAGCTAGGTCCGAGGTGC | 500 | 60°C |

### 2.2.2.4 Primer optimisation and efficiency testing:

The optimal concentration of a primer pair was determined using a primer matrix, keeping the annealing temperature at 60°C (*Figure 2.2A*). The lowest concentration of primer pair yielding the best Ct value (lower number of cycles to reach the detection threshold) was selected. The selected primer pair was checked for the formation of primer dimer using a no template control (NTC). When amplification was detected in NTC, the primer pair was further optimised by selecting a higher annealing temperature. The specificity of amplification was determined by a melt curve analysis and checking the size of the amplicon using agarose gel electrophoresis (see section 2.2.3) and Sanger sequencing of the product. The efficiency of each primer pair was determined using a ten-fold dilution series (*Figure 2.2B*). A standard curve analysis was performed to determine the reaction efficiency with a goal of 90-110% reaction efficiency and $R^2$ values >0.98.

As a representative example, the optimisation of primer pair amplifying *HK1* (NM_033496.3) is shown in *Figure 2.2*. The optimal primer concentration was determined using a primer matrix (*Figure 2.2A*). The primer efficiency was subsequently determined by plotting the standard curve of the dilution series and reaction efficiency was found to be 93% and $R^2$ value of 0.99.

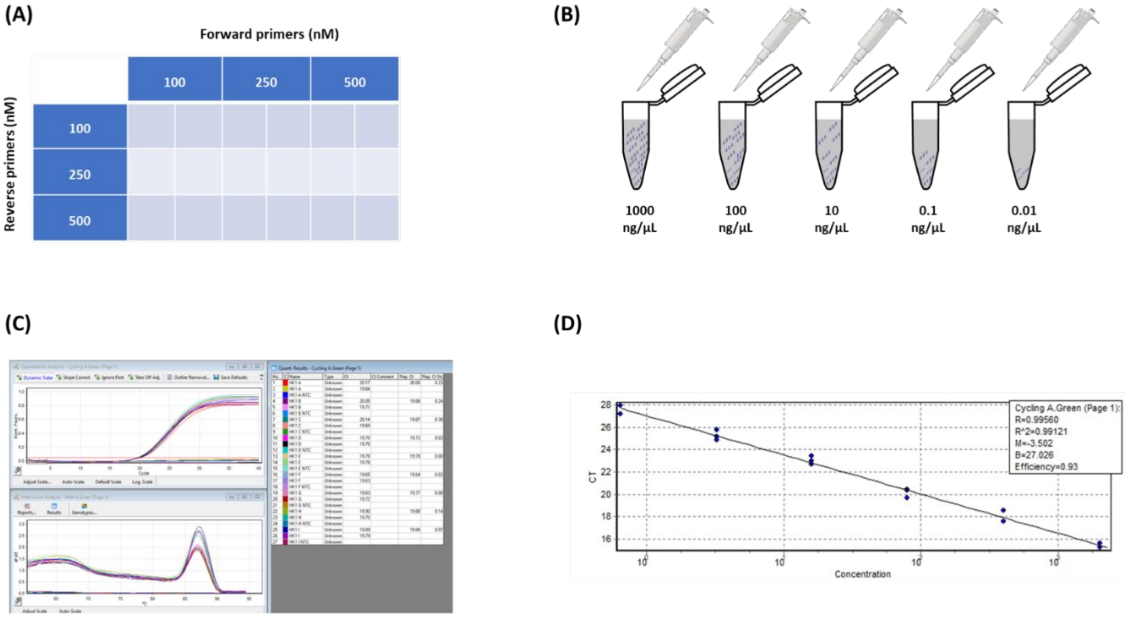


**Figure 2.2: Schematics and representative example for primer optimisation and efficiency testing. (A)** Primer optimisation matrix. **(B)** Schematics of 10-fold dilution series to plot standard curve for primer efficiency testing. **(C)** Optimisation of primer pair to amplify *HK1* using primer matrix. **(D)** Representative standard curve to determine primer efficiency for the primer pair to amplify *HK1*.

## 2.2.3 Agarose gel electrophoresis:

Agarose gel electrophoresis was used in the project for the following applications: resolving the DNA fragments following PCR amplification, 5’ Rapid amplification of cDNA ends (RACE), and restriction endonuclease digestion. The required percentage of gel (depending on the size of the amplicon) was cast in Tris-acetate-EDTA (TAE) buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA) supplemented with SYBR safe DNA gel stain (1:10,000). The sample was mixed with DNA gel loading dye (1:6) (ThermoFisher) before loading onto the gel. An appropriate size marker (1Kb or 100bp) was also used alongside the sample. Electrophoresis was carried out at 100V for 1 hour and visualised using G:BOX Chemi XX9 gel imaging system (Syngene).

## 2.2.4 SDS-PAGE and Western blotting:

### 2.2.4.1 Preparation of cell lysate:

The cells were lysed in radio-immunoprecipitation assay (RIPA) buffer [50 mM (v/v) Tris-Cl (pH 8.0), 1% (v/v) NP-40, 1% (v/v) sodium deoxycholate, 0.1% (v/v) SDS, 150 mM (v/v) NaCl], containing protease inhibitors (Roche). The lysate was then incubated on ice for 30 minutes and sonicated with ice-cooling for 3 × 5 sec pulses at a frequency of 5 microns using a Soniprep 150 plus (MSE) followed by centrifugation at 13,000 x g for 10 min at 4°C. The lysates were used immediately for protein estimation or stored at -20°C until further use.

### 2.2.4.2 Protein quantification:

The concentration of protein in the lysate was determined using the BCA Protein Assay Kit (ThermoFisher) based on the bicinchoninic acid assay by Smith *et al.* (1985). Bovine serum albumin (BSA) was used to produce a standard curve and the protein concentration was determined using a straight-line equation. Protein standards were prepared using dilutions of BSA as detailed in *Table 2.5*. 20 µL of standard and samples (1:10 diluted) were added to a well in a 96-well plate (Sarstedt) in triplicate and 160 µL of BCA reagent (1:50, reagent A and B) was then added to each well. The plate was then incubated at 37°C for 30 mins and absorbance was measured at 562 nm. The standard curve having an $R^2$ value (linearity) > 0.98 were used for protein quantitation.

**Table 2.5: BSA standard curve.**

| Protein concentration (mg/mL) | 2mg/mL BSA (µl) | RIPA buffer (µL) |
|---|---|---|
| 0 | 0 | 20 |
| 0.2 | 2 | 18 |
| 0.4 | 4 | 16 |
| 0.6 | 6 | 14 |
| 0.8 | 8 | 12 |
| 1.2 | 12 | 8 |

| | | |
|---|---|---|
| 1.6 | 16 | 4 |
| 2.0 | 20 | 0 |

### 2.2.4.3 SDS-PAGE (Sodium dodecyl sulphate-polyacrylamide gel electrophoresis):

The SDS-PAGE was performed in reducing conditions. Typically, 50 µg of protein were mixed with 4X Laemmli buffer [2% SDS, 10% (v/v) glycerol, 50 mM Tris-HCl (pH 6.8), bromophenol blue 0.02% (w/v), 1% β-mercaptoethanol (v/v)] and incubated at 95°C for 5 minutes. Total protein was separated by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) consisting of 10% resolving gel [9.6 mL of dH2O, 6 mL of resolving buffer (1.5 M Tris base, 0.4% w/v SDS, pH 8.4), 8 mL of 30% (w/v) acrylamide/bis solution, 8 µL of TEMED, and 90 µL of freshly prepared 10% (w/v) ammonium persulphate] and 5% stacking gel [2.1 mL of dH2O, 0.5 mL of stacking buffer (0.5 M Tris base, 0.4% w/v SDS, pH 6.8), 0.38 mL of 30% (w/v) acrylamide/bis solution, 30 µL 10% SDS, 3 µL of TEMED, and 30 µL of freshly prepared 10% (w/v) ammonium persulphate]. Electrophoresis of 25 µL sample/well was carried out at 100 V for 2 hours in running buffer (3 g Tris base, 14.4 g Glycine, 1 g SDS). Precision Plus Protein Dual Colour Standards (BioRad) was used as a molecular weight marker.

### 2.2.4.4 Immunoprobing:

Proteins were subsequently transferred to Amersham Protran 0.2 µm nitrocellulose membrane (GE Healthcare) at 100 V for 90 mins in transfer buffer (3 g Tris base, 14.4g Glycine, 200 mL methanol). The membrane was then blocked with 5% (w/v) skimmed milk powder in TBST (2.42 g Tris base, 8.78 g NaCl, 300 µL Tween20) for 1 hour at room temperature with agitation. Subsequently, the membrane was incubated with primary antibody (*Table 2.4*) diluted in 5% milk in TBST overnight at 4°C with agitation. After this, the membrane was washed with TBST three times for 10 min at room temperature followed by incubation with the appropriate secondary antibody (*Table 2.4*). Next, the membrane was washed as previously described and signals were detected by incubating the membrane with 1 mL Clarity ECL Western Blot Substrate (BioRad). Exposure and imaging were carried out using a luminescent image analyser LAS-4000 (GE healthcare).

**Table 2.6: Antibodies used for Western blotting.**

| Antibody | Type | Specificity | Host | Concentration | Supplier |
|---|---|---|---|---|---|
| Anti-AGAP2 | Polyclonal | Detect amino acid sequence CQASLDSIREAVINSQ specific to PIKE-A/isoform 2 | Goat | 1:1000 | Sigma-Aldrich Cat#SAB2501250; RRID:AB_10620617 |
| Anti-HK1 | Monoclonal | Detect amino acids 316-410 of HK1 of human origin | Mouse | 1:100 | Santa Cruz Biotechnology |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Cat#sc-46695; RRID:AB_627721 |
| Anti-DNA/RNA G-quadruplex (clone BG4) | Monoclonal | DNA/RNA G-quadruplex | Mouse | 3 µg (GRIP) | Absolute Antibody Cat#Ab00174-1.1 |
| IgG Isotype Control antibody | | | Mouse | 3 µg (GRIP) | ThermoFisher Cat# 31903, RRID:AB_10959891 |
| Anti-Ubiquitin | Polyclonal | Detects ubiquitin, polyubiquitin and ubiquitinated proteins | Rabbit | 1:1000 | Cell Signaling Technology Cat#3933; RRID:AB_2180538 |
| Anti-Caveolin 1 | Polyclonal | Detect Peptide with sequence C-DELSEKQVYDAH specific to Caveolin 1 | Goat | 1:1000 | Everest Biotech Cat# EB06817, RRID:AB_2072197 |
| Anti-PON2 (clone D-12) | Monoclonal | Detects amino acids 61-113 mapping to an internal region of PON2 of human origin | Mouse | 1:100 | Santa Cruz Biotechnology Cat# sc-373981, RRID:AB_10917573 |
| Anti-β-Tubulin | Monoclonal | Detects β tubulin, types I, II, III, and IV of bovine, rat, mouse and human | Mouse | 1:1000 | Sigma-Aldrich Cat#T8328; RRID:AB_1844090 |
| Anti-β-Actin | Monoclonal | Detects N-terminal end of the b-isoform of actin | Mouse | 1:5000 | Sigma-Aldrich Cat#A2228; RRID:AB_476697 |
| Anti-eIF4A (clone C32B4) | Monoclonal | Detects residues surrounding Met316 of human eIF4A protein | Rabbit | 1:1000 | Cell Signaling Technology Cat#2013; RID:AB_2097363 |
| Anti-eIF4A1 | Polyclonal | Detects residues surrounding Gly12 of human eIF4A1 | Rabbit | 1:1000 | Cell Signaling Technology Cat#2490; RRID:AB_823487 |
| Anti-eIF4B | Polyclonal | Detects residues at the amino terminus of human eIF4B. | Rabbit | 1:1000 | Cell Signaling Technology |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Cat#3592; RRID:AB_2293388 |
| Anti-eIF4E (clone C46H6) | Monoclonal | eIF4AE protein | Rabbit | 1:1000 | Cell Signaling Technology Cat#2067; RRID:AB_2097675 |
| Anti-eIF4G (clone C45A4) | Monoclonal | Detects residues surrounding Gly188 of human eIF4G. | Rabbit | 1:1000 | Cell Signaling Technology Cat#2469; RRID:AB_2096028 |
| Anti-eIF4H (clone D85F2) | Monoclonal | eIF4AH protein | Rabbit | 1:1000 | Cell Signaling Technology Cat#3469; RRID:AB_2096038 |
| Anti-Rabbit IgG HRP | | Rabbit IgG | Goat | 1:2000 | Cell Signaling Technology Cat#7074; RRID:AB_2099233 |
| Anti-Mouse IgG HRP | | Mouse IgG | Horse | 1:2000 | Cell Signaling Technology Cat#7076; RRID:AB_330924 |
| Anti-Goat IgG HRP (A8919) | | Goat IgG | Rabbit | 1:25000 | Sigma-Aldrich Cat#A4174; RRID:AB_258138 |

### 2.2.4.5 Reprobing and stripping:

In order to detect another protein of different molecular weight, the previously immunoprobed membrane was treated with 30% hydrogen peroxide (Sigma Aldrich) for 15 minutes, followed by re-blocking and re-probing with another antibody (Sennepin *et al.,* 2009). To detect proteins having approximate molecular weight, the membrane was stripped by incubating with 0.5M NaOH for 8 minutes under gentle agitation. It was followed by washing with distilled water and TBST for three times (5 minutes each) and blocking and re-probing the membrane again, as described above.

#### 2.2.4.6 Densitometry analysis

For quantitative analysis of protein expression, the digital images obtained by chemiluminescence imaging systems were analysed using Image Studio Lite (Licor). The background was subtracted using a median intensity of pixel around the band area and the background-subtracted band intensity was normalised for protein loading by the corresponding loading control intensity values (β-Actin/β-tubulin).

### 2.2.5 5’ Rapid amplification of cDNA ends (5’ RACE):

The Rapid Amplification of cDNA Ends (RACE) is a technique for amplifying a nucleic acid sequence of mRNA between a defined internal site and the unknown region at either 5’ or 3’ end of the mRNA (Frohman *et al.,* 1988). The technique was initially employed to obtain full-length cDNA clones for novel transcripts (Han *et al.,* 2001) and has been optimised over the last two decades to facilitate precise identification of transcription start sites (TSS) and locating promoter region (Tillett *et al.,* 2000). The 5’ RACE utilises gene-specific and complementary adapter primers to amplify the 5’ end of the mRNA of interest which is sequenced to determine TSSs or to identify neighbouring promoter elements.

Two variations of 5’ RACE was used during the PhD project namely the homopolymeric tailing and ligation mediated oligo-capping.

#### 2.2.5.1 5’ RACE with homopolymeric tailing:

##### 2.2.5.1.1 Principle (*Figure 2.3A*):

The method has been pioneered by Frohman and colleagues (Frohman, 1993, Frohman *et al.,* 1988). The first strand cDNA is reverse transcribed from the total RNA using a gene-specific primer. The reverse transcriptase with reduced RNase H activity is used to capture the entire 5’ end of the mRNA. The original mRNA template is then removed by the RNase H enzyme which cleaves RNA in RNA:DNA duplex. A deoxyadenosine homopolymeric tail (dA-tail) is then added to the 3' end of the cDNA using terminal deoxynucleotidyl transferase (TdT). It is followed by second strand cDNA synthesis using oligo(dT) primer containing priming site for abridged universal amplification primer (AUAP). Subsequently, the cDNA is amplified using an abridged primer and a second gene-specific primer. If required, nested amplification is performed to detect rare mRNA transcripts. The PCR products are resolved by agarose gel electrophoresis and sequenced using Sanger sequencing (Source Bioscience).

##### 2.2.5.1.2 Methodology:

5' RACE System version 2.0 (ThermoFisher) was used as per the manufacturer’s guideline. Briefly, the first-strand cDNA was synthesised using a gene-specific primer 1 by reverse transcribing 3 µg of RNA with the supplied SuperScript II Reverse Transcriptase. The RNase mix was added to the reaction and incubated at the indicated temperature. The reaction mix was then purified using the provided SNAP column and a deoxyadenosine tail was added to the first strand cDNA using TdT. The second-strand synthesis was performed by SuperScript II RT using a 3' RACE Adapter primer and amplified using gene-specific primer 2 and supplied AUAP. The list of primers used is presented in *Table 2.7*. The steps of first-strand cDNA synthesis and amplification of second-strand cDNA were further optimised during the study (section 3.2).

**Table 2.7: List of primers used for 5' RACE**

| Primer | Sequence 5'→ 3' |
|---|---|
| Gene-specific primer 1 (AGAP2) | CAGGCGCAGTTCAGGAATG |
| Gene-specific primer 2 (AGAP2) | GAGCGGCTCAAAGTCCATTCCT |
| Abridged universal amplification primer (AUAP) | GGCCACGCGTCGACTAGTAC |
| RLM-RACE 5' Forward primer | CGACTGGAGCACGAGGACACTGA |
| RLM-RACE 5' Nested Forward | GGACACTGACATGGACTGAAGGAGTA |
| Gene-specific Nested Reverse | GCTATTGATCACAGCCTCTCGA |
| RLM-RACE Control Primer B.1 | GACCTGGCCGTCAGGCAGCTCG |

### 2.2.5.2 5' RNA ligase-mediated rapid amplification of cDNA ends (5' RLM-RACE):

#### 2.2.5.2.1 Principle (*Figure 2.3B*):

This method is based on the work of Maruyama and Sugano (1994) and Volloch *et al.* (1994). The technique involves the removal of 5' phosphate group using calf intestinal phosphatase (CIP) from truncated mRNA and non-mRNA that are not protected by 5' cap structures. The RNA is treated with tobacco acid pyrophosphatase (TAP) to remove the 5' cap from intact full-length mRNA, leaving a phosphate group at the 5' end that is required for ligation to the RNA oligo. Next, the RNA oligo containing a known priming site is ligated with the 5' end of the mRNA using a T4 RNA ligase. It is followed by reverse transcribing the mRNA of interest using a gene-specific primer and amplifying the first-strand cDNA using a primer complementary to the priming site on the RNA oligo and another gene-specific primer. If needed, nested amplification could be performed with nested primers. The PCR products are purified and inserted into an appropriate cloning vector for sequencing.

#### 2.2.5.2.2 Methodology:

The 5' RLM-RACE (ThermoFisher) was performed following the manufacturer's instructions. Briefly, 3 µg of total RNA were treated with CIP and TAP to dephosphorylate and remove the 5' mRNA cap structure, respectively. The RNA was then ligated to 250 ng of GeneRacer RNA adaptor by T4 RNA ligase. After each step, the RNA was precipitated using phenol/chloroform. The dephosphorylated, decapped, and ligated RNA was reverse transcribed using gene-specific primer 1. The cDNA was amplified using the RLM-RACE 5' forward primer and gene-specific primer 2. The PCR product was diluted 50-fold and amplified again using nested RLM-RACE and gene-specific nested primers and purified by 2% agarose gel electrophoresis. The purified product was cloned for sequencing using TOPO TA Cloning Kit (ThermoFisher) and transformed into TOP10 chemically competent E.coli (ThermoFisher) and cultured in LB medium containing 100 µg/mL ampicillin (Sigma-Aldrich); at least, ten independent clones were chosen and sequenced for each cell line by Sanger sequencing (Source Bioscience) and aligned using MultAlin (Corpet, 1988). The list of primers used is presented in *Table 2.7*.

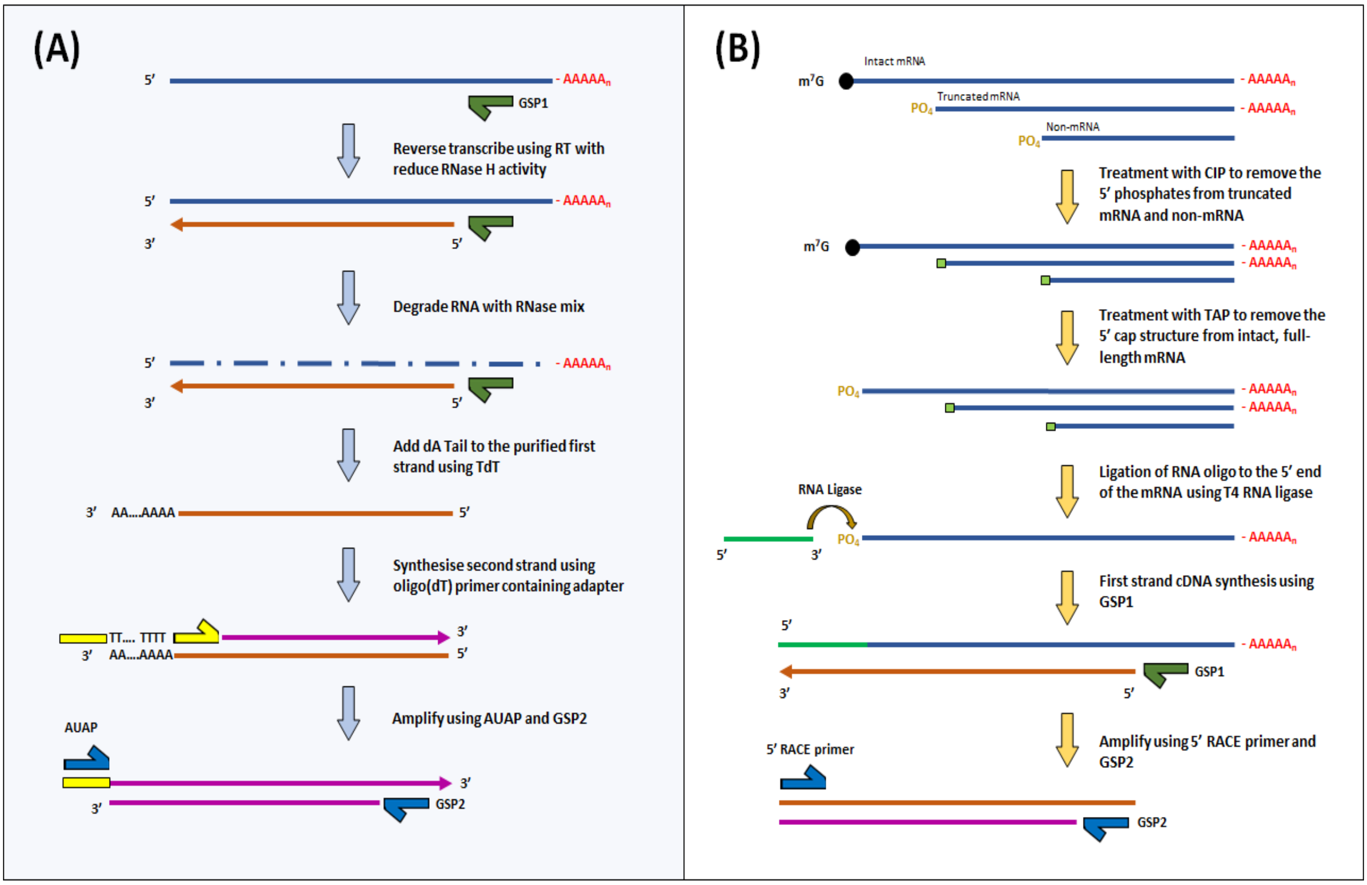

(A)
5'
- AAAAA$_n$
GSP1
Reverse transcribe using RT with reduce RNase H activity
5'
- AAAAA$_n$
3'
5'
Degrade RNA with RNase mix
5'
- AAAAA$_n$
3'
5'
Add dA Tail to the purified first strand using TdT
3' AA....AAAA
5'
Synthesise second strand using oligo(dT) primer containing adapter
TT.... TTTT
3'
3' AA....AAAA
5'
Amplify using AUAP and GSP2
AUAP
3'
3'
GSP2
(B)
Intact mRNA
m$^7$G
- AAAAA$_n$
Truncated mRNA
PO$_4$
- AAAAA$_n$
Non-mRNA
PO$_4$
Treatment with CIP to remove the 5' phosphates from truncated mRNA and non-mRNA
m$^7$G
- AAAAA$_n$
- AAAAA$_n$
Treatment with TAP to remove the 5' cap structure from intact, full-length mRNA
PO$_4$
- AAAAA$_n$
- AAAAA$_n$
RNA Ligase
Ligation of RNA oligo to the 5' end of the mRNA using T4 RNA ligase
PO$_4$
- AAAAA$_n$
5'
3'
First strand cDNA synthesis using GSP1
5'
- AAAAA$_n$
GSP1
3'
5'
5' RACE primer
Amplify using 5' RACE primer and GSP2
GSP2

**Figure 2.3: Overview of the 5' RACE Procedure.** Schematic representation of 5' RACE with homopolymeric tailing **(A)** and 5' RNA ligase-mediated RACE **(B)**. AUAP: Abridged universal amplification primer; CIP: calf intestinal phosphatase; dA: deoxyadenosine; GSP: gene-specific primer; RT: reverse transcriptase; TAP: tobacco acid pyrophosphatase.

## 2.2.6 Dual luciferase Reporter assay:

### 2.2.6.1 Plasmid Construction:

The plasmid (pcDNA3 RLUC POLIRES FLUC) was a gift from Nahum Sonenberg (Addgene pcDNA3; RRID: Addgene_45642). The bicistronic reporter plasmid expresses *Renilla* luciferase (Rluc) and *Firefly* luciferase (Fluc) (*Figure 2.4A*). Translation of the Rluc cistron is cap-dependent and mediated by an upstream cloned fragment under the control of the CMV promoter, whereas the Fluc cistron is directed by the poliovirus IRES (cap-independent) and serves as an internal control. The 5' UTR isoforms (shorter, longer and mutated longer) were designed and purchased from GeneScript (Hong Kong) (*Table 2.8*) and were inserted at the unique *NheI* restriction site proximal to Rluc ORF (*Figure 2.4B*). The 5' UTR isoforms and the plasmid were digested with the *Nhe*I restriction enzyme (Promega) and the products were separated on a 1% agarose gel, purified with Wizard SV kit (Promega), and treated with alkaline phosphatase (Promega) to prevent self-ligation. The purified digested plasmid and the 5’ UTR inserts were ligated using T4 DNA ligase (Promega). The constructs were transformed into DH5α competent cells (Thermo-Fisher) and cultured in LB medium containing 100 µg/mL ampicillin (Sigma-Aldrich). Positive clones were chosen, purified using NucleoSpin Plasmid Columns (Fischer Scientific), and confirmed by Sanger sequencing (Source Bioscience).

**Table 2.8: Sequence of 5’ UTR isoforms used in the study.**

| 5’ UTR | Sequence (5‘→ 3‘) |
|---|---|
| Shorter | AAAAGCTAGCGAAGGGGCCTTCTGAGGTTTGGGGGCTGTAGGGCCATGGGCCTCAGGGCCAGAGGTGGTTGTTAGCCTGGCAAGACAGGTCTGGGCAACGCTAGCAAAA |
| Longer | AAAAGCTAGCAAGGGCGGGCAGGGGCGGGGAGTTCTGGGCACAGCAGAAGGGGCCTTCTGAGGTTTGGGGGCTGTAGGGCCATGGGCCTCAGGGCCAGAGGTGGTTGTTAGCCTGGCAAGACAGGTCTGGGCAACGCTAGCAAAA |
| Mutated Longer (mutations represented by underline **G→ A**) | AAAAGCTAGCAAGAGCGAGCAGAGGCGAGGAGTTCTGGGCACAGCAGAAGGGGCCTTCTGAGGTTTGGGGGCTGTAGGGCCATGGGCCTCAGGGCCAGAGGTGGTTGTTAGCCTGGCAAGACAGGTCTGGGCAACGCTAGCAAAA |

## 2.2.6.2 Plasmid Transfection:

### 2.2.6.2.1 Transfection of Prostate cancer cell line (adherent):

High quality reporter constructs (with A260/A280 ratio ~ 1.8 and A260/A230 ratio > 2) were transfected into the PC cell line (DU145) using the JetPRIME transfection reagent (Polyplus) according to the manufacturer’s protocol. The quality of the transfected DNA was examined using Nanodrop 8000 spectrophotometer (ThermoFisher). Briefly, cells were seeded at a density of 2.5 x $10^5$ cells/well in 6-well plates for 24 hours before transfection. 2 µg of plasmid constructs were diluted into 200 µL of supplied jetPRIME buffer and mixed by vortexing. 4 µL of jetPRIME transfection reagent was then added to the mixture, briefly vortexed, and incubated for 10 minutes at room temperature to allow the formation of transfection complexes. After incubation, the transfection mixture was added dropwise to the cells in serum-containing medium and cells were incubated with the transfection mix at 37°C for 4 hours followed by replacement with a fresh growth medium. The cells were collected 48 hours after being transfected.

### 2.2.6.2.2 Transfection of Chronic Myeloid Leukaemia cell line (suspension):

The transient transfection of the CML cell line (KU812) was carried out using electroporation. The optimal conditions for electroporation that resulted in the highest transfection efficiency using the GFP plasmid and the lowest cell death rate were determined previously in our lab (Doush, 2015). The cells were prepared at a density of 2 X $10^6$ and resuspended in 100 µL Amaxa Cell Line Nucleofector solution V. Next, 1 µg of high-quality plasmid DNA was added to the transfection mixture. The mixture was transferred into an electroporation cuvette and was electroporated using the Nucleofector device (program X-001). Immediately after electroporation, 500 µL of complete medium was added to the cuvette and the cells were transferred to a well of 24-well plate. The cells were collected after 6 hours for dual luciferase reporter assay.

### 2.2.6.3 Luciferase assay:

After the indicated time points, cells were lysed with Passive Lysis Buffer (Promega) and the luciferase activity was measured using the Dual-Luciferase Reporter Assay System (Promega) on a CLARIOstar microplate reader (BMG Labtech). The plate reader uses two precision injectors to introduce reagents into the sample and record luminescence. Luciferase assay reagent – LAR II (to detect *Firefly* luciferase) and Stop & Glo® Reagent (to detect *Renilla* luciferase) was primed into the injectors. For measurements, 100 µL of LAR II was added to the lysate and *Firefly* luminescence was measured for 10 sec. After that, 100 µL of Stop & Glo® Reagent was dispensed and the *Renilla* luminescence was recorded for 10 seconds. The *Firefly* luciferase activity served as an internal normalising control.

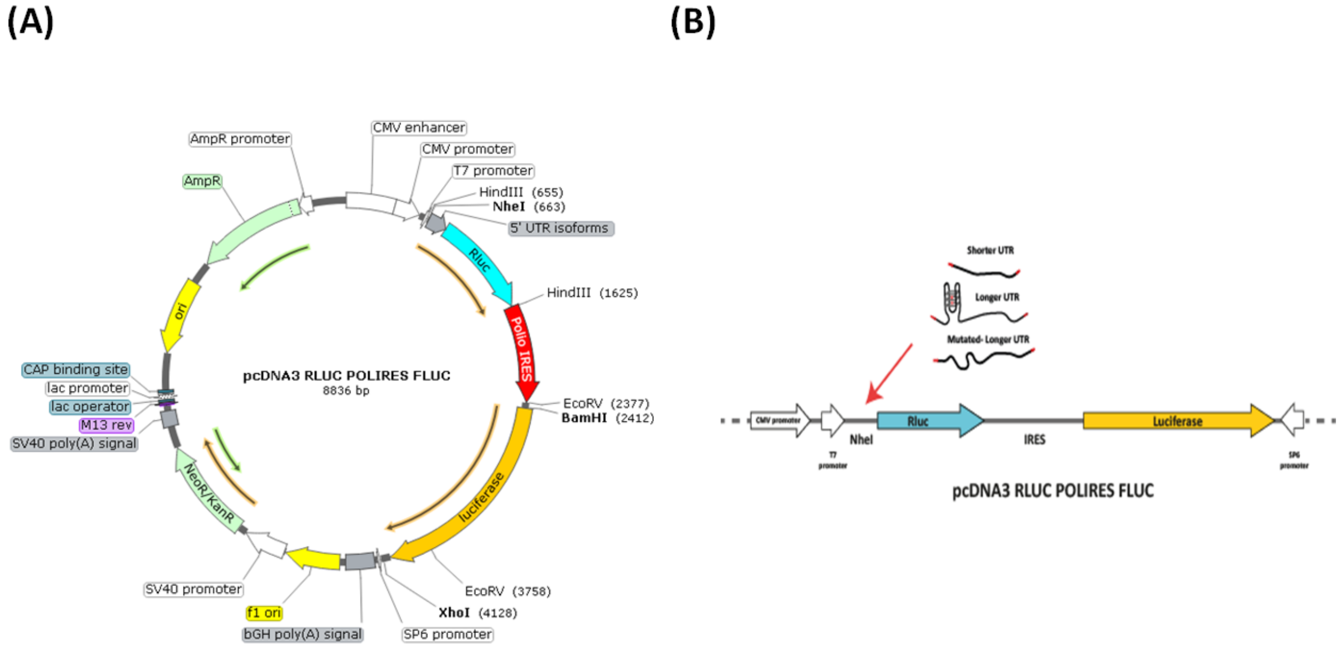


**Figure 2.4: Plasmid map of dual luciferase reporter plasmid (pcDNA3 RLUC POLIRES FLUC). (A)** Schematic representation of bicistronic Luciferase reporter (pcDNA3 RLUC POLIRES FLUC) plasmid. **(B)** The AGAP2 5' UTR fragments (shorter 5' UTR, longer 5' UTR with G4 consensus, and longer 5' UTR with G4 consensus destroyed) were inserted at the unique *NheI* restriction site proximal to *Renilla* luciferase ORF.

## 2.2.7 *In vitro* Transcription and Translation:

To analyse the influence of different 5' UTR isoforms (*Table 2.8*) in a cell-free system, the plasmid constructs were transcribed *in vitro* by T7 RNA polymerase using mMESSAGE mMACHINE Transcription Kit (Thermo Scientific) following the manufacturer's guidelines. Briefly, the plasmid was linearised using *XhoI* restriction endonuclease (Promega) by incubating with the restriction enzyme at 37°C for 2 hours. The linearised plasmid was then incubated with the supplied reaction mix at 37°C for 1 hour to produce a large amount of capped RNA using T7 RNA polymerase promoter in the plasmid. The lithium chloride precipitation was subsequently used to purify the cap RNA, separating proteins and unincorporated nucleotides including cap free analogues in the reaction mix. The resulting RNA was translated *in vitro* using Flexi Rabbit Reticulocyte Lysate Translation System (Promega). The RNA was incubated with the provided reaction mix and translated at 30°C for 90 min. The luciferase activity of the translation products was analysed using Dual-Luciferase Reporter Assay, as described above (section 2.2.6).

## 2.2.8 Polysome fractionation:

### 2.2.8.1 Principle:

Polysome fractionation is a technique to evaluate the density of polyribosomes that form on a given mRNA (Warner *et al.,* 1963). This technique has been extensively used to analyse the translation efficiency at the level of individual mRNA and the whole transcriptome to elucidate different factors responsible for modulating translational output (Reviewed in Kuersten *et al.,* 2013, Chassé *et al.,* 2016).  It is based on stalling the translating ribosome on mRNA using cycloheximide treatment that prevents translocation of elongating ribosomes. A sucrose density gradient is used to separate messenger ribonucleoprotein complexes (mRNP) based on the number of bound ribosomes, with transcripts associated with polyribosomes sediment at a higher density in the gradient relative to mRNP with lower ribosome density. The distribution of mRNA in each fraction could be determined by qRT-PCR and high throughput technology including microarray and RNA sequencing. The translational efficiency of mRNA is inferred from its distribution in the polysome fraction. The schematic of polysome fractionation is presented in *Figure 2.5*.

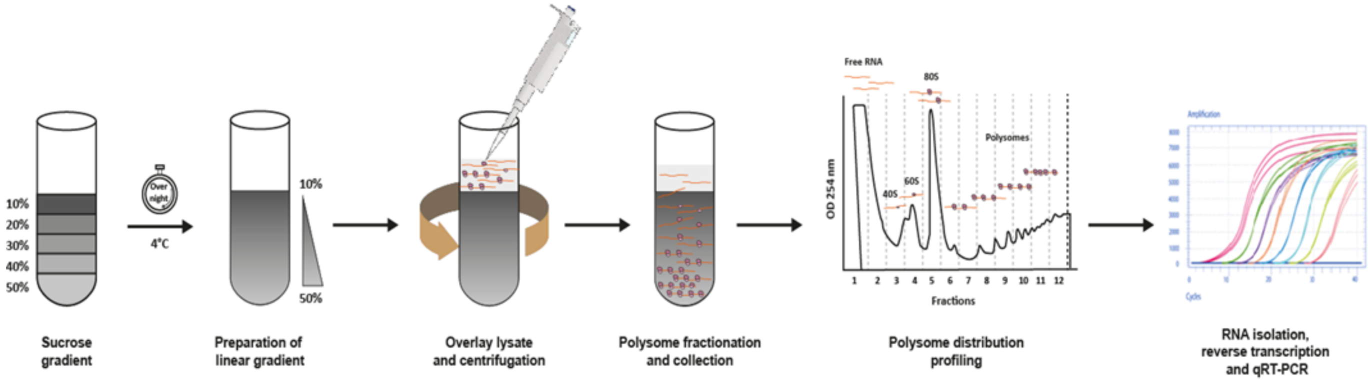

10%
20%
30%
40%
50%
Sucrose gradient
Over night
4°C
10%
50%
Preparation of linear gradient
Overlay lysate and centrifugation
Polysome fractionation and collection
Free RNA
80S
40S
60S
Polysomes
OD 254 nm
1 2 3 4 5 6 7 8 9 10 11 12
Fractions
Polysome distribution profiling
Amplification
Cycles
RNA isolation, reverse transcription and qRT-PCR

**Figure 2.5: Schematic of polysome fractionation.** A linear sucrose density gradient was prepared by underlaying 10-50% sucrose solutions and incubating it overnight at 4°C. The cytoplasmic lysate was then applied to the linear gradient and centrifuged to separate the polysomes based on their density. The gradients were fractionated while measuring absorbance at 254 nm to generate polysome distribution profiles. The RNA was isolated from each fraction and analysed using qRT-PCR.

## 2.2.8.2 Methodology:

### 2.2.8.2.1 Preparation of cell lysate:

The number of cells used per gradient was optimised for different cell line used in the study. The cell concentration yielding clear and distinct peaks in the polysome profile was selected. Typically, 2.5-5 x $10^7$ cells were used per gradient. The cells were treated with 100 µg/mL cycloheximide and incubated at 37°C and 5% $CO_2$ for 10 min. It was followed by washing and resuspending the cells in the lysis buffer (100 mM KCl, 5 mM MgCl2, 20 mM HEPES (pH 7.4), 0.5% NP-40, 100 µg/mL CHX, 2 mM DTT, 40 U/ml RNase inhibitor, and 1x protease inhibitor cocktail) and then incubating on ice for 10 minutes, vortexing briefly every 2 minutes. The cells were subsequently centrifuged at 2,000 x g for 5 minutes at 4°C to pellet the nuclei and larger debris. The supernatant was transferred to a fresh tube and again centrifuged at 13,000 x g for 5 minutes to collect smaller debris. The supernatant lysate was collected and used immediately for polysome fractionation or stored at -80°C. The lysates were prepared by Dr Cristina Montiel-Duarte.

### 2.2.8.2.2 Preparation of sucrose gradients and fractionation:

The steps of sucrose gradients preparation, loading, ultracentrifugation, polysome gradient profiling, and collection of fractions were performed by Dr Keith Spriggs from the University of Nottingham. Briefly, sucrose solutions of different concentrations ranging from 10% to 50% were prepared, underlaid in the order of increasing concentration, and left overnight at 4°C to yield a linear sucrose gradient. The lysate obtained in the above step was layered on top of the prepared 10-50% linear gradient and centrifuged at 190,000 x g for 90 minutes at 4°C. The gradients were then fractionated from top to bottom while measuring absorbance at 254 nm to generate polysome profiles. 500 µL of each sucrose fractions were collected and RNA was isolated using TRIzol extraction or stored at -80°C until use.

### 2.2.8.2.3 RNA purification from polysome fractions:

500 µL of TRIzol (ThermoFisher) and 200 µL of chloroform (Sigma-Aldrich) was added to each fraction and mixed by shaking vigorously for 10 seconds followed by centrifugation at 13,000 x g for 15 min at 4°C to separate the mixture into 3 layers. The RNA in the top aqueous layer was precipitated by adding 1 mL 100% ethanol and 40 µL of 3M sodium acetate (pH5.2), and incubated overnight at -20°C. Each fraction was spiked with 500 ng of *in vitro* transcribed luciferase RNA control to normalise for differences in RNA recovery. The exogenous luciferase RNA was prepared from pcDNA3 RLUC POLIRES FLUC plasmid (Addgene pcDNA3; RRID: Addgene_45642) using mMESSAGE mMACHINE Transcription Kit, as described above (section 2.2.7). After overnight incubation, the RNA was

precipitated by centrifugation at 12,000 x g for 30 minutes at 4°C. The supernatant was removed, and the pellet was washed with 800 µL of 75% ethanol and centrifuged again at 7,500 x g for 5 minutes at 4°C. The ethanol was carefully removed, and the pellet was air dried at room temperature for 2-3 minutes and resuspended in 50 µL of nuclease-free water. The RNA was further cleaned and concentrated using ReliaPrep RNA Miniprep Systems (section 2.2.2.1). The samples were reverse transcribed (section 2.2.2.2) and amplified using qPCR (section 2.2.2.3), as mentioned above.

#### 2.2.8.2.4 Analysis of polysome profiling data:

The polysome fractionation data was processed using the protocol published by Pringle *et al.* (2019). The Ct value of target mRNA transcript amplified in each fraction was subtracted from the luciferase RNA spike [$\triangle Ct = Ct_{luc} - Ct_{target}$]. The $\triangle\triangle Ct$ for each fraction was then calculated by subtracting the first (lightest) fraction from each fraction [$\triangle\triangle Ct = \triangle Ct_n - \triangle Ct_1$]. The expression of the target transcript in the given fraction was determined by $2^{\triangle\triangle Ct}$ [$E_n = 2^{\triangle\triangle Ctn}$]. Subsequently, the expression values for all the fractions were added together [$E_{total} = E_1 + E_2 + E_3 + ….. + E_n$]. Finally, the proportion of the target transcript in each fraction was determined by [$P = 100 \times E_n/E_{total}$] and plotted as a line graph illustrating the distribution of the proportions recovered in each fraction.

### 2.2.9 RNA sequencing:

RNA sequencing was performed to examine differentially expressed genes in PC and CML cell lines, and to identify genes with differential TSS distribution profiles. The sequencing was performed for the main cell lines included in the study (PC: DU145, PC3, LNCap; CML: KU812, TCCS, KCL-22), and the sequencing data for PC and CML cell lines were grouped together and considered as biological replicates.

For RNA sequencing, the RNA was extracted as described above (section 2.2.2.1). The RNA quality was assessed using a bioanalyser (Agilent) and samples with RNA integrity number (RIN) > 8 were considered optimal for RNA sequencing. 2 µg total RNA was used for library preparation with TruSeq Stranded mRNA Sample Prep Kit (Illumina). The library preparation and next generation sequencing were outsourced to the Edinburgh Clinical Research Facility. The RNA sequencing was performed using the HiSeq2500 instrument (Illumina), with over 45 million, paired-end, 75 base pair reads per sample.

The post-processing and alignment of the sequencing reads were performed using BaseSpace (Illumina) platform. Sequencing reads were mapped to the GRCh38 human reference genome assembly using STAR aligner (Dobin *et al.,* 2013). Differentially expressed genes were identified using the DESeq2 package in R (Love *et al.,* 2014). Heatmaps [based on FPKM values] were generated using the Complex heatmaps package in R (Gu *et al.,* 2016). A volcano plot was used to visualise significant genes and the magnitude of their fold change and was created using the Enhanced volcano package (Blighe *et al.,* 2020). The script to perform DESeq2 analysis and generate heatmap and volcano plot in R is presented below:

```
library(DESeq2)
library(ggplot2)
library(RColorBrewer)
library(EnhancedVolcano)
library(gplots)

# For DESeq2
data1 <- read.table("RNA-seq raw count.txt", header = TRUE, sep = "\t")
Data <- as.matrix(data1[ , -1])
rownames(Data) <- data1[ , 1]
colData <- DataFrame(condition=factor(c("ctrl","ctrl","ctrl", "treat", "treat",
"treat")))
dds <- DESeqDataSetFromMatrix(countData=Data,
                              colData=colData,
                              design=~condition)
dds <- DESeq(dds)
res <- results(dds)
head(results(dds, tidy=TRUE))
write.table(res,"DESEQ2 ARIF.txt",sep = "\t" ,col.names=T, row.names = F)
resOrdered <- res[order(res$padj),]
sig <- resOrdered[!is.na(resOrdered$padj) &
                    resOrdered$padj<0.10 &
                    abs(resOrdered$log2FoldChange)>=1,]
write.table(sig,"DESEQ2 sig.txt",sep = "\t" ,col.names=T, row.names = F)
```

```
# For heatmap
library(pheatmap)
library(ComplexHeatmap)
library(circlize)
my_Data <- read.table("Heatmap.tsv", sep="\t", stringsAsFactors=FALSE, header=TRUE)
my_data <- data.matrix(my_Data[ , -1])
rownames(my_data) <- my_Data[ , 1]
Heatmap(my_data)
col_fun = colorRamp2(c(-2, 0, 2), c("red", "black", "green"))
Heatmap(my_data, cluster_columns=TRUE,row_labels = rownames(my_data,
        row_names_side = "left",show_row_names = FALSE,column_dend_side ="top",
        column_names_side = "top", cluster_rows = TRUE, show_row_dend = FALSE,
        name = "Row Z-Score", col = col_fun, column_names_rot = 360,
        column_names_centered = TRUE,column_dend_height = unit(25, "mm"),
        heatmap_legend_param = list(legend_height = unit(4, "cm"),
                               color_bar = "continuous",
                               title_position = "leftcenter-rot"),
        column_names_gp= gpar(fontsize = 10,fontface = "bold"))
```

```
#Volcano plot
library(EnhancedVolcano)
library(magrittr)
library(rio)
final <- read.table("G4 DESEQ2 all .txt", header = TRUE, sep = "\t")
rownames(final) <- final[ , 1]
final$X = NULL
keyvals <- ifelse(
  final$log2FoldChange < -1 & final$padj <10e-2, 'blue',
  ifelse(final$log2FoldChange > 1 & final$padj <10e-2, 'red',
         'grey'))
keyvals[is.na(keyvals)] <- 'grey'
names(keyvals)[keyvals == 'blue'] <- 'DE in PC cell lines'
names(keyvals)[keyvals == 'grey'] <- 'NS'
names(keyvals)[keyvals == 'red'] <- 'DE in CML cell lines'
EnhancedVolcano(final, lab = rownames(final),
                x = 'log2FoldChange', y = 'padj',selectLab =
                c("CDKN2A","PXDN","FERMT2","ITGA3",
                  "CASK","ITGB5","DUSP3","HK1", "AGAP2"),
                xlab = bquote(~Log[2]~ 'fold change'),
                ylab = bquote(~-Log[10]~ "(adjusted"~~ italic("P")~ "value)"),
                axisLabSize= 14, pCutoff = 10e-2,FCcutoff = 1, ylim = c(0, 10),
                title = NULL, subtitle = NULL, gridlines.major = FALSE,
                gridlines.minor = FALSE, pointSize = 3.0, hline = 0.1,
                hlineCol = c("black"), vline = c(-1, 1),vlineCol = c("black"),
                labSize = 3.5,labCol = 'black', labFace = 'bold',
                boxedLabels = TRUE, colAlpha = 3/5, legendPosition = 'bottom',
                legendLabSize = 12,legendIconSize = 3, colConnectors = 'black',
                typeConnectors = "open", caption = NULL,
                colCustom = keyvals, cutoffLineType= "blank") +
                ggplot2::coord_cartesian(xlim=c(-12, 12))
```

### 2.2.10 Statistical analysis:

All statistical analysis was performed using GraphPad Prism 8.4 software (GraphPad, Inc, USA). For experiments where two groups were compared, a two-tailed Student's t-test was performed in case of normalised data and Mann-Whitney U-test was used for the analysis of non-parametric data. Normality was evaluated using the Shapiro-Wilk test. For comparison of three or more groups, a one-way ANOVA was performed followed by post-hoc Sidak's multiple comparison tests. For non-parametric data, Kruskal-Wallis followed by uncorrected Dunn's test was used. Unless otherwise stated, histogram columns represent the mean and error bars indicate the standard deviation (SD). The standard error of mean (SEM) was used to displayed variance in the polysome results. The number of replicates (n) for each experiment is stated in the figure legend. The data is considered to be statistically significant if $P < 0.05$ and are indicated in the figure legends by asterisks (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

# Chapter 3:

# Role of Transcription Start Site Selection in gene expression regulation

## 3.1 Introduction

### 3.1.1 AGAP2 regulation in PC and CML cell lines:

Our group, led by Dr Cristina Montiel-Duarte, has been studying the role of *AGAP2* (Arf GAP with GTP-binding protein-like domain, Ankyrin repeat and PH domain 2) (section 1.61) in different cancers and liver fibrosis (Doush *et al.,* 2019, Navarro-Corcuera *et al.,* 2020, Navarro-Corcuera *et al.,* 2019). Recently, we have elucidated the regulation of *AGAP2* expression in Prostate Cancer (PC) and Chronic Myeloid Leukaemia (CML) cell lines and have identified a novel role for SP1 and RARα on *AGAP2* transcription activation (Doush *et al.,* 2019).

To get insights into the factors that regulate *AGAP2* expression in PC (DU145) and CML (KU812) cell lines, we cloned and characterised *AGAP2* promoter region in the reporter vector to determine the minimal promoter region used in these cell lines. We noted a comparable minimal promoter region in both the cell lines and identified sequence elements sufficient to induce reporter activity (*Figure 3.1A*). The analysis of these sequence features revealed the presence of SP1 binding sites and a DR5 (a retinoic acid response element). Using siRNA (*Figure 3.1B*), specific inhibitors (*Figure 3.1C*), and chromatin immunoprecipitation (*Figure 3.1D, 1E*), we highlighted the role of these transcription factors in regulating *AGAP2* expression.

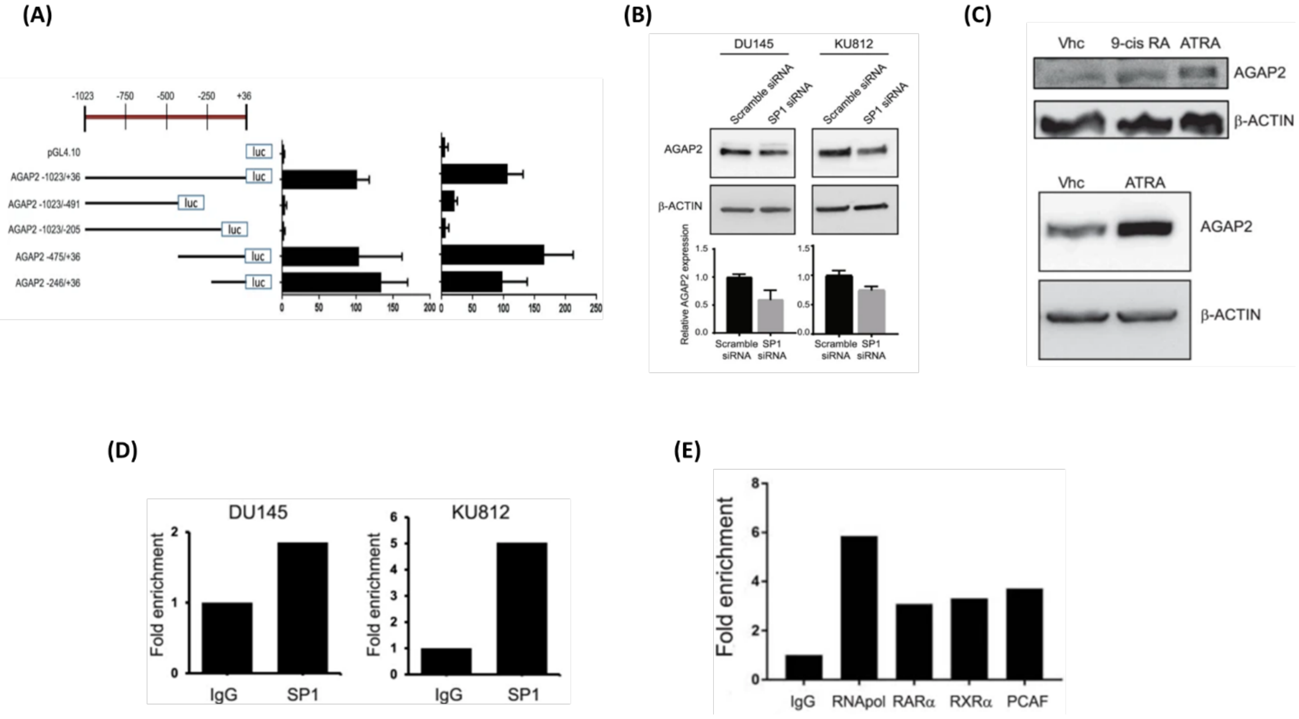


**Figure 3.1: The role of SP1 and RARα in regulating *AGAP2* expression in PC and CML cell line (Doush *et al.,* 2019). (A)** *AGAP2* promoter deletion fragments generated from human genomic DNA (Promega) and cloned into the promoter-less *Firefly* luciferase vector pGL4.10 (Promega). The AGAP2-luciferase constructs were transfected into DU145 (PC) and KU812 (CML) cell lines using CalPhos Mammalian Transfection kit (Clontech) and electroporation (Nucleofector 2b Device), respectively. The luciferase activity was measured using Dual-Light Luciferase & β-Galactosidase Reporter Gene Assay System (ThermoFisher) and normalised to the corresponding β-galactosidase values. The luciferase activity in KU812 (left) and DU145 (right) are presented relative to the values obtained for the full-length *AGAP2* −1023/+36 plasmid. **(B)** KU812 and DU145 cells were transfected with either scramble or SP1

siRNA. In the case of KU812, 67 nM of siRNA was transfected, and cells were collected 48 hours after transfection; for DU145 5 nM of siRNA was transfected followed by the collection of cells after 72 hours. After indicated time points, the cells were lysed and 10 µg of total protein were resolved using 10% SDS-PAGE followed by immunoblotting with SP1 isoform-specific antibody (Cell Signaling, RRID:AB_11220235). β-Actin levels were used as a loading control. Densitometry values (mean ± SD) for the relative protein expression are represented below the blots. **(C)** Detection of AGAP2 protein levels in DU145 (top) and KU812 (bottom) after 24 hours treatment with 1 µM ATRA and 9-cis RA in DU145 cells or 1 µM ATRA in KU812 cells, β-Actin was used as a loading control. **(D)** DU145 and KU812 cells were grown to 80% confluency under normal serum conditions and were fixed and sheared using the Ultra-sonicator (Covaris). The chromatin was immunoprecipitated with 1 µg of rabbit IgG or anti-SP1 antibody (Cell Signalling) and amplified with primers for the *AGAP2* promoter region. **(E)** DU145 cells grown, fixed and sheared as in (D) and immunoprecipitated using 2 µg of rabbit IgG (negative control) or anti-RNApol II (positive control) or anti-PCAF antibody and amplified with primers to detect *AGAP2* promoter region. Data in [D, E] are presented as fold enrichment relative to IgG Ct values. **Note:** Experiments for Figure B, C (below) and D were performed as part of the current PhD project. Experiments in (A) were performed by Dr Yegor Doush and experiment in (E) was performed by Dr Amaia Navarro-Corcuera.

Interestingly, we also noted large differences in *AGAP2* mRNA levels between PC and CML cell lines (*Figure 3.2A*). Preliminary analyses done by Dr Cristina Montiel-Duarte had also demonstrated inconsistency in AGAP2 mRNA and protein expression levels (*Figure 3.2B*). We noted a higher relative AGAP2 mRNA level in CML cell lines with lower protein expression, and the opposite was observed in PC cell lines: lower mRNA and higher protein expression. Intriguingly, we also noted differential usage of TSSs for *AGAP2* in these cell lines. In the CML cell line (KU812), the transcription starts 36 bp earlier compared to the PC cell line (DU145) and encoded an extra region in the 5’ UTR of *AGAP2* mRNA that contained a consensus for a G quadruplex. Given the differential TSS selection in the PC and CML cell lines, with earlier TSS selection in CML cell lines that encode G4 consensus in the 5’ UTR, we were curious to examine if it could explain the discrepancy observed in AGAP2 mRNA and protein levels noted in these cell lines. These preliminary findings formed the basis of the current work and drew our attention to the potential regulatory role of differential TSS selection in controlling *AGAP2* expression.

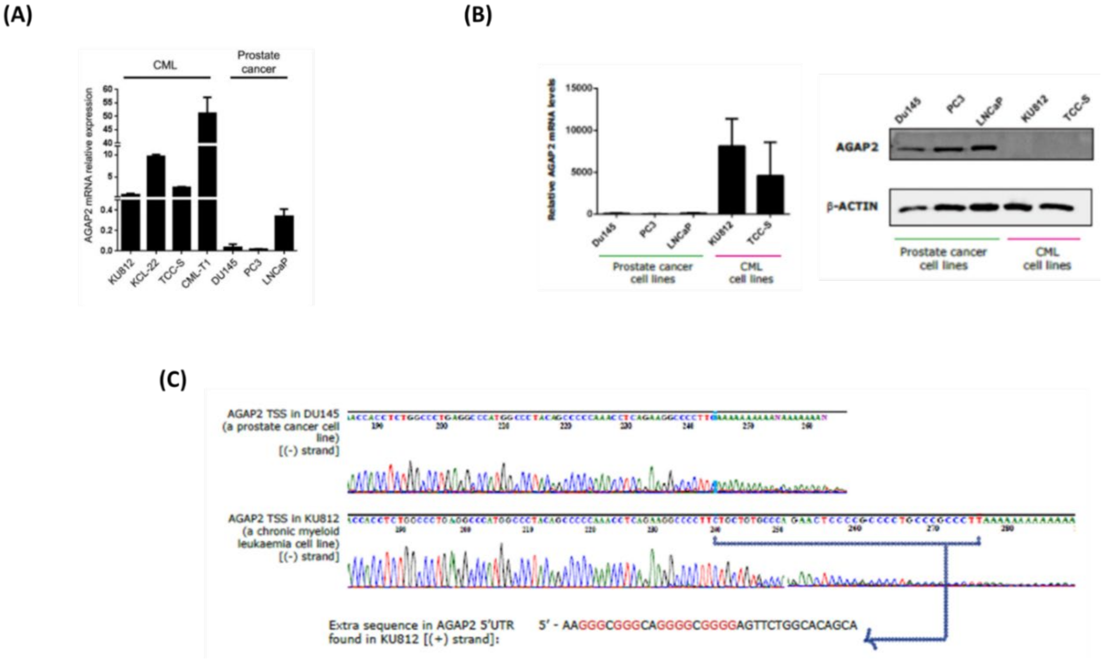


**Figure 3.2: AGAP2 expression in PC and CML cell lines. (A)** *AGAP2* mRNA basal levels were measured by qRT-PCR in CML (KU812, KCL, TCC-S, CML-T1) and PC cell lines (DU145, PC3 and LNCaP). The values presented are normalised against the levels of the housekeeping gene (HPRT and TBP) and shown relative to the CML cell line (KU812) (Doush *et al.,* 2019). **(B)** Preliminary experiment (n=1) showing the inconsistency in *AGAP2* mRNA and protein levels. The mRNA levels (left) were quantified by qRT-PCR and protein (right) were detected by immunoblotting with PIKE-A isoform-specific antibody. β-Actin was used as a loading control. **(C)** TSS (n =1) were identified using the 5′ RACE system with deoxyadenosine homopolymeric tailing (ThermoFisher). KU812 (CML) cell line exhibit an upstream TSS that produces a longer *AGAP2* mRNA 5′ UTR containing extra nucleotides (36bp) relative to DU145 (PC). **Note:** Preliminary experiments were conducted by Dr Cristina Montiel-Duarte.

## 3.1.2 TSS-mediated gene regulation:

The crucial step in the gene regulation process takes place at the level of transcription initiation and is modulated by alternative promoter usage, differential splicing, transcription start site selection, and epigenetic factors, producing a heterogeneous population of mRNA isoforms from a single gene locus ( Andersson and Sandelin, 2020, Baralle and Giudice, 2017, Danino *et al.,* 2015, Davuluri *et al.,* 2008). The complex transcriptional regulation also impacts the translation potential of the transcribed mRNA isoforms by encoding regulatory elements in the 5′ and 3′ UTR that variably influence mRNA maturity, stability, localisation, and translation efficiency (Culjkovic-Kraljacic and Borden, 2018, Curran and Weiss, 2016, Wilkie *et al.,* 2003). Given the impact of transcriptional and post-transcriptional processes on modulating protein translation, an inconsistency in mRNA and protein levels could be observed because the transcribed mRNA might not efficiently translate into protein.

A discrepancy in mRNA and protein levels has been observed in a variety of studies (Edfors *et al.,* 2016, Greenbaum *et al.,* 2003, Guo *et al.,* 2008, Maier *et al.,* 2009, Vogel and Marcotte, 2012). The contribution of mRNA concentration (transcription) and mRNA features (post-transcriptional) accounted for 36% variation in protein abundance (Vogel *et al.,* 2010). Taking into consideration the processes related to translation and protein degradation, about two-thirds of the variations in protein abundance could be explained overall (de Sousa Abreu *et al.,* 2009, Plotkin, 2010, Vogel *et al.,* 2010). In a study by Vogel *et al.* (2010), the author used experimental and computational approaches to measure cognate mRNA and protein levels of more than 1,000 genes and showed that regulatory features in the UTR, amino acid properties and coding sequence length were the strongest correlating factors for protein levels. Most of these features are dictated by the transcription start site selection (TSS) which determines the 5′ gene boundary and encode regulatory features in the 5′ leader sequence that could affect the translational potential of the mRNA transcript.

Our understanding of transcription initiation and selection of TSS is far from complete. It has been shown that the transcription does not initiate at a single discrete site and each gene, on average, displayed 4 robust TSSs (Forrest *et al.,* 2014). Alternative transcription initiation (ATI) has been shown to affect gene expression at multiple levels. It could change the 5′ gene boundary and encode regulatory features in the transcript leader that influence mRNA translation (Wang *et al.,* 2016). Studies have shown that alternative initiation and termination have a higher contribution to tissue dependent isoform-specific diversity compared to alternative splicing (Reyes and Huber, 2018). The ATI is driven by multiple promoter usage, and it has been reported by different studies that 30-50% of human genes are regulated by alternative promoters which are utilised depending on cell type, developmental stage, cellular environment, and diseased states including cancers (Reviewed in Davuluri *et al.,* 2008). The usage of alternative promoters for many genes does not change the principal ORF and thus do not impact the composition of protein products (Studies reviewed in Davuluri *et al.,* 2008, Landry *et al.,* 2003).

Given the role of ATI and its effects in modulating mRNA translation efficiency, several attempts have been made in the last decade to precisely identify TSS which would enable characterisation of core promoter features and defines the potential regulatory elements encoded in the 5' UTR of TSS isoforms. Notably, two databases namely FANTOM and DBTSS have comprehensively captured the dynamically changing landscape of TSS selection in different tissue, cell types, conditions, cancers, and development phases using mRNA cap-guided deep sequencing technologies (Forrest *et al.,* 2014, Yamashita *et al.,* 2010) (section 1.3.2). These and other relevant studies have also identified the presence of multiple TSSs within a given core promoter and highlighted the widespread use of differential TSS selection within a single core promoter region. (Carninci *et al.,* 2006, Forrest *et al.,* 2014, Karlsson *et al.,* 2017, Kawaji *et al.,* 2006, Suzuki *et al.,* 2001). The cluster of closely spaced TSSs within a core promoter is principally different from alternative promoters where the transcription initiation is separated by a wide genomic space.

The alternative TSS selection has been shown to significantly influence mRNA translation activity (Rojas-Duran and Gilbert, 2012). These heterogenous TSS isoforms exhibit differences in the length of 5'UTR and encode regulatory elements that might contribute towards the observed translation differences (Rojas-Duran and Gilbert, 2012, Suzuki *et al.,* 2001). Moreover, the distribution of start sites in a TSS cluster linked to a core promoter was found to differentially expressed in various cell lines (Ohmiya *et al.,* 2014), highlighting a novel layer of gene expression regulation. Despite their existence, the consequences of differential TSS selection have not been studied thus far. The dominant starting sites (major TSS) have usually been the focus of studies and was analysed in the context of ATI-mediated regulation, representing the translational impact of transcript isoforms derived from multiple closely situated promoters (Arribere and Gilbert, 2013, Dieudonné *et al.,* 2015, Li *et al.,* 2019, Rojas-Duran and Gilbert, 2012, Wang *et al.,* 2016, Zeitz *et al.,* 2019, other studies reviewed in Davuluri *et al.,* 2008). The data on the contribution of minor TSSs within the core promoter is currently lacking and the regulatory elements encoded by them are yet to be defined.

### 3.1.3 Aims of chapter 3:

Our previous published work focused on the regulation of AGAP2 expression in PC and CML cell lines, finding differences in mRNA transcription in these cell types (Doush *et al.,* 2019). Interestingly, the preliminary experiments have shown a discrepancy in AGAP2 mRNA and protein levels and a potential of differential TSSs usage within the core promoter region of *AGAP2* in PC and CML cell lines. If confirmed, the differential TSS selection could potentially modulate mRNA translatability by encoding regulatory features in the mRNA 5ʻ UTR and contributing to the observed inconsistency. In the current PhD project, using *AGAP2* gene as a model, the link between differential TSSs usage and inconsistency in mRNA translational output is further explored using PC and CML cell lines.

This chapter aims to:

- Examine the inconsistency in AGAP2 mRNA and protein levels and investigate the contribution of protein degradation through the Ubiquitin–Proteasome Pathway.
- Characterise the distribution of TSSs within the AGAP2 core promoter region of PC and CML cell lines.
- Analyse all genes with differentially distributed TSSs in PC and CML cell lines.

## 3.2 Method Optimisation

Initially, 5’ RACE with homopolymeric tailing was attempted to characterise TSSs in PC and CML cell line using the method detailed in 2.2.5.1.2. In order to capture the 5’ end of mRNA sequences rich in GC base pairs (e.g. *AGAP2*), the conventional protocol was modified as per the manufacture’s guidelines to amplify full length 5’ cDNA ends rich in GC nucleotides. The use of standard poly dC-tailing with mRNA rich in G-C base pairs in 5‘ end could result in truncated products because the deoxyinosine-containing anchor primer would non-specifically anneal to the GC rich region. Instead, dA-tailing of cDNA was performed as per the modified protocol followed by second-strand cDNA synthesis using the oligo(dT)-containing abridged amplification primer site. The modified protocol was previously used to generate the preliminary TSS data (*Figure 3.2C*).

However, in the current study, the suggested modifications and conditions were not able to produce any 5’ RACE products (*Figure 3.3A*). The control RNA template provided with the kit to check its performance was not suitable to use with the modified protocol as the control reaction was based on the standard poly dC-tailing.

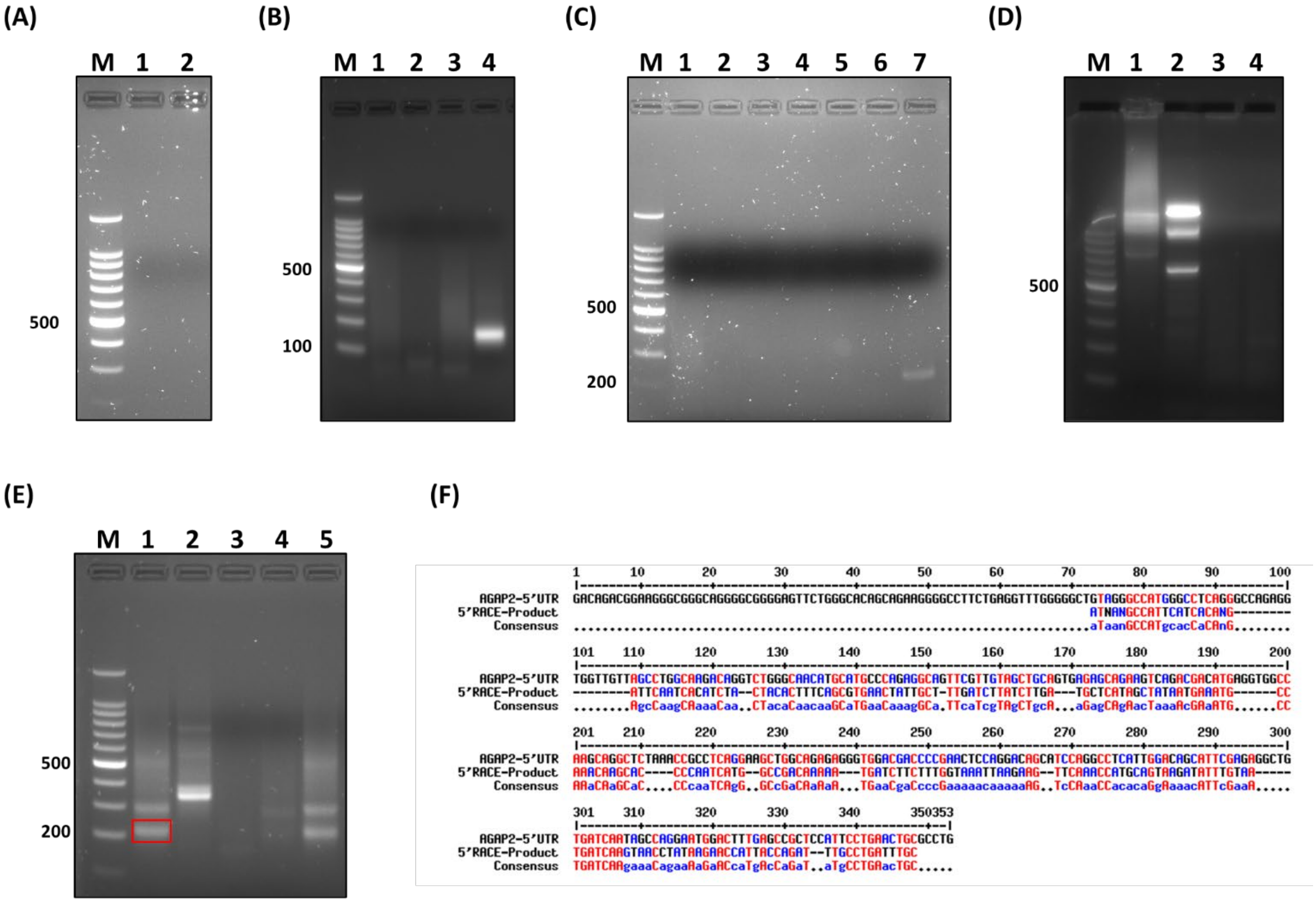
The conditions for the first-strand cDNA synthesis and nested amplification of dA-tailed cDNA were optimised to detect 5’ RACE products:

### 3.2.1 Optimisation of first-strand cDNA synthesis conditions:

The conversion of specific RNA sequences was carried by SuperScript II reverse transcriptase using a gene-specific primer 1 (*Table 2.7*). To optimise incubation timings and temperatures, four different conditions were selected: Incubation at 42°C for either 30 or 50 mins and incubating at 50°C for either 30 or 50 mins. We observed that longer incubation (50 mins) at a higher temperature (50°C) yielded a distinct cDNA product in the correct molecular weight range (~218 bp) (*Figure 3.3B*).

### 3.2.2 Optimisation of nested amplification conditions:

The PCR of the dA-tailed cDNA was performed using the provided AUAP primer and gene-specific primer 2 (*Table 2.7*). However, no or faint bands were detected after the electrophoretic separation of the amplified products, even after increasing the amount of cDNA for the PCR (*Figure 3.3C*). To enable detection of products, a nested PCR was carried out using AUAP and gene-specific nested primers (Table 2.7) using the PCR conditions (annealing at 63°C) outlined in the manufacturer’s protocol with a hot-start step. However, unexpected products of incorrect molecular weights were observed (*Figure 3.3D*). The nested amplification PCR conditions were optimised using a touch-down PCR starting with a higher initial annealing temperature (67°C) followed by a 0.1°C decrease in temperature after each cycle. Although the use of touchdown-conditions facilitated detection of amplified products in the correct molecular weight range (between 200-300 bp), multiple products were observed in some samples (*Figure 3.3E*) and Sanger sequencing of these products yield non-specific 5’ RACE products that did not correspond to known *AGAP2* 5’ UTR sequence (chr12:57742057-57742205) (Figure 3.3F).



```
                 1        10        20        30        40        50        60        70        80        90       100
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
   AGAP2-5'UTR   GACAGACGGAAGGGCGGGCAGGGGCGGGGAGTTCTGGGCACAGCAGAAGGGGCCTTCTGAGGTTTGGGGGCTGTAGGGCCATGGGCCTCAGGGCCAGAGG
5'RACE-Product                                                                            ATNANGCCATTCATCACANG--------
     Consensus   .........................................................................aTaanGCCATgcacCaCAnG........

               101      110       120       130       140       150       160       170       180       190       200
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
   AGAP2-5'UTR   TGGTTGTTAGCCTGGCAAGACAGGTCTGGGCAACATGCATGCCCAGAGGCAGTTCGTTGTAGCTGCAGTGAGAGCAGAAGTCAGACGACATGAGGTGGCC
5'RACE-Product   --------ATTCAATCACATCTA--CTACACTTTCAGCGTGAACTATTGCT-TTGATCTTATCTTGA---TGCTCATAGCTATAATGAAATG------CC
     Consensus   ........AgcCaagCAaaaCaa..CTacaCaacaaGCaTGaaCaaagGCa.TTcaTcgTAgCTgcA...aGagCAgAacTaaaAcGAaATG......CC

               201      210       220       230       240       250       260       270       280       290       300
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
   AGAP2-5'UTR   AAGCAGGCTCTAAACCGCCTCAGGAAGCTGGCAGAGAGGGTGGACGACCCCGAACTCCAGGACAGCATCCAGGCCTCATTGGACAGCATTCGAGAGGCTG
5'RACE-Product   AAACAAGCAC----CCCAATCATG--GCCGACAAAAA---TGATCTTCTTTGGTAAATTAAGAAG--TTCAAACCATGCAGTAAGATATTTGTAA-----
     Consensus   AAaCAaGCaC....CCcaaTCAgG..GCcGaCAaAaA...TGaaCgaCcccGaaaaacaaaaaAG..TcCAaaCCacacaGgAaaacATTcGaaA.....

               301      310       320       330       340       350353
                 |--------+---------+---------+---------+---------+--|
   AGAP2-5'UTR   TGATCAATAGCCAGGAATGGACTTTGAGCCGCTCCATTCCTGAACTGCGCCTG
5'RACE-Product   TGATCAAGTAACCTATAAGAACCATTACCAGAT--TTGCCTGATTTGC
     Consensus   TGATCAAgaaaCagaaAaGaACcaTgAcCaGaT..aTgCCTGAacTGC.....
```

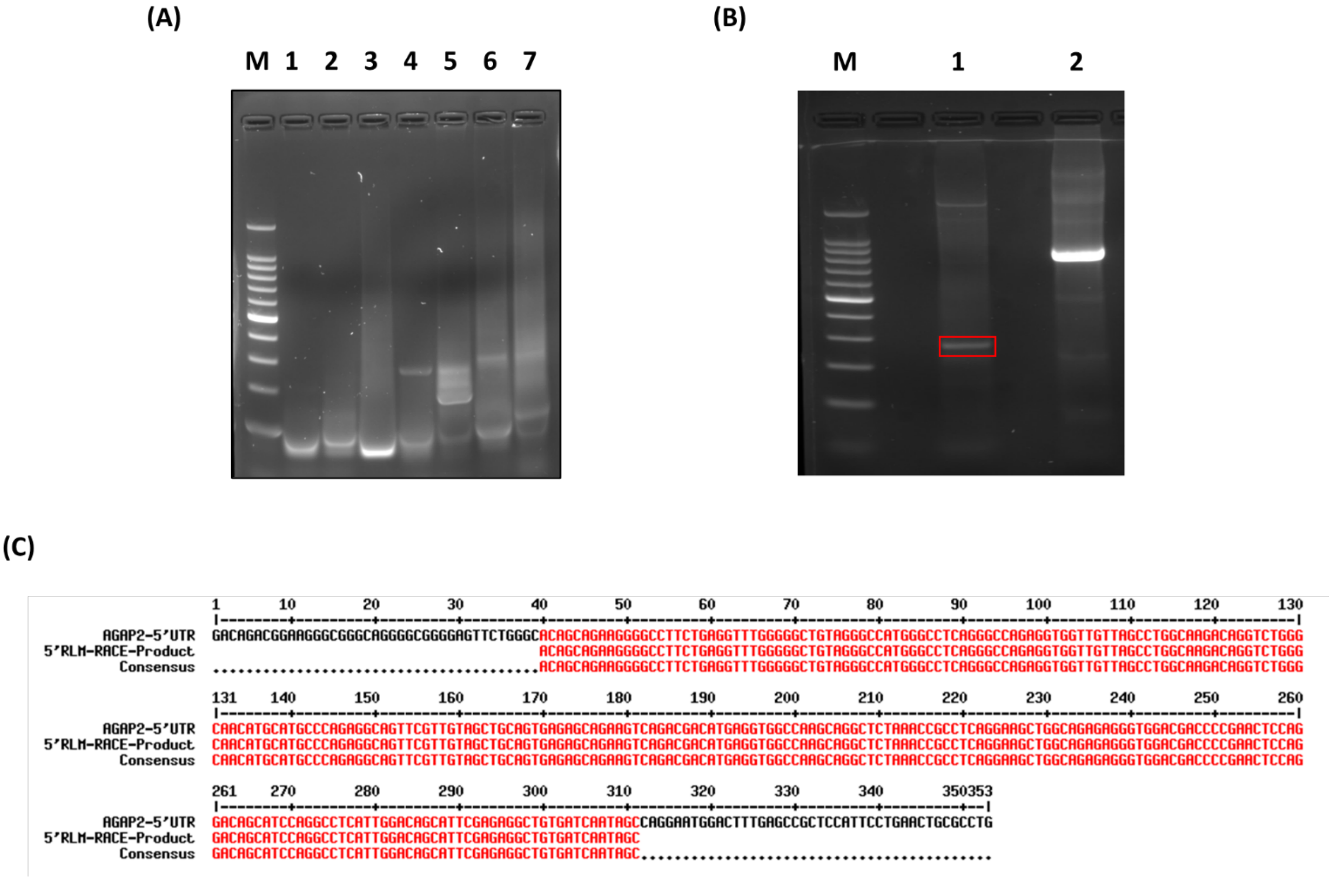
**Figure 3.3: Optimisation of 5' RACE with homopolymeric tailing. (A)** 2% Agarose gel electrophoresis to detect 5' RACE product following modified protocol and conditions used previously. Lane M:100 bp DNA ladder; Lane 1: KU812; Lane 2: DU145. **(B)** Reverse transcription of RNA using indicated incubation temperature and time and resolving the amplified first-strand cDNA product on the agarose gel. Lane M:100 bp DNA ladder; Lane 1: RT at 42°C for 30 mins; Lane2: RT at 42°C for 50 mins; Lane 3: RT at 50°C for 30 mins; Lane 4: RT at 50°C for 50 mins. The cDNA was amplified using AGAP2 forward primer (*Table 2.4*) and Gene-specific nested reverse primer (*Table 2.7*) using standard PCR condition (see 2.2.2.3). **(C)** Gel image showing PCR amplified products of dA-tailed cDNA using AUAP and gene-specific nested reverse primer (*Table 2.7*), different volumes of dA-tailed cDNA reaction mixture were tested. Lane M:100 bp DNA ladder; Lane 1: 2 µL dA-tailed reaction KU812; Lane 2: 2 µL dA-tailed reaction DU145; Lane 3: 5 µL dA-tailed reaction KU812; Lane 4: 5 µL dA-tailed reaction DU145; Lane 5: 7 µL dA-tailed reaction KU812; Lane 6: 7 µL dA-tailed reaction DU145; Lane 7: 10 µL dA tailed reaction mixture KU812. The concentrations of buffers, $MgCl_2$, and dNTP were adjusted accordingly if >5 µL of tailing reaction was used. **(D)** Agarose gel showing nested PCR amplification with AUAP and nested gene-specific reverse primer by setting the annealing temperature to 63°C. Lane M:100 bp DNA ladder; Lane 1: KU812; Lane 2: DU145; Lane 3: negative RT control; Lane 4: NTC. **(E)** Nested amplification performed using a touch-down PCR starting at 67°C followed by a 0.1°C decrease with each cycle. Lane M:100 bp DNA ladder; Lane 1: KU812; Lane 2: TCCS; Lane 3: DU145; Lane 4: PC3; Lane 5: KCL-22. The red box indicates the region of the gel that was excised and sequenced **(F)** Representative Multalin alignment of a sequenced PCR product in (E) lane 1 [red box]. The product was aligned to *AGAP2* 5' UTR region (chr12:57742057-57742205). The highly aligned bases are represented by red and bases that are poorly aligned are shown by blue.

## 3.2.3 Switching the approach to determine TSS:

Despite several attempts to optimise the 5' homopolymeric tailing kit, the genuine TSSs could not be captured. Additional optimisation experiments were also done including changing the amount of input RNA, increasing the incubation time for the tailing reaction, modifying the conditions for second-strand cDNA synthesis, adjusting $MgCl_2$

concentration, and modifying the primer concentrations used. However, none of these optimisation steps were successful. Examples of additional optimisations that were performed to detect the authentic 5’ RACE product using this kit are shown in *Figure 3.4A*. As noted in the figure, changing recommended primer concentrations for nested amplification increased the formation of primer-dimers (Lane: 1-3), and increasing the time for tailing reaction resulted in smearing (Lane: 6-7). It is also evident from the figure that changing the incubation time for second-strand cDNA synthesis did not detect specific 5’ RACE products (Lane: 4-5), the products amplified by both the conditions did not align to the *AGAP2* 5’ UTR sequence.

An alternative approach was therefore adapted to capture the 5’ end of the mRNA. A method based on 5' RNA ligase-mediated rapid amplification of cDNA ends (see section 2.2.5.2) was employed following the manufacture’s protocol. Using this approach, the 5’ RACE products were successfully amplified in the control RNA template and a test sample (PC3 cell line), and the product also aligned specifically to *AGAP2* 5’ UTR, detecting one of the FANTOM CAGE-verified TSS (*Figure 3.4B, 3.4C; Figure 3.12B*). The kit based on RNA ligase-mediated RACE was thereafter used in the project to characterise TSSs.



```
                    1       10        20        30        40        50        60        70        80        90        100       110       120       130
                    |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
       AGAP2-5’UTR  GACAGACGGAAGGGCGGGCAGGGGCGGGGAGTTCTGGGCACAGCAGAAGGGGCCTTCTGAGGTTTGGGGGCTGTAGGGCCATGGGCCTCAGGGCCAGAGGTGGTTGTTAGCCTGGCAAGACAGGTCTGGG
5’RLM-RACE-Product                                         ACAGCAGAAGGGGCCTTCTGAGGTTTGGGGGCTGTAGGGCCATGGGCCTCAGGGCCAGAGGTGGTTGTTAGCCTGGCAAGACAGGTCTGGG
         Consensus  .......................................ACAGCAGAAGGGGCCTTCTGAGGTTTGGGGGCTGTAGGGCCATGGGCCTCAGGGCCAGAGGTGGTTGTTAGCCTGGCAAGACAGGTCTGGG

                    131     140       150       160       170       180       190       200       210       220       230       240       250       260
                    |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
       AGAP2-5’UTR  CAACATGCATGCCCAGAGGCAGTTCGTTGTAGCTGCAGTGAGAGCAGAAGTCAGACGACATGAGGTGGCCAAGCAGGCTCTAAACCGCCTCAGGAAGCTGGCAGAGAGGGTGGACGACCCCGAACTCCAG
5’RLM-RACE-Product  CAACATGCATGCCCAGAGGCAGTTCGTTGTAGCTGCAGTGAGAGCAGAAGTCAGACGACATGAGGTGGCCAAGCAGGCTCTAAACCGCCTCAGGAAGCTGGCAGAGAGGGTGGACGACCCCGAACTCCAG
         Consensus  CAACATGCATGCCCAGAGGCAGTTCGTTGTAGCTGCAGTGAGAGCAGAAGTCAGACGACATGAGGTGGCCAAGCAGGCTCTAAACCGCCTCAGGAAGCTGGCAGAGAGGGTGGACGACCCCGAACTCCAG

                    261     270       280       290       300       310       320       330       340       350353
                    |--------+---------+---------+---------+---------+---------+---------+---------+---------+--|
       AGAP2-5’UTR  GACAGCATCCAGGCCTCATTGGACAGCATTCGAGAGGCTGTGATCAATAGCCAGGAATGGACTTTGAGCCGCTCCATTCCTGAACTGCGCCTG
5’RLM-RACE-Product  GACAGCATCCAGGCCTCATTGGACAGCATTCGAGAGGCTGTGATCAATAGC
         Consensus  GACAGCATCCAGGCCTCATTGGACAGCATTCGAGAGGCTGTGATCAATAGC..........................................
```

**Figure 3.4: Alternative approach to characterise TSSs. (A)** Additional optimisation of 5’ homopolymeric tailing kit without success. Lane M:100 bp DNA ladder; Lane 1: 5’ RACE produced amplified using the recommend concentration i.e. 400 nM AUAP and 400 nM GSP-2 without a hot start; Lane 2: PCR amplification using 400 nM AUAP and 400 nM GSP-2 with a hot start; Lane 3: PCR amplification using 600 nM AUAP and 600 nM GSP-2 with a hot start; Lane 4: Second-strand cDNA synthesis by incubating with the provided SuperScript II RT for 90 min at 50°C, the amplified product was sequenced and found to be non-specific. Lane 5: Second-strand cDNA synthesis by

incubating with SuperScript II RT for 90 mins at 42°C; Lane 6: Tailing of first-strand cDNA by incubating cDNA with TdT for 15 mins, the product amplified was non-specific; Lane 7: Tailing reaction incubated for 30 mins showing smearing. **(B)** 5’ RLM-RACE using the manufacture’s guidelines showing specific amplification of 5’ RACE products. Lane M:100 bp DNA ladder; Lane 1: Total RNA of PC3 cell line processed by following manufacturer’s protocol and amplified using nested primers (*Table 2.7*), the product amplified is in the correct molecular weight range (~300 bp: 30bp RNA oligo + 267 bp annotated *AGAP2* 5’ UTR); Lane 2: Control RNA template processed as above and the amplified using RLM-RACE 5’ forward primer and control primers B.1 and showed expected product (~872 bp), confirming optimal functioning of the kit. The red box indicates the region of the gel that was excised and cloned for sequencing **(C)** Alignment of the sequenced product from [B, Lane 1] showing specific alignment with *AGAP2* 5’ UTR, the TSS (first nucleotide) is a genuine CAGE-verified *AGAP2* TSS (*Figure 3.12B*).

## 3.3 Results and Discussion

### 3.3.1 Negative correlation between AGAP2 mRNA and protein expression levels:

To examine the discrepancy in AGAP2 mRNA and protein levels noted in the preliminary experiments, mRNA was extracted from PC cell lines (DU145, PC3 and LNCaP) and CML cell lines (KU812, TCCS, and KCL-22) and the *AGAP2* relative expression was analysed using qRT-PCR. The protein levels in these cell lines were measured by western blotting using isoform-specific AGAP2 antibody (Sigma-Aldrich). We noted a discrepancy in AGAP2 mRNA and protein levels in both groups of cell lines. In the CML cell lines, as noted previously, the relative mRNA expression was significantly higher compared to PC cell lines ($P < 0.0001$), however, the relative protein levels were significantly lower ($P = 0.012$). In contrast, the opposite occurred in PC cell lines (*Figure 3.5A*). The differences noted in the *AGAP2* mRNA levels in these two groups of cell lines matched the corresponding RNA-seq data in the Cancer Cell Line Encyclopaedia (Barretina *et al.,* 2012).

The discrepancy between AGAP2 mRNA and protein levels were not limited to PC and CML cell lines but also observed in different cancer cell lines such as HepG2, HuH7, MCF-7, PA-1, SKOV-3, U-2OS, RAJI, KG-1, and Kasumi-1. The pattern of inconsistent expression noticed in CML cell lines (high mRNA and low protein) was also observed in other cancer cell lines such as HepG2, PA-1, MCF-7, and Acute Myeloid Leukaemia cell lines (KG-1, KASUMI) *(Figure 3.1B)*, highlighting a cell line-specific regulation of *AGAP2* (*Figure 3.5B*).

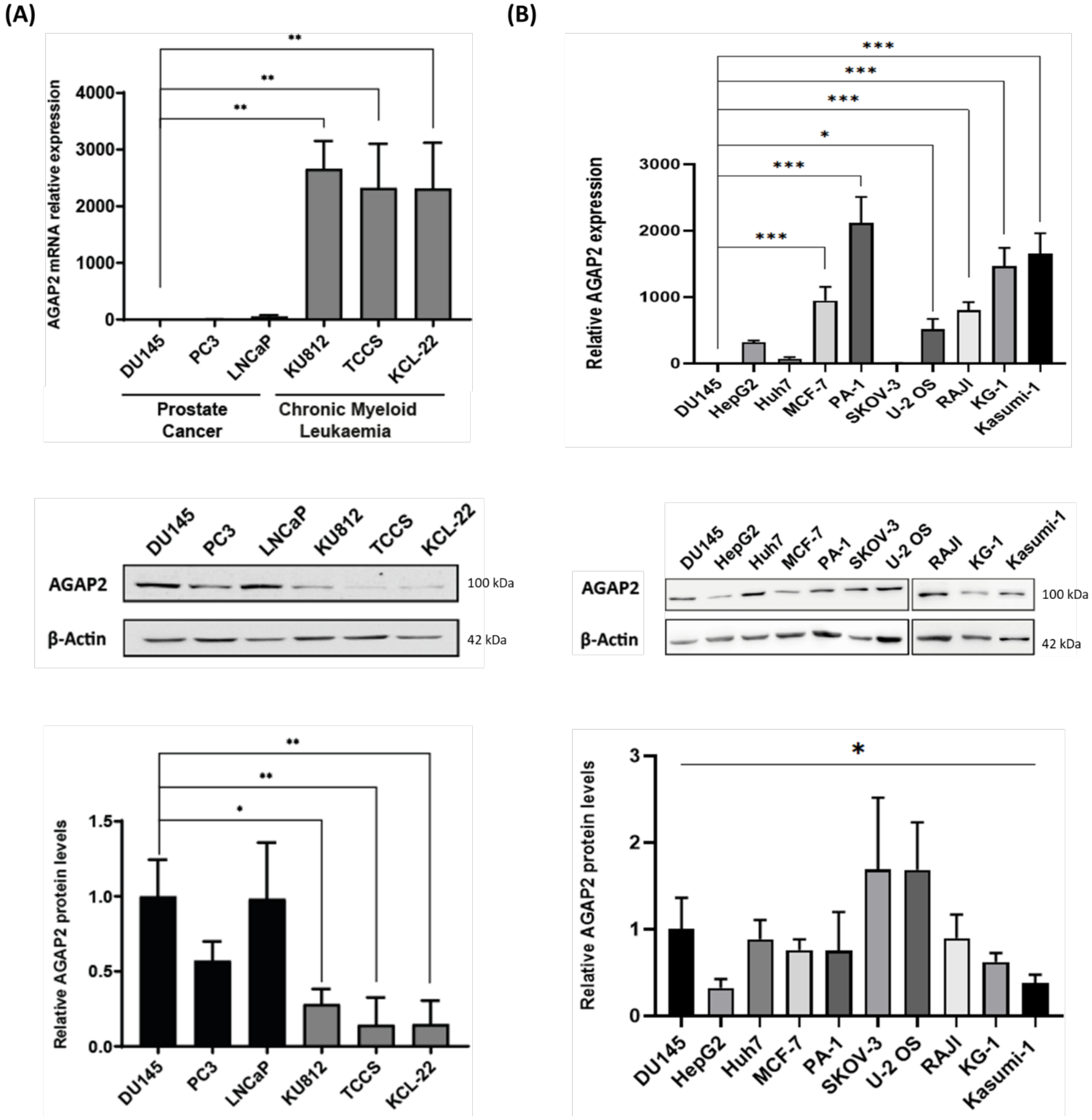


**Figure 3.5: Inconsistency in AGAP2 mRNA and protein.**AGAP2 mRNA basal levels (top) were measured in Prostate cancer cell lines (DU-145, PC3, LNCaP) and Chronic Myeloid Leukaemia cell lines (KU812, TCCS, KCL-22) by qRT-PCR. The values presented are normalised against the levels of the housekeeping gene (HPRT) and shown relative to the PC cell line (DU145). Statistical analyses were carried out by one-way ANOVA [F (5, 12) = 21.23, *P* < 0.0001)] with post-hoc Sidak's multiple comparison tests, *P*-values shown. AGAP2 protein levels (middle) were detected by resolving 50 µg of protein using 10% SDS-PAGE followed by immunoblotting with PIKE-A isoform-specific antibody. β-Actin was used as a loading control. Densitometry values (below) for the relative protein expression are presented below the blots. Differences were analysed using a Kruskal-Wallis [H (5) = 14.71, *P* = 0.012] followed by uncorrected Dunn's test, *P*-values shown. **(B)** AGAP2 relative mRNA levels in different cancer cell lines, normalised to HPRT and shown relative to DU145 (PC cell line). Statistical analyses for mRNA levels were carried out by one-way ANOVA [F (9, 20) = 41.30, *P* < 0.001)] with post-hoc Sidak's multiple comparison tests, *P*-values shown. AGAP2 protein levels are shown below with densitometry values for the relative protein expression. Differences between multiple samples were analysed using a Kruskal-Wallis test [H (9) = 20.85, *P* = 0.01]. All the data shown are the mean ± SD of three independent experiments (n=3), error bars represent S.D. (**P* < 0.05; ***P* < 0.01). Full representative immunoblots are presented in Appendix 1.

AGAP2 mRNA and protein levels showed an overall negative correlation. Combining AGAP2 mRNA and protein expression data from all the cell lines included in the study, we observed an anti-correlation between mRNA and protein levels (R=-0.64, *P* = 0.011) (*Figure 3.6A*). Restricting the data to PC and CML cell lines, a higher negative correlation was noted (R= -0.89, *P* = 0.016) (*Figure 3.6B*). As evident from *Figure 3.6A*, the leukaemia cell lines (KU812, TCCS, KCL-22, KG-1, RAJI, Kasumi-1), hepatocellular carcinoma cell line (HepG2), breast adenocarcinoma cell line (MCF-7), and ovary teratocarcinoma cell line (PA-1) predominately showed a discrepancy in AGAP2 expression, having higher levels of mRNA that did not correlate with protein abundance.

Other genes in the literature that followed the similar pattern of mRNA/protein discordance, as noted for *AGAP2* in CML cell lines, included *MX2*, *NES*, *TMC8*, *WISP1*, and *HLA-G* to name a few (Friedrich *et al.,* 2020, Swindell *et al.,* 2015, Yan *et al.,* 2018). In the case of *WISP1*, the discordant mRNA/protein expression in hepatocellular carcinoma cell lines was mediated by FAT10 (ubiquitin-like protein), tagging WISP1 for proteasomal degradation (Yan *et al.,* 2018). To evaluate the contribution of proteasomal degradation of *AGAP2* in CML cell lines, we used proteasomal inhibitors and examined the levels of AGAP2 protein in these cell lines.

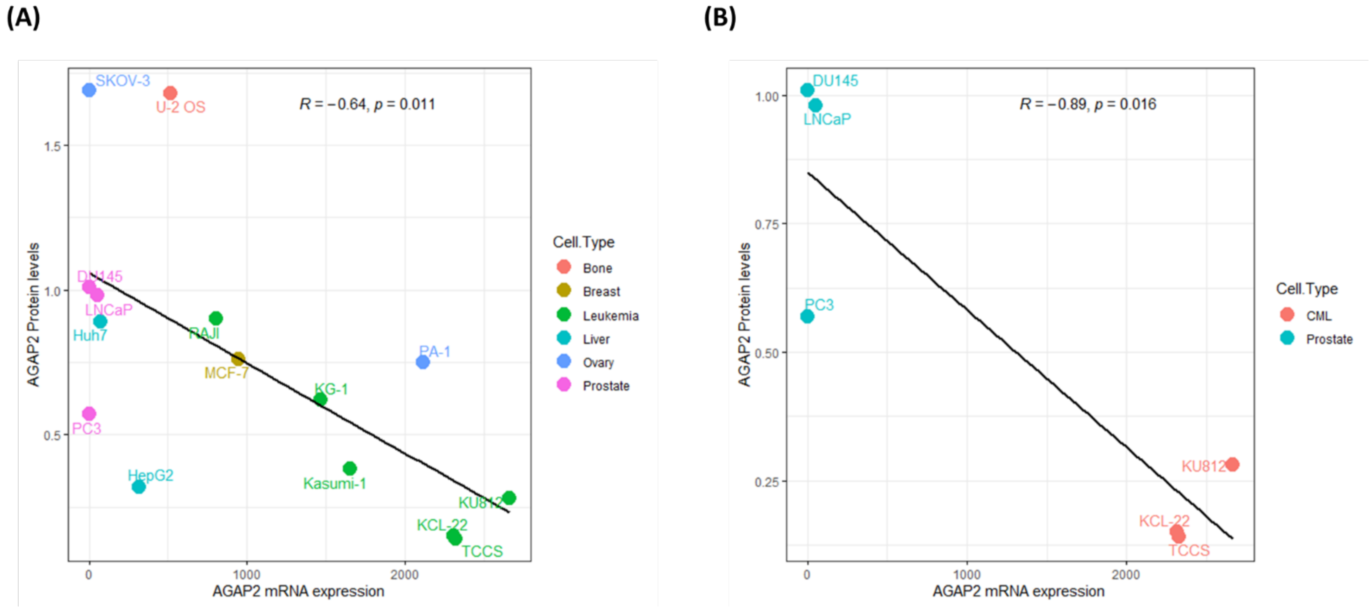


**Figure 3.6: Correlation between AGAP2 mRNA and protein. (A)** Correlation between mRNA (x-axis) and protein expression levels (y-axis) for AGAP2 in different cell types included in the study. An overall negative correlation (Pearson's R= -0.64, *P* = 0.011) is noted. **(B)** Correlation between AGAP2 mRNA and protein in PC and CML cell lines, a strong negative correlation (Pearson's R= -0.89, *P* = 0.016) is observed. The data in (A) and (B) are presented relative to DU145 (PC cell line).

### 3.3.2 Role of the proteasomal degradation in AGAP2 expression levels:

Protein degradation plays an important role in maintaining steady-state protein abundance (Reviewed in Vogel and Marcotte, 2012). Ubiquitin (Ub)–proteasome pathway (UPP) is responsible for degrading the majority of intracellular proteins (Rock *et al.,* 1994). The degradation of protein by UPP is carried out by covalently tagging ubiquitin molecules to a protein substrate which is recognised by 26S proteasome complex, breaking down the protein into constituent amino acids with the release of reusable ubiquitin (Reviewed in Ciechanover and Schwartz, 1998, Lecker *et al.,* 2006). The catalytic core of the proteasome complex facilitates three distinct proteolytic activities namely chymotrypsin-like, trypsin-like, and caspase-like proteasomal activity (Dick *et al.,* 1998, Kisselev *et al.,* 2003, Nussbaum *et al.,* 1998). The chymotrypsin-like proteolytic site has been shown to be the most important in protein breakdown (Heinemeyer *et al.,* 1997) and could be target using specific inhibitors that modulate its proteolytic activity. The proteasomal inhibitors MG132 and Bortezomib have been shown to selectively inhibit the chymotrypsin-like activity of the catalytic core of the proteasome complex (Kisselev *et al.,* 2006).

The degradation of AGAP2 protein by the UPP is currently unknown. The role of UPP in AGAP2 protein regulation has not been reported yet in the literature. To evaluate the contribution of UPP in AGAP2 protein degradation and to account for the discrepancy in AGAP2 mRNA and protein levels noted in the CML cell lines, two proteasomal inhibitors (MG132 and Bortezomib) were used to evaluate AGAP2 protein in the CML cell lines. It is expected that the AGAP2 protein levels in the CML cell lines would increase if the proteins were ubiquitinated and degraded by UPP.

### 3.3.2.1 Optimisation of MG132 and Bortezomib treatments:

For MG132, the literature cited the use of 1-50 µM concentration with a treatment duration ranging from 2 to 24 hours (Chui *et al.,* 2019, Estève *et al.,* 2009, Jung *et al.,* 2019, Noels *et al.,* 2009, Xu *et al.,* 2009). The treatment period of 4 hours was frequently used by various studies (Ahuja *et al.,* 2017, Chui *et al.,* 2019, Jung *et al.,* 2019, Noels *et al.,* 2009). The following concentrations for a 4-hour MG132 treatment were selected: 5 µM for KU812 and TCCS, and 50 µM for KCL-22.

For Bortezomib, treatments with a concentration range of 10 to 250 nM for 2-48 hours were reported in the literature (Alam *et al.,* 2017, Fan and You, 2020, Pitcher *et al.,* 2015, Van Herck *et al.,* 2009). A study by Yang *et al.* (2016) has tested different bortezomib concentrations on CML cell lines (KU812, K562) for the treatment duration of 6 hours. The following bortezomib concentrations was selected for a 6-hour treatment in CML cell lines: 200 nM for KU812, 10 nM for TCCS, and 50 nM for KCL-22.

The anti-ubiquitin antibody was used to optimise the selected proteasomal inhibitors concentrations. The relative levels of ubiquitinated proteins would increase under the optimal condition as the treatments would inhibit the proteolysis of the ubiquitin-tagged proteins, increasing their cellular levels compared to vehicle control. The selected treatment concentrations were evaluated using an anti-ubiquitin antibody (Table 2.6). As demonstrated in *Figure 3.7*, the treatment conditions for proteasomal inhibitors were optimal and resulted in an increase in the levels of ubiquitinated protein relative to the control. These concentrations were then used to analyse the proteasomal-mediated degradation of AGAP2.

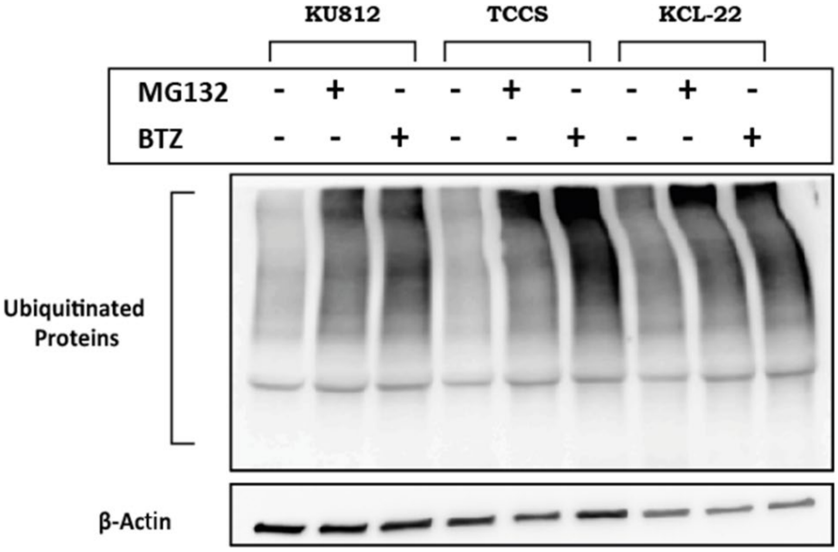


**Figure 3.7: Levels of ubiquitinated protein after treatment with proteasomal inhibitors.** Immunoblotting with ubiquitin antibody after treatment (+) with proteasomal inhibitors: MG132 [KU812 (5µM), TCCS (5µM), KCL-22 (50µM) for 4 hours] and Bortezomib [KU812 (200nM), TCCS (10nM), KCL-22 (50nM) for 6 hours] or untreated DMSO control (-). β-Actin was used as a loading control.

### 3.3.2.2 Examining AGAP2 proteasomal degradation in CML cell lines:

To evaluate the association between AGAP2 protein degradation by UPP and lower protein abundance noted in CML cell lines, the cells were treated with proteasomal inhibitors as optimised above. The analysis of western blotting data showed a limited role of UPP in modulating AGAP2 protein levels in CML cell lines, no statistically significant differences were observed in the cells treated with proteasomal inhibitors (MG132 or bortezomib) compared to untreated control (*Figure 3.8*). The data shows that ubiquitination has a limited contribution in AGAP2 degradation and could not account for the inconsistency in the mRNA and protein levels noted in the CML cell lines.

However, protein degradation is a highly elaborate and complex process that is influenced by many different factors. In addition to UPP, targeted protein degradation is also mediated by autophagy which is an important protein quality control mechanism in the context of misfolded and aggregated proteins (Kruse *et al.,* 2006). Other protein degradation mechanisms could be also responsible for controlling AGAP2 abundance and needs further evaluation.

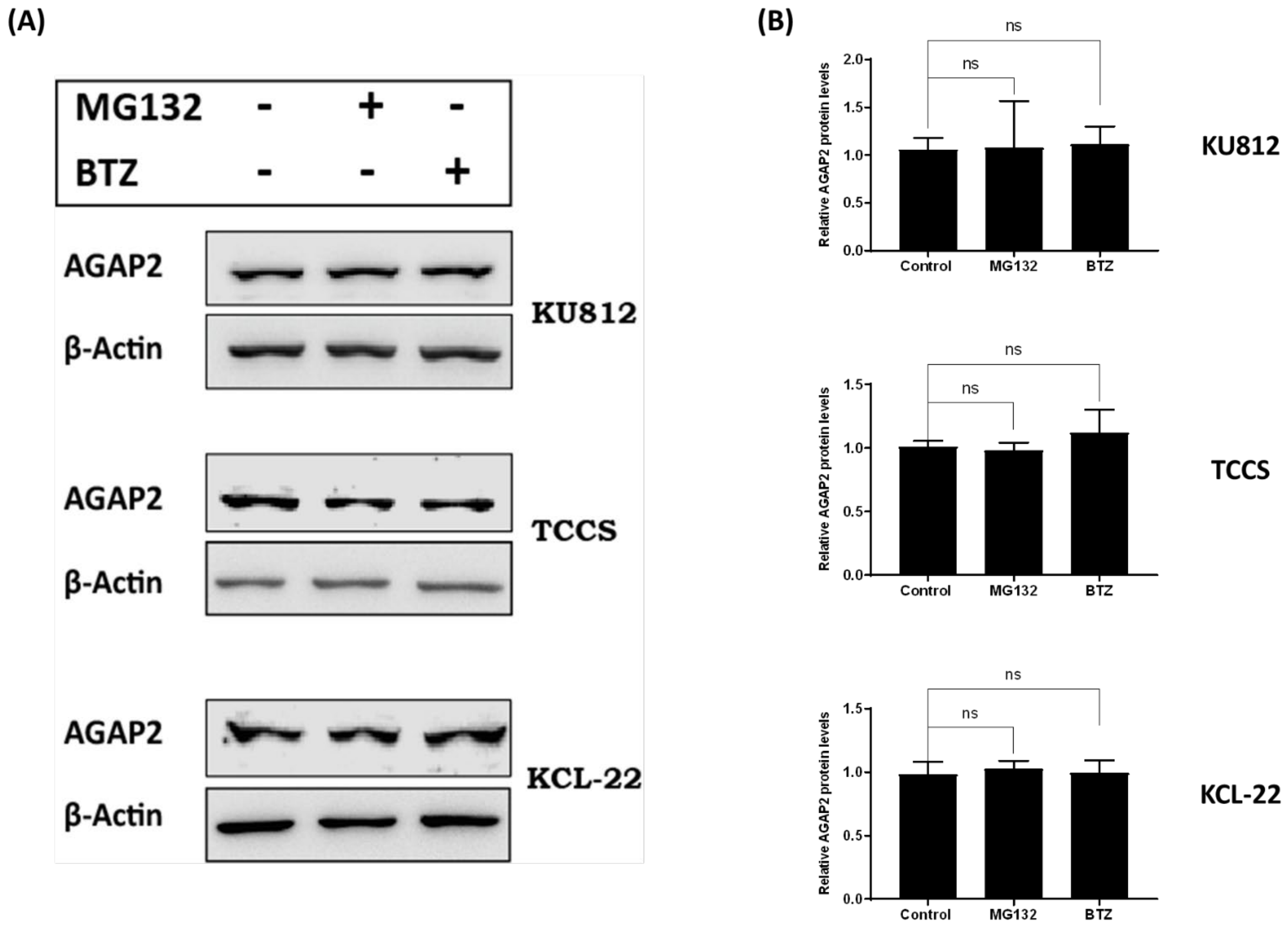


**Figure 3.8: Inhibiting AGAP2 proteasomal degradation in CML cell lines. (A)** Western blot analysis of AGAP2 in CML cell lines treated (+) with proteasomal inhibitors: MG132 [KU812 (5µM), TCCS (5µM), KCL-22 (50µM) for 4 hours] and Bortezomib [KU812 (200nM), TCCS (10nM), KCL-22 (100nM) for 6 hours] or untreated DMSO control (-). Densitometry values are presented in **(B)**. β-Actin was used as a loading control. The data shown are the mean ± SD of three independent experiments (n=3), Error bars represent S.D. Differences between the treated samples and untreated control were analysed by Mann–Whitney U test. BTZ: bortezomib.

### 3.3.3 Role of translation initiation factors in *AGAP2* expression discrepancies:

The levels of intracellular translation initiation factors are critical in regulating protein abundance, particularly in response to environmental stressors (Reviewed in Crawford and Pavitt, 2019, Spriggs *et al.,* 2010). A variety of eukaryotic initiation factors (eIFs) play an important role in facilitating cap-dependent translation initiation (section 1.1.2.1). The activity of these eIFs could be controlled by their phosphorylation state and the phosphorylation of eIF4E, eIF4B, eIF4G has been shown to positively correlate with translation (Hershey *et al.,* 2000).

In the current study, we evaluated the relative cellular levels of key eIFs in PC and CML cell lines to determine their contribution in regulating AGAP2 protein abundance. Our results showed no statistically significant differences in the basal levels of the different eIFs studied (*Figure 3.9*).

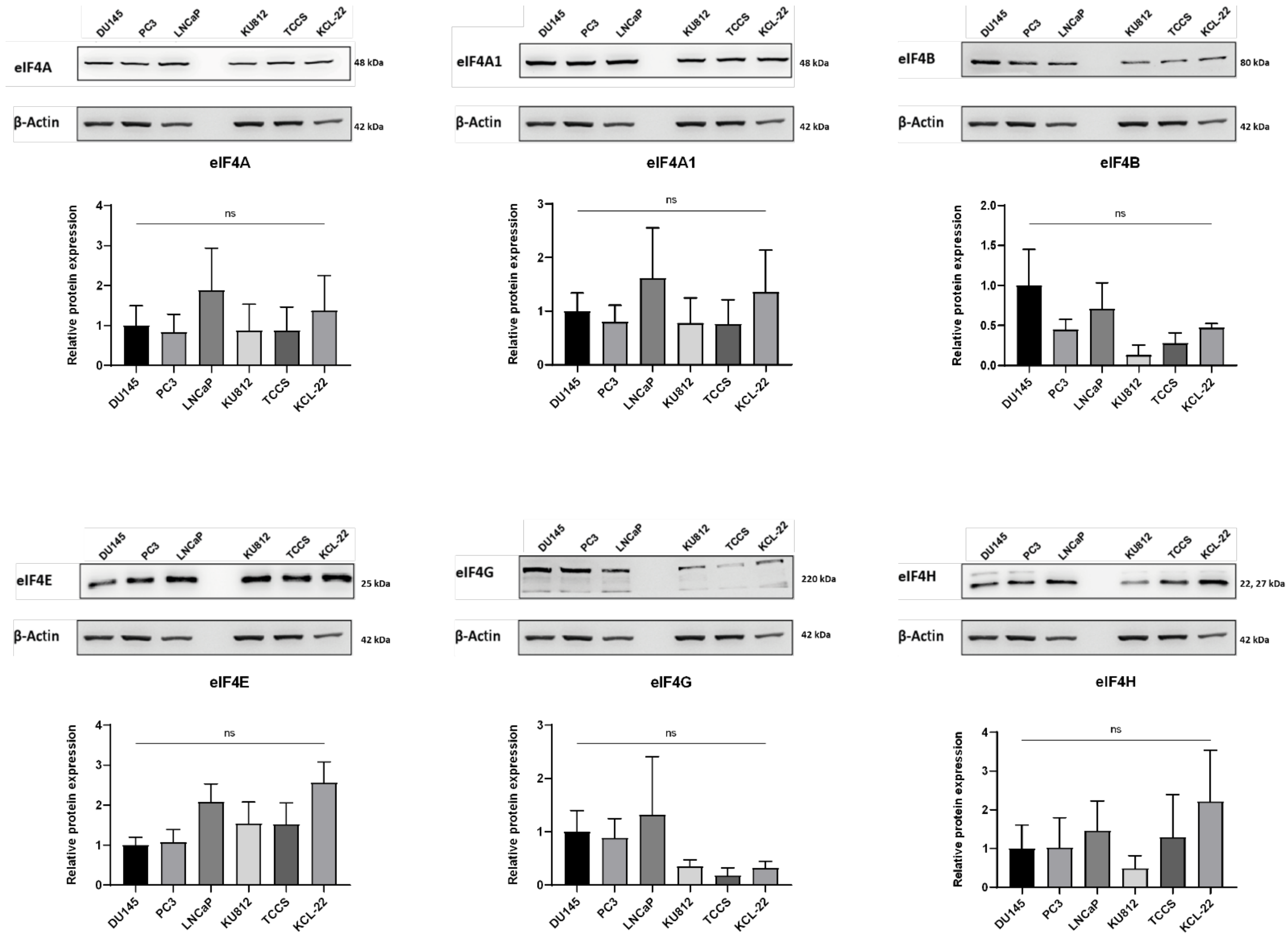


**Figure 3.9: Levels of selected translation initiation factors in PC and CML cell lines.** Western blotting of selected rate-limiting translation initiation factors (eIF4A, eIF4A1, eIF4B, eIF4E, eIF4G, eIF4H) detected by resolving 50 µg of protein using 4–20% precast protein gel followed by immunoblotting with specific antibodies. β-Actin was used as a loading control. All the proteins in the representative blot were detected by probing a single membrane. Densitometry values for the relative protein expression are represented below the blots. Differences were analysed using a Kruskal-Wallis followed by uncorrected Dunn's test, no statistically significant difference observed (*ns*: not significant). All the data shown are the mean ± SD of three independent experiments (n=3), error bars represent S.D.

We also analysed the phosphorylation state of selected eIFs such as eIF4E, eIF4G, eIF4B in different cell types included in our study (*Figure 3.10*). The analysis revealed no statistically significant differences in the phosphorylated eIFs (eIF4B, eIF4E and eIF4G) studied displayed (*Figure 3.10*).

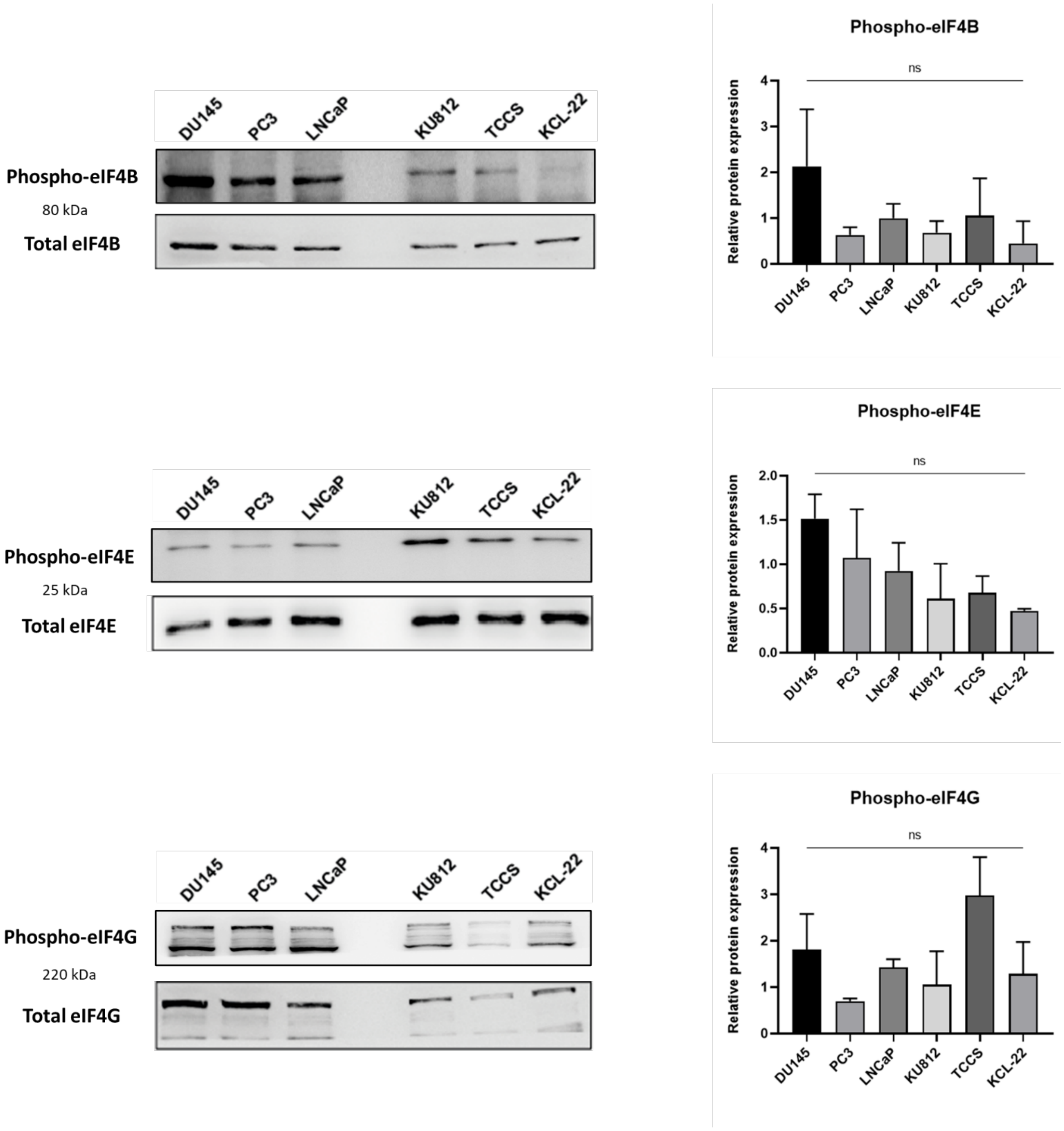


**Figure 3.10: Levels of selected phosphorylated translation initiation factors in PC and CML cell lines.** Western blotting of selected phosphorylated translation initiation factors (phospho-eIF4B, phospho-eIF4E, phospho-eIF4G) detected by resolving 50 µg of protein using 4–20% precast protein gel followed by immunoblotting with specific antibodies. The proteins were normalised using corresponding total eIF levels. Densitometry values for the relative protein expression are presented to the right of blots. Differences were analysed using a Kruskal-Wallis followed by uncorrected Dunn's test, *P*-values shown and is presented relative to DU145 (PC cell line). The data shown are the mean ± SD of three independent experiments (n=3), error bars represent S.D. (ns: not significant).

The absence of any significant differences in the relative basal levels of eIFs do not rule out their role in modulating differential protein abundance. The process of translation requires a sophisticated interplay between various eIFs, cofactors, signalling cascades, and ncRNA (Reviewed in Sonenberg and Hinnebusch, 2009). The features in the mRNA 5' and 3' UTR such as secondary structures, RBP (RNA binding protein), and Poly(A)-binding protein (PABP) are also involved in controlling translation activation and efficiency (Wilkie *et al.,* 2003). Furthermore, the helicases activity of eIF4A and different DExH-Box protein members are required to scan mRNA with highly structured 5' UTR and govern its translation (Pisareva *et al.,* 2008). These complex and coordinated interactions shape mRNA-specific translational profile. Further studies are required to comprehensively understand the complex translation initiation regulation and dissect the possible contributions of different eIFs in mediating inconsistency in mRNA and protein levels.

## 3.3.4 TSS usage for *AGAP2* in PC and CML cell lines:

During the transcription initiation, transcription factors bind to the specific sequences within the promoter region of the gene and recruit the RNA polymerase (RNAP) holoenzyme, followed by the unwinding of promoter DNA and selection of a TSS (section 1.1.1). The selection of the TSS by RNAP within the core promoter region is dynamic and variable in both prokaryotic and eukaryotic organisms (Qiu *et al.,* 2020, Vvedenskaya *et al.,* 2016). The diversity in TSS selection has been previously demonstrated for a large number of human cell lines and tissues. (Carninci *et al.,* 2006, Suzuki *et al.,* 2001). Our preliminary work has identified a potential of differential TSS usage for *AGAP2* in PC and CML cell lines (*Figure 3.2C*). In the current project, we have characterised the distribution of *AGAP2* TSSs within the core promoter region of PC and CML cell lines.

### 3.3.4.1 Characterising AGAP2 TSSs distribution:

The *AGAP2* TSSs in PC and CML cell lines were determined using 5' RLM-RACE (section 2.2.5.2). Following the manufacturer's instructions, the RNA was dephosphorylated, decapped, ligated with an RNA oligo, reverse transcribed using *AGAP2* gene-specific primer 1, and amplified using nested primers (*Table 2.7*). The amplified 5' RACE products were resolved using agarose gel electrophoresis showing the expected product (*Figure 3.11A*). The bands were excised, purified and cloned into a vector, transformed into chemically competent bacteria, and were grown in the presence of selecting medium. Single clones from each cell line were sequenced to provide a snapshot of *AGAP2* TSS distribution patterns in PC and CML cell lines (*Figure 3.11B*). It could be noted from the figure that the cloned sequence consistently showed an upstream (>100 bp from ATG start codon) TSSs in CML cell lines relative to PC cell lines. To characterise the distribution of *AGAP2* TSSs in detail, ten clones from KU812 (CML) and DU145 (PC) were sequenced, and the results are presented relative to the ATG start codon (*Figure 3.11C*). As evident from *Figure 3.11C*, the TSSs in DU145 were largely clustered around 90 bp upstream of the ATG start codon. On the contrary, the TSSs distribution in KU812 was relatively broad with prominent transcription starting around 90 bp from ATG, similar to DU145. However, earlier TSSs (130 bp upstream of ATG start codon) were also noted in KU812.

Our experiment showed that there is a population of TSSs which are differentially distributed in the core promoter region of *AGAP2* in PC and CML cell lines. In contrast to one optimal TSS in PC and CML cell lines, there are multiple starting sites that are differentially selected in these cell lines and produce a heterogenous 5' UTR population. The TSSs in the CML cell line (KU812) is broadly distributed compared to PC (DU145) cell line with several starting positions including earlier starting sites that encode a longer 5' UTR containing the G4 forming consensus sequence

(*Figure 3.11C, Figure 3.12B*). Such different patterns of TSS distribution confirm the cell-specific expression of TSSs, as also noted by other studies (Kawaji *et al.,* 2006, Ohmiya *et al.,* 2014).

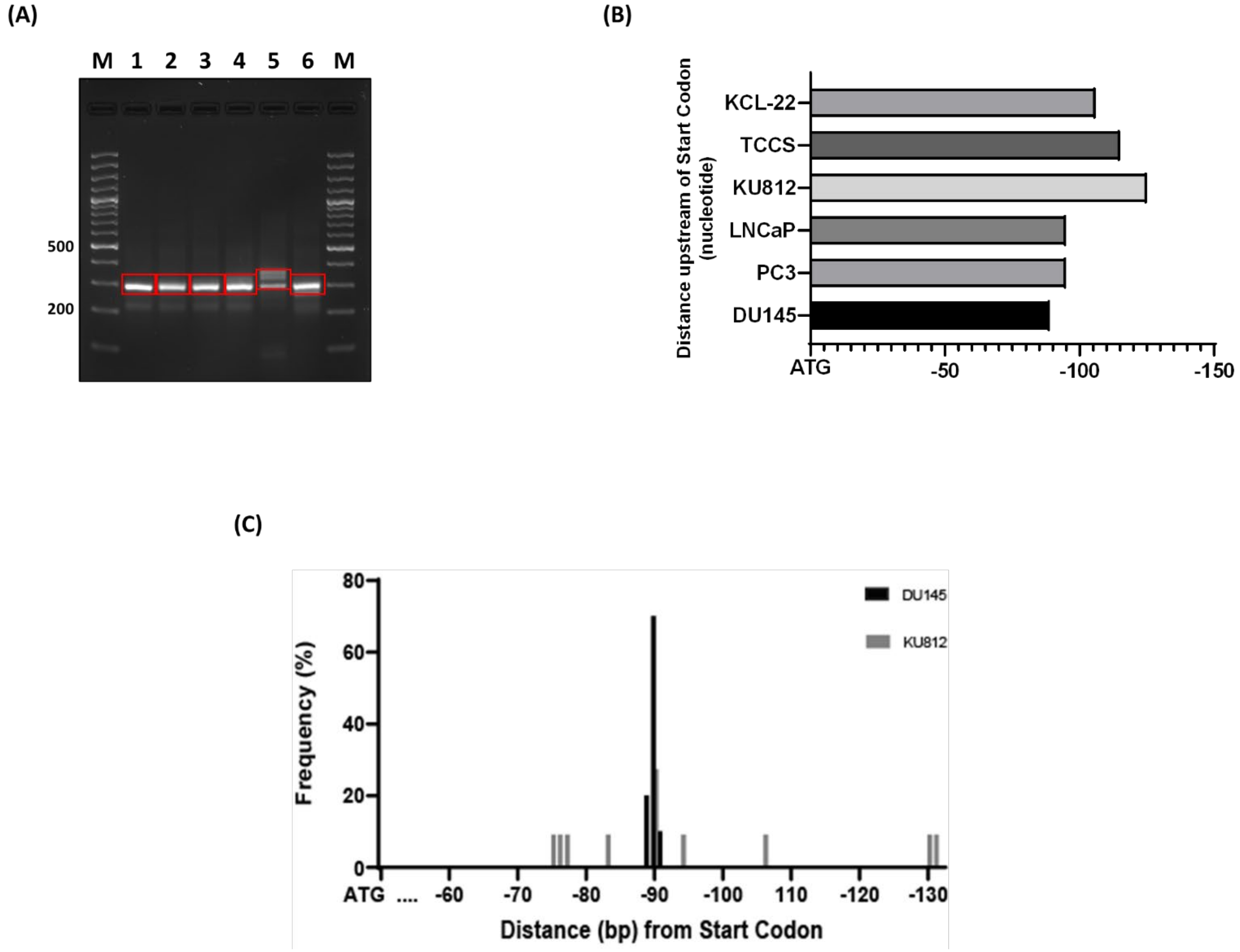


**Figure 3.11: Distribution of AGAP2 TSSs in PC and CML cell lines. (A)** 5' RACE products from PC and CML cell lines resolved on 2% agarose gel and showing expected bands at the correct molecular weight (~300 bp: 30bp RNA oligo + 267 bp annotated *AGAP2* 5' UTR). The red box indicates the region of the gel that was excised and cloned for sequencing. Lane M:100 bp DNA ladder; Lane 1-6: DU145, PC3, LNCaP, KU812, TCCS, KCL-22. **(B)** The position of the *AGAP2* TSSs in PC and CML cell lines relative to ATG start codon, the TSSs were obtained by cloning the excised bands in (A) and sequencing a single purified clone (n=1) of the transformed bacteria cultured on selecting agar plate. The sequence was aligned to the *AGAP2* 5' UTR region (chr12:57742057-57742205) and the position of starting nucleotide was considered as the transcription starting site. **(C)** The array of TSSs in DU145 (PC) and KU812 (CML) cell line determined by 5' RLM-RACE. The distribution frequency is plotted as the percentage at nucleotide position relative to ATG start codon (n=10).

### 3.3.4.2 Validation of TSSs obtained by 5' RLM-RACE:

The TSSs obtained in the current study were validated by using the TSS data of the corresponding cell line in the FANTOM database. The FANTOM database is the largest collection of transcription initiation data, consisting of more than 1,000 human and mouse primary cells, tissues, and cancer cell lines (Lizio *et al.,* 2015). The TSSs in the FANTOM database have been determined using HeliScope Cap Analysis of Gene Expression (CAGE) technology. It uses a biotinylated cap-trapping approach to capture the 5' ends of the cDNAs, which are subsequently converted to short tags (20-27nt long) and directly sequenced using single-molecule sequencing (Kanamori-Katayama *et al.,* 2011). The CAGE tag starting sites represent the actual mRNA TSSs with base pair-level accuracy (Kawaji *et al.,* 2014).

The TSS distribution characterised by 5’ RLM-RACE displayed a good concordance with the TSS distribution pattern defined in the FANTOM CAGE database (*Figure 3.12A*). As evident from the figure, there is a good overlap between TSSs identified in our study and TSSs detected by the high throughput HeliScopeCAGE sequencing approach. Our study also identified some novel TSSs that were not annotated in the FANTOM database. It is worth mentioning that the TSS distribution characterised by our study and in the FANTOM CAGE database are associated with a single core promoter region of *AGAP2*. These TSSs associated with the core promoter region were clustered together and separated from other clusters by a clear genomic space.

The cumulative distribution profile of *AGAP2* TSS in the FANTOM database is presented in *Figure 3.12B*. It could be noted that AGAP2 uses a variety of TSSs with prominent transcription initiating around 90 bp upstream of ATG start codon, as also noted in our study. The TSS with the highest frequency is usually categorised as the major TSS and is commonly used as the starting position of *AGAP2* mRNA transcript in different RefSeq databases like NCBI (NM_014770.4) and Ensembl (ENST00000257897.7). As also highlighted in *Figure 3.12B*, earlier TSSs (~130 bp upstream of ATG start codon) were frequently used in KU812 and consequently increased the length of the 5’ UTR of *AGAP2* mRNA.  We noticed that the extra nucleotides incorporated in the 5’ UTR by the earlier TSS selection had runs of guanine nucleic acid that followed the consensus for a G quadruplex structure ($G_{3+}N_{1\text{-}7}G_{3+}N_{1\text{-}7}G_{3+}N_{1\text{-}7}G_{3+}$). This alternatively transcribed G quadruplex forming sequence could have a regulatory role and it is further explored in Chapter 4.

Our study, for the first time, provided evidence that alternative TSS selection within a single cluster could encode a regulatory element in the 5’ UTR region. Previous studies have identified a variety of regulatory elements such as upstream open reading frame (uORF) and upstream start codon (uAUG), RNA secondary structures, RNA G quadruplexes, and internal ribosomal entry site (IRES) that are encoded in the mRNA 5’ UTR isoforms by alternative TSS selection (Blaschke *et al.,* 2003, Hollerer *et al.,* 2019, Pickering and Willis, 2005, Pozner *et al.,* 2000). However, the alternative TSSs encoding these features belonged to different clusters and were separated by greater than 500 bp. On the contrary, the alternative TSS selection in KU812 that yield a G quadruplex forming sequence in the longer 5’ UTR isoform was only 36 bp upstream to the TSS in DU145 (*Figure 3.12B*). This novel layer of gene expression regulation has not been described yet and is further delineated in Chapter 4.

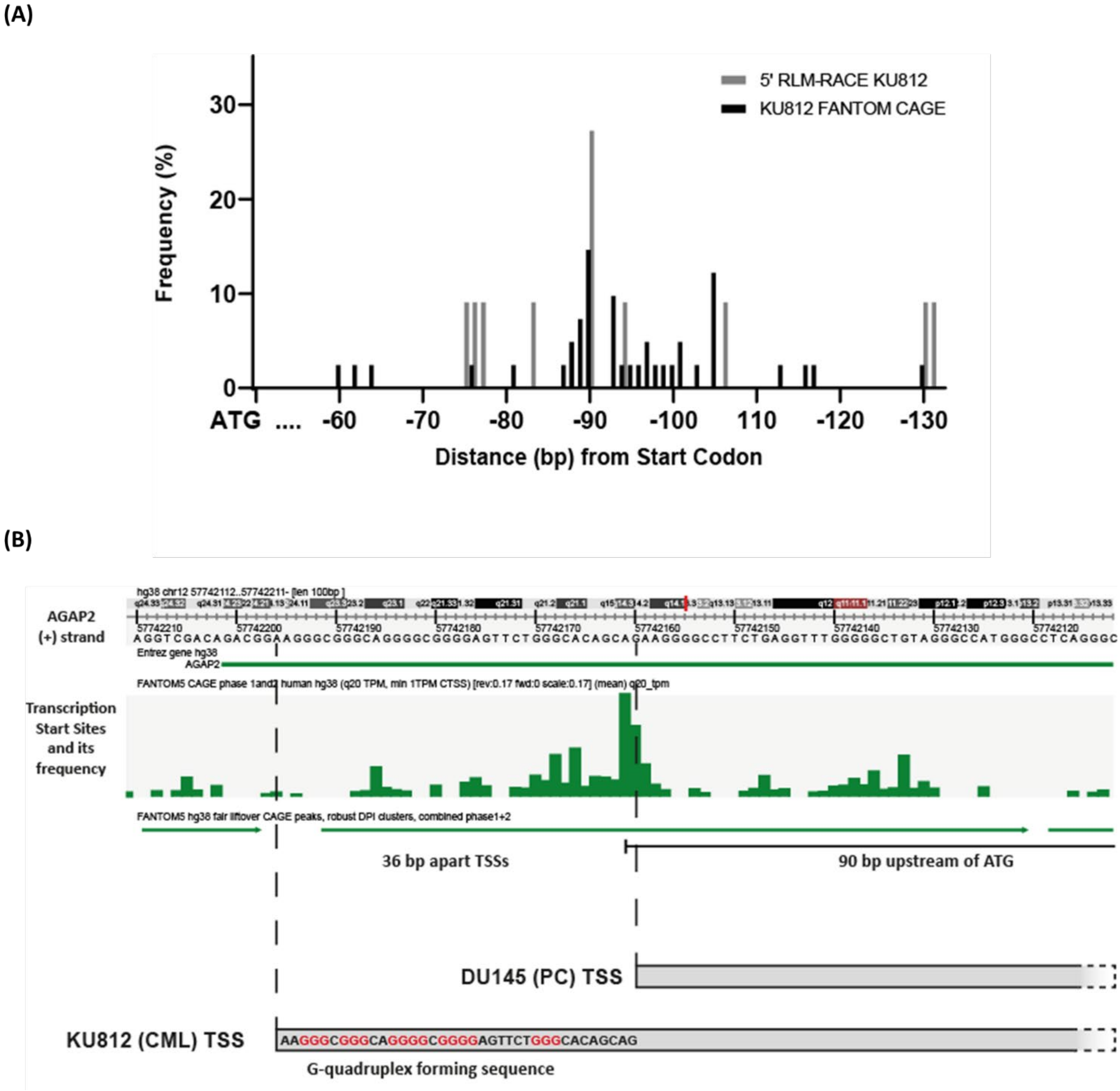


**Figure 3.12: Overlay between TSSs obtained by RLM-RACE and FANTOM database. (A)** TSSs for AGAP2 in KU812 obtained by 5' RLM-RACE is plotted alongside the TSSs reported by the HeliScopeCAGE technology used in the FANTOM database. The TSS distribution profiles determined in our study showed a good agreement with FANTOM CAGE TSSs. The relative frequencies are shown as a percentage at the nucleotide positions from the start codon. **(B)** Image derived from ZENBU genome browser, a data exploration tool for FANTOM database (Lizio *et al.,* 2015), showing the distribution and usage of different *AGAP2* TSSs within a core promoter region. The alternative TSSs mapped by 5' RLM-RACE in KU812 (CML) cell line that encoded extra nucleotides in the 5' UTR of KU812 containing the G quadruplex forming sequence is shown below.

### 3.3.5 Expression of longer (G4-containing) 5' UTR isoforms in PC and CML cell lines:

We investigated the relative abundance of the longer 5' UTR isoforms corresponding to the earlier TSSs in PC and CML cell lines. For this purpose, all the TSS isoforms encoding a G4 consensus sequence in the *AGAP2* mRNA 5' UTR were classified as longer 5' UTR. We designed a nested PCR reaction to amplify the longer 5' UTR, originating from upstream TSSs, in both the cell lines (*Figure 3.13A*). A nested PCR was used because the usage of upstream TSSs were relatively lower compared to the major starting sites in both the cell lines (*Figure 3.11C*, *3.12B*), and also the level of *AGAP2* mRNA in PC cell lines was significantly lower and didn't produce any amplification signals during the first PCR (data not shown). The levels of longer 5' UTR isoforms obtained by the nested PCR amplification were normalised by the entire 5' UTR population. The nested PCR product was verified using Sanger sequencing to confirm the expected amplicon. The results are presented in *Figure 3.13B* and reveal significantly higher levels of longer G4-containing 5' UTR isoforms in CML cell lines relative to PC cell lines (*P* < 0.0001). For validation, the nested PCR was also performed on the 5' RLM-RACE amplified cDNA which likewise displayed a significant increase in the level of longer 5' UTR in the CML cell lines (*P* < 0.001) (*Figure 3.13C*).

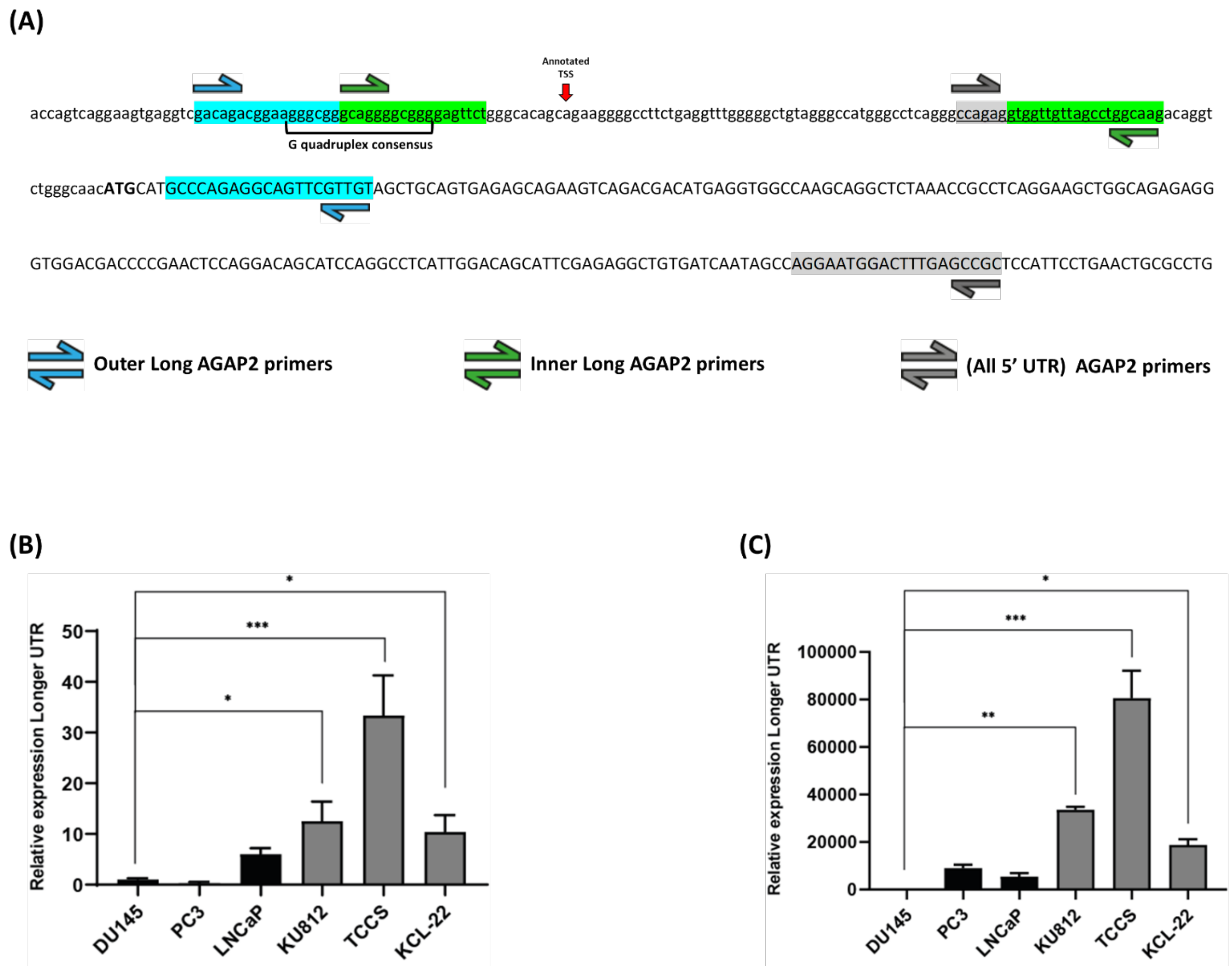


**Figure 3.13: Levels of longer G4-containing *AGAP2* 5' UTR isoforms in PC and CML cell lines. (A)** Nested PCR amplification to detect longer *AGAP2* 5' UTR isoform in PC and CML cell lines. The first PCR was performed using outer long AGAP2 primers. The PCR product was diluted 50 folds and amplified again using inner long AGAP2 primers. To detect the entire UTR population, the forward primer was designed closer to the coding sequence. The sequences of the primers are listed in *Table 2.4*. The figure shows the first 373 nucleotides of *AGAP2* mRNA transcribed from the furthest upstream TSS annotated in the FANTOM database. The coding sequence is presented in UPPERCASE (with ATG marked in bold) and the 5' UTR is presented in the lowercase. The annotated TSS in NCBI and Ensembl is shown by the red arrow. Primer sequences are shown by respective coloured box and the AGAP2 primer in the overlapped region is shown by underline. **(B)** Relative levels of *AGAP2* mRNA with longer 5' UTR in PC and CML cell lines. The levels are normalised by the entire 5' UTR population and presented relative to DU145 (PC cell line). The data is the mean ± SD of three independent experiments (n=3). Statistical differences

were analysed by one-way ANOVA [F (5, 12) = 29.35, *P* < 0.0001)] with post-hoc Sidak's multiple comparison tests, *P*-values shown. **(C)** Validation of the levels of longer *AGAP2* 5' UTR in PC and CML cell lines using RACE- amplified cDNA (n=2). The levels are normalised by the entire 5' UTR population and presented relative to DU145. Statistical differences were analysed by one-way ANOVA [F (5, 6) = 73.65, *P* < 0.001)] with post-hoc Sidak's multiple comparison tests, *P*-values shown. (*P < 0.05; **P < 0.01; ***P < 0.001).

The higher levels of longer 5' UTR isoforms in the CML cell lines also confirm the relatively higher usage of upstream TSSs in these cell lines. However, these upstream, G4 encoding TSSs, represent a fraction of all the TSSs used in the cluster. The transcription in both the PC and CML cell line is predominately driven from the TSSs located around -90 bp upstream of the ATG start codon, also termed as major TSS, that do not incorporate the G4 forming sequence in the 5' UTR. Even though the G4 forming TSSs are not the major TSS, they could still render a large portion of the mRNA population susceptible to TSS-G4 mediated regulation.

The distribution shape and usage of TSSs with the core promoter region defines the distinct group of promoters and promoter context: single dominated peak class of promoters are represented by concentrated transcription start positions with a single dominant TSS and are associated with TATA box; whereas the remaining classes are categories of broadly distributed TSSs with dominant or multimodal peak distributions and are shown to be associated with CpG islands (Carninci *et al.,* 2006, Carninci *et al.,* 2005). In our study, we noted comparable usage of *AGAP2* minimal promoter region in PC and CML cell lines and also documented the relevance of SP1 and ATRA on *AGAP2* transcription activation in both groups of cell lines (*Figure 3.1*). Taking these results into account, the alternative promoter usage in PC and CML cell lines is unlikely to explain the observed heterogeneity in the 5' UTR.

The selection of TSS within a cluster is a dynamic process dictated by a variety of factors including sequence features, ncRNA, various stimuli, and epigenetic factors (Javahery *et al.,* 1994, Jiang and Pugh, 2009, Leenen *et al.,* 2016, Turowski and Tollervey, 2020), also see section 3.3.6. It is plausible that a TSS selection could switch under the influence of these factors and change the relative proportions of 5' UTR isoforms. The work by Leenen *et al.* (2016) has verified this microvariability in the TSS location induced by different stimuli. In the case of *AGAP2*, the selection of earlier G4 forming TSSs could be also induced by different factors and require further evaluation.

### 3.3.6 Differentially expressed TSSs in PC and CML cell lines:

The TSS distribution within a core promoter region varies among different tissues and exhibits cell-specific distribution profiles (Kawaji *et al.,* 2006, Ohmiya *et al.,* 2014). We have noted a distinct TSS distribution profile in PC and CML cell lines using 5' RLM-RACE (*Figure 3.11C*). The differential TSS distribution patterns produce heterogeneity in the 5' UTR isoforms transcribed from the gene (*Figure 3.12B*). Differentially transcribed 5' UTR isoforms could contain regulatory features that might impact mRNA translation potential. To identify other genes with differentially distributed TSSs, we tested a publicly available bioinformatics tool (SEASTAR) to generate a list of genes with a differential distribution pattern of TSSs in PC and CML cell lines.

### 3.3.6.1 Systematic Evaluation of Alternative Start site in RNA (SEASTAR) Algorithm:

The SEASTAR is a computational pipeline developed by Qin *et al.* (2018) to identify alternative TSS and quantify their expression levels using the RNA sequencing data. The SEASTAR pipeline is implemented by initially producing a processed transcript assembly by annotating the aligned sequencing reads. Then, a non-redundant set of the first exon is generated, and a logistic regression model is applied to identify the bona fide first exon. The SEASTAR subsequently compares the usage of the alternative first exon (AFE) across the distinct biological conditions and applies rMATS statistical methods (Shen *et al.,* 2014) to identify differential first exon usage. Using a DaPars change point statistical model (Xia *et al.,* 2014), it also tests whether tandem TSSs exists within the first exon and detect TSSs that significantly differ between two sample groups. The technical overview of the algorithm is illustrated in *Figure 3.14*.

The SEASTAR pipeline is written in Bash and R script and can be accessed using Github (https://github.com/Xinglab/SEASTAR). The transcription start sites identified by the SEASTAR package showed a good agreement with the FANTOM CAGE starting sites and also bears the hallmark of active promoters (Qin *et al.,* 2018). In the current project, the SEASTAR pipeline was employed to leverage the RNA sequencing data to gain insights into alternative TSSs usage and distribution profiles in PC and CML cell lines.

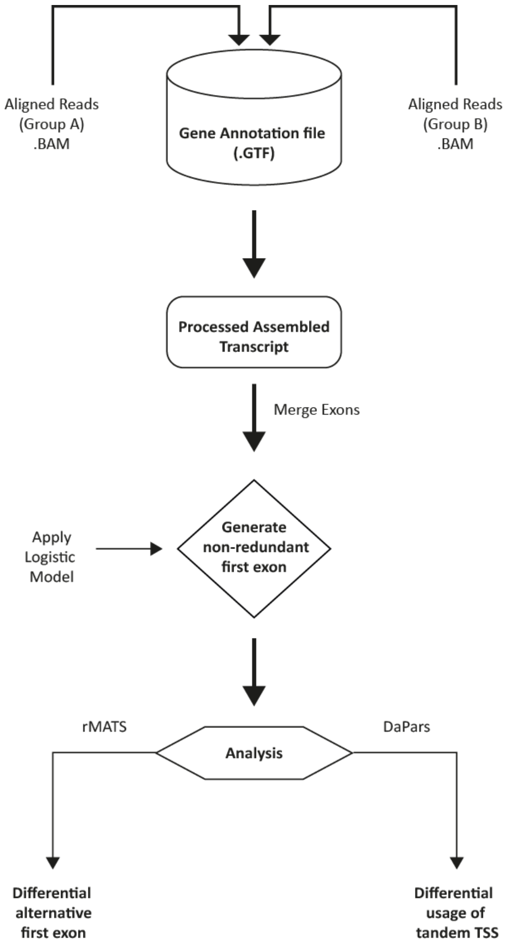


**Figure 3.14: Flowchart of SEASTAR algorithm.** The SEASTAR computational pipeline to identify differentially expressed alternative transcription initiation sites. The algorithm input the aligned reads in the (.bam) format and

use the existing transcriptome annotation (.GTF) to construct a processed transcript assembly. The overlapping putative first exons are merged to generate a non-redundant set of first exons. The bona fide first exon is then determined using a logistic regression model and the differential usage of the alternative first exon is determined using the rMATS statistical method (Shen *et al.,* 2014). The alternative tandem TSSs within the first exon is detected by the DaPars statistical model (adapted from Xia *et al.,* 2014) and the relative proportions of significantly different TSSs are computed.

### 3.3.6.1.1 Executing SEASTAR script:

The aligned RNA sequencing data in the form of binary alignment matrix (.bam) format was used as an input for the SEASTAR package. The RNA-seq data of PC cell lines (DU145, PC3, LNCaP) was assigned to group A and that of CML cell lines (KU812, TCCS, KCL-22) was assigned to group B. The human gene annotation (GRCh38.p13, release 33) was used to annotate the aligned transcripts. The comprehensive gene annotation was downloaded from (https://www.gencodegenes.org/human/release_33.html) in the Gene Transfer Format (.GTF). The package was executed in Ubuntu 18.04 using the following command:

```
bash ./SEASTAR.sh
-A ./DU145.alignments.bam,./PC3.alignments.bam,./LNCap.alignments.bam
-B /KU812.alignments.bam,./TCCS.alignments.bam,./KCL22.alignments.bam
-o ./OUTPUT
-G ./annotation/gencode.v33.annotation.gtf
-i ./annotation/hg38.chrom.sizes
-s /home/arif/SEASTAR/bowtieindex/hg38.fa
-c 0.1 -p 1 -b U -d 100 -S U
```

The annotation of genes and transcripts was carried out in the reference mode (-G), skipping reference annotation-based transcript assembly (RABT) to identify novel TSSs. The distance (-d) among the TSSs derived from the same promoter region was set to 100 bp (default values). The other parameters including the splicing difference cut off were also set to default values.

### 3.3.6.1.2 SEASTAR Output:

The output generated by the SEASTAR pipeline is comprised of two distinct tables analysing significantly different AFE (*Table 3.1*) and alternative tandem TSSs (*Table 3.2*) between PC and CML cell lines. The differentially expressed first exon was analysed by counting the RNA-seq reads that map to the first exon relative to the other exons in the gene. The differential usage of the first exon was then evaluated by rMATS statistical analysis. For analysis of alternative tandem TSS using DaPars statistical model, the length of the first exons were recorded in PC and CML cell lines and a switch point between the shortening and lengthening region was predicted, splitting the first exon into two regions. Using the rMATS statistical method as above, the significant differences in the relative usage of two split regions were computed.

The following output columns were generated:

- **TSS ID:** The ID of each TSS. Each ID represents one clustering of the raw TSSs with a distance less than 100bp.

- **IC SAMPLE 1:**
  - Table 1: counts of the first exons for Group A.
  - Table 2: counts of the RNA-seq reads in longer region of split exon for Group A, replicates are separated by a comma.
- **SC SAMPLE 1:**
  - Table 1: counts of other first exons in the respective gene for Group A.
  - Table 2: counts of the RNA-seq reads in shorter region of split exon for Group A, replicates are separated by a comma.
- **IC SAMPLE 2:**
  - Table 1: counts of the first exons for Group B.
  - Table 2: counts of the RNA-seq reads in longer region of split exon for Group B, replicates are separated by a comma.
- **SC SAMPLE 2:**
  - Table 1: counts of other first exons in the respective gene for Group B.
  - Table 2: counts of the RNA-seq reads in shorter region of split exon for Group B, replicates are separated by a comma.
- **IncFormLen:**
  - Table 1: length of the first exon understudy, used for normalization
  - Table 2: length of the longer region of split exon, used for normalization
- **SkipFormLen:**
  - Table 1: average length of other first exons in the respective gene, used for normalization
  - Table 2: length of shorter region of split exon, used for normalization
- **IncLevel1**: usage ratio for Group A replicates (comma separated), calculated from normalized counts using formula: (ICSAMPLE1/incformLen)/(ICSAMPLE1/IncformLen + SC_SAMPLE1/SkipformLen)
- **IncLevel2**: usage ratio for Group B replicates (comma separated) calculated from normalized counts using formula: (ICSAMPLE2/incformLen)/(ICSAMPLE2/IncformLen + SC_SAMPLE2/SkipformLen)
- **IncLevelDifference:** average difference between IncLevel1 and IncLevel2
- **FDR:** adjusted *P*-value using the Benjamini-Hochberg method.

### 3.3.6.1.3 Interpretation of the SEASTAR analysis and relevance to the current study:

The analysis of differentially utilised first exons demonstrated 7 genes with significantly different AFE usage (*Table 3.1*). *Table 3.1* also presents the list of the top 30 gene transcripts showing heterogenous first exon in the PC group relative to the CML group. The list of top 15 tandem TSSs that showed differences in PC and CML cell are shown in *Table 3.2*. The analysis of differential tandem TSSs usage revealed no statistically significant differences between PC and CML cell lines. It could be attributed to combining the sequencing data from different related cell lines into two groups for analysis. The PC or CML group had sequencing data from three different but connected cell lines instead of technical replicates, resulting in data variability and loss of any significant differences. The DaPars analysis was repeated using only DU145 cell line for the PC group and KU812 cell line for the CML group and the output generated a list of 4 tandem TSSs (TSS ID: TSS66290, TSS101970, TSS110127, TSS109986; *Table 3.2*) that showed significant differential distribution.

However, the analysis of AFE and alternative tandem TSSs by the SEASTAR analysis pipeline had limited applicability in the current study. The type of TSSs examined by the SEASTAR pipeline does not fit the context of the current study. The alternative TSSs examined by the SEASTAR are in the form of alternative first exons and alternative tandem TSSs (*Figure 3.15A*). The TSSs generated by alternative first exon usage are separated by a large genomic space and would likely change the ORF (*Figure 3.15A*). On the contrary, the focus of the current project is to study the differential distribution of TSSs that change the length of 5’ UTR by a few nucleotides (< 100 bp) without changing the ORF.

For tandem TSSs, the SEASTAR analyses the use of different TSSs within a given first exon that are separated by >100 bp and could also change the principal ORF in some instances (*Figure 3.15A*). This grouping of tandem TSSs that are within 100 bp of each other might overlook the contribution of TSS variants separated by just a few nucleotides. The SEASTAR analysis was also rerun by changing the default clustering of TSS to 50 bp, but it did not change the output generated possibly because of the sequencing depth in the 5’ UTR region. Although the RNA sequencing in the current project was carried out at a sufficient depth (48-53 million reads per sample), the coverage depth at the 5’ UTR region was lower (2-3X).

Therefore, to analyse relevant TSSs that fit within the focus of the study, a customised in-house bioinformatics pipeline was designed that analysed the differential distribution of TSSs within the core promoter region that does not change the ORF (see below and Chapter 5).

**Table 3.1: List of differential alternative first exons between PC and CML cell lines curated by the SEASTAR pipeline.**

| ID | Gene name | Locus | Length | Exon start | Exon end | IC SAMPLE 1 | SC SAMPLE 1 | IC SAMPLE 2 | SC SAMPLE2 | Inc-Form-Len | Skip-Form-Len | IncLevel1 | IncLevel2 | IncLevel-Difference | PValue | FDR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TSS27274 | SEPTIN9 | chr17:77281498-77500593 | 3733 | 77281437 | 77281554 | 26,73,14 | 55,96,17 | 1,0,0 | 150,88,74 | 118 | 118 | 0.321,0.432,0.452 | 0.007,0.0,0.0 | 0.399 | 0.00E+00 | 0.00 |
| TSS43024 | ATP2C1 | chr3:130893967-131003107 | 5067 | 1.31E+08 | 1.31E+08 | 22,14,15 | 0,0,4 | 13,12,10 | 62,62,22 | 370 | 370 | 1.0,1.0,0.789 | 0.173,0.162,0.313 | 0.714 | 8.91E-12 | 4.84E-08 |
| TSS1367 | MACF1 | chr1:39409502-39428092 | 5767 | 39409503 | 39414477 | 10,371,2 10,327 | 389,291,392 | 205,201,152 | 13,831,264,670 | 4975 | 4975 | 0.727,0.806,0.455 | 0.129,0.137,0.185 | 0.512 | 4.98E-11 | 1.80E-07 |
| TSS43023 | ATP2C1 | chr3:130850594-131003150 | 4888 | 1.31E+08 | 1.31E+08 | 0,0,4 | 20,13,14 | 56,56,20 | 12,11,9 | 334 | 334 | 0.0,0.0,0.222 | 0.824,0.836,0.69 | -0.709 | 1.05E-10 | 2.84E-07 |
| TSS103723 | TFDP2 | chr3:141944427-142149521 | 9857 | 1.42E+08 | 1.42E+08 | 16,6,11 | 8,4,12 | 4,6,3 | 32,190,25 | 362 | 362 | 0.667,0.6,0.478 | 0.111,0.031,0.107 | 0.499 | 2.01E-09 | 4.36E-06 |
| TSS94677 | SBNO2 | chr19:1127709-1169146 | 914 | 1168694 | 1169146 | 15,23,63 | 42,33,93 | 2,4,7 | 61,56,98 | 453 | 453 | 0.263,0.411,0.404 | 0.032,0.067,0.067 | 0.304 | 2.55E-06 | 0.004617549 |
| TSS121548 | EXOSC3 | chr9:37779713-37785064 | 1813 | 37784721 | 37785064 | 60,33,52 | 40,24,40 | 23,66,17 | 85,150,45 | 344 | 344 | 0.6,0.579,0.565 | 0.213,0.306,0.274 | 0.317 | 5.11E-06 | 0.007928907 |
| TSS41870 | MAPKAPK3 | chr3:50611519-50613900 | 558 | 50611520 | 50611737 | 1,3,3 | 30,86,43 | 14,44,49 | 50,48,48 | 218 | 218 | 0.032,0.034,0.065 | 0.219,0.478,0.505 | -0.357 | 3.46E-05 | 0.046926539 |
| TSS52494 | UTRN | chr6:144344100-144426328 | 572 | 1.44E+08 | 1.44E+08 | 2,0,1 | 27,48,118 | 10,4,16 | 2,13,7 | 204 | 204 | 0.069,0.0,0.008 | 0.833,0.235,0.696 | -0.562 | 9.77E-05 | 0.117928509 |
| TSS3215 | PDE4DIP | chr1:148952340-148986477 | 8824 | 1.49E+08 | 1.49E+08 | 207,167,72 | 0,15,0 | 44,49,7 | 61,11,11 | 1565 | 1565 | 1.0,0.918,1.0 | 0.419,0.817,0.389 | 0.431 | 1.11E-04 | 0.120431507 |
| TSS27278 | SEPTIN9 | chr17:77320312-77402415 | 553 | 77320313 | 77320348 | 10,11,1 | 15,41,8 | 33,20,17 | 13,7,5 | 36 | 36 | 0.4,0.212,0.111 | 0.717,0.741,0.773 | -0.503 | 1.40E-04 | 0.137794634 |
| TSS36847 | LRRFIP1 | chr2:237692215-237765915 | 4370 | 2.38E+08 | 2.38E+08 | 25,30,13 | 18,29,56 | 63,52,24 | 7,9,7 | 377 | 377 | 0.581,0.508,0.188 | 0.9,0.852,0.774 | -0.416 | 2.74E-04 | 0.247708893 |
| TSS121855 | MLLT3 | chr9:20341668-20622499 | 6724 | 20622245 | 20622518 | 9,1,12 | 0,0,3 | 5,8,5 | 38,23,5 | 274 | 274 | 1.0,1.0,0.8 | 0.116,0.258,0.5 | 0.642 | 5.41E-04 | 0.393269159 |
| TSS82787 | NEDD4 | chr15:55826921-55917131 | 7235 | 55915284 | 55917131 | 308,14,12 | 78,31,0 | 6,14,0 | 63,86,8 | 1848 | 1848 | 0.798,0.311,1.0 | 0.087,0.14,0.0 | 0.627 | 5.43E-04 | 0.393269159 |
| TSS87831 | FAM117A | chr17:49710994-49788592 | 1522 | 49788502 | 49788989 | 14,17,15 | 6,6,10 | 31,30,19 | 94,179,22 | 488 | 488 | 0.7,0.739,0.6 | 0.248,0.144,0.463 | 0.395 | 4.96E-04 | 0.393269159 |
| TSS75525 | NFYB | chr12:104119354-104138017 | 1727 | 1.04E+08 | 1.04E+08 | 20,13,7 | 6,19,13 | 5,3,5 | 38,51,57 | 642 | 642 | 0.769,0.406,0.35 | 0.116,0.056,0.081 | 0.424 | 8.74E-04 | 0.593341246 |
| TSS41873 | MAPKAPK3 | chr3:50617150-50649291 | 2537 | 50617131 | 50617226 | 13,38,19 | 0,1,1 | 22,21,21 | 6,19,22 | 96 | 96 | 1.0,0.974,0.95 | 0.786,0.525,0.488 | 0.375 | 1.07E-03 | 0.680241565 |
| TSS16948 | ARHGAP5 | chr14:32077303-32159728 | 9589 | 32077074 | 32077435 | 63,26,11 | 145,40,33 | 4,7,9 | 0,3,0 | 362 | 362 | 0.303,0.394,0.25 | 1.0,0.7,1.0 | -0.584 | 1.32E-03 | 0.796244285 |
| TSS105386 | SHISA5 | chr3:48467797-48504050 | 2385 | 48504019 | 48504431 | 13,28,32 | 50,87,77 | 2,12,0 | 44,114,106 | 413 | 413 | 0.206,0.243,0.294 | 0.043,0.095,0.0 | 0.202 | 1.62E-03 | 0.925194179 |
| TSS100031 | FRG1CP | chr20:28563849-28602773 | 2344 | chr20 | 28602609 | 28602843 | 4,22,12 | 0,0,0 | 7,5,9 | 0,0,0 | 235 | 235 | 1.0,1.0,1.0 | 1.0,1.0,1.0 | 0 | 1.00E+00 |

**Table 3.2: List of alternatively used tandem TSSs in PC and CML cell line generated by the SEASTAR pipeline.**

| TSS ID | IC SAMPLE 1 | SC SAMPLE 1 | IC SAMPLE 2 | SC SAMPLE 2 | IncFormLen | SkipFormLen | IncLevel1 | IncLevel2 | IncLevelDifference | PValue | FDR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **TSS66290.1** | 233,17,144 | 64,100,152 | 15,51,16 | 115,186,136 | 678 | 202 | 0.52,0.048,0.22 | 0.037,0.076,0.034 | 0.214 | 0.297 | 1 |
| **TSS66290.2** | 64,100,152 | 233,17,144 | 115,186,136 | 15,51,16 | 202 | 678 | 0.48,0.952,0.78 | 0.963,0.924,0.966 | -0.214 | 0.297 | 1 |
| **TSS101970.1** | 375,361,326 | 148,139,125 | 461,373,152 | 188,172,75 | 337 | 203 | 0.604,0.61,0.611 | 0.596,0.566,0.55 | 0.038 | 1 | 1 |
| **TSS101970.2** | 148,139,125 | 375,361,326 | 188,172,75 | 461,373,152 | 203 | 337 | 0.396,0.39,0.389 | 0.404,0.434,0.45 | -0.038 | 1 | 1 |
| **TSS104426.1** | 107,78,112 | 49,25,49 | 130,126,85 | 34,48,44 | 202 | 164 | 0.639,0.717,0.65 | 0.756,0.681,0.611 | -0.014 | 1 | 1 |
| **TSS104426.2** | 49,25,49 | 107,78,112 | 34,48,44 | 130,126,85 | 164 | 202 | 0.361,0.283,0.35 | 0.244,0.319,0.389 | 0.014 | 1 | 1 |
| **TSS104755.1** | 8,526,261,987 | 476,327,882 | 154,032,072,050 | 7,651,544,849 | 466 | 417 | 0.616,0.631,0.668 | 0.643,0.65,0.684 | -0.021 | 1 | 1 |
| **TSS104755.2** | 476,327,882 | 8,526,261,987 | 7,651,544,849 | 154,032,072,050 | 417 | 466 | 0.384,0.369,0.332 | 0.357,0.35,0.316 | 0.021 | 1 | 1 |
| **TSS104913.1** | 355,218,448 | 109,328,108 | 337,764,484 | 96,211,139 | 279 | 357 | 0.806,0.46,0.841 | 0.818,0.822,0.817 | -0.117 | 1 | 1 |
| **TSS104913.2** | 109,328,108 | 355,218,448 | 96,211,139 | 337,764,484 | 357 | 279 | 0.194,0.54,0.159 | 0.182,0.178,0.183 | 0.117 | 1 | 1 |
| **TSS108247.1** | 214,260,214 | 57,58,54 | 300,205,449 | 81,66,120 | 202 | 82 | 0.604,0.645,0.617 | 0.601,0.558,0.603 | 0.035 | 1 | 1 |
| **TSS108247.2** | 57,58,54 | 214,260,214 | 81,66,120 | 300,205,449 | 82 | 202 | 0.396,0.355,0.383 | 0.399,0.442,0.397 | -0.035 | 1 | 1 |
| **TSS109087.1** | 340,247,441 | 50,48,76 | 537,228,161 | 82,38,19 | 360 | 172 | 0.765,0.711,0.735 | 0.758,0.741,0.802 | -0.03 | 1 | 1 |
| **TSS109087.2** | 50,48,76 | 340,247,441 | 82,38,19 | 537,228,161 | 172 | 360 | 0.235,0.289,0.265 | 0.242,0.259,0.198 | 0.03 | 1 | 1 |
| **TSS109986.1** | 8,291,091,218 | 8,591,189,184 | 214,138,201 | 236,128,211 | 202 | 407 | 0.66,0.649,0.705 | 0.646,0.685,0.657 | 0.009 | 1 | 1 |
| **TSS109986.2** | 8,591,189,184 | 8,291,091,218 | 236,128,211 | 214,138,201 | 407 | 202 | 0.34,0.351,0.295 | 0.354,0.315,0.343 | -0.009 | 1 | 1 |
| **TSS110127.1** | 109,198,117 | 28,69,39 | 89,102,77 | 29,39,29 | 202 | 159 | 0.754,0.693,0.703 | 0.707,0.673,0.676 | 0.031 | 1 | 1 |
| **TSS110127.2** | 28,69,39 | 109,198,117 | 29,39,29 | 89,102,77 | 159 | 202 | 0.246,0.307,0.297 | 0.293,0.327,0.324 | -0.031 | 1 | 1 |
| **TSS115354.1** | 128,188,198 | 20,20,23 | 337,519,335 | 40,63,43 | 299 | 232 | 0.832,0.879,0.87 | 0.867,0.865,0.858 | -0.003 | 1 | 1 |
| **TSS115354.2** | 20,20,23 | 128,188,198 | 40,63,43 | 337,519,335 | 232 | 299 | 0.168,0.121,0.13 | 0.133,0.135,0.142 | 0.003 | 1 | 1 |
| **TSS118917.1** | 173,514,021,552 | 627,596,098,385,549 | 167,815,351,811 | 553,687,237,884,617 | 202 | 721 | 0.09,0.076,0.061 | 0.098,0.07,0.071 | -0.004 | 1 | 1 |
| **TSS118917.2** | 627,596,098,385,549 | 173,514,021,552 | 553,687,237,884,617 | 167,815,351,811 | 721 | 202 | 0.91,0.924,0.939 | 0.902,0.93,0.929 | 0.004 | 1 | 1 |
| **TSS118918.1** | 173,314,021,546 | 629,246,112,985,759 | 167,415,271,810 | 555,557,252,084,781 | 202 | 721 | 0.09,0.076,0.06 | 0.097,0.07,0.071 | -0.004 | 1 | 1 |
| **TSS118918.2** | 629,246,112,985,759 | 173,314,021,546 | 555,557,252,084,781 | 167,415,271,810 | 721 | 202 | 0.91,0.924,0.94 | 0.903,0.93,0.929 | 0.004 | 1 | 1 |
| **TSS118919.1** | 405,034,163,925 | 128,591,017,313,271 | 409,737,504,521 | 126,601,309,616,907 | 660 | 263 | 0.112,0.118,0.105 | 0.114,0.102,0.096 | 0.008 | 1 | 1 |
| **TSS118919.2** | 128,591,017,313,271 | 405,034,163,925 | 126,601,309,616,907 | 409,737,504,521 | 263 | 660 | 0.888,0.882,0.895 | 0.886,0.898,0.904 | -0.008 | 1 | 1 |
| **TSS118920.1** | 172,714,011,540 | 630,616,144,785,878 | 167,015,191,808 | 559,287,263,684,980 | 202 | 721 | 0.089,0.075,0.06 | 0.096,0.069,0.071 | -0.004 | 1 | 1 |
| **TSS118920.2** | 630,616,144,785,878 | 172,714,011,540 | 559,287,263,684,980 | 167,015,191,808 | 721 | 202 | 0.911,0.925,0.94 | 0.904,0.931,0.929 | 0.004 | 1 | 1 |
| **TSS118968.1** | 123,124,108 | 15,16,9 | 239,162,91 | 38,17,16 | 312 | 148 | 0.795,0.786,0.851 | 0.749,0.819,0.73 | 0.045 | 1 | 1 |
| **TSS118968.2** | 15,16,9 | 123,124,108 | 38,17,16 | 239,162,91 | 148 | 312 | 0.205,0.214,0.149 | 0.251,0.181,0.27 | -0.045 | 1 | 1 |

### 3.3.6.2 In-house bioinformatics pipeline to analyse differential TSS distribution:

A custom-made bioinformatics pipeline was developed that combined different packages to analyse the differential distribution of TSSs within a region of interest in the 5’ UTR (*Figure 3.15B*). Our pipeline utilised the TSS data from the FANTOM CAGE database and mapped the TSSs within the gene’s 5’ UTR. An area of interest in the 5’ UTR was selected and the number of TSSs within that region were subsequently counted, normalised and the differences between different samples were calculated using linear modelling and the empirical Bayes approach. The details on the development of the bioinformatics pipeline, its technical aspects, and the results related to differential TSS distribution is presented in chapter 5.

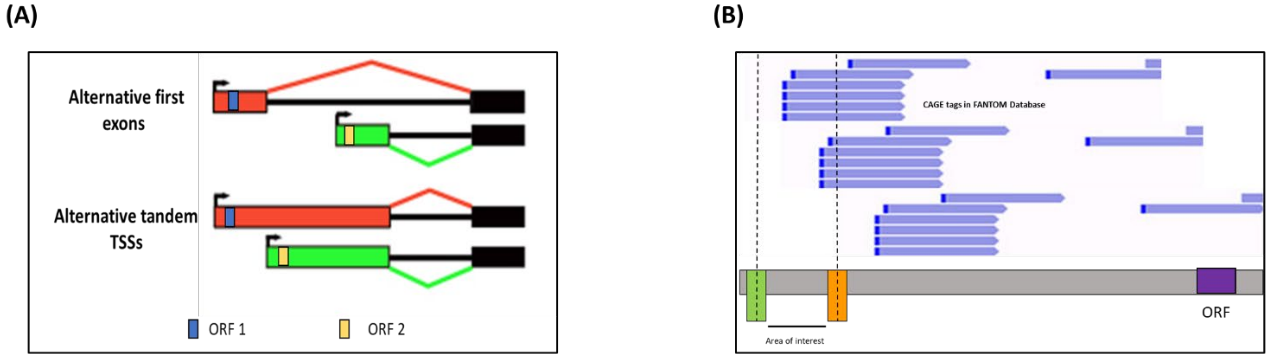


**Figure 3.15: Relevance of the SEASTAR analysis in the current project.** The SEASTAR examines alternative TSSs that in the form of alternative first exons and alternative tandem TSSs. The TSSs interrogated by the SEASTAR pipeline would likely change the ORF and are spread across a larger genomic region **(A)**. The bioinformatics pipeline designed in the current PhD project evaluated TSSs within the area of interest that does not change the ORF and the distance among the alternative TSS is less than 100 bp **(B)**.

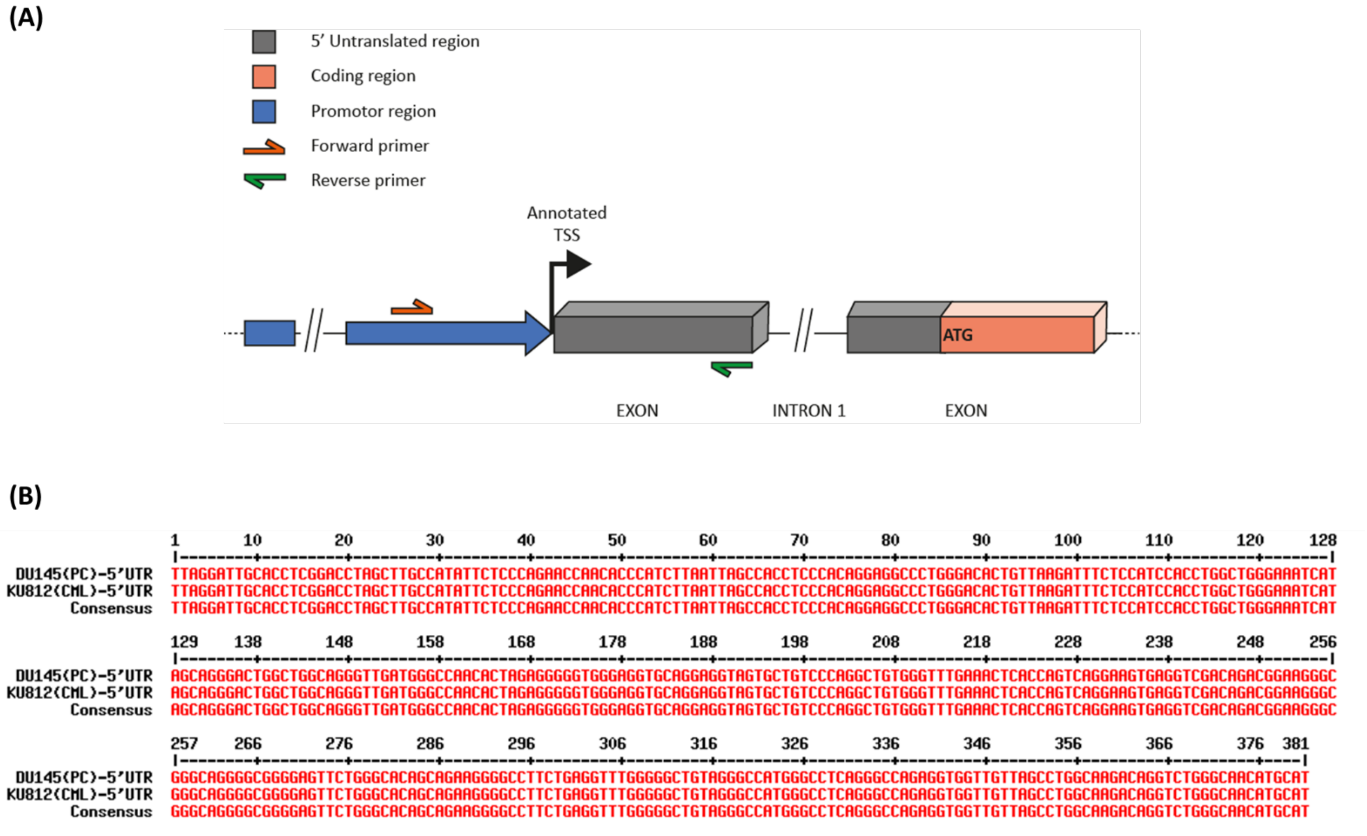
## 3.3.7 Nucleotide sequence around the TSS in PC and CML cell lines:

The selection of TSS by the RNA Pol II is a complex process dictated by transcription factors, DNA sequence elements, ncRNA, and other epigenetic factors (Javahery *et al.,* 1994, Jiang and Pugh, 2009, Kugel and Goodrich, 2017, Pardee *et al.,* 1998, Turowski and Tollervey, 2020). Studies have shown that about 50% of human core promoters contain an initiator (Inr) element that encompasses the TSS (Gershenzon and Ioshikhes, 2005, Yang *et al.,* 2007). The consensus sequence flanking the TSS, as shown by mutagenesis studies, was found to be Y-Y-A(+1)-N-W-Y-Y from −2 to +5 [where, Y = pyrimidine (C/T), W = (A/T), N = (A/C/G/T), and +1 is the TSS] (Kadonaga, 2012) and consensus of B-B-C-A(+1)-B-W for focused TSSs [where, B = (C/G/T), W = (A/T)] (Vo Ngoc *et al.,* 2017). Conversely, the analysis of the FANTOM CAGE database has identified a considerably shorter mammalian initiator consensus with transcription starting preferentially with a purine at +1 and pyrimidine at -1 position: YR (+1) [where, R = Purine (A/G) and Y = pyrimidine (C/T)] (Carninci *et al.,* 2006). The majority of *AGAP2* TSS obtained by 5’ RLM-RACE in the current project followed the YR consensus as reported by Carninci *et al.* (2006) for the FANTOM CAGE dataset (*Figure 3.12B*).

A study by Neininger *et al.* (2019) has demonstrated significant enrichment of SNPs (single nucleotide polymorphisms) and indels (insertion–deletion mutations) in the +/-200 bp area around the TSS. The authors have also assessed the SNP pattern in the direct vicinity (−15 to +12) of TSSs and found a significant SNP density peak at position -1 relative to TSS.  Mutations in the Inr sequence have shown to modulate the TSS usage and the transcriptional levels (Kugel and Goodrich, 2017, Vo Ngoc *et al.,* 2017).

In the current study, we analysed the 5' UTR genomic sequence of AGAP2 in the PC and CML cell lines to detect any relative mutations (SNP and indels) in the vicinity of the TSSs that could explain differential TSS usage of AGAP2 in PC and CML cell lines. Genomic DNA from DU145 (PC) and KU812 (CML) was extracted and a +/-200 bp region around the annotated TSS was amplified using *AGAP2* genomic forward and outer long *AGAP2* reverse primer (*Table 2.4, Figure 3.16A*). The amplified product was resolved on a gel and cloned for sequencing. The alignment of sequences displayed no differences in the sequences in the TSS region of PC and CML and ruled out the role of DNA sequence features in mediating differential TSS selection in these cell lines (*Figure 3.16B*). The genomic sequence corresponding to the *AGAP2* 5' UTR region in PC and CML cell line was similar to the *AGAP2* sequence in the RefSeq database (Appendix 2).

However, other epigenetic factors such as CpG methylation, histone modifications, and chromatin organisation could also potentially influence TSS selection by modulating the accessibility of a particular TSS and the recruitment of trans-acting factors that shield or advocate a specific TSS. The role of these epigenetic factors has been extensively studied in regulating gene expression in physiological and pathological conditions including cancers (Cholewa-Waclaw *et al.,* 2016, Kagohara *et al.,* 2017). Moreover, different studies have highlighted their enrichment/activity in the region around the TSS (Ando *et al.,* 2019, Luo *et al.,* 2018), making them relevant as a potential determinant of TSS selection. Further studies are required to elucidate their role in influencing TSS selection.



```
                 1       10        20        30        40        50        60        70        80        90       100       110       120     128
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+-------|
 DU145(PC)-5’UTR TTAGGATTGCACCTCGGACCTAGCTTGCCATATTCTCCCAGAACCAACACCCATCTTAATTAGCCACCTCCCACAGGAGGCCCTGGGACACTGTTAAGATTTCTCCATCCACCTGGCTGGGAAATCAT
KU812(CML)-5’UTR TTAGGATTGCACCTCGGACCTAGCTTGCCATATTCTCCCAGAACCAACACCCATCTTAATTAGCCACCTCCCACAGGAGGCCCTGGGACACTGTTAAGATTTCTCCATCCACCTGGCTGGGAAATCAT
       Consensus TTAGGATTGCACCTCGGACCTAGCTTGCCATATTCTCCCAGAACCAACACCCATCTTAATTAGCCACCTCCCACAGGAGGCCCTGGGACACTGTTAAGATTTCTCCATCCACCTGGCTGGGAAATCAT

                 129    138       148       158       168       178       188       198       208       218       228       238       248     256
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+-------|
 DU145(PC)-5’UTR AGCAGGGACTGGCTGGCAGGGTTGATGGGCCAACACTAGAGGGGGTGGGAGGTGCAGGAGGTAGTGCTGTCCCAGGCTGTGGGTTTGAAACTCACCAGTCAGGAAGTGAGGTCGACAGACGGAAGGGC
KU812(CML)-5’UTR AGCAGGGACTGGCTGGCAGGGTTGATGGGCCAACACTAGAGGGGGTGGGAGGTGCAGGAGGTAGTGCTGTCCCAGGCTGTGGGTTTGAAACTCACCAGTCAGGAAGTGAGGTCGACAGACGGAAGGGC
       Consensus AGCAGGGACTGGCTGGCAGGGTTGATGGGCCAACACTAGAGGGGGTGGGAGGTGCAGGAGGTAGTGCTGTCCCAGGCTGTGGGTTTGAAACTCACCAGTCAGGAAGTGAGGTCGACAGACGGAAGGGC

                 257    266       276       286       296       306       316       326       336       346       356       366       376  381
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+----|
 DU145(PC)-5’UTR GGGCAGGGGCGGGGAGTTCTGGGCACAGCAGAAGGGGCCTTCTGAGGTTTGGGGGCTGTAGGGCCATGGGCCTCAGGGCCAGAGGTGGTTGTTAGCCTGGCAAGACAGGTCTGGGCAACATGCAT
KU812(CML)-5’UTR GGGCAGGGGCGGGGAGTTCTGGGCACAGCAGAAGGGGCCTTCTGAGGTTTGGGGGCTGTAGGGCCATGGGCCTCAGGGCCAGAGGTGGTTGTTAGCCTGGCAAGACAGGTCTGGGCAACATGCAT
       Consensus GGGCAGGGGCGGGGAGTTCTGGGCACAGCAGAAGGGGCCTTCTGAGGTTTGGGGGCTGTAGGGCCATGGGCCTCAGGGCCAGAGGTGGTTGTTAGCCTGGCAAGACAGGTCTGGGCAACATGCAT
```

**Figure 3.16: Analysis of flanking DNA sequence around annotated TSS of PC and CML cell lines. (A)** The genomic DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen) following the manufacturer’s guidelines. The +/- 200 bp region around the annotated TSS was amplified using forward primer (AGAP2 genomic) and reverse primer (outer long reverse) (*Table 2.4*) and sequenced using Sanger sequencing. Figure not drawn to scale. **(B)** Multalin alignment of the amplified sequence showed no differences in the sequence of *AGAP2* TSS region in PC and CML cell lines.

## 3.4 Summary of findings (Chapter 3)

In summary, the results presented in this chapter indicated a discrepancy in AGAP2 mRNA and protein levels primarily in PC and CML cell lines, also noted in other cell types, with higher mRNA expression and lower protein abundance found in the CML cell lines. The reduced AGAP2 protein yield observed in CML cell lines was not associated with protein degradation by UPP. No differences were also observed in the basal level of selected rate-limiting translational initiation factors, limiting their relevance in mediating discordant mRNA and protein levels. This chapter also characterised the TSS usage for *AGAP2* in PC and CML cell lines by 5' RLM-RACE and demonstrated distinct patterns of TSSs distribution in both the cell group. The TSSs identified in our study were in agreement with the TSSs obtained by the highly sophisticated deep CAGE technology in the FANTOM database and highlighted the efficacy of the 5' RLM-RACE technique in detecting genuine TSSs. Additionally, a SEASTAR computational pipeline to classify differentially distributed TSSs using RNA-seq data was described in this chapter, interrogating its applicability to the current study, and also emphasizing the gaps in the SEASTAR tool that forms the basis for developing our bioinformatic pipeline. Furthermore, the evidence in this chapter also ruled out any mutations in the TSS region that could influence differential TSS selection in PC and CML cell lines.

# Chapter 4:

# Alternative transcription of G quadruplex structures and its consequence on Translation

## 4.1 Introduction

In the previous chapter, we have discovered unique TSS distribution profiles in PC vs. CML cell lines. We also evidenced that the differential TSS selection in these cell groups produces heterogeneity in the 5' UTR isoforms and add extra nucleotides in the 5' UTR of *AGAP2* mRNA in KU812 (CML cell line) due to the upstream TSS selection. The earlier TSS selection increases the length of 5' UTR and could incorporate regulatory features that might affect the mRNA translation efficiency and could possibly explain the observed inconsistency in AGAP2 mRNA and protein expression (*Figure 4.1*). As indicated in chapter 3, we have noted an upstream TSS (-130 bp relative to ATG start codon) in the KU812 cell line that would transcribe 36 extra base pairs in the 5' UTR relative to the major TSS. Interestingly, these extra nucleotides contained a G quadruplex (G4) forming sequence which is not present in the 5' UTR isoform derived from other downstream starting sites and also the major TSS (*Figure 4.1*). As pointed out in the earlier chapter, the TSSs examined in the current PhD project belong to the same TSS cluster and are within 50 bp of each other. In this chapter, we have explored the significance of the G4 consensus sequence in the alternatively transcribed longer TSS isoform and its impact on mRNA translation.

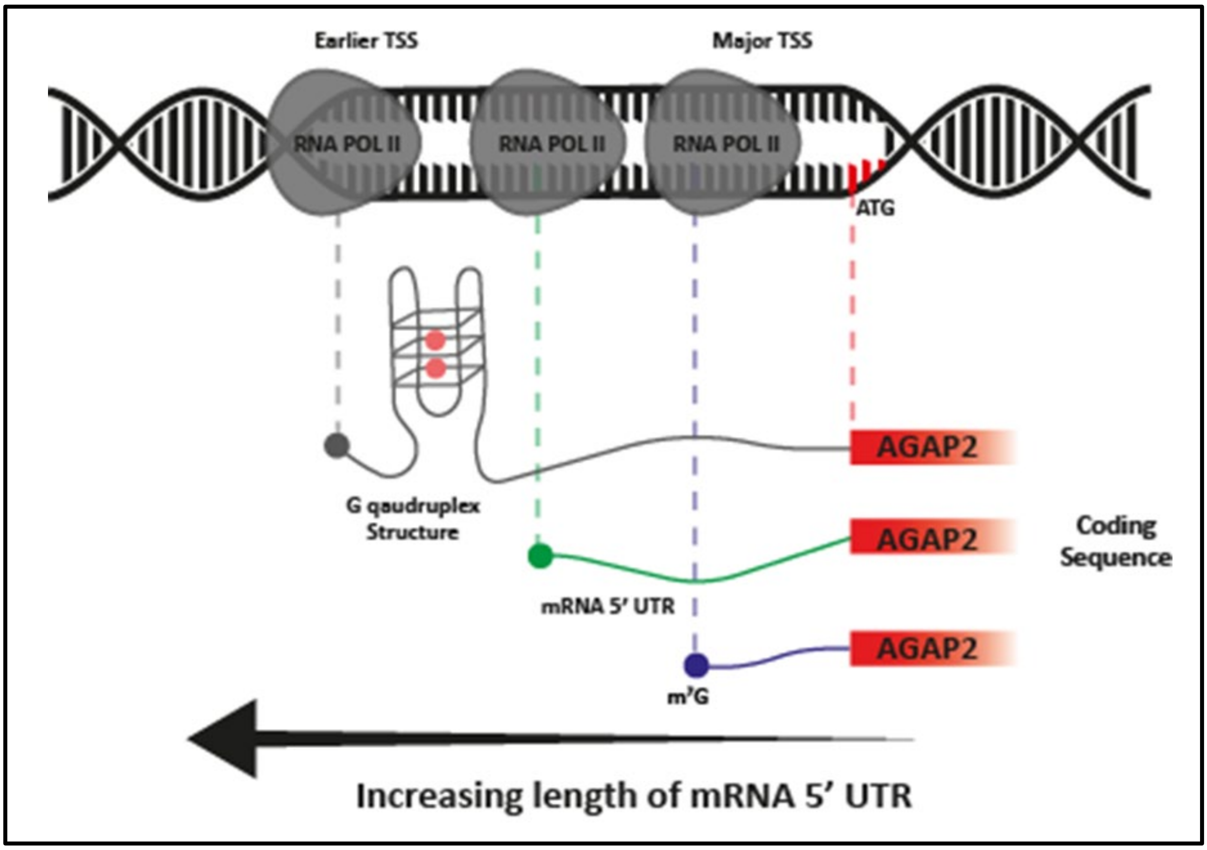
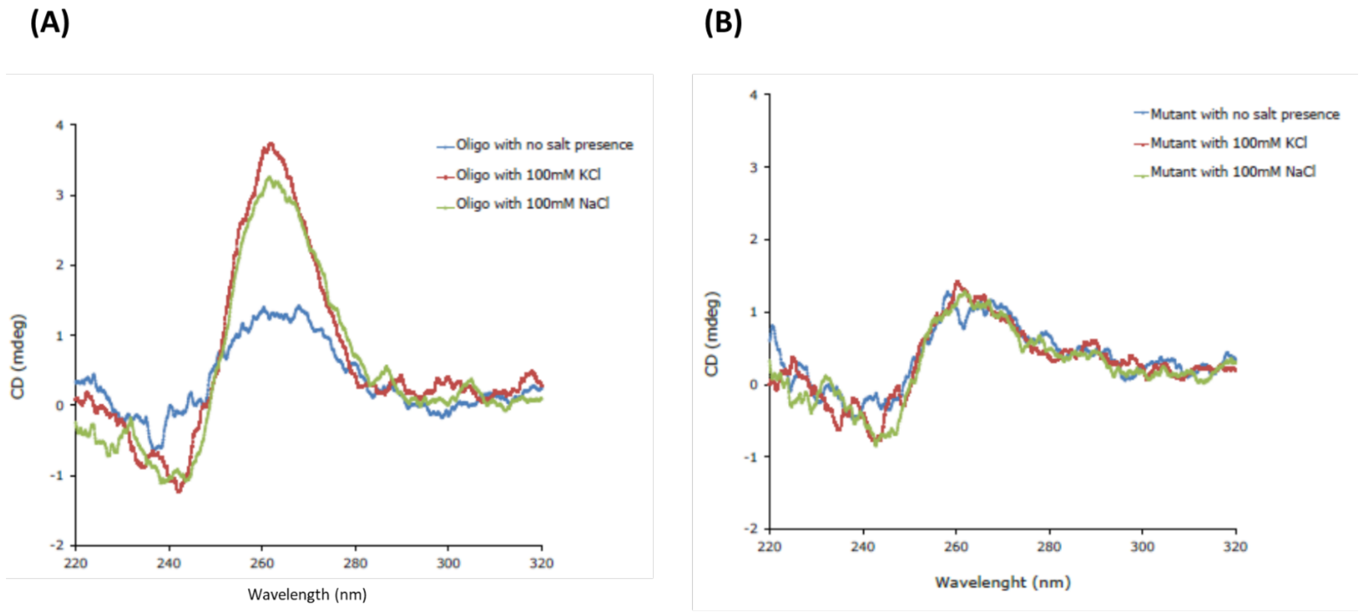


**Figure 4.1: TSS selection and 5' UTR heterogeneity.** The TSS selection within the core promoter region generates diverse 5' UTR isoforms with the length of the 5' UTR depending on transcript starting position. The selection of earlier TSS increases the length of 5' UTR and incorporate additional nucleotides that could have a regulatory role. In the previous chapter, we noted that the selection of earlier TSS selection encoded a G quadruplex forming sequence in the longer 5' UTR isoform which was absent in other isoforms derived from downstream TSSs.

G-quadruplexes are non-canonical, four-stranded, secondary structures which are formed by sequences rich in guanine nucleic acid (section 1.4). The consensus sequence capable of forming a G4 could be described as $G_X$-$N_{1-7}$-$G_X$-$N_{1-7}$-$G_X$-$N_{1-7}$-$G_X$, where x is 3–6 and N corresponds to any nucleotide (A, G, T, C, or U) (Reviewed in Fay *et al.,* 2017). However, the existence of imperfect G4s that do not follow the canonical consensus sequence has been also reported (Reviewed in Puig Lombardi and Londoño-Vallejo, 2020). The G4 motifs are often enriched in the 5' UTR of mRNA, suggesting an important role in regulating mRNA translation (Huppert *et al.,* 2008). Numerous studies have highlighted the association of G4 structures in the 5' UTR with translational suppression (Arora *et al.,* 2008, Reviewed in Beaudoin and Perreault, 2010). However, in some instances, an activating role for G4 has also been noted (Agarwala *et al.,* 2013).

The effect of G4 motifs on mRNA translation is complex and depends on the location of G4 structures, G4-binding proteins, and G4-resolving helicases. A study by Kumari *et al.* (2008) has reported that the G4 motif significantly represses translation when they are located within the first 50 bp of the mRNA 5' UTR. The steric effects of these motifs likely interfere with serial translation processes including the formation of PIC, scanning, and elongation (Reviewed in Bugaut and Balasubramanian, 2012). Moreover, the formation of G4s by the 5′ UTR (CGG)n repeat has been implicated in reduced polysome formation and stalled ribosomes on *FMR1* mRNA, resulting in decreased translational efficiency (Feng *et al.,* 1995, Primerano *et al.,* 2002). Furthermore, a range of proteins interacts with the G4 structure through specialised RNA recognition motifs, facilitating the formation of specialised eIF4A-dependent translational processes (Masuzawa and Oyoshi, 2020, Wolfe *et al.,* 2014). In spite of a body of literature on G4 and its translational effects, little is known about the significance and consequences of alternatively transcribed G4 motifs by TSS selection within a single cluster.

The formation of G4 structures has been extensively studied *in vitro* using circular dichroism (CD) spectroscopy. These G4s exhibit unique CD spectral signatures depending on G4 particular topology (Reviewed in Del Villar-Guerra *et al.,* 2018). Previously, we employed CD spectroscopy to analyse the formation of G4 by the extra nucleotides encoded in the longer 5' UTR of *AGAP2* mRNA (Doush, 2015). We showed that the longer UTR extra sequence folds into a parallel G quadruplex *in vitro* (*Figure 4.2A*) and the characteristic CD spectrum disappeared by mutating the consensus (*Figure 4.2B*).



5’ – GGGCGGGCAGGGGCGGGG – 3’

5’ – GAGCGAGCAGAGGCGAGG – 3’

**Figure 4.2: CD spectroscopy of G4 forming sequence in the longer 5’ UTR of *AGAP2* mRNA (Doush, 2015). (A)** The Circular dichroism (CD) experiments were performed on 10 µM of RNA oligos in Tris–HCl (pH 7.5) buffer containing either 100 mM of NaCl or KCl or no salts. The measurements were performed using a JASCO J-715 Spectropolarimeter (JASCO). Quartz cell cuvettes of 0.1 cm path length were used, and wavelengths were recorded between 220 - 320 nm at a scan speed of 50 nm/min with a response time of 2 sec. All CD spectra were generated at 25°C. **(A)** The CD spectra RNA oligo of extra nucleotides in the longer *AGAP2* mRNA 5’ UTR folded in presence of 100 mM NaCl, 100 mM KCl, or no salts. The characteristics formation of parallel G4 is noted, exhibiting a positive peak at ~260 nm and a negative peak at ~240nm in the presence of salts. **(B)** Loss of the characteristic G4 formation in RNA oligo with mutations (G→A). The sequences of RNA oligos are presented below the CD spectra with the mutations shown in RED. **Note:** The experiment for CD spectroscopy was performed by Dr Christoph Ufer from Charité University, Berlin.

Despite the possibility of determining the formation of G4 structures *in vitro*, the techniques to demonstrate their formation inside the cell are still under development. Different approaches have been used to detect the RNA G4 structures inside the cell including the use of G4-stabilising ligands/ions (Biffi *et al.,* 2014, Kwok *et al.,* 2016), small molecule probes (Yang *et al.,* 2018), RNA structural mapping (Guo and Bartel, 2016), reverse transcription stalling (Kwok and Balasubramanian, 2015), RNA G4-protein interactions (Herdy *et al.,* 2018), ligands with fluorescence activity (Chen *et al.,* 2018), self-biotinylation methodology (Einarson and Sen, 2017), and G4-structure specific antibody (Biffi *et al.,* 2013). However, most of the methodologies described above used specific ligands and/or reactive small molecules that could shift the equilibrium in the favour of G quadruplex formation and might not be representative of actual RNA G4 conformations in the living cells. Further studies are required to explore novel techniques to capture these motifs in their native state which would provide direct evidence linking functional consequences associated with these motifs to their formation inside the cell.

### 4.1.1 Aims of chapter 4:

The upstream *AGAP2* TSS selected in KU812 (CML cell line) produces a longer mRNA 5’ UTR containing the G4 sequences. Our previous CD spectroscopy experiment confirmed the additional nucleotides incorporated in the longer 5’ UTR formed a G4 *in vitro*. This G4 motif could potentially contribute to the discordant mRNA and protein expression profile observed in the CML cell group by affecting the mRNA translation efficiency.

This chapter aims to:

- Validate the *in cellula* formation of the G4 structure in the longer 5’ UTR isoforms of *AGAP2* mRNA using an in-house developed immunoprecipitation technique.
- Study the relative presence of G4-containing 5’ UTR isoforms in PC and CML cell lines.
- Investigate the translational consequences of the 5’ UTR G4 motif using reporter assay and polysome profiling.

## 4.2 Method Development (G4-RNA-Immunoprecipitation)

Currently, most studies analysing the G4-containing mRNA population within a cellular system utilise transcriptome-wide sequencing approaches based on either reverse transcriptase stalling or the use of small molecules (Reviewed in Kamura *et al.,* 2020, Yang *et al.,* 2018). However, the ligands and/or small molecules used in these techniques could induce the formation of G4 structures and might not be representative of actual RNA G quadruplex folding in the living cells. As part of the current PhD project, we have developed an in-house immunoprecipitation technique termed GRIP (G4-RNA-Immunoprecipitation) to selectively enrich and pulldown RNA containing G quadruplex structures that could be quantitively analysed by real-time PCR. Compared to the current approaches, our technique is capable to capture these structures in their native states and selectively enrich mRNA containing them.

### 4.2.1 Principle:

This protocol is used to enrich RNA with G4 structure using a structure-specific antibody. The antibody (BG4) was generated by Biffi *et al.* (2013) to selectively bind DNA G4. The antibody was selected from the Sanger phage display library using a panel of intramolecular DNA G quadruplex structures (Biffi *et al.,* 2013). The BG4 antibody was also used by the same group to visualise the RNA with G4 motifs (Biffi *et al.,* 2014). To achieve this, the authors selectively stabilised RNA G4 using a ligand (carboxypyridostatin) and then fixed the cells using formaldehyde followed by detection with the BG4 antibody and a secondary fluorescently tagged antibody against the BG4 to detect the RNA G4. However, the use of a stabilising ligand in the study could influence RNA structure, metabolism, and impact native G quadruplex conformations. The stabilising small-molecule ligand (BioTASQ) has been also shown to prefer specific RNA topology (Yang *et al.,* 2018). Moreover, BG4 antibody binds with stronger affinity to DNA G4 (Kd: 1.1-2.0 nM) compared to RNA G4 (Kd 5.5-18.0 nM) (Biffi *et al.,* 2014, Biffi *et al.,* 2013) and, therefore, would require an optimised method to facilitate their use to detect only RNA G4 within the cellular environment.

To overcome these limitations and to exploit the potential of the BG4 antibody to selectively bind RNA G4 structure, a method was designed to enrich the cytosolic RNA, remove contaminating genomic DNA, and capture the G4 structure in the mRNA in their native conformation without the use of a fixative or stabilising ligand. To specifically isolate cytosolic RNA, cells were permeabilised with the weak non-ionic detergent 'digitonin' which at lower concentration selectively permeabilises the plasma membrane, leaving the nuclear envelope and other major membrane-bound organelles intact (Adam *et al.,* 1990). The preferential permeabilization of the plasma membrane is due to its higher cholesterol content which forms complexes with digitonin, creating pores in the membrane (Schulz, 1990, Colbeau *et al.,* 1971).

The cytosolic RNA is extracted in a buffer containing $K^+$ ions at a concentration similar to intracellular $K^+$ levels (150 mM). These potassium ions are required for stabilising G4 structures (Wang and Liu, 2017), enabling their detection with the structure-specific BG4 antibody. To specifically promote the detection of RNA G4, the protocol incorporates a DNase treatment step to remove any contaminating genomic DNA. Moreover, the method integrates steps to decrease non-specific background signals by repeated washing and releasing the bound RNA from the antibody-beads complex by incubating at a higher temperature which unfolds these structures.

### 4.2.2 Protocol:

In the GRIP protocol, the longer G4-containing 5' UTR isoforms of *AGAP2* mRNA were pull down using the TCCS cell line. The TCCS cell line were specifically selected as it exhibited higher levels of G4-containing longer *AGAP2* 5' UTR isoform compared to other CML cell lines used in the study (Figure 3.13B, 3.13C).

The GRIP was performed with a structure-specific G quadruplex (BG4) antibody (Absolute Antibody). TCCS cells (15 X $10^6$) were collected, washed with ice-cold PBS, and resuspended in ice-cold lysis buffer (150 mM KCL, 50 mM HEPES, 25 µg/mL Digitonin, 100 U/mL RNase inhibitor). The cells were incubated with lysis buffer for 10 minutes at 4°C using end over end rotation and centrifuged at 2,000 x g for 5 minutes at 4°C. The supernatant (cytosolic fraction) was saved and 10 % was removed to be used as input control. When transfections were required, 1 x $10^6$ DU145 cells were seeded in a 100 mm dish, transfected using JetPRIME transfection reagent (Polyplus), trypsinised after 48 hours, and processed as above.

The lysate was precleared by incubating with 100 µL protein G magnetic beads for 1 hour at 4°C in an end over end rotator to remove non-specific binding to the empty beads. The protein G magnetic beads were prepared for pre-clearing and antibody binding by thoroughly washing the beads thrice with PBS-T. In case of antibody binding, the beads were incubated with either 3 µg of BG4 antibody or an equivalent isotype-matched negative antibody control. The negative antibody is an isotype-matched antibody from the same species that maintain the similar property to the primary (target) antibody but lacks specific target binding and is used to measure the level of non-specific background signals. The precleared lysate was incubated overnight with BG4/control antibody bound to protein G magnetic beads (Biorad). After incubation, the beads were magnetised, washed thrice with the lysis buffer, and incubated at 65°C for 15 minutes to release the bound nucleic acids. The eluent was treated with 2U of RNase-free DNase I (ThermoFisher) for 15 minutes at 37°C to remove contaminating DNA. The RNAs from input and IP fractions were then isolated through TRIzol (ThermoFisher) extraction followed by isopropanol precipitation.

For RNA extraction, 1 mL or 500 µL TRIzol reagent was added to the DNase treated sample or the input controls, respectively, followed by the addition of 200 µL of chloroform per 1 mL of TRIzol. The tubes were shaken vigorously for 15 seconds, incubated at room temperature for 5 minutes, and subsequently centrifuged at 12,000 x g for 15 minutes at 4°C to separate the solution into 3 layers: clear aqueous layer containing RNA, middle interphase layer containing mostly the DNA and lower pink organic phase containing protein. The top aqueous phase (60% of the volume of TRIZOL Reagent used) was removed and 20 µg glycogen and 500 µL of 100% room temperature isopropanol per 1 mL TRIzol reagent was added to the aqueous phase. The mixture was then incubated at room temperature for 10-15 minutes followed by centrifugation at 12,000 x g for 30 minutes at 4°C to precipitate the RNA. The RNA pellet was washed by 1 mL 75% ice-cold ethanol per 1 mL TRIzol reagent and centrifuged twice at 7,500 x g for 5 minutes at 4°C to remove ethanol. The washed pellet was air-dried for 2-3 minutes at room temperature and resuspend in 15 µL of nuclease-free water. The RNA was then converted to cDNA (section 2.2.2.2) and amplified using qRT-PCR (section 2.2.2.3).

The steps of the GRIP protocol are illustrated in *Figure 4.3*.

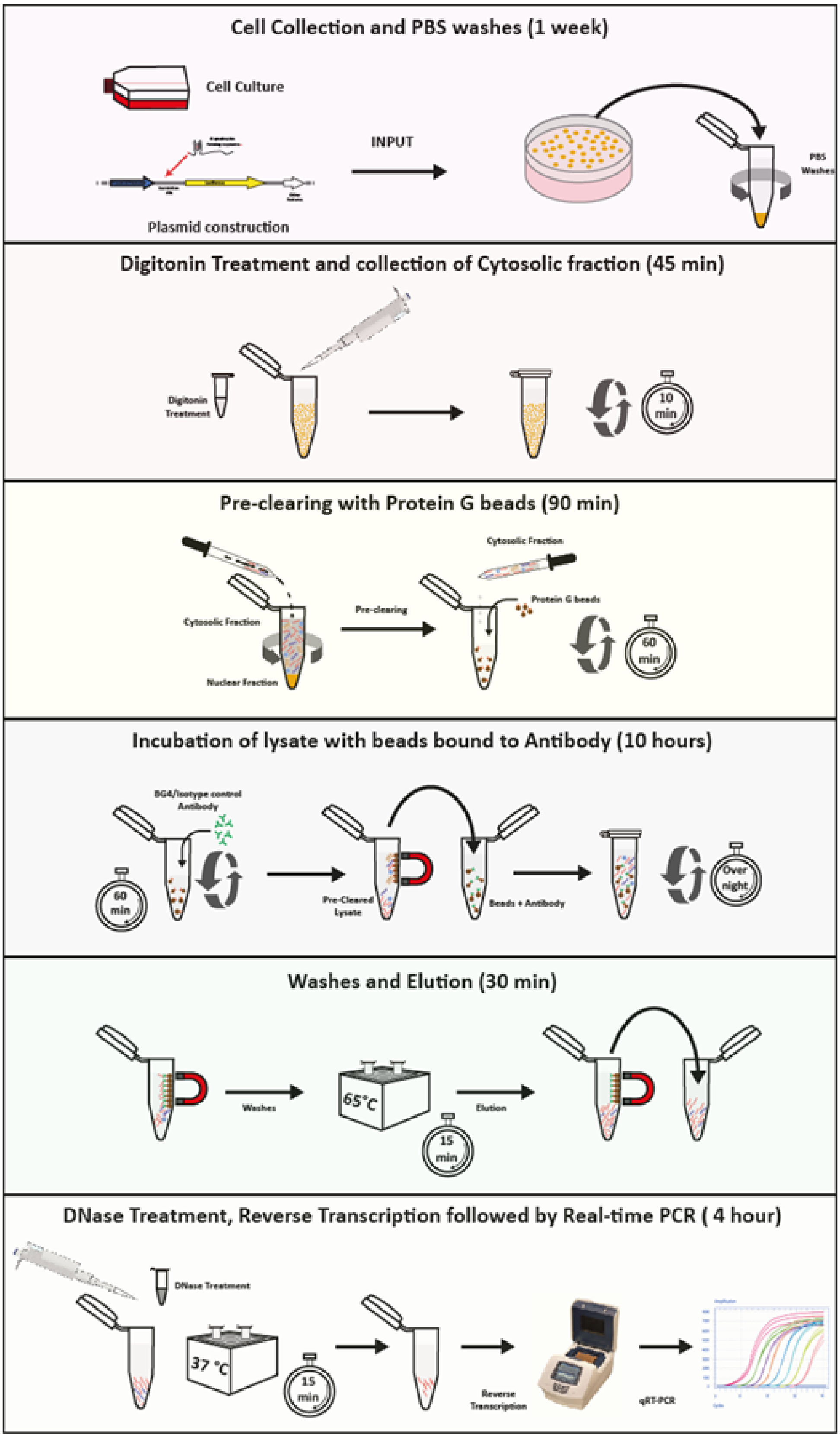

Cell Collection and PBS washes (1 week)
Cell Culture
INPUT
Plasmid construction
PBS
Washes
Digitonin Treatment and collection of Cytosolic fraction (45 min)
Digitonin
Treatment
10
min
Pre-clearing with Protein G beads (90 min)
Cytosolic Fraction
Nuclear Fraction
Pre-clearing
Cytosolic Fraction
Protein G beads
60
min
Incubation of lysate with beads bound to Antibody (10 hours)
BG4/Isotype control
Antibody
60
min
Pre-Cleared
Lysate
Beads + Antibody
Over
night
Washes and Elution (30 min)
Washes
65°C
15
min
Elution
DNase Treatment, Reverse Transcription followed by Real-time PCR ( 4 hour)
DNase Treatment
37 °C
15
min
Reverse
Transcription
qRT-PCR

**Figure 4.3: Schematics of GRIP protocol.** Overview of G quadruplex RNA immunoprecipitation (GRIP). The cells are treated with 25 µg/mL digitonin, and the extracted cytoplasmic fraction is precleared and incubated overnight with structure-specific G quadruplex (BG4) antibody bound to protein G magnetic beads. After incubation, the complex is washed, and the bound RNA is eluted by unfolding the G4 by heating at 65 °C for 15 minutes. The eluent is treated with DNase I treatment followed by qRT-PCR. The time required for each step is shown in (parenthesis).

## 4.2.3 Optimisation of GRIP method:

The rationale for including different steps along with optimisation of a key step is presented below:

### 4.2.3.1 Optimisation of digitonin concentration:

The final concentration of digitonin in the GRIP lysis buffer needs to be optimised for different cell types. The lowest concentration of digitonin yielding satisfactory levels of mRNA of interest and lower levels of genomic contamination should be selected. For TCCS, different concentrations (25 µg/mL, 50 µg/mL, and 100 µg/mL) were tested relative to lysis buffer without digitonin (0 µg/mL). The levels of *AGAP2* mRNA were detected using AGAP2 forward and reverse primers (*Table 2.4*). The genomic contamination was analysed using primers designed to amplify the promoter region using AGAP2 genomic (-425) forward and AGAP2 genomic (-218) primers (*Table 2.4*) (*Figure 4.4A*). The results indicated that 25 µg/mL digitonin concentration yield sufficient mRNA levels with the lowest genomic contamination and was selected as the final concentration of digitonin in the GRIP lysis buffer (*Figure 4.4B*).

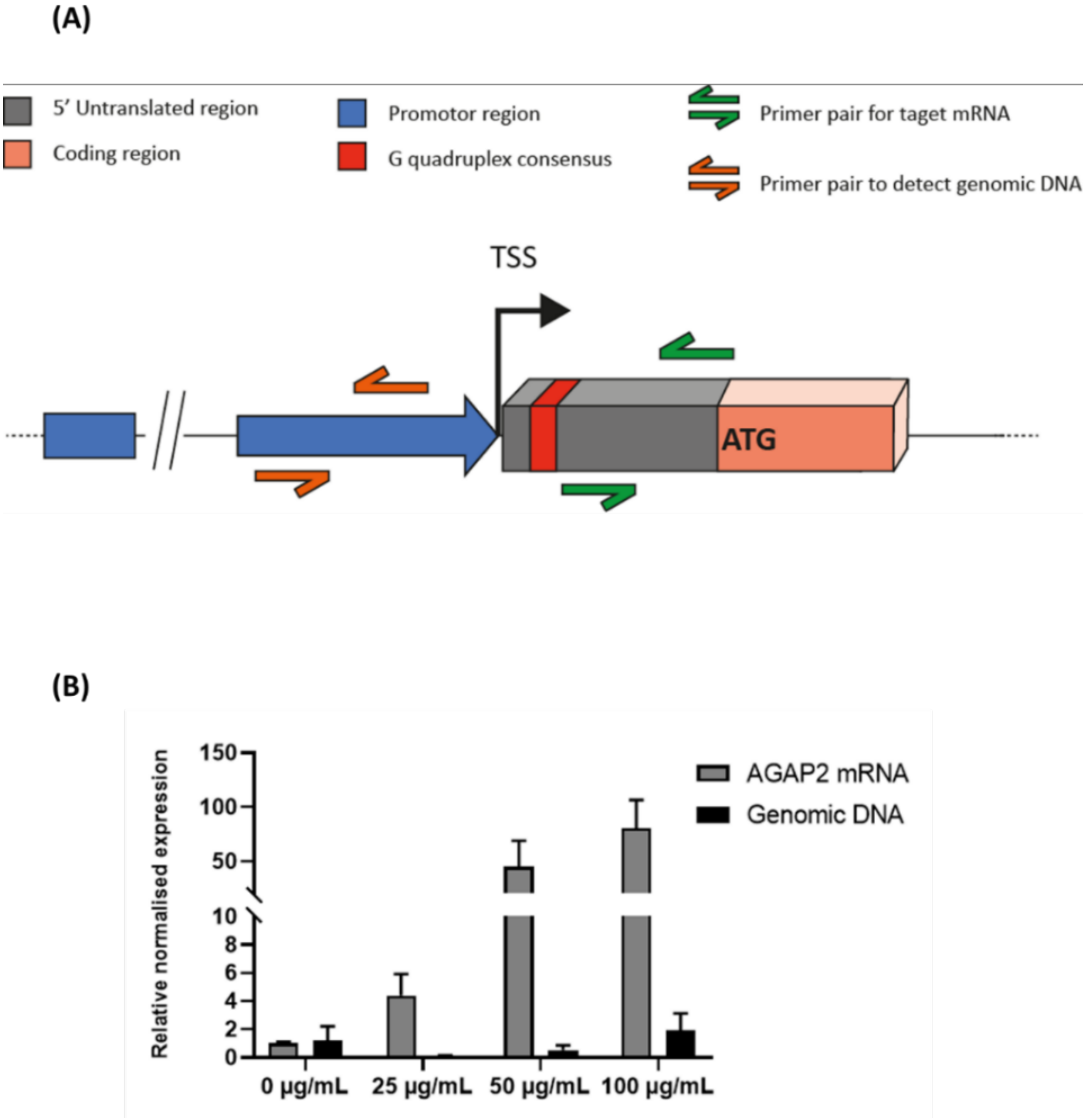


**Figure 4.4: Optimisation of digitonin concentration. (A)** Primer designing consideration for the target mRNA (*AGAP2*) and genomic DNA. Figure not drawn to scale. **(B)** TCCS cells were lysed with GRIP lysis buffer containing varying concentration of digitonin (x-axis). After lysis and extraction of cytosolic fraction, the RNA was isolated, and the expression was analysed using qRT-PCR. The expression level of the *AGAP2* mRNA and genomic DNA (primer amplifying the promoter region) were normalised using housekeeping gene *HPRT* and presented relative to 0 µg/mL. The data shown are the mean ± SD of two independent experiments (n=2).

### 4.2.3.2 Rationale for including preclearing and DNase treatment step:

The preclearing step was included to decrease the background associated with non-specific binding of RNA to the beads. To examine the relevance of preclearing, GRIP was performed for a positive control mRNA (*NRAS*) with or without the preclearing step. The formation of the G4 motif in the 5′ UTR of *NRAS* mRNA has been already established in the literature (Kumari *et al.,* 2007). The enrichment was evaluated relative to the negative isotype antibody control. As shown in *Figure 4.5A*, the GRIP without the preclearing step showed lower differences in the *NRAS* enrichment between BG4 and negative antibody compared to the GRIP with the preclearing step. It could be attributed to the increase in the background signals in GRIP without the preclearing step owing to nonspecific RNA binding to the beads.

The DNase treatment step was added to prevent amplification of signals resulting from the binding of BG4 antibody to G4 motifs in the DNA. The G4 consensus in the non-template (+) DNA strand could also theoretically form a G4 structure if the DNA is in single-stranded conformation. Since the BG4 antibody binds with a stronger affinity to the DNA compared to RNA G4 (Biffi *et al.,* 2014, Biffi *et al.,* 2013), the detection of G4 motifs in the target of interest could be amplified from both the mRNA and DNA. To avoid this bias, a DNase step was performed to remove any contaminating genomic DNA. The relevance of incorporating the DNase treatment is depicted in *Figure 4.5B*. As shown in the figure, the GRIP performed without the DNase treatment showed a significantly higher amplification of the genomic DNA compared to the GRIP with DNase added ($P < 0.01$) (*Figure 4.5B*). Alternatively, a negative RT control could be also used to highlight the genomic contamination of the GRIP eluent.

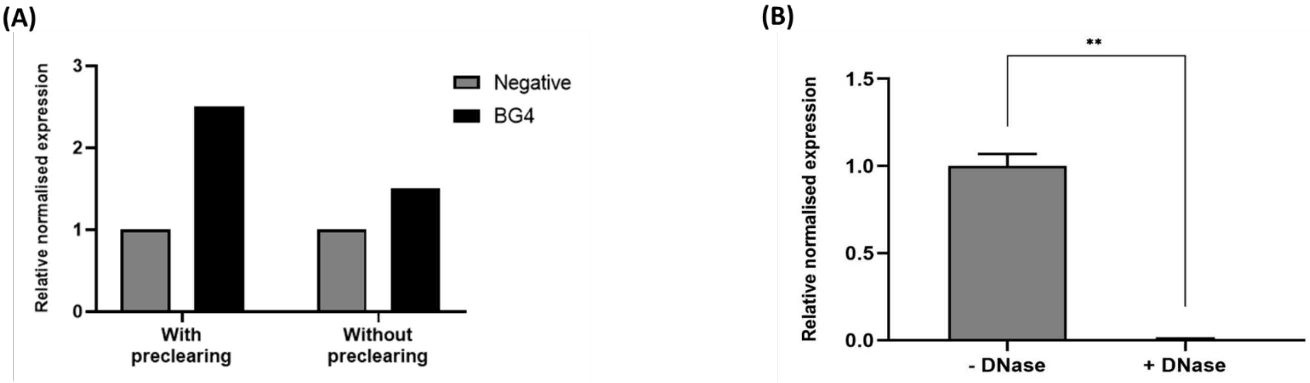


**Figure 4.5: Relevance of the preclearing and DNase treatment steps in the GRIP method. (A)** The levels of *NRAS* mRNA after GRIP enrichment with or without the preclearing step. The levels are normalised by the input control and presented relative to the negative control antibody. A lower difference between the BG4 and negative antibody enrichment is noted compared to the GRIP with the preclearing step (n=1). **(B)** Levels of *AGAP2* genomic region in the eluant after BG4 enrichment. The -DNase samples are processed identically except for the DNase treatment step. The levels are normalised by the input control and presented relative to -DNase condition. The data shown are the mean ± SD of two independent experiments (n=2) and differences are analysed by unpaired t-test [t(2) = 20.49, *P* = .0024] . ***P* < 0.01.

# 4.3 Results and Discussion

## 4.3.1 Demonstration of G4 formation inside the cell using GRIP method:

We used our GRIP method to pulldown G4 structures in the *AGAP2* mRNA. Briefly, the CML TCCS cells were lysed using digitonin treatment to isolate the cytoplasmic cellular fraction which was precleared and incubated with BG4 or negative isotype control antibody followed by washing, elution, DNase treatment, and amplified using qRT-PCR (*Figure 4.6A*). Employing GRIP, we noted a significant enrichment of *AGAP2* in the BG4 antibody fraction relative to the negative isotype control ($P < 0.001$) (*Figure 4.6B*). Significant enrichment was also noted for *NRAS* ($P < 0.01$) and *MM16* ($P < 0.001$) (*Figure 4.6B*). The formation of the G4 motif in *NRAS* and *MM16* has already been established in the literature and were used as a positive control in the GRIP experiment (Kumari *et al.,* 2007, Morris and Basu, 2009). The TBP was used as a negative control since it lacks a G4 consensus sequence in its entire mRNA as determined computationally using the pqsfinder web application (Labudová *et al.,* 2019). No statistically significant differences were noted for *TBP* ($P = 0.419$), confirming the specificity of our technique to pulldown only mRNA containing-G4.

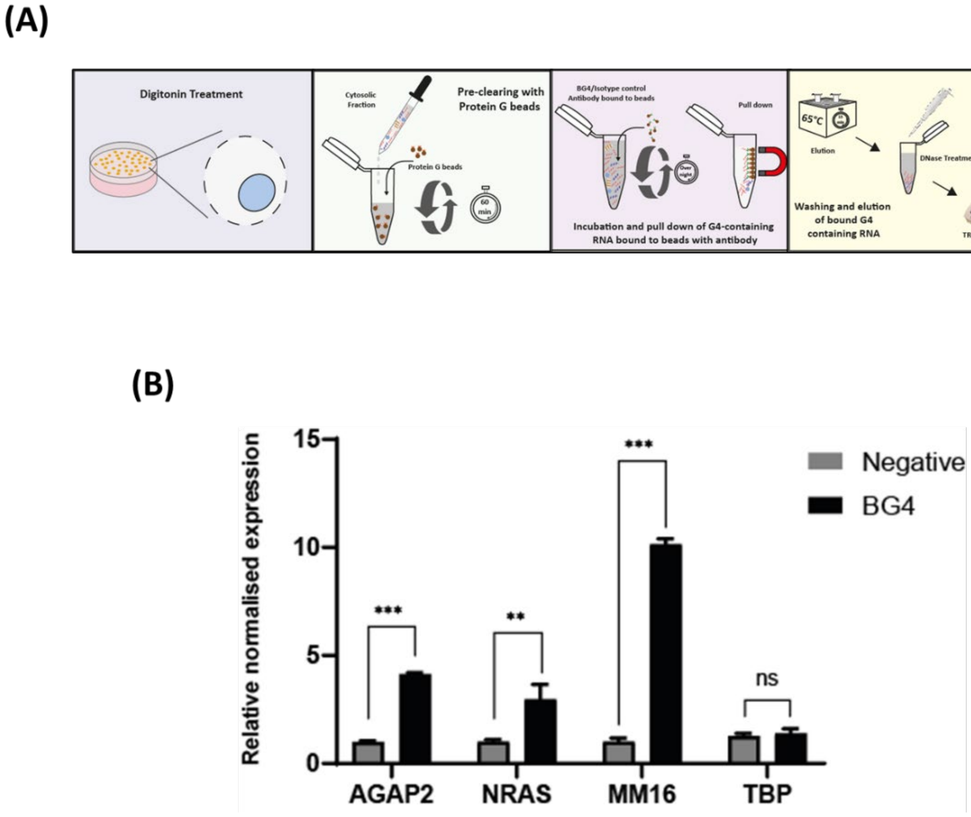
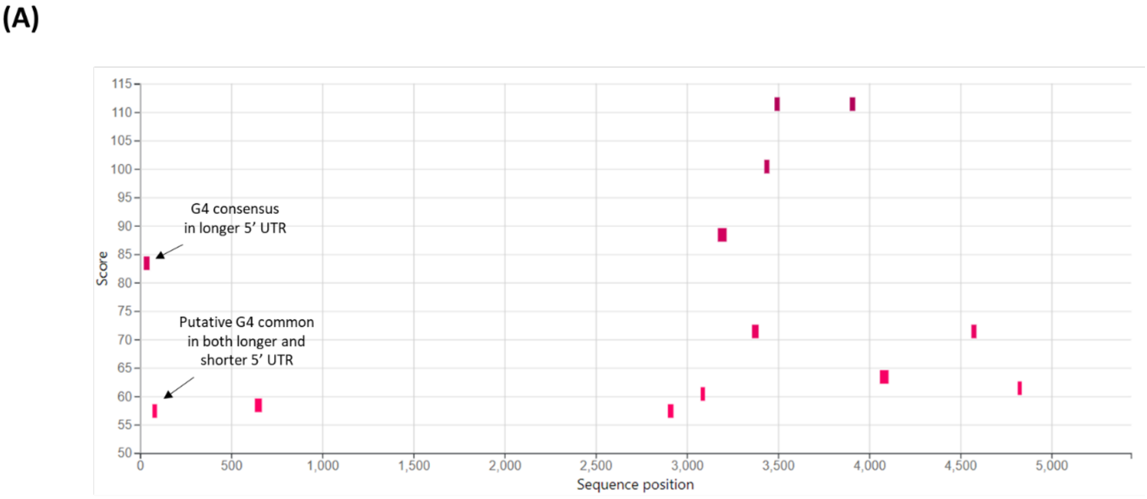


**Figure 4.6: *In cellula* formation of G4 motif in *AGAP2* mRNA using GRIP. (A)** A brief overview of the GRIP method. **(B)** GRIP performed in the TCCS and the immunoprecipitated samples were normalised by their input controls. *NRAS* and *MM16* mRNAs were used as a positive control for the presence of G4 structures, as documented in the literature. *TBP* mRNA (NM_003194.5) was used as a negative control as it lacks G4 consensus sequences in its full-length mRNA. Data shown correspond to three independent immunoprecipitations (n=3) and the error bars denote standard deviation. Differences between samples were analysed with unpaired t-test, *P*-values shown (***$P < 0.001$, **$P < 0.01$, ns: not significant).

However, *AGAP2* mRNA contains several putative G4 forming sequences along its entire length (*Figure 4.7A*). The analysis of *AGAP2* transcript (NM_014770.4) using the pqsfinder web application (https://pqsfinder.fi.muni.cz/) (Labudová *et al.,* 2019) exhibited 14 G4 consensuses in the *AGAP2* mRNA transcript. To specifically detect the G4 formation in the longer 5' UTR of *AGAP2* mRNA, we transfected either an empty vector or a plasmid vector containing the longer 5' UTR (*Table 2.8*) cloned proximal to *Renilla* luciferase (2.2.6.1). The presence of G4 consensus in the *Renilla* luciferase mRNA was ruled out using the pqsfinder package that showed no putative G4 sequences in either sense or antisense strand. The results showed significant enrichment of *Renilla* luciferase mRNA in the lysate of cells transfected with the cloned 5' UTR compared to the empty vector, or the pull down performed with the negative isotype control (*P* = 0.0117), indicating the formation of G4 in the longer 5' UTR of *AGAP2* mRNA (*Figure 4.7B*). The shorter UTR cloned to the *Renilla* luciferase mRNA was not used as a control because the shorter UTR also contained a putative G4 consensus (Figure 4.7A) that is present in both the longer and shorter 5' UTR isoforms and could be potentially enriched following GRIP. Instead, empty luciferase mRNA that did not have any G4 consensus sequences would serve as good negative control.



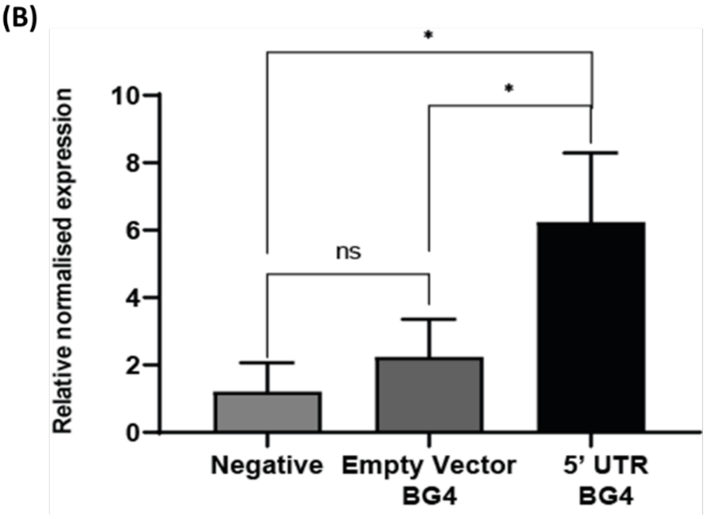


**Figure 4.7: Enrichment of G4 motif in the 5' UTR of *AGAP2* mRNA using GRIP. (A)** Image derived from the pqsfinder browser (https://pqsfinder.fi.muni.cz/) showing the position of 13 G4 consensus sequences along the entire length of *AGAP2* mRNA transcript. The G4 forming sequences are depicted by pink box and displayed according to their sequence positions in *AGAP2* mRNA transcript, the width of the box indicates the length of the G4 consensus. **(B)** GRIP performed in DU145 cells transfected with either an empty vector (No 5' UTR) or the same vector with *AGAP2* longer 5' UTR in front of the *Renilla* luciferase gene. The levels of *Renilla* mRNA in the immunoprecipitated samples are normalised by their input controls. A nonspecific isotype antibody (IgG) was used as a negative control. Differences between samples were analysed by unpaired two-tailed t-tests, *P*-values shown. All the data shown correspond to three independent immunoprecipitations (n=3) and the error bars denote standard deviation. (**P* < 0.05; ns: not significant).

The results highlighted the effectiveness of our GRIP technique to pulldown mRNAs with G4 motifs and also confirmed the formation of a G4 structure in the 5' UTR of *AGAP2* mRNA. Recently, Maltby *et al.* (2020) attempted RNA immunoprecipitation using the BG4 antibody and demonstrated enrichment of G4 motifs in the 5' UTR of *Task3* mRNA. The authors used a sonicated homogenate for immunoprecipitation and the RNA was incubated at a higher temperature (70°C) for a longer duration (1 hour) to elute the bound nucleic acid. However, these steps (homogenisation and longer incubation at high temperature) could enhance the detection of DNA G4 instead of RNA G4. To address these caveats and successfully immunoprecipitate G4 in RNA only, our GRIP method employed digitonin to selectively enrich cytosolic RNA and utilised RNase-free DNase I to degrade any trace amount of genomic DNA. We also adopted a unique elution and RNA precipitation strategy that would facilitate selective detection of G4-containing RNA.

Despite the effectiveness of our GRIP method to successfully enrich G4-containing mRNA population, the dynamic of G4 formation and its folding and unfolding patterns are still a matter of debate. The formation of intracellular G4 structures is dependent on different factors including binding of RBP, levels of helicases, and ion concentrations (Reviewed in Cammas and Millevoi, 2017). The complex interplay between these factors determines the formation of the RNA G4 motif inside the cell. A study by Guo and Bartel (2016) using complementary approaches (Dimethyl sulphate treatment and selective 2'-hydroxyl acylation analysed by primer extension) revealed that a large number of predicted RNA G4 are overwhelmingly unfolded in the cells. Moreover, a study by Chen *et al.* (2018) reported that a rapid dynamic transition occurs between folding and unfolding states for some RNA G4s. The transient nature of RNA G4 folding is also exhibited by Yang *et al.* (2018), who used crosslinking in their study to capture these structures. Taking these findings together, the RNA G4 exist in an equilibrium between transiently folded and unfolded state and are influenced by a variety of competing factors. Unlike other studies that use formaldehyde and G4 stabilising ligands to 'fix' these structures (Yang *et al.,* 2020, Yang *et al.,* 2018), our GRIP method captures these structures in their native state. Our GRIP method has proven to be an effective technique to enrich mRNA with G4 structures and has successfully demonstrated the pulldown of *AGAP2* 5' UTR G4 using appropriate controls.

## 4.3.2 Translational consequences of 5’ UTR G4 motif in *AGAP2*:

5’ UTR G4 structures have been previously shown to suppress mRNA translation in most cases by disrupting key processes including the formation of PIC and ribosome scanning and translocation (Bugaut and Balasubramanian, 2012, Beaudoin and Perreault, 2010). In our study, we utilised two different approaches to examine the translation consequences: dual-luciferase reporter assays and polysome profiling. These distinct techniques would facilitate a broader understanding of the impact of 5’ UTR G4 on mRNA translation efficiency and output.

### 4.3.2.1 Reporter assay to evaluate the impact on translational output mediated by 5’ UTR G4:

The effects of 5’ UTR G4 structures on mRNA translation were evaluated using a dual-luciferase reporter assay with a bicistronic plasmid. The plasmid expresses *Renilla* and *Firefly* luciferase reporter with the translation of *Renilla* cistron mediated by an upstream cloned fragment. The Firefly cistron undergoes cap-independent translation through the poliovirus IRES and serves as an internal control. We generated dual-luciferase reporter constructs comprising either shorter 5’ UTR without G quadruplex forming sequences (as noted in PC cell lines), longer 5’ UTR containing G quadruplex forming sequences (observed in CML cell lines) and mutated longer 5’ UTR with mutations in G quadruplex consensus (section 2.2.6.1; *Table 2.8*; *Figure 2.4A, 2.4B*). These 5’ UTR variants were fused to *Renilla* luciferase ORF. Since only one unique restriction site (*NheI*) was used in the plasmid to insert the fragment proximal to the *Renilla* luciferase, several attempts were made to correctly insert the fragment into the reporter vector. The Sanger sequencing was used to verify the required positive clones (*Figure 4.8*).

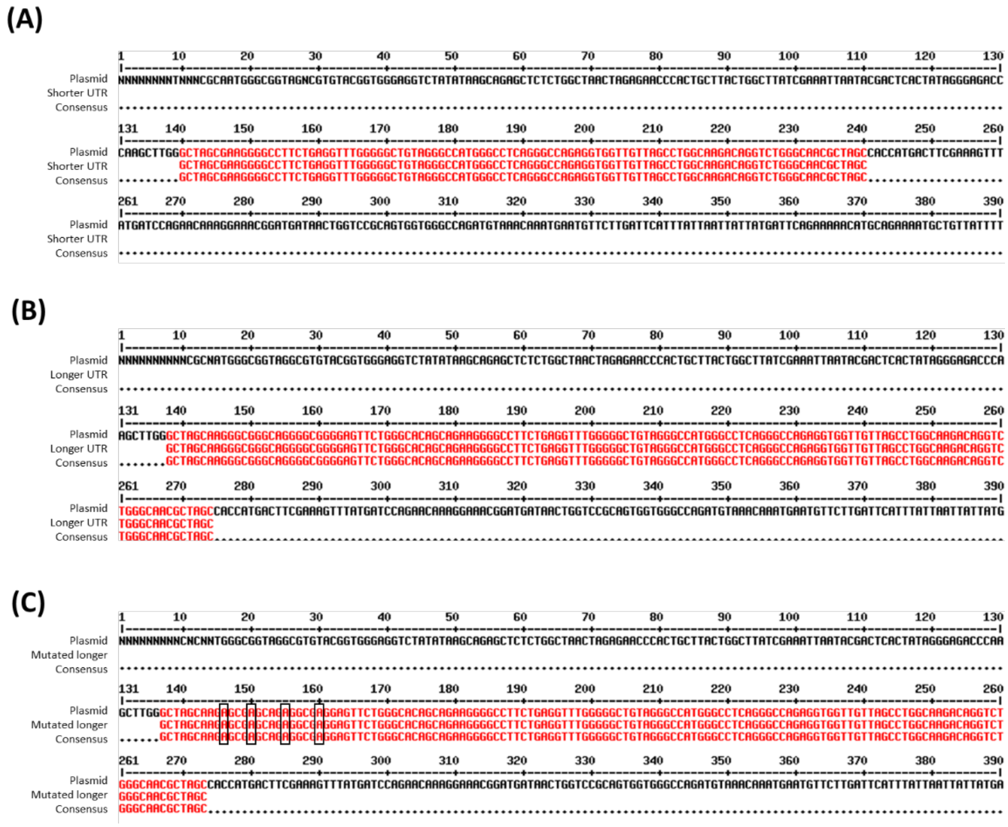


**Figure 4.8: Alignment of sequenced clones to verify insertion of 5' UTR fragment.** The 5' UTR fragments (*Table 2.8*) were inserted at the unique *NheI* restriction site (section 2.2.6.1). The constructs were transformed into DH5α competent cells, cultured in selecting LB medium, and positive clones were sequenced using Sanger sequencing. The alignment of sequenced clones shows correct insertion of shorter 5' UTR **(A)**, longer 5' UTR **(B)**, and mutated longer 5' UTR fragment **(C)**. The (G→ A) mutations in the mutated longer 5' UTR is highlighted using a black box.

These inserted fragments influence the translation of *Renilla* luciferase mRNA. The impact on the mRNA translation was examined by measuring the levels of Renilla luciferase protein using the DLR assay, which estimate the levels of protein by analysing the luciferase enzymatic activity (section 2.2.6.3). Using the *in vitro* transcription and translation system, we observed that the longer 5' UTR isoform containing the G4 forming sequence induced a significant decrease in the normalised reporter activity relative to the shorter 5' UTR. This effect was reversed by mutating the G quadruplex consensus that prevented the formation of these secondary structures (*Figure 4.9A*). We also transfected these reporter constructs into DU145 (PC) and KU812 (CML) cell lines and observed similar shifts in relative reporter activity (*Figure 4.9B, 4.9C*). Together, these results suggested that the G4-containing longer 5' UTR decreases the translation potential of mRNA which could be rescued by mutating the G4 consensus.

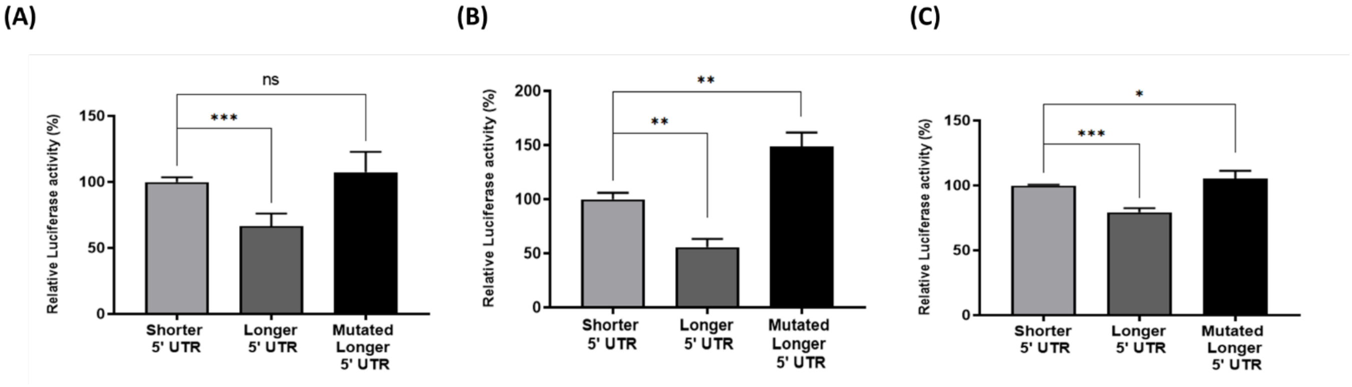


**Figure 4.9: Impact of 5' UTR G4 on the translational output using the dual-luciferase reporter assay. (A)** The dual-luciferase reporter activity of the reporter constructs using the *in vitro* transcription and translation system. The graph is the mean +/- SD of 4 independent experiments (n =4) and expressed as relative Rluc/Fluc ratio. Differences were analysed using a Kruskal-Wallis [H (2) = 47.13, *P* =< 0.001] followed by uncorrected Dunn's test, *P*-values shown. **(B)** Reporter activity after transfecting the plasmids in DU145 (PC cell line) using JetPRIME transfection reagent (section 2.2.6.2.1) and analysing the luciferase activity 48 hours post-transfection. The graph represents the mean of 3 independent experiments (n=3) and expressed as relative Rluc/Fluc ratio. Differences were analysed using a Kruskal-Wallis [H (2) = 23.79, *P* =< 0.001] followed by uncorrected Dunn's test, *P*-values shown. **(C)** Reporter activity measured in KU812 (CML cell line) after 6 hours of electroporation (section 2.2.6.2.2). The graph represents the mean of 3 independent experiments (n=3) and is expressed as relative Rluc/Fluc ratio. Differences were analysed using a Kruskal-Wallis [H (2) = 40.49, *P* =< 0.001] followed by uncorrected Dunn's test, *P*-values shown. (**P* < 0.05; ***P* < 0.01; ****P* < 0.001; ns: not significant). The *Firefly* luciferase was used as an internal control. The Rluc/Fluc ratio of the shorter 5' UTR for *in vitro* transcription and translation assay **(A)** and for the transfected cells in **(B)** and **(C)** were 18.84, 2.01 and 3.27, respectively, and this was equalled to 100).

Our reporter assay results from the *in vitro* transcription and translation system showed about ~ 45% decrease in the relative luciferase activity compared to shorter 5' UTR due to the presence of G4 motif in the longer *AGAP2* 5'

UTR isoform. Other studies evaluating the effect of 5′ UTR G4 on the luciferase activity using the cell-free system reported ~70% decrease for *NRAS* (Kumari *et al.,* 2007) and *ADAM10* (Lammich *et al.,* 2011), and ~85% for *ERS1* (Balkwill *et al.,* 2009)*.* In our study, the transfection of the reporter constructs into DU145 and KU812 cell lines exhibited ~35% and ~21% decrease, respectively. Other comparable studies reported ~80% decrease for *ZIC1* (Arora *et al.,* 2008), ~55% for *MMP16* (Morris and Basu, 2009), ~45% for *EBAG9* (Beaudoin and Perreault, 2010)*,* ~50% for *BCL-2* (Shahid *et al.,* 2010), ~60% for *FZD2* (Beaudoin and Perreault, 2010), ~70% for *TRF2* (Gomez *et al.,* 2010), and ~35% for *NCAM2* and *THRA* (Beaudoin and Perreault, 2010)*.* The studies cited above used different cell lines for the transfection of reporter constructs. The HELA cells were used by Arora *et al.* (2008) and Morris and Basu (2009). The HEK 293 cells were used by Beaudoin and Perreault (2010) and Gomez *et al.* (2010). All the studies cited above used a chemical-based transfection method (Lipofectamine). In our study, we used PC cell line (DU145) and CML cell line (KU812) for transfection using chemical method and electroporation, respectively.

We also noted a significant increase in the relative luciferase level of the mutated longer 5′ UTR compared to the shorter 5′ UTR isoform in the cell-free system and after transfecting the plasmids into the KU812 cell line (*Figure 4.9A, 4.9C*). The increase in the relative translation could result from certain sequence features that are introduced while mutating the runs of guanine to destroy the G4 consensus and requires further evaluation. Nonetheless, our reporter experiments underlined the relevance of the *AGAP2* 5′ UTR G4 motif in suppressing translation. In contrast to other studies above which examined the 5′ UTR G4 in different gene transcripts, our study emphasised the relevance of the G4 motif in alternatively transcribed 5′ UTR isoforms that are less than 50 bp apart and uncovered this novel layer of gene expression regulation.

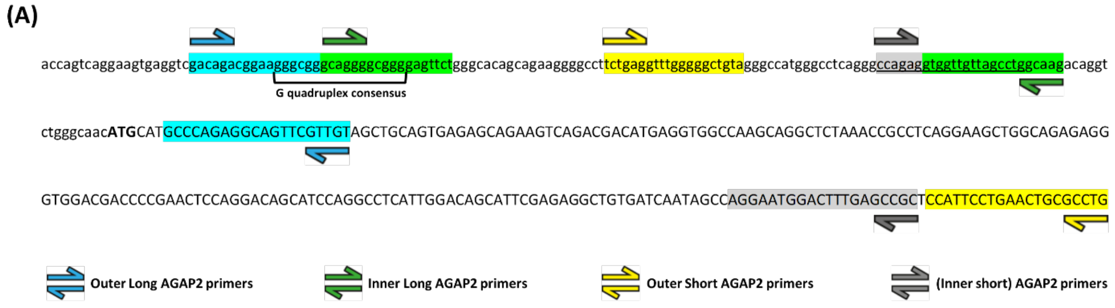
In our reporter experiments, we noted that the impact of 5′ UTR G4 was less profound in KU812 (~21% decrease) compared to DU145 (~35% decrease), even though, the patterns of relative luciferase activity were similar. This could be explained by the differences in the method of transfection (see Appendix 3). Due to the differences in the transfection method (chemical-based method vs. electroporation), the post-transfection timings for the collection of the cells were selected accordingly. For cells transfected using electroporation, an earlier timepoint was selected as the electro-permeabilization of the membrane results in the direct and rapid transfer of the foreign DNA and hence the earlier expression compared to chemical-based methods (Kim and Eberwine, 2010).

### 4.3.2.2 Polysome association profile of longer and shorter 5′ UTR isoforms:

The polysome profiling provides valuable information about the translational status of specific mRNA by determining the ribosome density on the given mRNA (Chassé *et al.,* 2016). The polysome loading and occupancy have been frequently used as a proxy for translational efficiency in the literature and have been previously used to examine the effects of different mRNA structural features including G4s on protein translation (Faye *et al.,* 2013, Murat *et al.,* 2018, Thandapani *et al.,* 2015). A variety of studies have also used polysome profiling to study the translational potential of TSS isoforms derived from alternative promoters (Li *et al.,* 2019, Wang *et al.,* 2016). In our study, we performed polysome profiling to evaluate the association of poly-ribosome with mRNA transcripts having shorter or longer versions of the 5′ UTR. For the purpose of this experiment, the 5′ UTR isoform containing the full

G4 consensus sequence were classified as longer 5' UTR whereas the remaining others that do not contain any part of the consensus were grouped as shorter 5' UTR isoforms.

The longer and shorter *AGAP2* 5' UTR isoforms were detected using the nested PCR amplification (*Figure 4.10A*) and normalised using luciferase RNA spike to control for differences in recovery. The PCR data were processed as described in section 2.2.8.2.4. A representative example of calculation for longer and shorter *AGAP2* 5' UTRs in different fractions of TCCS cell line is displayed in *Figure 4.10B*.


(A)
accagtcaggaagtgaggtcgacagacggaagggcgggcaggggcggggagttctgggcacagcagaaggggccttctgaggtttgggggctgtagggccatgggcctcagggccagaggtggttgttagcctggcaagacaggt
G quadruplex consensus
ctgggcaacATGCATGCCCAGAGGCAGTTCGTTGTAGCTGCAGTGAGAGCAGAAGTCAGACGACATGAGGTGGCCAAGCAGGCTCTAAACCGCCTCAGGAAGCTGGCAGAGAGG
GTGGACGACCCCGAACTCCAGGACAGCATCCAGGCCTCATTGGACAGCATTCGAGAGGCTGTGATCAATAGCCAGGAATGGACTTTGAGCCGCTCCATTCCTGAACTGCGCCTG
Outer Long AGAP2 primers
Inner Long AGAP2 primers
Outer Short AGAP2 primers
(Inner short) AGAP2 primers

(B)

| Longer 5' UTR AGAP2 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Fraction | Ct luc | Ct target | △Ct =Ct luc -Ct target | △△Ctn= △Ctn -△Ct1 | En=2^^Ctn | E total | Proportion = 100 X En/Etotal |
| 1 | 14.59 | 17.56 | -2.97 | 0.00 | 1.00 | 35.39 | 2.8 |
| 2 | 15.54 | 14.69 | 0.85 | 3.82 | 14.12 | | 39.9 |
| 3 | 15.92 | 16.84 | -0.92 | 2.05 | 4.14 | | 11.7 |
| 4 | 15.57 | 16.71 | -1.14 | 1.83 | 3.56 | | 10.0 |
| 5 | 16.29 | 17.55 | -1.26 | 1.71 | 3.27 | | 9.2 |
| 6 | 16.26 | 17.81 | -1.55 | 1.42 | 2.68 | | 7.6 |
| 7 | 16.67 | 18.65 | -1.98 | 0.99 | 1.99 | | 5.6 |
| 8 | 17.65 | 19.18 | -1.53 | 1.44 | 2.71 | | 7.7 |
| 9 | 16.10 | 18.97 | -2.87 | 0.10 | 1.07 | | 3.0 |
| 10 | 15.91 | 19.12 | -3.21 | -0.24 | 0.85 | | 2.4 |
| **Shorter 5' UTR AGAP2** | | | | | | | |
| Fraction | Ct luc | Ct target | △Ct =Ct luc -Ct target | △△Ctn= △Ctn -△Ct1 | En=2^^Ctn | E total | Proportion = 100 X En/Etotal |
| 1 | 14.59 | 15.08 | -0.49 | 0.00 | 1.00 | 3585.57 | 0.0 |
| 2 | 15.54 | 13.08 | 2.46 | 2.95 | 7.73 | | 0.2 |
| 3 | 15.92 | 11.06 | 4.86 | 5.35 | 40.79 | | 1.1 |
| 4 | 15.57 | 9.78 | 5.79 | 6.28 | 77.71 | | 2.2 |
| 5 | 16.29 | 9.98 | 6.31 | 6.80 | 111.43 | | 3.1 |
| 6 | 16.26 | 7.82 | 8.44 | 8.93 | 487.75 | | 13.6 |
| 7 | 16.67 | 7.95 | 8.72 | 9.21 | 592.22 | | 16.5 |
| 8 | 17.65 | 8.85 | 8.80 | 9.29 | 625.99 | | 17.5 |
| 9 | 16.10 | 6.94 | 9.16 | 9.65 | 803.41 | | 22.4 |
| 10 | 15.91 | 6.69 | 9.22 | 9.71 | 837.53 | | 23.4 |

**Figure 4.10: Primer designing and representative analysis of polysome profiling data. (A)** Primers for nested PCR amplification to detect longer and shorter *AGAP2* 5' UTR isoform. For longer 5' UTR, the first PCR was performed using outer long AGAP2 primers. The PCR product was diluted 50 folds and amplified again using inner long primers. To amplify shorter 5' UTR, the PCR was carried out with outer short primers and the diluted PCR product was reamplified using AGAP2 primers. The sequences of the primers are listed in *Table 2.4*. The figure shows the first 373 nucleotides of *AGAP2* mRNA transcribed from the furthest upstream TSS annotated in the FANTOM database. The coding sequence is presented in UPPERCASE (with ATG marked in bold) and the 5' UTR is presented in lowercase. Primer sequences are shown by respective coloured box and the inner short forward primer in the overlapped region is shown by underline. **(B)** Example calculation for polysome profiling data of TCCS (replicate 1) using qRT-PCR with luciferase (luc) spike in. The calculations are shown for longer and shorter UTR isoforms amplified using nested PCR in (A). The values of proportions were used to plot the graphs.

Interestingly, we noted a decreased polysome association of *AGAP2* mRNA with the longer 5’ UTR compared to shorter 5’ UTRs (*Figure 4.11C, 4.11D*). The polysome profiling was performed with two CML cell lines (KU812 and TCCS) in which the higher relative levels of longer 5’ UTR have been established (*Figure 3.13B, 3.13C*). As evident in *Figure 4.11C* and *4.11D*, the longer UTR was primarily enriched in the non-polysomal ribonucleoprotein (RNP) fraction (Fraction 1-3). On the other hand, the shorter UTR was mainly enriched in the polysome fraction (6-10). We also did not observe any changes in the global RNA polysome profiles in these cell lines (*Figure 4.11A, 4.11B*). By pooling the data of non-polysomal (fraction 1-5) and polysomal (fraction 6-10) from both the cell lines, we noted significantly increased levels of longer 5’ UTR in the non-polysomal fraction compared to mRNA with a shorter 5’ UTR length ($P < 0.001$) (*Figure 4.11E*). It implies inefficient translation of mRNA population with longer 5’ UTRs. The polysomal profiling was not performed in the PC cell lines because the primary objective of the experiment was to compare polysome association between the longer vs. shorter 5’ UTR isoforms; since the levels of longer 5’ UTR were very low in PC cell lines, these were not ideal for comparing longer and shorter 5’ UTR isoforms levels.

However, the primers designed for the polysomal profiling experiment could specifically amplify the longer 5’ UTR isoforms only. The sequences amplified in the shorter 5’ UTR isoforms are also shared by the longer version (*Figure 4.10A*). Therefore, the nested PCR for the shorter isoforms would in effect amplify the entire 5’ UTR population including the longer isoforms. Since most of the shorter transcript isoforms are derived from the annotated major TSS that does not contain the G4 consensus (*Figure 3.12B*), the amplified product would mostly contain the transcripts generated from the major TSS and would have lower levels of longer transcript variants. This would not affect the interpretation of our results as the primers for the shorter UTR would amplify products representative of the major TSS. Even excluding shorter UTR isoforms from the analysis, the result of the polysome fractionation would still remain valid, i.e., the longer version of the transcript poorly associated with polysome and are enriched in the non-polysomal fraction.

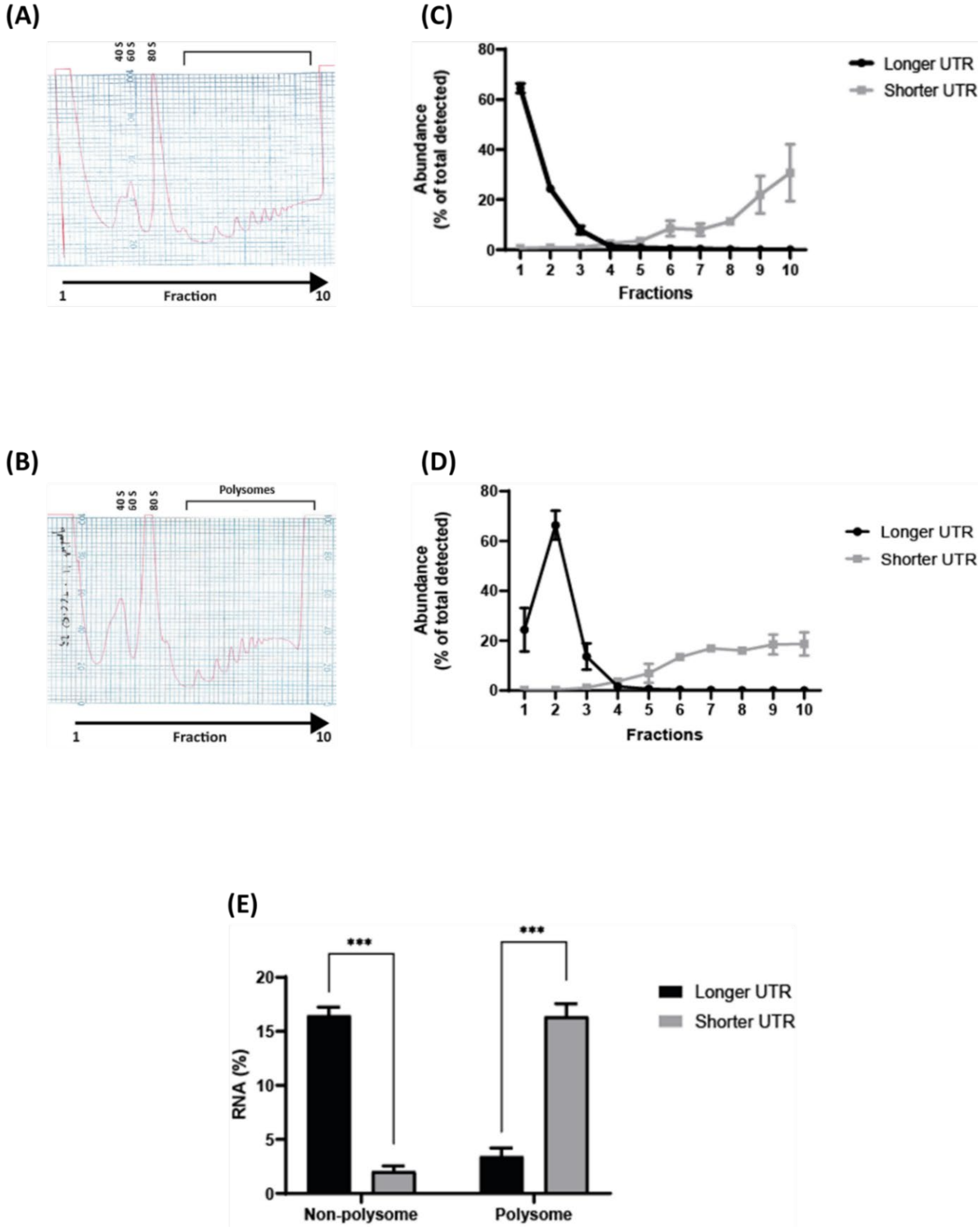


**Figure 4.11: Polysome association of longer and shorter 5' UTR isoforms of *AGAP2* mRNA.** Lysates for polysome profiling were prepared from KU812 and TCCS (CML cell lines) and fractionated through a sucrose gradient. The profiles were monitored by measuring the absorbance at 254 nm ($A_{254nm}$). The representative polysome profiles are shown for KU812 **(A)** and TCCS **(B)**. The relative distribution of *AGAP2* mRNA with longer and shorter 5' UTR are shown in polysome fractions 1-10 of KU812 **(C)** and TCCS **(D)**. The RNA distribution is presented as the fraction of the total RNA recovered. The mRNA levels were normalised to the exogenous spike-in luciferase control mRNA. The graphs above represent the means ± SEM of 2 independent experiments (n=2). **(E)** Relative levels of *AGAP2* mRNA with longer and shorter 5' UTR in non-polysomal (Fraction 1 – 5) and polysomal (Fraction 6 –10) segments pooled from both the cell lines. The data represent the means ± SD of the fraction of the total RNA recovered and *P*-values were calculated by an unpaired students t-test, ***$P$ < 0.001. **Note:** the polysome profiles [A, B] were generated by Dr Keith Spriggs from the University of Nottingham.

Other studies in the literature have also pointed out the poor translational efficiency of the longer 5′ UTR (Arrick *et al.,* 1991, Davuluri *et al.,* 2000, Sobczak and Krzyzosiak, 2002, Wang *et al.,* 2016). A study by Wang *et al.* (2016) has conducted a systematic analysis of isoform-specific translation and have identified different cis-regulatory features in the longer 5′ UTR that contributed to poor translation capability. The authors also conducted a non-linear regression modelling, integrating a variety of regulatory 5′ UTR features which together explain 57% of the variance in the observed translation efficiency differences between TSS isoforms. The two single best predictors of translation efficiency differences, as identified by the authors, were uORFs and 5′ UTR length. The authors did not analyse the contribution of 5′ UTR G4 motifs which could account for some of the remaining unexplained variations. An earlier study by Rojas-Duran and Gilbert (2012) has also demonstrated large differences in the translational efficiency for mRNA transcribed from alternative starting sites in multiple promoters. The alternate TSS selection within a core promoter could also potentially mediate divergent translational profiles due to alternatively transcribed G4 structures which, as evidenced through our study, could significantly impact mRNA translational output (*Figure 4.9*).

The longer UTR provides a favourable platform to harbour a variety of regulatory elements and structural features that could modulate mRNA translation (Reviewed in Leppek *et al.,* 2018). Most of the longer UTR isoforms reported in the literature above were derived from alternative promoters and are usually > 500 bp longer compared to their shorter counterpart (Landry *et al.,* 2003, Kimura *et al.,* 2006). However, a recent study by Palavecino *et al.* (2020) has reported the inhibitory effects of a longer 5′ UTR, encoded by alternative TSS selection, which contained an uAUG in the extra ~100 bps incorporated in the longer 5′ UTR. Similarly, in our study, we have identified a G4 structure in the alternative longer 5′ UTR that differs from the shorter isoform by only 36 additional nucleotides. To our knowledge, ours is the first study that has determined physiologically small changes in the length of the 5′ UTR isoform that significantly impact on polysome seeding.

G4 motifs in the 5′ UTR have been reported to associate poorly with the ribosome. Reduced ribosomal occupancy has been noted for G4-containing mRNAs by a variety of studies (Cammas *et al.,* 2015, Murat *et al.,* 2014, Yang *et al.,* 2020). In a study by Murat *et al.* (2014), the authors noticed a reduced abundance of *VEGF* mRNA in the polysomal fractions after treating the HeLa cells with a G4 stabilising ligand. A review of the literature has shown that the 5′ UTR G4, like other secondary structures, does not interact directly with the polysome but exerts its effects indirectly by impairing assembly of the translation initiation machinery and perturbation of ribosome scanning toward the start codon (Babendure *et al.,* 2006, Bugaut and Balasubramanian, 2012, Jenkins *et al.,* 2010, Koromilas *et al.,* 1992, Kozak, 1989, Kozak, 1986). In our study, we also found a decreased association of the polysomes with G4-containing longer 5′ UTR isoforms which resulted in poor mRNA translation exhibited by the enrichment of the longer isoforms in the non-polysomal fraction. This form of regulation in which the mRNA translational potential could change drastically by the incorporation of a few extra nucleotides (<50 bp) that form a stable G4 motif has not been defined before.

## 4.4 Summary of findings (Chapter 4)

In summary, this chapter builds upon the findings from the previous chapter and confirmed the formation of G4 structures by the alternatively transcribed G4 consensus sequences encoded in the longer 5' UTR of *AGAP2* mRNA by an upstream TSS selection, as noted in the CML cell lines. Using an in-house developed immunoprecipitation technique (GRIP), we demonstrated the *in cellula* formation of the G4 motif in the 5' UTR of *AGAP2* mRNA. Combining the GRIP results with our earlier *in vitro* CD spectroscopy experiments, we provided conclusive evidence for the formation of an alternatively transcribed G4 structure in the longer 5' UTR of *AGAP2* mRNA. We have also exhibited differential expression of the longer G4-containing *AGAP2* 5' UTR isoform with higher levels noted in the CML cell lines and validated it using 5' RLM-RACE amplified cDNA. This chapter also highlighted the functional consequences of the 5' UTR G4 motif using reporter assays and polysome profiling. Our dual-luciferase reporter assay data showed a significant reduction in the luciferase activity exhibited by the G4-containing longer UTR which was reversed by destroying the G4 structure by mutating the consensus sequence. These G4-containing isoforms were also enriched in the non-polysomal fraction indicating inefficient translation. Together, these results suggest that earlier TSSs are frequently selected in CML relative to PC cell lines resulting in mRNA with longer 5' UTR containing G4 forming sequences. These sequences fold into a stable G4 and influence the translation ability of *AGAP2* mRNA, contributing to the observed discrepancy between mRNA and protein expression.

# Chapter 5:
# Bioinformatics analysis to identify alternatively transcribed G4 forming sequences in the human genome

## 5.1 Introduction

Based on our work presented in the previous chapters, we have elucidated a novel mechanism that contributed to *AGAP2* regulation in the CML cell lines. The results presented in chapter 3 described the discrepancy in AGAP2 mRNA and protein levels in PC and CML cell lines, with higher mRNA and lower protein output noted in the CML cell lines. We also noted the differential distribution of TSSs in these cell lines with earlier TSSs found in the CML cell lines that incorporated a G4 forming sequence in the resulting longer 5′ UTR isoforms. Our experiments in Chapter 4 confirmed the formation of a stable G4 structure by these extra nucleotides. We also validated the increased levels of these G4-containing *AGAP2* 5′ UTR isoforms in the CML cell lines using qRT-PCR. Moreover, in chapter 4, we highlighted the translational impact of these alternatively transcribed 5′ UTR G4 and observed a substantial decrease in the translation efficiency mediated by these structures. *Figure 5.1* summarises the understanding acquired through the experiments conducted in the previous chapters. Together, these findings reveal a novel mechanism that controls *AGAP2* expression in the CML cell lines and could contribute to the observed mismatch in mRNA and protein levels.

The discrepancy in the RNA and protein expression has been frequently reported in the literature and previous studies linking differential TSS selection to the impaired translational output usually analysed TSSs originating from alternative promoters (section 1.3.1, section 3.1.2). Previous studies have reported that a gene displays a distinct distribution of TSSs within a core promoter region in different cell lines (Carninci *et al.,* 2005, Ohmiya *et al.,* 2014) which could contribute to the cell-specific gene regulation. In our study, using *AGAP2* as a model, we evidenced the impact of alternate TSSs on controlling the translational efficiency of the mRNA isoforms. To our knowledge, the consequences of alternative TSS selection, which are separated by only a few nucleotides, have not been studied before.

Earlier studies have demonstrated significant enrichment of G4 forming sequences in the 5′ UTR with 2,034 putative G quadruplex sequences found in annotated major 5′ UTR of protein-coding genes (Bedrat *et al.,* 2016, Huppert *et al.,* 2008). However, the presence of G4s in the alternatively transcribed 5′ UTR isoforms have not been studied before. The differential TSS selection generates a heterogeneous mRNA 5′ UTR population that could contain different regulatory features influencing mRNA translation potential. Considering only the major TSS isoforms for discovering 5′ UTR regulatory elements would not provide a complete picture and would miss key regulatory elements. The transcript isoform derived from the major TSSs have been the focus of attention and minor TSSs were largely considered as nonadaptive and products of molecular errors (Xu *et al.,* 2019). However, as documented through our work using *AGAP2*, alternatively transcribed G4 motifs by upstream TSS selection significantly affect mRNA translation potential. We believe that the TSS-G4 mediated mechanism which we discovered for *AGAP2* might be also implicated in regulating the expression profile of other genes.

Given the potential of alternatively transcribed G4 motifs in influencing mRNA translation, it could lead to the discrepancy in mRNA and protein level, where higher levels of mRNA do not translate into protein. Finding the genes with alternatively transcribed G4 and analysing their translational impact would produce new knowledge regarding the contribution of the TSS-G4 mechanism in mediating divergent translation profiles. It would also delineate the relevance of this novel mechanism in regulating the expression of other genes in addition to *AGAP2*.

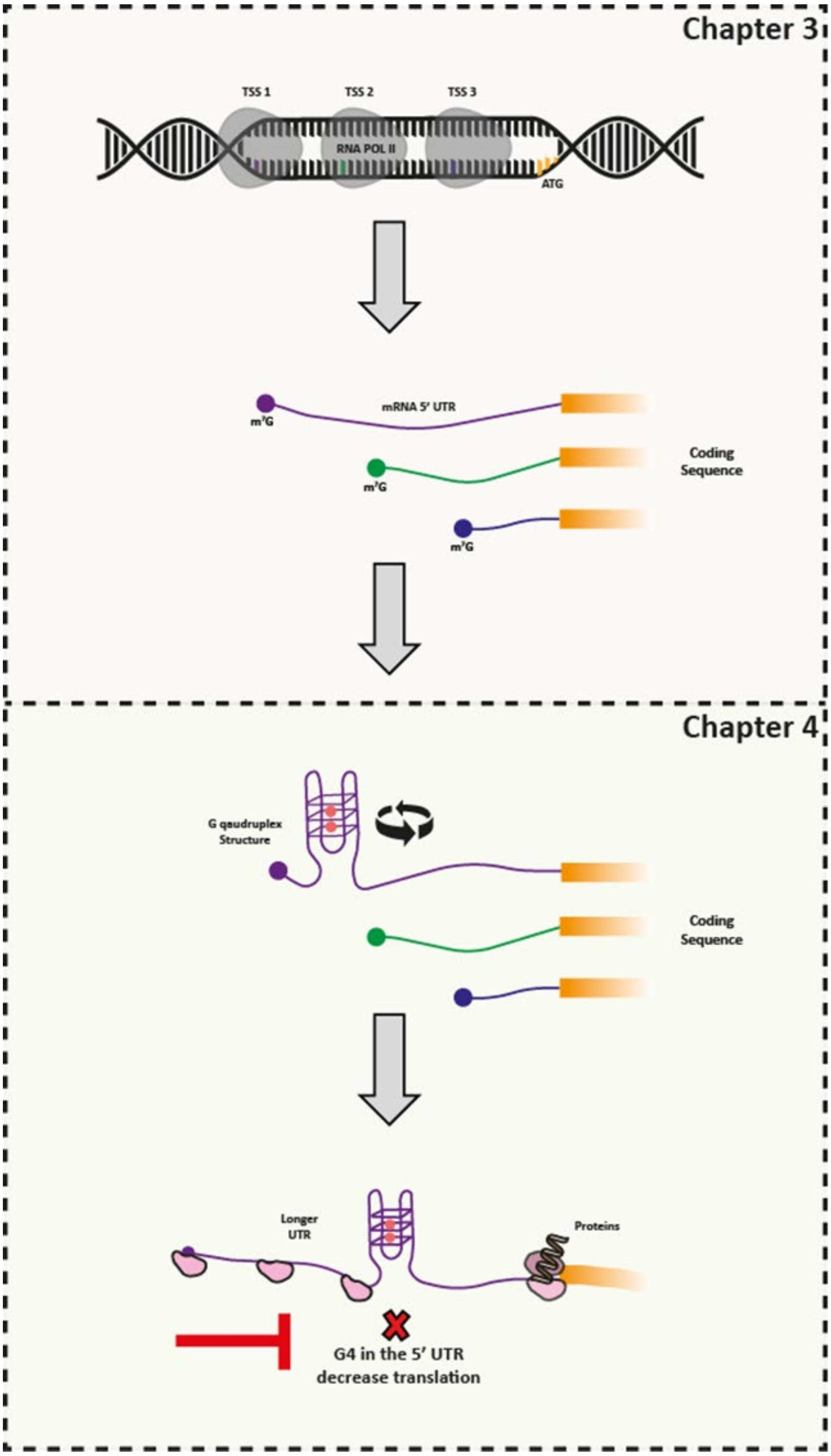


**Figure 5.1: Summary of the results generated in the previous chapters.** Brief overview of the data produced in chapter 3 (top) and chapter 4 (bottom). Through our work in the previous chapters, we noted differential TSS distribution in PC and CML cell lines that produced a heterogeneous mRNA population with varying length of 5' UTRs. The longer 5' UTR contain the G4 consensus that folds into stable G4 and decreased the polysome association, resulting in reduced translation efficiency.

### 5.1.1 Aims of chapter 5:

To address the gaps in the literature and identify alternatively transcribed G4 forming sequences in the human genome, we designed a bioinformatics pipeline to generate a list of genes that encode a G4 consensus in the divergent TSS isoforms by the selection of upstream TSSs. It would facilitate the identification of other genes that are susceptible to TSS-G4 mediated regulation, as noted for *AGAP2*. For these genes, we also analysed the differential expression of G4-containing transcript variants in PC and CML cell lines. To find relevant genes for validating this novel mechanism, we used different datasets to identify genes that demonstrated inconsistency in mRNA and protein expression and showed higher usage of G4 forming TSS.

This chapter aims to:

- Identify other genes in the human genome with alternatively transcribed G4 forming sequence in the variant 5' UTR isoforms, derived from an upstream TSS selection within the same TSS cluster.
- For the genes with alternatively transcribed G4, map the differential distribution of G4 forming TSSs in PC and CML cell lines, that will encode the G4 consensus sequences in the longer 5' UTR isoforms.
- Identify other genes in PC and CML cell lines that display a discordant mRNA and protein profile with higher relative mRNA level and lower protein abundance, as reported for *AGAP2*.
- Generate a list of genes for validation that display a higher level of mRNA, lower protein abundance, and higher relative expression of G4 forming transcript variants.
- Validate the TSS-G4 mechanism in other genes and evaluate its relevance in modulating mRNA translation output.

## 5.2 Designing the Bioinformatics pipeline

A bioinformatics pipeline was designed to meet the aims outlined above. The pipeline was designed combining various packages and was executed in different platforms including R, Python, and Microsoft Excel. The pipeline is principally divided into three steps as depicted in *Figure 5.2*. Briefly, in step 1, the TSS data was downloaded from the FANTOM database followed by converting the data to a relevant format for analysis. The converted TSS data were subsequently mapped to the +/- 50 bp region around the genes annotated 5' UTR regions. In step 2, the sequences between alternative TSSs were extracted and analysed for the presence of G4 forming sequences using the pqsfinder package. The list of genes with alternatively transcribed G4 was curated and put into pathway maps using Metacore to identify enriched pathways, network, and processes. In step 3, the differential distributions of these G4 encoding TSSs were determined in PC and CML cell lines that would incorporate the G4 consensus in the differential 5' UTR isoforms. The NCI-60 microarray and SWATH-MS data were also processed to identify genes in PC and CML cell lines that displayed higher mRNA expression and lower protein abundance. The genes with discordant mRNA and protein levels were then combined with genes showing the differential expression of G4 TSS to generate a list of genes in PC and CML cell lines that had higher mRNA and lower protein levels and increased relative expression of G4 forming TSS. This gene list was then used to validate the TSS-G4 mechanism.

### 5.2.1 STEP 1: Downloading and annotating FANTOM TSSs:

#### 5.2.1.1 Downloading FANTOM database:

The TSS data was downloaded from the FANTOM (Functional Annotation of the Mammalian Genome) database. The FANTOM database (version 5) is the largest collection of human TSS profiles and contain a diverse range of cells and tissues, sequenced using a single platform, to provide a comprehensive map of all the TSSs used in the human (Forrest *et al.,* 2014, Lizio *et al.,* 2015). The FANTOM TSS database for human contains the TSS profiles for 573 primary cell lines, 152 human post-mortem tissues and 250 different cancer cell lines covering distinct cancer subtypes.

The 5' end of the mapped CAGE reads (tags) were counted at a single bp resolution, and the counts were normalised as tags per million (TPM) after scaling by normalization factors calculated by Relative Log Expression (RLE) – an expression of 5 TPM means that out of a million transcripts, 5 transcripts corresponded to the TSS in question. The normalised counts were downloaded from the ZENBU browser, a data exploration and mining tool that displays the TSS data from all the CAGE experiments in the FANTOM database (Severin *et al.,* 2014). The data was download in the browser extensible data (.bed) format which presents the genomic coordinates (chromosome location - start and end positions) of the normalised collective count of tag-starting sites. The raw (.bed) data was processed and the TSSs with less than 2 TPM were removed to select robust starting sites. The data was split up into plus (+) and minus (-) strands for easier handling and analysis. The data were processed and analysed using a high-power computer available at the John van Geest Cancer Research Centre located in Nottingham Trent University.

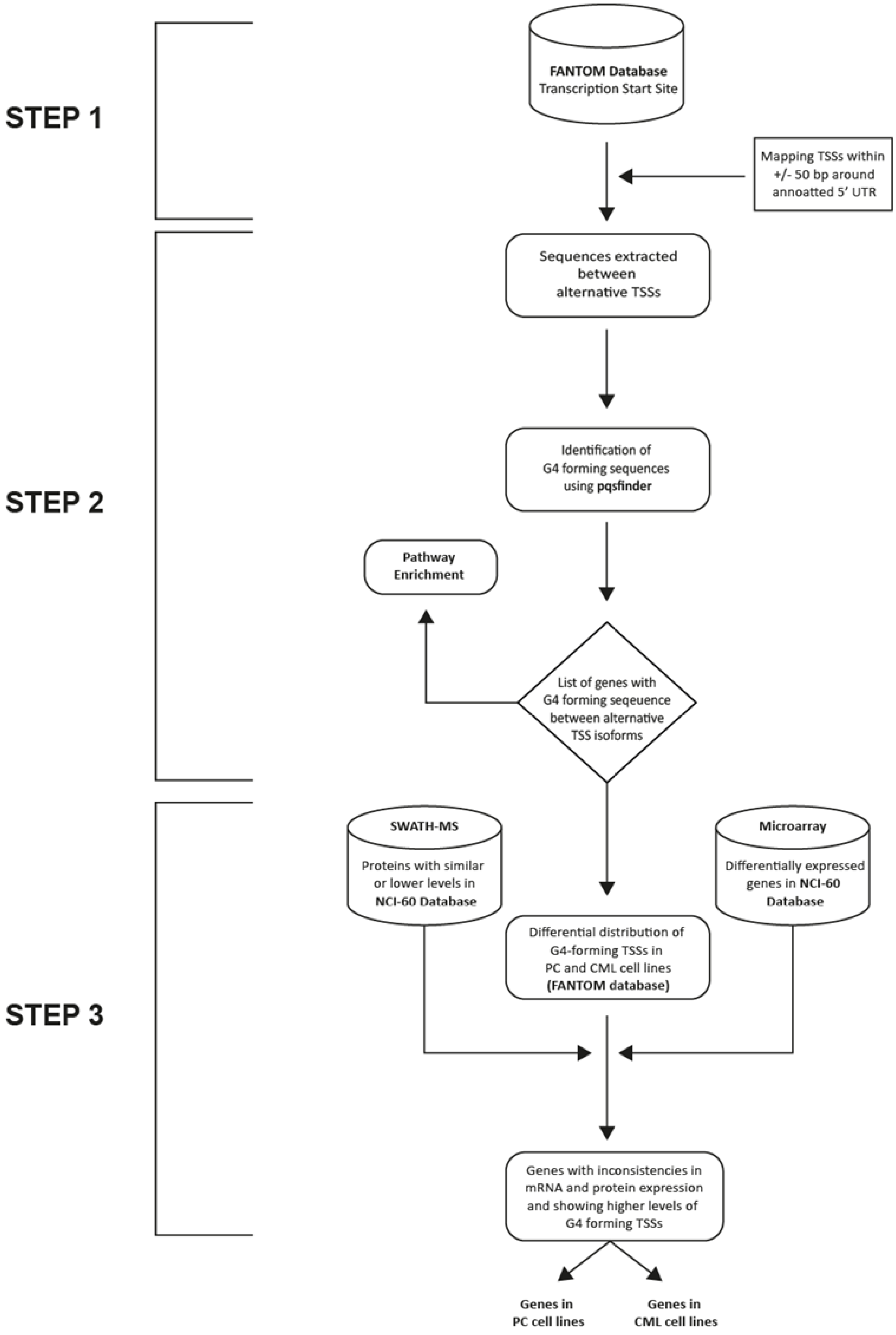

STEP 1
STEP 2
STEP 3
FANTOM Database
Transcription Start Site
Mapping TSSs within +/- 50 bp around annoatted 5' UTR
Sequences extracted between alternative TSSs
Identification of G4 forming sequences using pqsfinder
Pathway Enrichment
List of genes with G4 forming seqeuence between alternative TSS isoforms
SWATH-MS
Proteins with similar or lower levels in NCI-60 Database
Microarray
Differentially expressed genes in NCI-60 Database
Differential distribution of G4-forming TSSs in PC and CML cell lines (FANTOM database)
Genes with inconsistencies in mRNA and protein expression and showing higher levels of G4 forming TSSs
Genes in PC cell lines
Genes in CML cell lines

**Figure 5.2: Overview of the bioinformatics pipeline.** Bioinformatics pipeline to identify a list of genes in PC and CML cell lines that exhibit inconsistencies in mRNA and protein expression levels with higher levels of G quadruplex forming TSSs which encode alternatively transcribed G4 motifs in the 5' UTR isoforms. G4: G quadruplex; TSS: Transcription Start Site; UTR: Untranslated region; PC: Prostate cancer; CML: Chronic Myeloid Leukaemia.

#### 5.2.1.2 Mapping the TSSs to gene 5' UTR:

The processed TSSs were mapped within the +/- 50 bp region around the annotated transcription starting site. The gene TSS annotations were downloaded from Ensembl biomart based on GRCh38.p13 (release 98) (Yates *et al.,* 2019). The annotated TSS usually correspond to the major transcription starting position and a 50 bp region on either side was selected to identify all the alternative starting sites within this area. The mapping was performed using the LOOKUP function in Excel that matches the TSSs to a gene 5' UTR region based on genomic coordinates. The genes on plus (+) and minus (-) strands were mapped separately. The gene transcripts having a common starting position were grouped and overlapping transcripts from different genes were omitted from the analysis. The overlapping transcripts from different genes were removed to specifically associate a TSS to a single gene transcript/s.  For mapping purposes, the genomic coordinates that were considered included chromosome strand, chromosome number, the starting and ending positions of gene 5' UTR and TSS position (if it is within the defined range).

### 5.2.2 STEP 2: Detecting G4 forming sequences between alternative starting sites:

#### 5.2.2.1 Extracting sequence between major and upstream TSSs:

All the TSSs mapped within the defined region with overlapping CAGE tags were included for downstream analysis. The overlapping tags were considered because they would form part of the single TSS cluster. The TSS with the highest TPM in the cluster was considered as the major starting site. The major TSSs were identified for all the gene transcripts and in most of the cases, they were within +/-10 bp of the annotated transcript starting sites.

The upstream TSS relative to the major TSS in the cluster was then identified. For our bioinformatics analysis, the upstream TSSs was defined as the earliest CAGE tag starting site in the defined +/- 50 bp region around the annotated TSS that is more than 15 bp upstream to the major TSS. The 50 bp region around the annotated (major) TSS was selected to identify nearby TSSs that would encode for alternate 5' UTR isoforms containing extra G4 forming consensus sequence.  The 15 nucleotides distance between the major and upstream TSSs was selected because the minimum length of the G4 reported in the literature is 12 bp (Bakalar *et al.,* 2019). The sequences between the two alternative starting positions (major TSS and upstream minor TSS) were extracted using Bio. Entrez module in Biopython (Cock *et al.,* 2009). The following script was written to output sequences between major and upstream TSSs for all the genes included for analysis. The script was executed in the PyCharm community edition (version 2019.3.1). In the script below, the strand value of 1 or 2 could be used for plus and minus strand, respectively.

```
import Entrez, SeqIO
import pandas as pd

data = pd.read_csv("C:/Users/suran/Desktop/Bioinformatics Analysis Revised Jan 2020/TSS
Negative/Seq analysis/Combine final for analysis neg textfile.txt", header ="infer",
delimiter="\t")

def get_dna_sequence(startCoordinate, stopCoordinate, chromosome):
    Entrez.email = "A.N.Other@example.com"
    Start = (startCoordinate)
    Stop = (stopCoordinate)
    Chr = (chromosome)
    handle = Entrez.efetch(db="nucleotide",
                           id=Chr,
                           rettype="fasta",
                           strand=2,
                           seq_start=Start,
                           seq_stop=Stop)
    record = SeqIO.read(handle, "fasta")
    handle.close()
    return record.seq

def get_coordinates(data):
    coordinates = []
    textFile = (data)
    for line in textFile:
        coordinates.append(line)
    return coordinates

startCoordinates = get_coordinates(data["Start"])
stopCoordinates = get_coordinates(data["Stop"])
chromosome = get_coordinates(data["Chrom"])
transcript_id = get_coordinates (data["Transcript_id"])
strand= get_coordinates(data["Strand"])
sequenceTextFile = open("C:/Users/suran/Desktop/Bioinformatics Analysis Revised Jan
2020/TSS Negative/Seq analysis/TSSnegativeseq.txt", "w")
sequenceFastaFile= open("C:/Users/suran/Desktop/Bioinformatics Analysis Revised Jan
2020/TSS Negative/Seq analysis/TSSnegativeseqFASTA.txt", "w")

for i in range(len(startCoordinates)):
    sequenceText = ""
    sequence = get_dna_sequence(startCoordinates[i], stopCoordinates[i], chromosome[i])
    sequenceTextFile.write(str(startCoordinates[i]).rstrip('\n'))
    sequenceTextFile.write(" " + str(stopCoordinates[i]).rstrip('\n'))
    sequenceTextFile.write(" " + str(sequence).rstrip('\n'))
    sequenceTextFile.write(" " + str(chromosome[i]).rstrip('\n'))
    sequenceTextFile.write(" " + str(transcript_id[i]).rstrip('\n'))
    sequenceTextFile.write(" " + str(strand[i]).rstrip('\n'))
    sequenceTextFile.write("\n")
    sequenceFastaFile.write(">" + str(transcript_id[i]).rstrip('\n'))
    sequenceFastaFile.write("\t" + str(sequence).rstrip('\n'))
    sequenceFastaFile.write("\n")
print("Output Completed!")
```

### 5.2.2.2 Identifying G4 consensus in the extracted sequence:

The extracted sequences were analysed for the presence of a G quadruplex consensus using the pqsfinder package in R (Hon *et al.,* 2017). The pqsfinder is an intensive and imperfection-tolerant computational tool to detect potential G4 forming sequences. It is trained using the currently known and experimentally existing G4 structures and accommodates possible divergences (mismatches, bulges, loop length) from the ideal G4 consensus sequence. It also computes a score based on the G-tetrad stacking and the presence of mismatches and bulges. The score generated has shown to be closely related to the G4 stability (Hon *et al.,* 2017). The authors have reported that the package algorithm has superior accuracy compared to other existing tools. Moreover, this package could be adapted for batch analysis which was suitable for analysing all the extracted sequences generated in the above step. The following script was executed in R using the pqsfinder package:

```
library(pqsfinder)
library(qdapTools)

dna <- readDNAStringSet(file="sequences.fa")

pqs <- lapply(dna,pqsfinder, strand = "+")

pqsdf1 <- list_df2df(lapply(lapply(pqs,ranges),as.data.frame))
pqsdf2 <- list_df2df(lapply(lapply(pqs,score),as.data.frame))
pqsdf3 <- list_df2df(lapply(lapply(pqs,strand),as.data.frame))
pqsdf4 <- list_df2df(lapply(lapply(pqs,DNAStringSet),as.data.frame))
pqsdfcombine <- cbind (pqsdf1, pqsdf2$`X[[i]]`, pqsdf3$X[[i]], pqsdf4$x)

colnames(pqsdfcombine)[1] <- "Transcript_id"
colnames(pqsdfcombine)[5] <- "G-Score"
colnames(pqsdfcombine)[6] <- "Strand"
colnames(pqsdfcombine)[7] <- "G4 Seq"

colnames(pqsdfcombine)
write.table(pqsdfcombine,file="Gqaudtable.txt", sep = "\t", row.names = FALSE, col.names
= TRUE)

#generate output to text file
#set maxprint to avoid removing entries

options(max.print=1000000)
sink('pqsG4-output.txt')
print(pqs)
sink()
```

In the script above, the G4 was detected in the sense sequence only and the G4 score threshold to label a sequence as G4 forming sequence was set to the default value of 52, as it exhibited a balanced accuracy on the human G4 sequencing data (Hon *et al.,* 2017). The score assignment by the package has been shown to positively correlate with the propensity of G4 formation. The maximum score generated by the pqsfinder package is 395 and is produced by a sequence input only comprising of runs of Gs. The default values to label a putative G4 sequence was optimised by the authors using human G4 sequencing data generated by Chambers et al (2015). The genes with alternatively transcribed G4 were characterised according to their biotype using the attributes defined in the Ensembl biomart based on GRCh38.p13 (release 98) (Yates *et al.,* 2019).

#### 5.2.2.3 Pathway maps of genes with alternatively transcribed G4:

Functional pathway maps of genes that contain alternatively transcribed G4 were created using Metacore, version 21.1 (Clarivate Analytics) [https://portal.genego.com/]. The ranked hypergeometric test was used to determine the enriched pathways maps and GO (gene ontology) processes. The GO processes were presented as a histogram and pathway maps were presented as dot plot generated using ggplot2 package in R using the following script.

```
library(ggplot2)
data <- read.csv("Enrichment_analysis truncated.txt", sep ="\t", header = TRUE,
        stringsAsFactors = FALSE)
S1<- ggplot(data, aes(x=GeneRatio, y=reorder(Maps, as.numeric(-p.adjust)),
     color=p.adjust, size=Count)) + geom_point(alpha = 0.8) + theme_classic() +
     ylab(NULL) + guides(color = guide_colourbar(order=1),size = guide_legend(order=2))
S1 = S1+scale_color_gradient(low = "red2",  high = "mediumblue", guide = "colourbar",
     space = "Lab") + xlab("Gene Ratio") + theme(axis.title.x = element_text(face="bold",
     size = 10)) +theme(axis.text.y = element_text(size = 8, face = "bold"))
ggsave("Outpout.tiff", width = 8, height = 5, device='tiff', dpi=700)
```

### 5.2.3 STEP 3: Identifying genes for validation:

The objective of this step of the bioinformatics pipeline was to integrate the matched TSS and mRNA and protein expression data to determine the genes which showed higher mRNA expression, lower protein abundance and significantly increased level of G4 forming TSS isoforms, as also reported for *AGAP2* gene. The list would be used to validate the TSS-G4 mechanism in which the selection of G4 forming TSSs reduce the mRNA translational output, producing mismatched mRNA and protein expression profiles.

#### 5.2.3.1 Determining the differential distribution of G4 TSSs in PC and CML cell line:

To determine the differential distribution of G4 forming TSSs in PC and CML cell lines, the TSS count data for PC cell lines [DU145 (10490-107B4), PC3 (10439-106E7)] and CML cell lines [replicates for K562 (10454-106G4, 10824-111C5)] were downloaded from the FANTOM database. The TSS data, available in the format of (.bam) file, were downloaded from the link: https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/basic/human.cell_line.hCAGE/.

The (.bam) files were converted to (.bed) format to construct the TSS information according to the genomic coordinates (chromosome location - start and end position). The conversion was done following the guidelines provided by the authors (Lizio *et al.,* 2015). Briefly, the (.bam) files were processed by samtool package (Li *et al.,* 2009), and the high quality reads corresponding to an accuracy of 99% (Phred Quality Score 20) were retained. The bam files were then converted to bed coordinates, strand-selected, and sorted. The 5’ ends of the mapped CAGE tags were then aggregated by groupBy command in the bedtool package (Quinlan and Hall, 2010). The command-line codes are shown below and executed separately for the plus (+) and minus (-) strand.

```
# For Plus (+) Strand:
samtools view -uq 20 BAMfile \
| bamToBed -i stdin \
| awk 'BEGIN{OFS="\t"}{if($6=="+"){print $1,$2,$5}}' \
| sort -k1,1 -k2,2n \
| groupBy -i stdin -grp 1,2 -opCols 3 -ops count \
| awk 'BEGIN{OFS="\t"}{print $1,$2,$2+1,  $1":"$2".."$2+1",+"  ,$3,"+"}'

#For minus (-) Strand:
samtools view -uq 20 BAMfile \
| bamToBed -i stdin \
| awk 'BEGIN{OFS="\t"}{if($6=="-"){print $1,$3,$5}}' \
| sort -k1,1 -k2,2n \
| groupBy -i stdin -grp 1,2 -opCols 3 -ops count \
| awk 'BEGIN{OFS="\t"}{print $1,$2-1,$2,  $1":"$2-1".."$2",-"  ,$3,"-"}'
```

The command generates a (.bed) file containing the raw TSS count data in the standard chromosomal coordinates. For the analysis of G4 forming TSSs, the gene list generated in section 5.2.2.2 was used and the proportions of G4 forming TSSs were estimated by dividing the numbers of cumulative TSS tags within a 21 bp subregion upstream of the G4 starting position and the total tags in the selected TSS cluster (G4 + non G4 TSSs) (*Figure 5.3*). The 21 bp subregion was selected to maintain uniformity and also because the CAGE tags are about 21 bp long and any upstream overlapping tags within this region would belong to the same cluster (Takahashi *et al.,* 2012). The differential proportions were analysed by linear modelling and empirical Bayes approach using the Limma package in R (Ritchie *et al.,* 2015). The genes with significantly higher proportions of G4 forming TSSs were selected in PC and CML cell lines. The G4 forming TSSs were representative of the 5’ UTR population containing G4 consensus sequences.

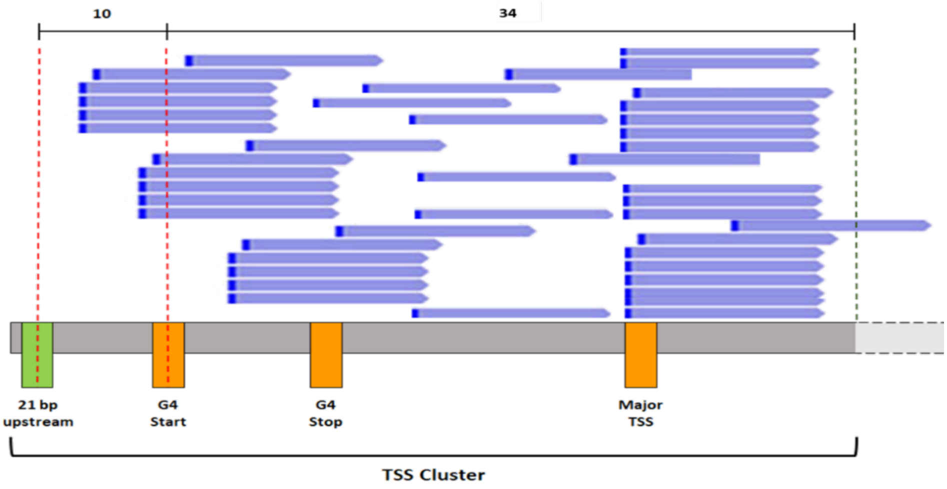


**Figure 5.3: Diagrammatic representation of estimating differential G4 TSS distribution.** Overlapping CAGE tag starting sites within the defined TSS cluster. The proportion of G4 forming TSSs, that would encode a G4 forming sequence in the corresponding 5' UTR, is calculated by counting CAGE tag starting sites within the 21 bp region upstream of the G4 start site and divided by the total number of starting sites within the TSS cluster. The proportions were computed for PC and CML cell lines and the significantly different proportions were determined using linear modelling and empirical Bayes approach. As an example, the number of G4 forming (10) and non G4 (34) TSSs in the cluster are shown at the top and were used to calculate proportions.

The script for processing (.bed) file, mapping to 21 bp region upstream of G4 start sites, counting the TSSs within the region, merging the TSSs from different cell lines, and performing Limma analysis to determine differential expression is illustrated below:

```
library(limma)
library(edgeR)

# Read table positive
Gene <-
read.table("chronic%20myelogenous%20leukemia%20cell%20line%3aK562%20ENCODE%2c%20biol_rep1
           .CNhs12334.10824-111C5.hg38.nobarcode.ctss.bed", header = FALSE, sep = "\t",
stringsAsFactors = FALSE)
colnames(Gene) <- c("chrom","V", "position", "chrom-loc", "TPM", "Strand")
Gene$V = NULL
plus <- read.table("Pos coordinates G4 -20 to major.txt", header = TRUE, sep = "\t",
        stringsAsFactors = FALSE)
pos <- Gene[Gene$Strand == "+",]
# True false argument positive

pos$found <- ifelse(sapply(seq_along(pos$position), function(i) {inds <-plus$Start <=
             pos$position[i] & plus$Stop >= pos$position[i] & pos$chrom[i]==
             plus$Chromosome any(inds)}), "YES", "NO")
table(pos$found)["YES"]
posfound <- pos[pos$found == "YES",]
write.table(posfound,"K562rep1posall.txt",sep ="\t",col.names=T, row.names = F)

# Load back gene table positive
posgene <- read.table("K562rep1posall.txt", header = TRUE, sep = "\t", stringsAsFactors =
           FALSE)
poslist <- aggregate(posgene$TPM, by=list(Category=posgene$Gene), FUN=sum)
colnames(poslist) <- c("Gene","TPMsum")
```

```
# Read table negative
neg <- read.table("Neg coordinates G4 +20 to major.txt", header = TRUE, sep = "\t",
       stringsAsFactors = FALSE)
KUneg<- Gene[Gene$Strand == "-",]

# True false argument negative
KUneg$found <- ifelse(sapply(seq_along(KUneg$position), function(i) {inds <-neg$Start >=
               KUneg$position[i] & neg$Stop <= KUneg$position[i] & KUneg$chrom[i]==
               neg$Chromosome any(inds)}), "YES", "NO")
table(KUneg$found)["YES"]
KUnegfound <- KUneg[KUneg$found == "YES",]
write.table(KUnegfound,"K562rep1negall.txt",sep = "\t",col.names=T, row.names = F)

#Load back gene table negative
KUneggene <- read.table("K562rep1negall.txt", header = TRUE, sep = "\t", stringsAsFactors
             = FALSE)
KUneglist <- aggregate(KUneggene$TPM, by=list(Category=KUneggene$Gene), FUN=sum)
colnames(KUneglist) <- c("Gene","TPMsum")
```

```
#Merge both table
KUgene <- rbind(poslist, KUneglist)
write.table(KUgene,"K562rep1all.txt",sep = "\t",quote = F,col.names=T, row.names = F)
datag4 <- read.table("K562rep1G4.txt", header = TRUE, sep = "\t")
dataall <- read.table("K562rep1all.txt", header = TRUE, sep = "\t")
Mergedgene <- merge(x=datag4, y=dataall, by="Gene", all= TRUE)
Mergedgene[is.na(Mergedgene)] <- 0
colnames(Mergedgene) <- c("Gene","G4", "ALL")
write.table(Mergedgene,"K562rep1GeneMerged.txt",sep = "\t",col.names=T, row.names = F)

#Merge Cell lines
Mergedallcelline <- merge(x=MergedPC, y=MergedCML, by="Gene", all= TRUE)
Mergedallcelline[is.na(Mergedallcelline)] <- 0
colnames(Mergedallcelline) <- c("Gene","DU145","PC3", "K562","K562.rep1")
write.table(Mergedallcelline,"DUPCK562andrep1Merged.txt",sep = "\t",quote =
F,col.names=T, row.names = F)
```

```
#Limma Analysis for differential distribution of G4 TSSs

library(limma)
library(edgeR)
counts <- read.delim("DUPCK562andrep1Merged.txt", sep = "\t", row.names = 1, header = T)
d1 <- DGEList(counts = counts[,1:4], group = c("PC", "PC", "CML", "CML"))
design1 <- model.matrix(~group, data = d1$samples)
fit1 <- lmFit(d1$counts, design1)
fit1 <- eBayes(fit1)
output1 <- topTable(fit1, sort.by = "p", n = Inf)
output1$Significant <- ifelse(sapply(seq_along(output1$adj.P.Val), function(i) {inds <-
output1$adj.P.Val[i] <= 0.05 any(inds)}), "Significant", "NS")
length(which(output1$adj.P.Val< 0.05))
write.table(output1,"Output DUPCK562rep1.txt",sep = "\t",col.names=T, row.names = T)
```

### 5.2.3.2 Characterising genes with a discrepancy in mRNA and protein levels in PC and CML cell lines:

To identify gene showing inconsistency in mRNA and protein level as noted for *AGAP2*, the NCI-60 microarray (GSE32474) and NCI-60 SWATH-MS databases were used (Guo *et al.,* 2019, Reinhold *et al.,* 2015). The required data was downloaded using the CellMinerCDB website (https://discover.nci.nih.gov/cellminercdb/) (Rajapakse *et al.,* 2018). The differentially expressed RNAs in PC (DU145, PC3) and CML (K562) cell lines were analysed by GEO2R (NCBI) and verified in R using the Limma package (Ritchie *et al.,* 2015). The significant differences in protein mass spectral intensity values were evaluated by linear modelling and empirical Bayes statistics using Limma. The genes with a differential RNA expression of 2-folds or greater (RNA logFC >=1) and no statistically significant differences in protein levels and/or significantly lower protein levels were considered to have a discrepancy in mRNA and protein levels. The significance was defined at an adjusted *P*-value threshold < 0.05. The discordant genes were determined separately for PC and CML cell lines.

The script to identify differentially expressed RNA and protein using R is exhibited below:

```
########### Differential RNA expression analysis with limma #######################

library(GEOquery)
library(limma)
library(umap)

# load series and platform data from GEO
gset <- getGEO("GSE32474", GSEMatrix =TRUE, AnnotGPL=TRUE)
if (length(gset) > 1) idx <- grep("GPL570", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

# log2 transformation
ex <- exprs(gset)
qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
LogC <- (qx[5] > 100) || (qx[6]-qx[1] > 50 && qx[2] > 0)
if (LogC) { ex[which(ex <= 0)] <- NaN
  exprs(gset) <- log2(ex) }

# assign samples to groups and set up design matrix
gs <- factor(sml)
groups <- make.names(c("PC","CML"))
levels(gs) <- groups
gset$group <- gs
design <- model.matrix(~group + 0, gset)
colnames(design) <- levels(gs)

# fit linear model
fit <- lmFit(gset, design)
cts <- paste(groups[1], groups[2], sep="-")
cont.matrix <- makeContrasts(contrasts=cts, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)

# compute statistics and table of top significant genes
fit2 <- eBayes(fit2, 0.01)
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=250)
tT <- subset(tT,
select=c("ID","adj.P.Val","P.Value","t","B","logFC","Gene.symbol","Gene.title"))
write.table(tT, file=stdout(), row.names=F, sep="\t")
```

```
########## Differential protein expression analysis with limma ###################

library(limma)
library(edgeR)

counts <- read.delim("Proteinlog2normalised.txt", sep = "\t", row.names = 1, header = T)
d1 <- DGEList(counts = counts[,c(1,2,3,4,7,8)],
      group = c("PC","PC","PC","PC","CML","CML"))
design1 <- model.matrix(~group, data = d1$samples)
fit1 <- lmFit(d1$counts, design1)
fit1 <- eBayes(fit1)
output1 <- topTable(fit1, sort.by = "p", n = Inf)
output1$Significant <- ifelse(sapply(seq_along(output1$adj.P.Val), function(i) {inds <-
                       output1$adj.P.Val[i] <= 0.05
any(inds)}), "Significant", "NS")
length(which(output1$adj.P.Val< 0.05))
write.table(output1,"DEprotein.txt",sep = "\t",col.names=T,row.names = T)
```

### 5.2.3.3 Integrating the data to find the genes for validation in PC and CML cell lines:

The genes with significantly higher levels of G4 forming TSSs, higher mRNA expression and lower protein abundance were found using InteractiVenn (Heberle *et al.,* 2015). It is a web-based tool to find and illustrate the relationship between different sets. Using this tool, the common genes between three analysed datasets (FANTOM, NCI-60 microarray, NCI-60 SWATH-MS) were identified and represented as a Venn diagram.

## 5.3 Results and Discussion

### 5.3.1 Alternatively transcribed G4 consensus sequences in the FANTOM database:

To identify genes with alternatively transcribed G4 forming sequence, we performed a bioinformatics analysis to find G4 consensus between alternative TSS isoforms within a defined region of the cluster (section 5.2.2). The analysis was performed using all the available human samples (~1,000) in the FANTOM database. The bioinformatics pipeline was designed to detect genes with putative G4 sequences between the major and upstream TSSs within the same TSS cluster. Our analysis identified 4,920 transcripts associated with 3,888 genes that contain G4 forming sequences between the two transcription start positions in the TSS cluster, upstream of major TSS.

The average length of the TSS cluster between the major and earliest TSSs in the cluster was 58.72 (+/-19.46) bp. The mean length of the G4 consensus was found to be 27.67 (+/-8.17) bp. The average distance between the earliest TSSs in the cluster and the G4 start position was 16.66 (+/-15.84) bp. In most cases, a single G4 was noted between the alternative TSSs and multiple G4 were only noted for 107 transcripts (2.1%). Interestingly, a large majority of the transcripts identified by our analysis were protein-coding genes (91.1%) followed by lncRNA (7.2%). The biotypes of the genes containing alternatively transcribed 5' UTR G4 forming sequences are presented in *Figure 5.4*. The selective enrichment of alternatively transcribed G4 in the protein-coding genes indicates a potential regulatory role of these structures, making the genes susceptible to regulation by the TSS-G4 mechanism.

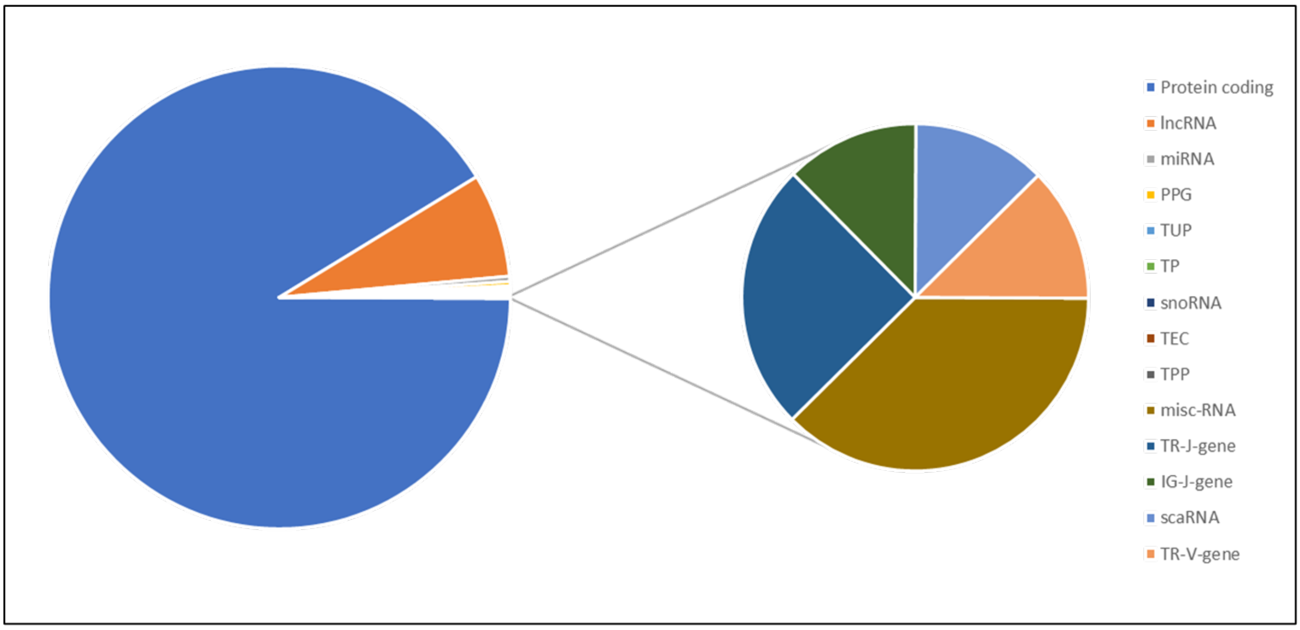


**Figure 5.4: Biotypes of the genes with alternatively transcribed G4. (A)** Biotypes of the genes that contain the G4 consensus sequence between the two transcription start positions. lncRNA: long noncoding RNA; miRNA: microRNA; PPG: Processed pseudogene; TUP: Transcribed unprocessed pseudogene; TP: Transcribed unitary pseudogene; TEC: To be experimentally confirmed; TPP: Transcribed processed pseudogene; TR-J: Joining chain T cell receptor; IG-J: Joining chain immunoglobulin; scaRNA: Small Cajal body-specific RNA; TR-V: Variable chain T cell receptor.

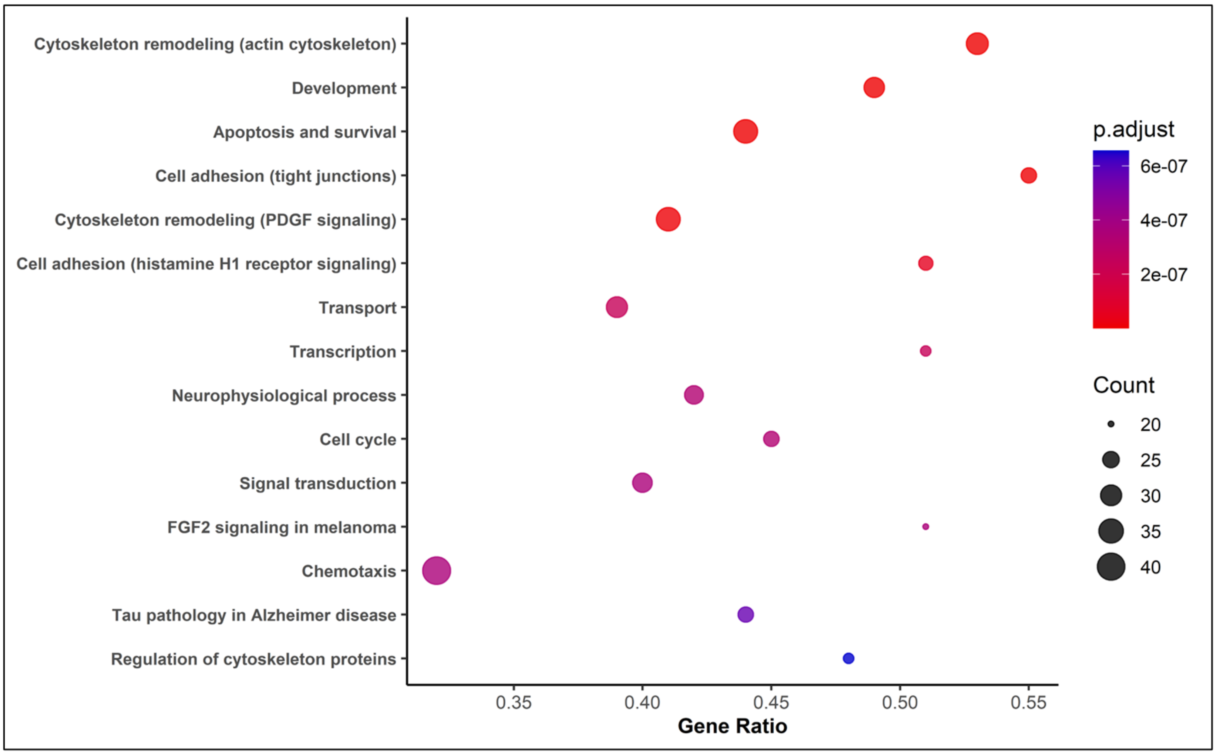
To identify important regulatory pathways that could be modulated by alternatively transcribed G4 motifs, we assigned the above genes to the pathway maps and GO processes using Metacore (section 5.2.2.3). The top three significantly enriched pathways included cytoskeleton remodelling, development, and apoptosis and survival (*Figure 5.5A*). To explore the functional processes involved, we performed a gene ontology analysis using Metacore as described above and found that these genes were frequently involved in regulating cellular and metabolic processes and the organisation of cellular components (*Figure 5.5B*).

Our study, for the first time, reported the incorporation of alternatively transcribed G4 forming sequences by the selection of nearby upstream TSSs within the cluster. Other studies on 5’ UTR have reported the G4 structures in the annotated 5’ UTR (Huppert *et al.,* 2008) and alternative UTR of protein-coding transcripts (splice variants) (Lee *et al.,* 2020). The number of protein-coding genes with alternatively transcribed G4 forming sequences within a core promoter region, as noted in our study, was 3491 genes. The results obtained by our study is in alignment with another study that analysed G4 forming sequences between splice variants and reported 2967 unique protein-coding genes that encoded for at least one transcript isoform containing a G4 sequence within the 5’ UTR Lee *et al.,* 2020). The bioinformatics analysis carried out by Huppert *et al.* (2008) on 5’ UTR G4s has reported a higher incidence of putative G4 sequences near the 5’ end of the UTR. Likewise, our study also showed the presence of these sequences within the first 30 nucleotides from the 5’ end. The enriched pathways and process identified by our studies were also noted by other studies on 5’ UTR G4s (Huppert *et al.,* 2008, Lee *et al.,* 2020). Some of the common enriched pathway and processes included signal transduction, transcription regulation, and regulation of cell metabolism. It highlights that the G4 formation in 5’ UTR regulate specific biological pathways and processes and could have functional implications in the cell. Since the TSS selection is dynamic, as defined in the earlier chapters, the production of alternative 5’ UTRs with or without these structures could act as an additional layer of regulation for the cells to respond to different stimuli and extrinsic and intrinsic factors.

In the current study, we restricted the alternatively G4 analysis to a smaller (max 100 bp) region between the upstream and the major TSSs in the cluster. Kawaji *et al.* (2006), using the FANTOM database, found the mean length of the TSS tag cluster to be more than 130 bp with the majority of reported tag cluster having less than 250 bp. Taking this into consideration, our restrictive analysis might overlook other potential G4 forming sequence between different alternative sites separated by more than 100 bp. However, in the current study, we aimed to analyse the translation impact induced by minor shifts in the TSS selection and designed our bioinformatics pipeline to specifically detect alternative G4 encoded by TSSs separated by only a few nucleotides. In-depth bioinformatic analysis to detect all the possible (overlapping/non-overlapping) G4 consensus sequences in the entire TSS cluster would characterise the G4 landscape of the cluster and generate novel insights.

In the current study, the G4 forming sequence was determined between the major and upstream TSSs in the cluster. The major TSS in the cluster was identified by the tag with the highest TPM. However, this method of determining a major TSS would not be valid in the case of bi or multimodal TSS peak distribution profiles, where multiple major TSSs exists (Carninci *et al.,* 2006). For a small number of cases in our studies (bi or multi-modal peak), the G4 consensus might not be upstream to the major TSS and may be between two major TSS or downstream to major starting sites. An advanced bioinformatics analysis would be required to tailor the analysis according to the TSS distribution context. Nonetheless, our bioinformatics analysis has robustly identified a G4 forming sequence between two alternative starting sites within the same TSS cluster.

(A)



(B)

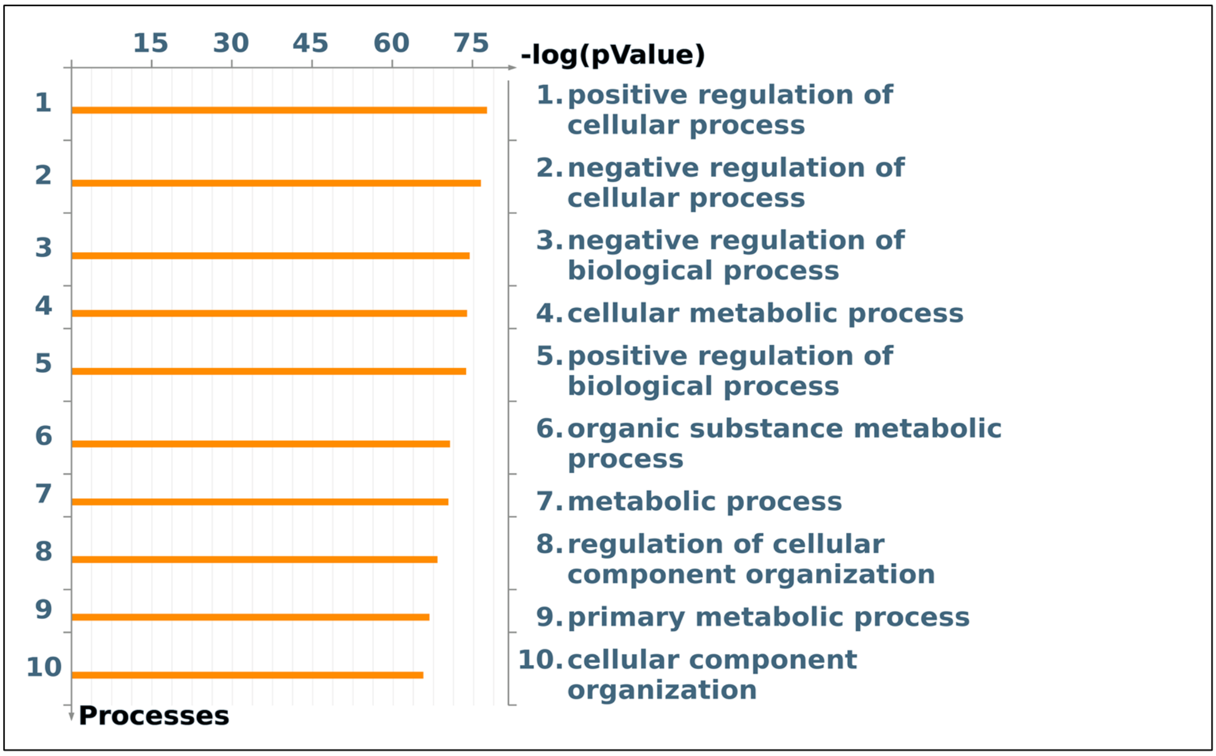


**Figure 5.5: Pathway maps and Go processes of the genes with alternatively transcribed G4. (A)** Metacore pathway enrichment analysis of transcripts containing alternatively transcribed 5' UTR G4. The dot plot shows the top 15 enriched pathways with the largest gene ratio, the size of the dots represents the number of genes in each pathway and the colour of the dots represents the adjusted p values (BH). **(B)** GO-process enrichment was performed using Metacore and enrichment was determined by meeting a BH adjusted *P*-value cutoff of 0.05. The histogram showing the top 10 GO processes enriched, the results are ranked by -log(*P*-values).

### 5.3.2 Genes with discordant mRNA and protein expression and significantly higher levels of G4 forming TSSs:

We analysed the proportions of G4 forming TSSs in PC and CML cell lines to determine differential expression of G4-containing 5' UTR isoforms (section 5.2.3). Using linear modelling and the empirical Bayes approach, we identified 1,007 genes out of 3,888 that showed differential cell line-specific expression of G4 TSSs. Next, we determined differential mRNA and protein expression in the matched PC and CML cell lines using the NCI-60 microarray data and NCI-60 SWATH-MS data, respectively. We found that 6,585 genes displayed differential relative RNA expression profiles in PC and CML cell lines and also noted that 605 proteins were significantly different in these cell line groups.

We were interested in identifying genes that had discrepancies in mRNA and protein expression with higher mRNA and lower protein expression, as observed for *AGAP2* in the CML cell line. To find mismatched genes, we examine their mRNA and protein levels and selected those genes that had 2-folds or greater relative RNA expression (RNA logFC >=1) and no statistically significant differences in protein levels and/or significantly lower proteins levels. We then merged the discordant expression data with differential G4 TSS data to identify genes having discordant mRNA and protein expression with higher levels of G4 forming TSSs. Our analysis identified 2 genes (*HK1* and *TMEM263*) in the CML cell lines and 35 genes in the PC cell lines (*ALDH7A1*, *ARHGAP29*, *BAG5*, *CASK*, *CAV1*, *CD9*, *CDKN2A*, *CKAP4*, *CNN3*, *COL4A1*, DUSP3, *ERMP1*, *FAM114A1*, *FARP1*, *FERMT2*, *IFT27*, *ITGA3*, *ITGB4*, *ITGB5*, *LMO7*, *LRRC1*, *LRRFIP1*, *MEST*, *MLKL*, *NDRG1*, *NT5C2*, *PGM2L1*, *PODXL*, *PXDN*, *RRBP1*, *SDC1*, *SYNCRIP*, *TBL1XR1*, *TPD52*, and *WASL*) (*Figure 5.6*). Surprisingly, the bioinformatics pipeline did not identify *AGAP2* as a gene displaying discrepancy in mRNA and protein levels in the CML cell line (K562). It could be explained by the use of K562 cell lines in the CML group which in contrast to other CML cell lines included in the project do not exhibit relatively higher levels of *AGAP2* mRNA. No statistically significant differences were noted in *AGAP2* mRNA expression levels between the PC (DU145, PC3) and CML (K562) cell lines in the NCI-60 microarray database.

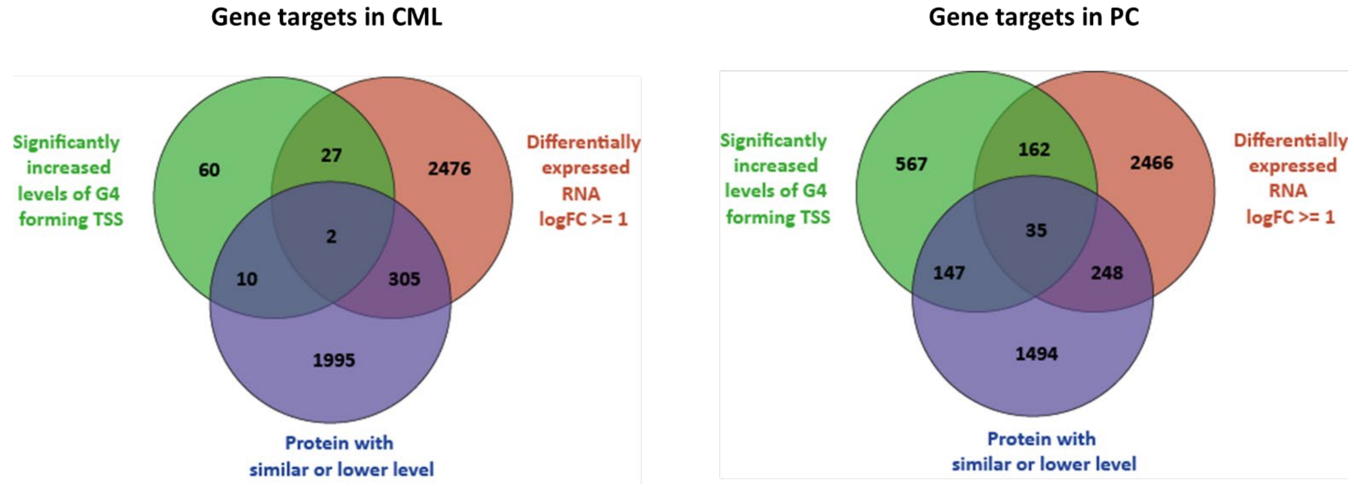

Gene targets in CML
Significantly increased levels of G4 forming TSS
Differentially expressed RNA logFC >= 1
Protein with similar or lower level
60
27
2476
2
10
305
1995
Gene targets in PC
Significantly increased levels of G4 forming TSS
Differentially expressed RNA logFC >= 1
Protein with similar or lower level
567
162
2466
35
147
248
1494

**Figure 5.6: Genes displaying inconsistency in mRNA and protein level and higher levels of G4 forming TSS.** Venn diagrams illustrating the overlapping genes in FANTOM and NCI-60 database that show differential levels of G4 TSS, differentially expressed mRNA (>=1 log FC), and no statistically significant differences in protein levels and/or significantly lower proteins levels in CML cell lines (left) and PC cell lines (right). The differential expressions were computed by linear modelling followed by empirical Bayes statistics using the Limma package.

Our analysis identified genes that demonstrated an inconsistency in mRNA and protein levels that could be mediated by alternatively transcribed G4. These structures, as shown by our study, could modulate the translational efficiency of their respective mRNA transcript, leading to inconsistent RNA and protein profiles. However, the criteria used in our study to define the discordant mRNA and protein would detect only those genes with larger significant differences in translation efficiency. The alternative 5' UTR G4 might not always mediate such larger differences in the mRNA translation output. As a result, our approach could possibly miss many other relevant genes that do not display such larger relative differences in the protein output. Moreover, our defined criteria only select genes that display a 2-fold or greater relative increase in mRNA, consequently, it would ignore the genes with similar mRNA level but differences in protein abundance. Furthermore, the protein expression data for the PC and CML cell lines used for the analysis was inadequate. The NCI-60 SWATH-MS dataset had protein expression levels for only about 3,100 genes that were common to all the cell lines in the NCI-60 database. The lack of protein data for a large number of genes in PC and CML cell lines resulted in a lower than expected numbers of genes detected for validation.

The NCI-60 mRNA and protein expression database had the data for only one of the CML cell lines (K562). The K562 is one of the commonly used CML cell lines but was not part of the CML cell line group in the current study. Instead, the CML group included other related cell lines such as KU812, TCCS, and KCL-22. Although these cell lines would be grossly similar, the RNA and protein expression and the TSS distribution profiles might not be representative of the CML cell lines included in our study. To our knowledge, there were no other datasets, except NCI-60, that analysed the RNA and protein levels in PC and CML cell lines using the same platform. Despite some evident limitations, our analysis was able to detect relevant genes which could be used to demonstrate the relevance of the TSS-G4 mechanism in controlling the expression of gene output.

### 5.3.3 RNA sequencing of the cell lines included in our study:

To make the bioinformatics analysis relevant to the cell lines included in our study, we performed RNA sequencing of the PC (DU145, PC3, LNCaP) and CML (KU812, TCCS, KCL-22) cell lines used in the current project (section 2.2.9). A single sequencing reaction was performed for each cell line and the expression profiles of the PC and CML cell lines were grouped together and considered as biological replicates. The principal component analysis (PCA) was performed to cluster the samples based on the patterns of gene expression and evaluate the level of similarity/dissimilarity between the two groups. The PCA analysis showed clear differences between the two cell line groups. The first two principal components (PC1 and PC2) explained more than 70% of the variability among the samples with distinct clustering of the PC and CML groups (*Figure 5.7A*).

The differential expression analysis of the RNA-seq data identified 2,350 genes that had significantly different relative levels in PC and CML cell lines. The heatmap of the top 1,500 differentially expressed genes is presented in Figure *5.7B*. As depicted in the heatmap, the PC and CML cell lines samples have clustered into distinct cell line groups using hierarchical clustering analysis and validate discrete clustering of the biological replicates used for each group. We then examined the differential RNA expression of the genes selected for validation (section 5.3.2) and noted that 8 out of 37 genes selected for validation in PC and CML cell lines showed similar differential expression patterns as reported in NCI-60 microarray data. The relative profiles of the genes showing higher relative expression in PC and CML cell lines are shown using the volcano plot (*Figure 5.7C*).

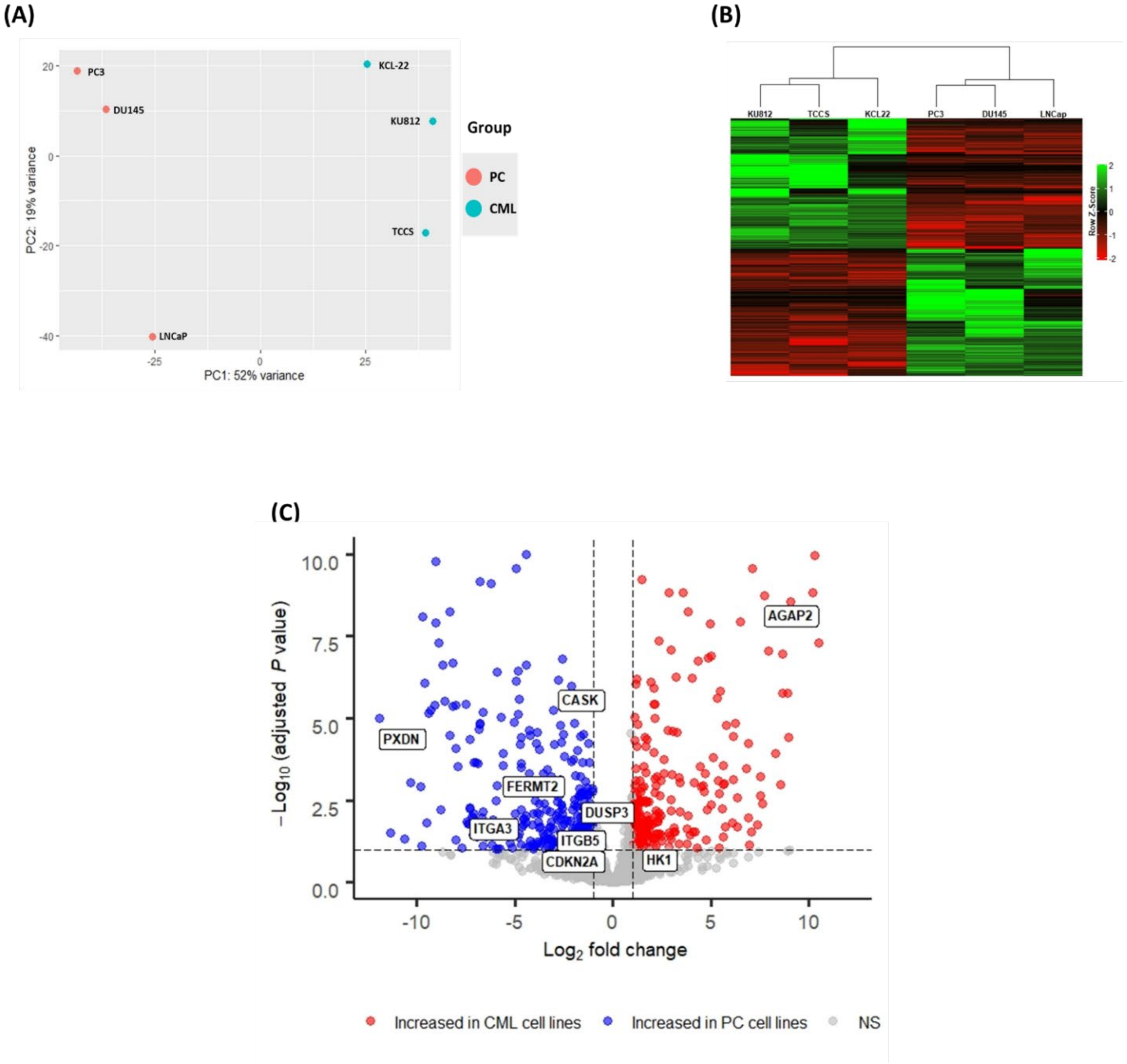


**Figure 5.7: Expression profile of differentially expressed genes. (A)** PCA of RNA sequencing data. The analysis was performed using the top 500 differentially expressed genes in PC and CML cell lines. The percentages on each axis represent the % variation explained by the respective principal component. **(B)** Heatmap showing the z-scores based on FPKM values of the top 1,500 differentially expressed genes. Each row in the heatmap represents a gene and the column represents a sample. The green colour indicates highly expressed genes and red shows genes with low expression. The dendrogram representing hierarchical clustering based on Euclidean distance is presented above the heatmap. **(C)** Volcano plot highlighting differentially expressed (DE) genes in PC and CML cell lines with log2 fold change >1 and log10 adjusted *P*-value < 0.01. The expression profiles of the genes that showed similar patterns of expression in the NCI-60 database and our study are shown in the box. The differential expression of *AGAP2* is also shown. The red and blue indicates genes with higher relative expression in CML and PC cell lines used in our study, respectively.

The expression pattern of most of the genes that were selected for validation using the NCI-60 microarray database did not match the expression profiles of the PC and CML cell lines included in our study. Out of 37 genes selected for validation, only 8 showed similar expression profiles. It is attributed to differences in the cell lines used for analysis that led to contrasting expression pattern. The LNCaP from the PC group and all the CML cell lines used in the study were not present in the NCI-60 database. Moreover, the available CML cell line (K562) in the NCI-60 database was not part of the current study. Although being part of the same cell line group, these cell lines are established from different patients, sites, and disease levels and could demonstrate variable expression for some genes. (Jiang *et al.,* 2016)

Moreover, grouping RNA sequencing data of different related cell line into either PC or CML group could also create biological noise and variation, as noted for LNCaP and TCCS in PCA analysis (*Figure 5.7A*). The use of technical replicates, instead of biological replicates, for one of the PC or CML cell lines, preferably a cell line whose TSS data is available in the FANTOM database (CML: KU812, KCL-22; PC: DU145, PC3) and performing both RNA sequencing and mass spectrometry analysis on these cell lines could precisely identify relevant genes for validation that could be used to examine the relevance of TSS-G4 mechanism.

### 5.3.4 Validation of the TSS-G4 mechanism using another gene:

Thus far, our study has established the contribution of the TSS-G4 mechanism in regulating AGAP2 expression. To demonstrate the universality of this mechanism, we performed a bioinformatics analysis of NCI-60 and FANTOM databse to generate a list of genes that exhibited a discrepancy in RNA and protein expression and had a significantly higher level of G4-containing TSSs. We selected some of these genes and analysed their mRNA and protein expression profiles in the PC and CML cell lines included in our study to identify genes that displayed a similar level of mismatches as noted for *AGAP2* (high mRNA and lower protein).

We analysed the RNA and protein expression profiles of *HK1* genes from the CML cell line group as it demonstrated higher proportions of G4 forming TSSs (*Figure 5.8*). We also selected other genes (*PON2 and CAV-1*) from the PC cell line group that were identified using an earlier bioinformatics analysis without linear modelling and empirical Bayes (Appendix 4). However, we didn't note a discrepancy in mRNA and protein levels in these cell lines (*Figure 8.4*) possibly due to no statistically significant differences in TSS distribution, which was not accounted for in the earlier bioinformatics analysis.

The analysis of the *HK1* RNA and protein levels revealed a similar pattern of inconsistency as observed for the *AGAP2* gene. *Figure 5.8A* and *5.8B* show the relative RNA and protein expression for HK1 in PC and CML cell lines included in our study and highlight significantly higher relative level of the mRNA in two CML cell lines (KU812 and TCCS) with relatively lower protein expression in these cell lines (*Figure 5.8B*). The KCL-22, unlike other CML cell lines, did not show inconsistency in mRNA and protein expression. Hence, we decided to evaluate the contribution of the TSS-G4 mediated mechanism in regulating *HK1* expression in KU812 and TCCS cell lines.

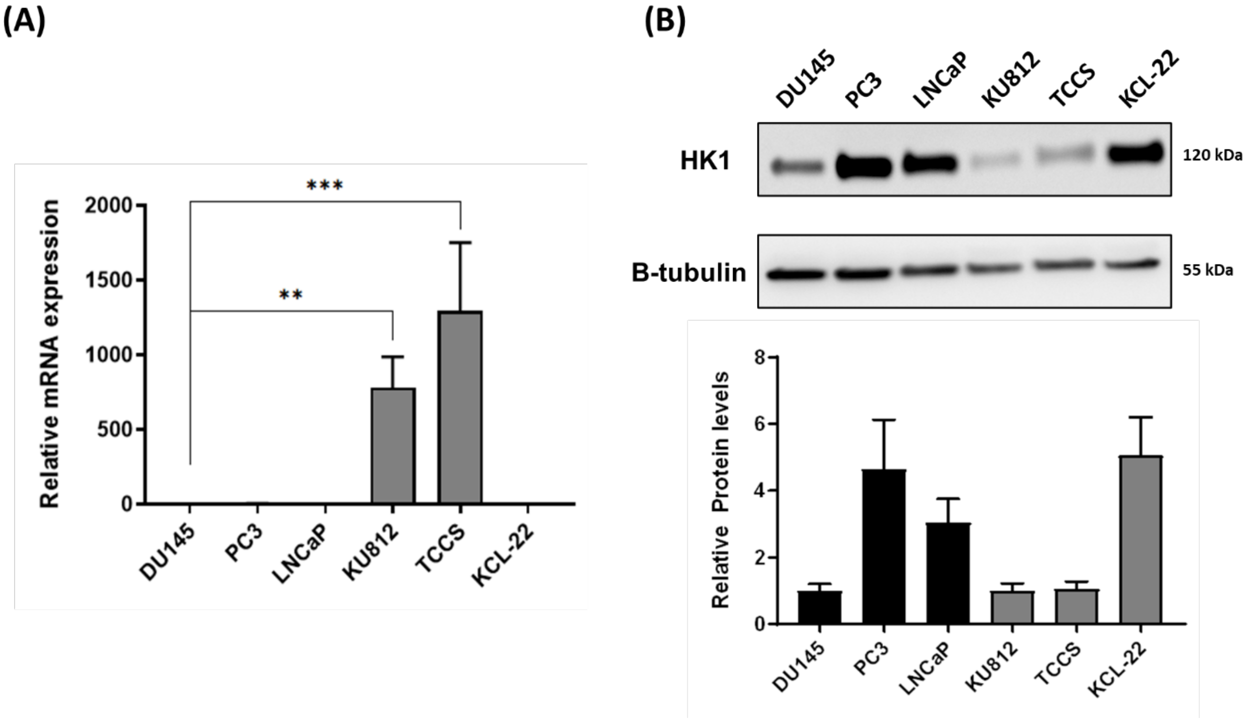


**Figure 5.8: RNA and protein expression profiles of HK1 and CDKN2A. (A)** The relative mRNA and protein expression profiles of HK1 in PC (DU145, PC3, LNCaP) and CML cell lines (KU812, TCCS, KCL-22). The mRNA measured by qRT-PCR, normalised using the *HPRT* housekeeping gene and shown relative to DU145. The difference in RNA expression was analysed using one-way ANOVA [F (5, 12) = 22.25, *P* < 0.001)] with post-hoc Sidak's multiple comparison tests, *P*-values shown. **(B)** *HK1* protein levels were detected by resolving 50 µg of protein using 10% SDS-PAGE followed by immunoblotting with *HK1* isoform-specific antibody. The representative blot is shown and normalised for protein loading using β-tubulin. Densitometry values for the relative protein expression are presented below the blots and displayed relative to DU145. Full immunoblot for HK1 is presented in Appendix 5.

Hexokinase 1 (*HK1*) gene is located on chromosome 10 and is one of the key genes in glucose metabolism and implicated in neurodevelopmental abnormalities (Okur *et al.,* 2019). The distribution of TSSs for the *HK1* transcript (NM_033496.3) in the FANTOM database is shown in *Figure 5.9*. As depicted in the figure, the alternative TSS selection encodes 45 extra nucleotides in the longer 5' UTR of the *HK1* gene which contained the consensus sequence for a G4. As observed for *AGAP2,* the upstream G4 forming starting site is a minor TSS and the majority of transcription in the FANTOM database is initiated from the downstream position, mainly at -130 bp from the ATG start codon. Our bioinformatics analysis (section 5.3.1) revealed that the G4 forming upstream TSS was significantly higher in CML relative to the PC group.

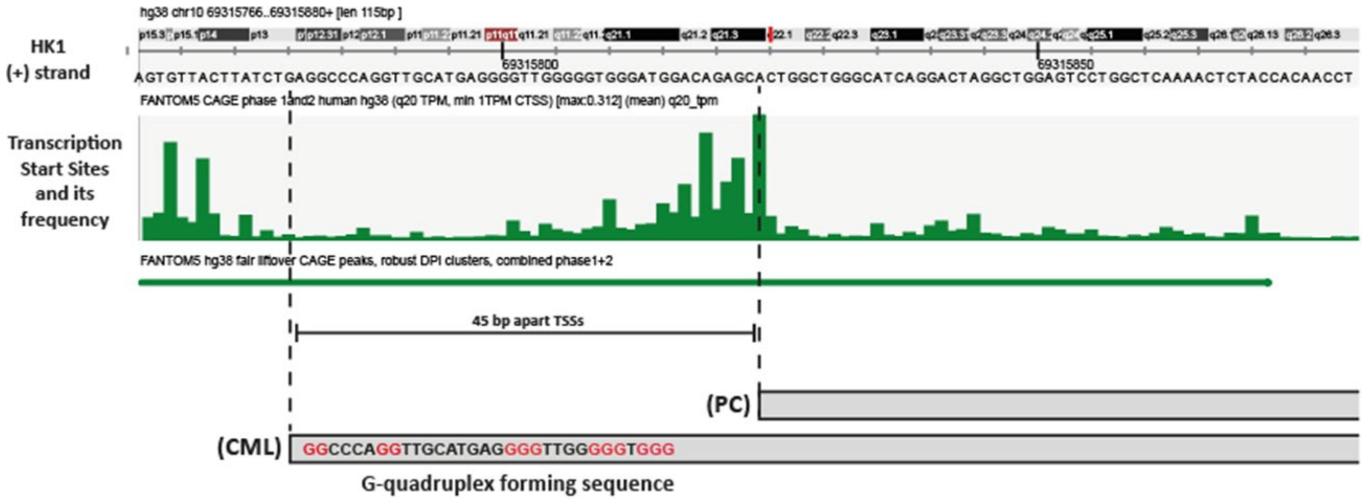


**Figure 5.9: TSS distribution profile of HK1 gene transcript (NM_033496.3).** Image derived from ZENBU genome browser showing the distribution and usage of different *HK1* TSSs. The alternative TSS in the CML cell line that encodes extra nucleotides in the longer 5' UTR and contained the G quadruplex forming sequence is shown below.

The formation of the G4 motif by the extra nucleotide in the longer 5' UTR of *HK1* was verified using the GRIP method in the TCCS cell line (Figure 5.10A, also section 4.2). We have noted statistically significant enrichment of *HK1* in the BG4 fraction which signifies the formation of G4 structures within the *HK1* mRNA ($P < 0.001$). In the current GRIP experiment, the pulldown of longer G4-containing 5‘ UTR of *HK1* mRNA was analysed using samples generated previously to examine *AGAP2* G4 pulldown (section 4.3.1). It will be relevant to use a PC cell line in the current experiment as a control to analyse differences in the relative pulldown between CML and PC cell lines. Unlike AGAP2, the PC cell lines have detectable levels of longer G4-containing 5' UTR of *HK1* (*Figure 5.10B*) and could serve as a control in this experiment. An experiment with a PC GRIP control would be required in future to further validate the G4 formation by the longer 5' UTR of *HK1*.

We have also analysed the levels of longer G4-containing 5' UTR of *HK1* mRNA using a qRT-PCR and exhibited a significant increase in the relative levels of G4 forming 5' UTR population in the KU812 and TCCS cell lines, consistent with relatively higher usage of G4 forming TSSs in these cell lines (*Figure 5.10B*). As shown in *Figure 5.10B*, about 2-3 folds relative increase was noted in KU812 and TCCS compared to other cell lines. No significant increase was noted for the third CML cell line (KCL-22) which also did not display a mismatched RNA and protein expression profile. It indicates that a lower usage of upstream G4 TSSs in these cell lines could explain the higher protein abundance despite lower relative RNA levels.

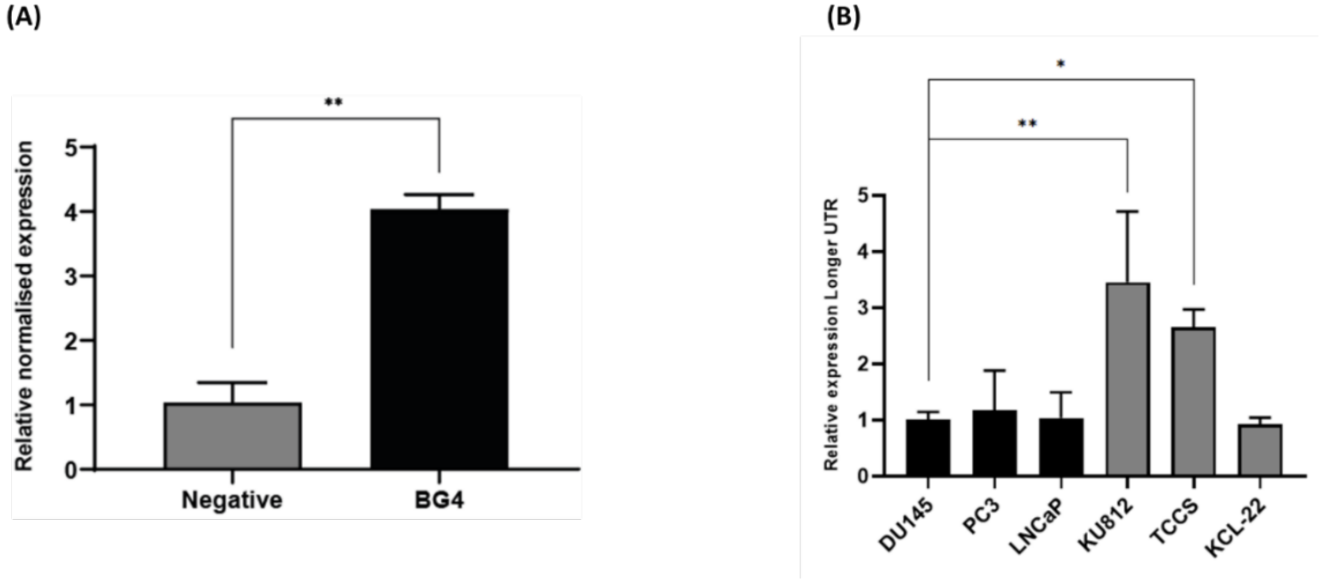


**Figure 5.10: GRIP to detect G4 in the *HK1* gene and relative levels of the *HK1* longer 5’ UTR isoforms in PC and CML cell lines. (A)** GRIP for *HK1* normalised by input control, the data correspond to three independent immunoprecipitations (n=3) in TCCS cell line, and the error bars denote standard deviation. Differences between samples were analysed by unpaired two-tailed t-tests [t(2) = 10.99, *P* = 0.0082], *P*-values shown. **(B)** Relative levels of *HK1* mRNA with longer 5’ UTR containing the G quadruplex forming sequence in PC and CML cell lines determined using qRT-PCR. The data is the mean ± SD of three independent experiments (n=3). The longer UTR levels were normalised by the levels by the entire 5’ UTR population and presented relative to DU145. Statistical differences were analysed by one-way ANOVA [F (5, 12) = 8.6, *P* < 0.001)] with post-hoc Sidak's multiple comparison tests, *P*-values shown. (**P* < 0.05; ***P* < 0.01).

We also noted that the longer 5’ UTR isoforms of *HK1* that contained the G4 forming sequence poorly associate with polysomes and were enriched in the non-polysomal fraction (*Figure 5.11*). As noted in *Figure 5.11A*, the longer 5’ UTR isoforms were mainly amplified in the RNP fraction, while the shorter 5’ UTRs without the G4 were enriched in the polysomal fraction, signifying efficient translation of the shorter 5’ UTR. Pooling the data from replicates of KU812 and TCCS cell lines into non-polysomal (fraction 1-5) and polysomal (fraction 6-10) showed a significant increase in the longer 5’ UTR isoforms in the non-polysomal fraction (*P* < 0.0001) (*Figure 5.11B*).

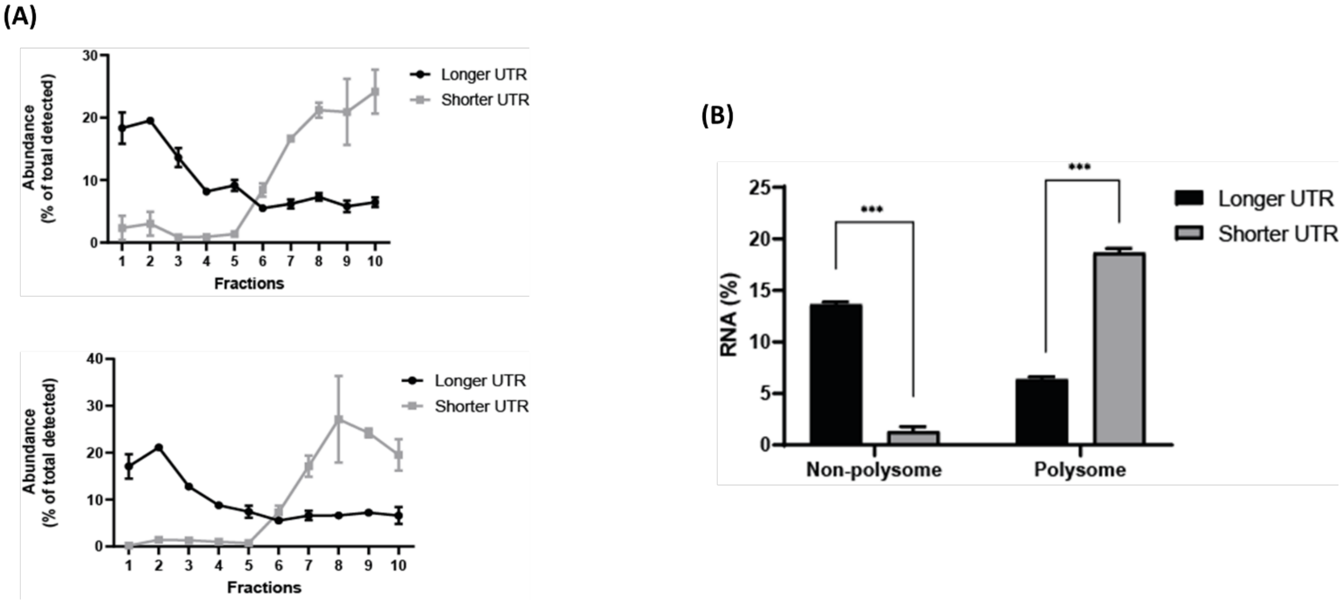


**Figure 5.11: Polysome association of the longer and shorter 5' UTR isoforms of *HK1* mRNA. (A)** The relative distribution of *HK1* mRNA with shorter and longer 5' UTR are shown in polysome fractions 1-10 of KU812 (top) and TCCS (bottom) cell line. The RNA distribution is presented as the fraction of the total RNA recovered. The mRNA levels were normalised to the exogenous spike-in luciferase control mRNA. The graphs above represent the means ± SEM of 2 independent experiments (n=2). **(B)** Relative levels of *HK1* mRNA with longer and shorter 5' UTR in non-polysomal (Fraction 1 – 5) and polysomal (Fraction 6 –10) segments pooled from both the cell lines. The data represent the means ± SD of the fraction of the total RNA recovered and *P*-values were calculated by an unpaired students t-test, ****P* < 0.001.

Our validation experiments with *HK1* demonstrated that the TSS-G4 mediated gene regulation mechanism contribute to the inconsistency observed for *HK1* in PC and CML cell lines. It also highlights the widespread existence of this mechanism and is not just limited to the *AGAP2* but is also relevant in controlling the expression of other genes. In the current study, PC and CML cell lines were used to study this mechanism using the *AGAP2* gene as a model. However, this mechanism could be also implicated in other cell lines as noted by the discrepancy in AGAP2 mRNA and protein levels in other cell lines (*Figure 3.5B*). Further studies are required to understand the relevance of this mechanism in other cell lines and primary cells.

The selection of earlier TSSs that encode regulatory elements and results in a translationally inefficient mRNA has been previously reported for many different genes including *BRCA1*, *SHOX*, and *RUNX1* (Blaschke *et al.,* 2003, Pozner *et al.,* 2000, Sobczak and Krzyzosiak, 2002). However, the alternative 5' UTRs in these studies are separated by a large genomic space and originate from multiple promoters. On the contrary, in our study, the alternative 5' UTRs are derived from TSSs that are less than 50 bp apart that incorporate a G4 forming sequence by the selection of TSSs separated by only a few nucleotides. In addition to an alternatively transcribed G4 sequence by the selection of variant TSSs in the same cluster, other shorter regulatory elements such as uORFs, uAUGs, and hairpins could be also alternatively transcribed in a similar fashion and modulate mRNA translation potential.

Previous studies have shown a significant translation impact by uORF and uAUG (Reviewed in Chatterjee and Pal, 2009, Leppek *et al.,* 2018), and it will be interesting to analyse the impact of these elements when they are incorporated as a result of minor changes in the TSS selection. Recently, Palavecino *et al.* (2020) had reported the impact of alternatively transcribed uAUG by the selection of ~100 bp upstream TSSs. These alternatively transcribed motifs by selecting nearby TSSs in the core promoter could encompass a novel untapped layer of gene regulation. Further studies analysing these extra nucleotides sequences incorporated by upstream TSSs would enable the identification of different alternatively transcribed regulatory features that could modulate mRNA translation potential based on transcript starting position in the cluster.

The relevance of the TSS-G4 mediated mechanism in regulating genes other than *AGAP2* adds a novel layer of gene expression control, which in the case of *AGAP2* and *HK1* also contributed to discrepant mRNA and protein expression profile. Other regulatory features in the 5' UTR, 3' UTR, ncRNA landscape and cell-specific factors need further evaluation to understand their contribution in mediating mRNA and protein discrepancies. In addition to identifying factors that impair mRNA translational efficiency, further studies are also required to examine the factors responsible for promoting mRNA translation. In our study, we have noted increased translation of *AGAP2* mRNA in the PC cell line in spite of the significantly lower relative mRNA level compared to the CML cell line. Likewise, we also noted a similar pattern for *HK1* in KCL-22 which also showed higher protein abundance, albeit significantly lower mRNA expression. Understanding these factors would create new insights into the dynamic modulation of mRNA translation and would facilitate the discovery of novel elements that increase translation efficiency.

## 5.4 Summary of findings (Chapter 5)

In summary, this chapter presented the development of the bioinformatics pipeline to identify alternatively transcribed G4 forming sequence in the human genome using the FANTOM database. Our analysis identified 4,920 transcripts associated with 3,888 genes that contained alternatively transcribed G4 consensus by the selection of upstream TSSs within the defined cluster. These G4-containing genes were enriched in key pathway maps including cytoskeleton remodelling, apoptosis, and cell adhesion. This chapter also described the bioinformatics analysis to identify genes with differential distribution of G4 forming TSSs. Additionally, we performed analysis on the NCI-60 microarray and SWATH-MS database and curated a list of genes that displayed discrepancies in RNA and protein expression. We integrated the data with the G4 TSS data to identify genes that had higher mRNA expression, lower protein abundance, and a significantly higher level of G4 forming TSS. We then examined the gene list to find a suitable target for validation and identified the *HK1* gene for further evaluation as it showed higher relative proportions of G4 TSSs and demonstrated a similar pattern of RNA/protein inconsistency as noted for *AGAP2*. Two CML cell line (KU812, TCCS) with discordant RNA/protein profile exhibited higher relative levels of G4-containing longer 5' UTR isoforms with the formation of G4 structure verified using our GRIP method. We also showed that these alternatively transcribed G4 in *HK1*, like *AGAP2*, decrease the translation efficiency as shown by decreased polysome association. This indicated that the TSS-G4 mediated mechanism we discovered for *AGAP2* regulation is also implicated in regulating the expression of other genes.

# Chapter 6:

# General Discussion, Future Work, and Conclusion

## 6.1 General Discussion

Our study has discovered a novel TSS-G4 mediated mechanism that modulates mRNA translation by incorporating a G4 forming sequence through alternative TSS selection, decreasing mRNA translation efficiency (*Figure 6.1*). These alternative TSSs are selected within a single core promoter region and are separated by only a few nucleotides (< 50 bp). The drastic change in the mRNA translational potential mediated by alternatively transcribed G4 structures has not been reported before in the literature. Our study exhibited the relevance of this mechanism in the *AGAP2* gene and using a variety of approaches have demonstrated the formation and functional consequences of these structures on translation. Our bioinformatics analysis has identified many potential genes that encode a G4 consensus sequence in the alternative 5′ UTR isoforms and are susceptible to regulation by TSS-G4 mediated mechanism. We also validated this mechanism using the *HK1* gene, indicating that it is not just implicated in controlling *AGAP2* expression but is more universal and involved in regulating the expression of other genes.

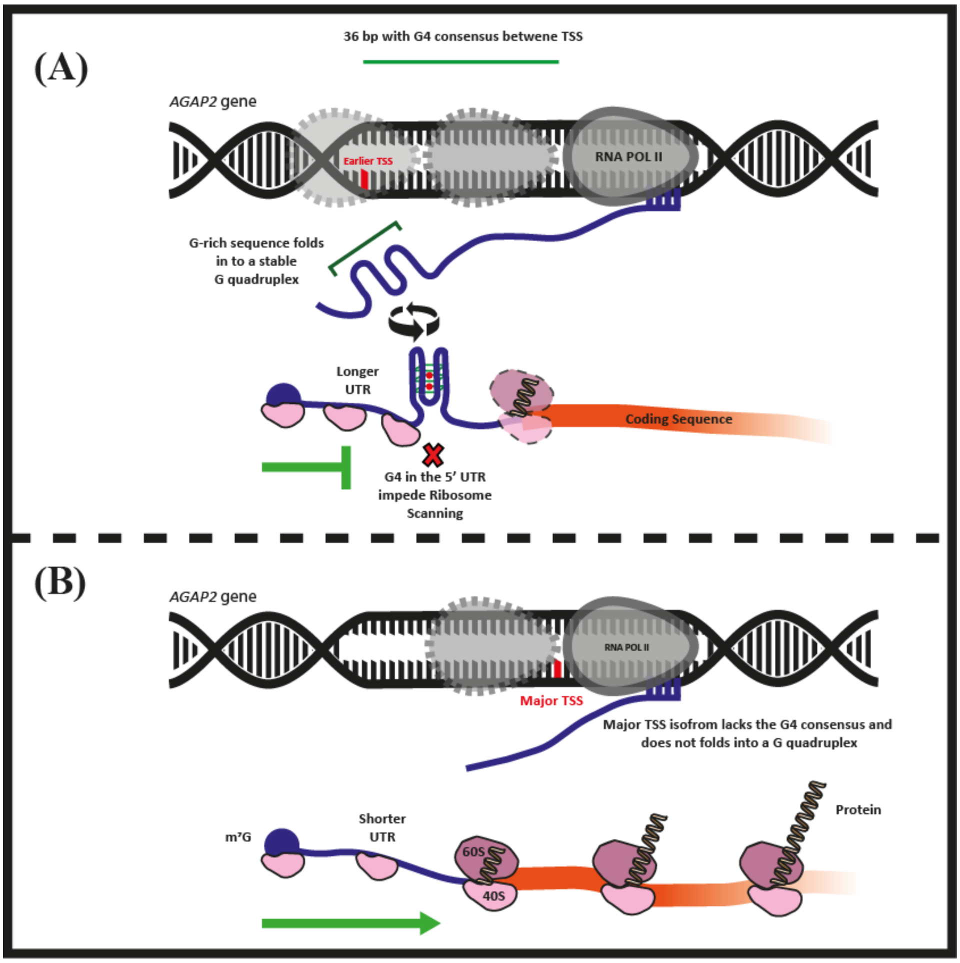

36 bp with G4 consensus betwene TSS
(A)
AGAP2 gene
Earlier TSS
RNA POL II
G-rich sequence folds in to a stable G quadruplex
Longer UTR
Coding Sequence
G4 in the 5′ UTR impede Ribosome Scanning
(B)
AGAP2 gene
RNA POL II
Major TSS
Major TSS isofrom lacks the G4 consensus and does not folds into a G quadruplex
m⁷G
Shorter UTR
60S
40S
Protein

**Figure 6.1: TSS-G4 mediated mechanism.** The selection of alternative TSSs in the cluster incorporates G4 regulatory elements in the 5′ mRNA UTR. **(A)** The selection of earlier TSS results in a longer 5′ UTR that contain a G4 forming sequence. The formation of G4 structure by the extra nucleotides decreases the translational efficiency, possibly by impeding ribosome scanning and decreasing the association with polysomes and result in reduced translational output. **(B)** The selection of downstream TSS yields a shorter 5′ UTR without the G4 forming sequence and prominently associate with polysomes, showing increased translation efficiency.

Alternative transcription initiation (ATI) contributes to the transcriptomic diversity of eukaryotic organisms. It produces a multitude of transcript isoforms from a single gene that qualitatively and quantitatively differs in their ability to produce proteins (Davuluri *et al.,* 2008, Landry *et al.,* 2003). The usage of a TSS determines the nature of the 5' UTR and the regulatory elements it encompasses. The selection of earlier, upstream TSSs yield a longer 5' UTR region, harbouring regulatory elements which modulate mRNA translation potential (Reviewed in Hinnebusch *et al.,* 2016). It is reasonable to assume that the longer the length of 5' UTR, the more regulatory features it could incorporate. A previous study that modelled the isoform-specific translation efficiency differences identified 5' UTR length as one of the two single best predictors in the model (Wang *et al.,* 2016). These and a variety of other studies have exhibited and characterised the cis-acting elements in the longer version of transcript isoform and their impact on mRNA translation potential (Arrick *et al.,* 1991, Davuluri *et al.,* 2000, Sobczak and Krzyzosiak, 2002, Wang *et al.,* 2016). So far, studies investigating the regulatory role of ATI have focused on the transcript isoforms that are derived from alternative promoters and are separated by a clear genomic space. In our study, we have elucidated the impact on translational output mediated by the selection of nearby TSSs in the cluster and showed that the longer 5' UTR isoforms derived from earlier TSSs in the cluster demonstrated a significant difference in the amount of protein produced. Unlike previous studies, our study has shown that even a smaller difference (< 50 bp) between alternative TSSs could lead to a large difference in translation efficiency.

A study by Xu *et al.* (2019) proposed that there is only one optimal TSS per gene and alternative TSS are nonadaptive and largely reflect molecular errors in transcription initiation. However, these assumptions are only relevant in the context of conservation genetics and underestimate the presence and contribution of divergent TSSs on gene output (Karlsson *et al.,* 2017, Rojas-Duran and Gilbert, 2012). Contrary to the finding reported by Xu *et al.* (2019), a study by Karlsson *et al.* (2017), measuring the TSS usage in a single cell, showed that the expression of major and minor TSSs are coregulated and these TSSs are not stochastically expressed. Additionally, it has been shown that polysomes associate differentially with distinct 5' UTR isoforms (Dieudonné *et al.,* 2015, Wang *et al.,* 2016), supporting the regulatory role of divergent TSS isoforms. A dominant (major) TSSs is usually considered by different RefSeq databases (NCBI and Ensembl) to annotate the 5' UTR boundary. This approach identifies the representative 5' UTR region of the transcript but provides limited information about the TSS diversity which generates a heterogenous 5' UTR population. The selection of upstream minor TSSs in the cluster could result in a relatively longer UTR and could encode regulatory features. At the outset, the contribution of minor TSSs to the mRNA 5' UTR diversity is seemingly low as the majority of transcription are derived from the major TSS. However, given the dynamic nature of TSS selection which could be influenced by a variety of factors detailed in section 3.3.6, the minor TSSs are equally important and could incorporate cis-acting features by switching the starting sites.

The analysis of TSS profiles using the FANTOM database, which is the largest repository of human TSS data, has revealed that gene transcription is spread across a region of multiple start sites in discrete clusters, with the distribution profiles depending on the promoter context (Kawaji *et al.,* 2006). Some of the distribution profiles, particularly those that are associated with the CpG based promoters display broad, bi or multi-modal TSSs peak distribution profiles. The presence of multi-dominant TSSs also subverts the claims made by Xu *et al.* (2019) about the existence of only one optimal TSS per gene. In addition to dominant TSSs, the minor TSSs are equally important, notably, the upstream TSSs as these produce TSS isoforms with variant 5′ UTR containing additional nucleotides that could have a regulatory role. In our study, the selection of an upstream TSS encoded a stable G4 structure in the 5′ UTR of *AGAP2* mRNA. Although that G4 forming TSSs was minor starting sites, the collective impact of all the G4 yielding TSSs could be substantial. In the case of *AGAP2*, we noted that any TSSs located 120 nucleotides upstream of the start codon in exon 1 would yield a G quadruplex consensus sequence in the 5′ UTR isoforms. Taking together with dynamic changes in TSS, the selection of upstream TSSs could significantly impact the overall gene output and result in a discrepancy in mRNA expression and protein abundance.

The distribution profiles of TSS are found to be tissue specific (Kawaji *et al.,* 2006, Ohmiya *et al.,* 2014). Kawaji *et al.* (2006) analysed the TSS clusters in the FANTOM database and showed a positional (median TSS) and regional (within 21 bp subregion) bias in the TSS distribution depending on the tissue types. These varying distribution profiles produce cell-specific TSS diversity and generate a heterogenous 5′ UTR population for the same gene transcript in different cells that only varies by a small number of nucleotides in the 5′ UTR. Our study also showed distinct distribution profiles of *AGAP2* TSSs in PC and CML cell lines with upstream (G4 forming) TSSs frequently noted in the CML cell lines. It could be argued that lower TSS diversity in PC cell lines could be due to reduced relative mRNA levels. However, the expression level of a gene was found to have an inverse correlation with the TSS diversity (Xu *et al.,* 2019). The distinct distribution profiles of TSS have been previously associated with promoter context (Carninci *et al.,* 2006, Carninci *et al.,* 2005). However, in our previous study, we have noted a comparable minimal promoter region for *AGAP2* in both the cell lines with a prominent role of SP1 and ATRA in transcriptional activation of *AGAP2* in these cell lines (Doush *et al.,* 2019) (section 3.1.1). Thus, ruling out the role of differential promoter usage in the observed TSS distribution pattern. A variety of cell-specific factor as highlighted in Chapter 3 (section 3.3.6) could also influence TSS selection and could lead to distinct distribution profiles in these cell lines. Further studies are required to expand our understanding of the determinants for TSS selection.

Our bioinformatics analysis has identified many genes in the human genome (3,888) that are susceptible to regulation by the TSS-G4 mechanism (section 5.3.1). These genes contained a G4 forming sequence in the 5′ UTR isoforms derived from alternative TSSs between major and upstream starting sites that are less than 100 bp apart. Analysing the TSS distribution profiles of these genes in PC and CML cell lines revealed significantly different distribution for more than 1,000 genes for G4 forming TSSs (section 5.3.2). Our bioinformatics analysis not only supports the finding reported by earlier studies on the cell-specific distribution of TSSs but also highlighted the potentially consequential distribution of TSSs that would encode a G4 motif in the alternative 5′ UTR and significantly impact the mRNA translation output. The pathway maps of these genes indicated enrichment of pathways and processes that are implicated in key regulatory mechanisms affecting the cellular phenotype and fate. We believe that the G4 forming sequences between alternative starting sites are non-random and are enriched in genes that have a regulatory role. Our analysis also showed that most of the genes (91.1%) that contained the alternative G4 sequence were protein-coding, confirming their non-random distribution and emphasizing their potential role in controlling gene output.

The 5‘ UTR G4 has been previously shown to modulate mRNA translation efficiency (Agarwala *et al.,* 2013, Arora *et al.,* 2008, Beaudoin and Perreault, 2010). These structures like other secondary structures produce a steric hindrance, disrupting serial translational processes including the formation of PIC and ribosome scanning and translocation (Bugaut and Balasubramanian, 2012, Beaudoin and Perreault, 2010). However, the previous studies on RNA G4s have reported their formation in the annotated 5’ UTR (major TSS) or the alternative splice isoform (Huppert *et al.,* 2008, Lee *et al.,* 2020). Our study, for the first time, reported the formation of a stable G4 motif in the alternative 5’ UTR isoforms by the selection of earlier TSSs in the cluster. It forms a distinct layer of regulation in which the transcript isoforms of a gene only differ by the presence of an additional G4 putative sequence in the 5’ UTR region that directs their translatability. The alternatively transcribed G4 creates variability in translational efficiency of different transcript isoforms and results in a mismatched profile where higher mRNA levels do not correlate with protein abundance. A study by Wang *et al.* (2016) performed quantitative modelling integrating various features in the differential 5’ UTR isoforms that together explained 57% of the variance in the isoform-specific translation differences. Integrating alternatively transcribed G4 could improve the predictive performance of the model and could account for more variances.

Different computational tools are available to detect potential G4 forming sequences and are based on different algorithms (section 1.5.1). In the current study, the pqsfinder tool was used due to its ability to process a batch of sequences and superior accuracy compared to G4Hunter and QGRS Mapper (Hon *et al.,* 2017). However, the presence of a G4 forming sequence in the alternative 5’ UTRs does not guarantee their formation *in vivo*. A study by Guo and Bartel (2016) demonstrated that a large number of predicted G4 consensus sequences are unfolded in the cells. It implies that the mere presence of these sequences is not always associated with the formation of the G4 structure. The intracellular formation of G4 structures is dependent on different factors including monovalent ion concentration, RBP, and levels of helicases (Reviewed in Cammas and Millevoi, 2017) and the interaction of these factors determine the likelihood of G4 folding by the consensus sequences. It is, therefore, important to verify G4 formation inside the cell. The *in vitro* techniques including CD spectroscopy are useful and could provide some idea about the folding pattern of these structures (Kejnovská *et al.,* 2019). But CD spectroscopy analyses only a short isolated sequence and it is important to analyse the G4 formation in the context of the entire mRNA folding (Weldon *et al.,* 2016). Taking this into consideration, there is a need for a technique to established formation of the G4 motif by putative quadruplex sequences inside the living cell. In our study, we have addressed this gap by designing in-house an immunoprecipitation technique that has successfully detected G4 structure in the selected mRNA.

Different approaches have been used to validate the formation of G4 structures inside the cell (section 4.1). Due to the inherent limitations with some of these techniques, the equilibrium may shift in the favour of G4 formation. The use of fixative in some of these techniques may influence the RNA structure and detect G4 conformations which might not be representative of the native RNA G4. To overcome these limitations and to capture the RNA G4 motifs in their native state, our method uses the structure-specific BG4 antibody (section 4.2). Our GRIP method has demonstrated successful enrichment of these structures using the BG4 antibody (section 4.3.1). Recently, Maltby *et al.* (2020) described an immunoprecipitation of RNA with G4s using BG4 antibody. However, their technique does not ensure removal of the contaminating DNA and it is possible that their signal could originate from DNA instead of RNA G4. Conversely, in our method, we have employed digitonin to selectively enrich cytosolic RNA and remove the contaminating DNA using RNase-free DNase I. Our elution strategy also decreased the background signal and facilitated selective detection of G4-containing RNA. Using our GRIP method, we have not only demonstrated the presence of the G4 motif in *AGAP2* mRNA but also specifically showed the formation of G4 in the longer 5' UTR region using plasmid transfection. Currently, our GRIP method is based on PCR amplification to detect the relative enrichment of RNA G4 in selected genes and could be optimised to use with sequencing-based platforms.

Previously it has been shown that the alternative longer 5' UTR tends to weakly associate with the polysomes (Li *et al.,* 2019, Wang *et al.,* 2016). These longer UTR contained regulatory features that decrease the polysome occupancy and consequently the translational efficiency. Different studies have also reported reduced ribosomal association for the G4-containing mRNAs (Cammas *et al.,* 2015, Murat *et al.,* 2014, Yang *et al.,* 2020). A review of literature has shown that the 5' UTR G4, like other secondary structures, does not interact directly with the polysome but exerts its effects indirectly by impairing assembly of the translation initiation machinery and perturbation of ribosome scanning toward the start codon (Babendure *et al.,* 2006, Bugaut and Balasubramanian, 2012, Jenkins *et al.,* 2010, Koromilas *et al.,* 1992, Kozak, 1989, Kozak, 1986). In our study, we have also noted decreased polysome association of longer G4-containing 5' UTR of *AGAP2* and *HK1* (section 4.3.3.2, section 5.3.4). Together with the luciferase reporter assay in our study (section 4.4.3.1), it evidenced that the G4 in the 5' UTR decreased the mRNA translation efficiency. Such small changes in the length of 5' UTR isoforms that significantly impact polysome seeding and translation efficiency have not been previously described. It validates that the alternatively transcribed G4 motifs are functionally relevant and decreased the translational output of their mRNA transcript.

Our study noted a discrepancy in AGAP2 mRNA and protein expression level and attempted to elucidate the contribution of the TSS-G4 mechanism in mediating the observed discordant profile. Our dual-luciferase reporter assay demonstrated a significant decrease in the luciferase activity of longer G4-containing 5' UTR relative to the shorter UTR without the G4 and indicated about ~45% decrease using *in vitro* transcription and translation system and ~35% decrease after transfecting the reporter constructs into DU145 cell line (section 4.3.3.1). The extent of translation suppression, however, may not completely explain the magnitude of AGAP2 RNA/protein discrepancy observed in our study. Other regulatory features in the 5‘ UTR, 3' UTR, ncRNA landscape, and cell-specific factors needs further evaluation to elaborate their contribution in mediating mRNA and protein inconsistency. Our study also showed a mismatched profile for the *HK1* gene similar to that of *AGAP2* and confirmed a role of the TSS-G4 mechanism in mediating such profile. However, we did not perform the reporter assay to analyse the extent of the translational suppression caused by the alternative 5' G4 structure in the case of the *HK1*. It is likely that the TSS-

G4 mechanism might have a contributing role instead of a leading role in generating a discordant profile and further studies are required to define the effect of the TSS-G4 mediated mechanism on translation output. Additionally, the different regulatory features might also interact with each other to shape the gene output.

In our study, we have also performed bioinformatics analysis to identify genes with higher mRNA expression, lower protein abundance, and a higher level of G4 forming TSS. The criteria used in our study to classify discrepancy in mRNA and protein level would detect only those genes with large significant differences in translation efficiency. The alternative 5’ UTR G4 might not always mediate such large differences in the mRNA translation output. As a result, our approach could miss many other relevant genes that do not display larger relative differences in the protein output. Even though the TSS-G4 mechanism impacts the gene translation output in our study, the G4 forming starting sites are not the major TSS and would not always mediate very large differences in mRNA and protein levels as anticipated in our bioinformatics analysis. There is room for improvement in our bioinformatics pipeline to account for subtle differences in the mRNA and protein levels mediated by alternatively transcribed G4. Additionally, advanced bioinformatics modelling integrating different mRNA features would provide a system to predict mRNA translation potential based on alternatively transcribed features. It would also provide insights into the contribution of different regulatory elements that are incorporated into the 5’ UTR by upstream TSS selection and their impact in dictating mRNA translation.

Furthermore, the available CML cell line (K562) used in the bioinformatics analysis of the NCI-60 database was not part of the current study. Although it is one of the established CML cell lines, the results produced might not be fully applicable for use with the CML cell lines used in our study. For the same reasons, some of the genes selected from the validation list did not show the mRNA and protein expression profiles as noted in the NCI-60 database (section 5.3.4). Further studies that analyse the complete set (mRNA, protein and TSS distribution profile) using the same cell lines would identify pertinent genes that could be used for validation.

Our study analysed the differences in mRNA and protein level using *AGAP2* as a model and elucidated the contribution of the TSS-G4 mechanism in mediating discordant profiles in PC and CML cell lines. In our study, we have also evaluated the discrepancy in AGAP2 mRNA/protein levels in other cancer cell lines. However, differential selection of G4 TSSs that affect the translatability of mRNA isoforms is also relevant in normal cells and tissues. Additionally, the TSS-G4 mechanism is important during development where the tissue-specific and temporal-specific usage of TSSs has been previously defined (Zhang *et al.,* 2017). Further studies are required to examine the significance of the TSS-G4 mechanism in different physiological and pathological conditions. In this regard, an omics wide approach incorporating the TSS information would provide a broader understanding of how the minor shifts in TSS could encode regulatory elements, including G4, and affect translation efficiency. It is also important to see how the TSS-G4 fits in with other gene regulatory mechanisms and its interaction with other cell-specific factors that influence the selection of TSS. Identifying factors that significantly shift a gene TSS and encode regulatory elements in the 5’ UTR could be used as a therapeutic strategy to modulate the translational output of a gene. This strategy would be meaningful in cancer management to modulate the translation of oncogenes by increasing transcription of the transcript with inefficient translation potential.

In the case of HK1, one of the CML cell line (KCL-22) did not exhibit a discordant profile like other CML cell lines and had comparable levels as other PC cell lines (section 5.3.4). Additionally, like other PC cell lines, KCL-22 had lower levels of G4 forming TSSs. It highlights that all the CML cell line might not demonstrate similar TSS distribution profiles and due to subtle differences in cell-specific factors, the distribution of TSS and/or gene expression profiles could be significantly different compared to other cell lines in the group. Likewise, the cell line-specific expression and TSS distribution profiles might not be similar to the corresponding primary cell. A cell-specific rather than cell-group specific approach would be required to examine a relationship between TSS distribution profiles and gene expression in future studies. Moreover, further studies are also required to explore the basis for differential distribution of *HK1* TSSs in KCL-22 cell lines. Nonetheless, our study showed that HK1 demonstrated a discrepancy in mRNA and protein expression, as also observed for *AGAP2*, and is regulated by the novel TSS-G4 mediated mechanism.

## 6.2 Future Work

To extend the current body of work, the following recommendations are made for future work:

It will be an interesting avenue to identify factors that influences the selection of start sites within the TSS cluster and explain the tissue-specific TSS distribution profiles. Previous studies have identified different factors that could influence the selection of TSS including transcription factors, DNA sequence elements, ncRNA, and other epigenetic factors (Javahery *et al.,* 1994, Jiang and Pugh, 2009, Kugel and Goodrich, 2017, Pardee *et al.,* 1998, Turowski and Tollervey, 2020). The epigenetic modifications could potentially have an impact on tissue-specific TSS distribution profiles as these epigenetic changes are important in driving tissue-specific gene expression programs. Additionally, a recent study by Mao *et al.* (2018) has exhibited that the G4 sites in the promoter region are enriched for DNA methyltransferase 1 (DNMT1) occupancy and the G4 structures formed at these sites inhibit DNMT1 enzymatic activity leading to hypomethylation at CpG islands in the promoter region. In this regard, our group has performed BS-Seq (bisulphite sequencing) to identify differential methylation patterns in the TSS region and we are currently analysing the sequencing data. It will be appropriate to explore the contribution of CpG methylation in influencing the selection of TSS. In this regard, our group is also planning to explore the impact of differential methylation pattern using methylation (DNMT) inhibitors (5-aza-2′-deoxycytidine) and methylation induction through the CRISPR-Cas9 system. Likewise, other epigenetic factors including histone modifications and chromatin organisation could be also studied using nucleosome-scanning assay and ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing), respectively (Buenrostro *et al.,* 2015, Infante *et al.,* 2012). It will elucidate the role of these factors in shaping the TSS landscape.

Other cell-specific components might be also implicated in regulating *AGAP2* expression in PC and CML cell lines. To this end, we have previously noted an antisense lncRNA *AGAP2-AS1* near the 3′ end of the gene and observed an increase in the levels of *AGAP2-AS1* in the PC cell lines (Doush, 2015). The antisense lncRNAs have been previously shown to regulate the gene expression of their cognate sense transcripts (Csorba *et al.,* 2014, Deforges *et al.,* 2019). It is plausible that there is a link between the higher relative levels of *AGAP2-AS1* and a lower *AGAP2* mRNA expression in PC cell lines. Further studies would provide interesting insights into the regulation of sense mRNA translation by cis-antisense lncRNA. We are currently performing experiments to evaluate the contribution of the antisense transcript in controlling *AGAP2* expression. Other ncRNAs including miRNA could also impact *AGAP2* expression and we are also currently analysing the differential expression of miRNA in our RNA-Seq data. It would be also interesting to examine the role of 3′ UTR and poly-A tail length to explain the involvement of these post-transcriptional factors in regulating *AGAP2* discordant profiles.

Another area of interest would be to optimise the GRIP method for use with a sequencing platform. It would generate a transcriptome-wide map of G4-containing mRNA. Currently, two RNA-G4 seq data are available which were produced using reverse transcription stalling (Kwok and Balasubramanian, 2015) and G4-RNA-specific precipitation using small-molecule ligand (BioTASQ) (Yang *et al.,* 2018). However, as mentioned previously, these methodologies use G4 stabilising ligands and/or fixation which could also induce G4 formation and may not be representative of native RNA G4 conformations. Optimising the GRIP method for use with a sequencing platform would facilitate the identification of RNA G4 folding in their native state without the use of stabilising ligands and would supplement the existing sequencing data on RNA G4. Additionally, it is relevant to confirm the formation of *AGAP2* 5' UTR G4 in its long functional RNA. The 5' UTR G4 formation in the full-length mRNA could be in competition with other secondary structures, making the G4 unstable and non-functional (Weldon *et al.,* 2016). Further experiments analysing the formation of 5' UTR G4 using 7-deazaguanine substituted nucleotides in the full-length *AGAP2* RNA followed by RNA footprinting, as suggested by Weldon *et al.* (2017), could further validate the relevance of 5' UTR G4 in the context of the entire mRNA molecule. It will be also relevant to use well established RNA G4 stabilising ligands such as carboxypyridostatin (Di Antonio *et al.,* 2012), BRACO-19 (Moore *et al.,* 2006), and RHPS4 (Salvati *et al.,* 2007) to further confirm the formation of G4 structure in the alternative 5' UTR. The use of a G4 stabilising ligand could further strengthen the authenticity of the DLR and polysome profiling results. Furthermore, it will be interesting to analyse the luciferase activity of the longer G4 containing 5' UTR isoforms after treatment with a G4 stabilising ligand.

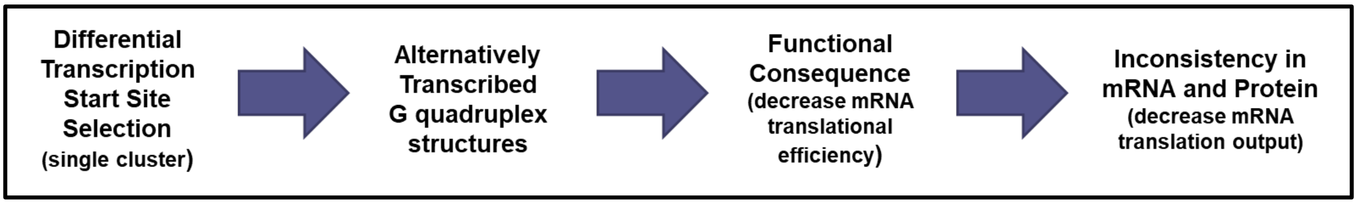
Future studies are also required to validate the TSS-G4 mechanism using other cell lines. In our study, we have also noted a discrepancy in AGAP2 mRNA and protein expression in other cell lines (section 3.3.1). Analysing the TSS distribution in these cell lines would underline the relevance of alternatively transcribed G4 structures in mediating the observed discordant mRNA and protein levels. Additionally, examining the TSS patterns in the corresponding normal and cancer cell would generate novel understandings about the changing TSS distribution with cancerous phenotype and would identify genes that display prominent changes in TSS distribution with cancer progression. It would enable the development of a novel diagnostic approach, where the changes in the distribution of TSSs could indicate underlying cancerous changes. In addition to cancer, characterising TSS profiles in other diseases and pathology would enhance the knowledge base relating to the complex gene expression changes and would identify therapeutic targets whose expression could be controlled by changing the selection of TSSs.

Another interesting area for future studies would be to map other alternatively transcribed regulatory elements in addition to G4. In a recent study by Palavecino *et al.* (2020), the authors have reported the inhibitory effects of alternatively transcribed uAUG by the selection of TSSs that were separated by only a few nucleotides. Advance bioinformatics analysis to map all the shorter sequence elements in the cluster that could be alternatively transcribed such as uORF, uAUG, i-Motifs, 5' TOP motif, and hairpin etc would uncover the complex regulatory potential of these elements. Future studies would be subsequently required to evaluate the impact of these incorporated elements on mRNA translational output. Unlike previous studies, these elements are integrated by selecting TSSs that are separated by only a few nucleotides in the cluster. Together, it could constitute an emerging domain for research in the broad field of gene regulation.

Lastly, it will be also intriguing to determine TSS profiles in a synchronised cell population which might show reduced variability in a gene's TSS distribution. An earlier study by Hwang *et al.* (1998) has demonstrated cell cycle-dependent usage of transcriptional start sites for the *Cyclin B1* gene. Other genes might also show cell cycle-dependent changes in TSS selection. It will be interesting to look at the TSSs distribution profile in the synchronised cell population and the effect of cell synchrony on TSSs usage pattern. To this end, we have optimised a cell synchronisation protocol for adherent and suspension cell lines using double thymidine block and nocodazole treatment (Surani *et al.,* 2021). Further studies are required to examine the TSS profile of a gene in a synchronised cell population compared to unsynchronised cells. Additionally, it will also be interesting to examine the TSS distribution profile at a single-cell level. A study by Karlsson *et al.* (2017) has measured the TSS activity in a single cell using the single-cell RNA-seq dataset. The single-cell analysis would overcome the stochastic nature of gene expression and would illustrate the functional consequences of alternative TSSs even if they are not differentially used at a population level.

## 6.3 Conclusion

In conclusion, our study highlighted a novel gene regulation mechanism involving alternative transcription initiation within a core promoter region that encodes a G4 forming sequence in the differential 5’ UTR isoforms, modulating mRNA translation efficiency. We used *AGAP2* gene as a model and demonstrated the differential distribution of TSSs in PC and CML cell lines with significantly higher usage of upstream TSSs in the CML cell lines that encode a G4 consensus sequence in the alternate 5’ UTR isoform of *AGAP2* mRNA. We have verified the formation of a stable G4 structure by these sequences using our in-house developed GRIP Method. Moreover, our study revealed that alternatively transcribed 5’ UTR G4 suppress translation and are poorly associated with polysomes. Furthermore, we demonstrated that the TSS-G4 mediated mechanism is not only limited to *AGAP2* expression regulation but is also implicated in controlling the expression of other genes. Using bioinformatics analysis, we identified other genes that could be potentially regulated by the TSS-G4 mediated mechanism and validated our results using the *HK1* gene.


Differential Transcription Start Site Selection (single cluster)

Alternatively Transcribed G quadruplex structures

Functional Consequence (decrease mRNA translational efficiency)

Inconsistency in mRNA and Protein (decrease mRNA translation output)

**Figure 6.2: Summary of the novel TSS-G4 mechanism.** The differential TSS usage encodes a G4 forming sequence in the alternative 5’ UTR isoforms that fold into a stable G4 inside the cell. These alternatively transcribed G4 structure decreases the mRNA translation efficiency and leads to a discrepancy in RNA and protein expression where higher RNA levels are not efficiently translated to proteins. This TSS-G4 mechanism is not only applicable for *AGAP2* regulation but could potentially regulate the expression of many other genes.

# Chapter 7:
# References

ADAM, S. A., MARR, R. S. & GERACE, L. 1990. Nuclear protein import in permeabilized mammalian cells requires soluble cytoplasmic factors. J Cell Biol, 111, 807-16.

AGARWALA, P., PANDEY, S., MAPA, K. & MAITI, S. 2013. The G-quadruplex augments translation in the 5' untranslated region of transforming growth factor β2. Biochemistry, 52, 1528-38.

AHN, J.-Y., HU, Y., KROLL, T. G., ALLARD, P. & YE, K. 2004a. PIKE-A is amplified in human cancers and prevents apoptosis by up-regulating Akt. Proceedings of the National Academy of Sciences of the United States of America, 101, 6993-6998.

AHN, J.-Y., RONG, R., KROLL, T. G., VAN MEIR, E. G., SNYDER, S. H. & YE, K. 2004b. PIKE (Phosphatidylinositol 3-Kinase Enhancer)-A GTPase Stimulates Akt Activity and Mediates Cellular Invasion. Journal of Biological Chemistry, 279, 16441-16451.

AHUJA, J. S., SANDHU, R., MAINPAL, R., LAWSON, C., HENLEY, H., HUNT, P. A., YANOWITZ, J. L. & BÖRNER, G. V. 2017. Control of meiotic pairing and recombination by chromosomally tethered 26S proteasome. Science, 355, 408-411.

ALAM, K., FARASYN, T., CROWE, A., DING, K. & YUE, W. 2017. Treatment with proteasome inhibitor bortezomib decreases organic anion transporting polypeptide (OATP) 1B3-mediated transport in a substrate-dependent manner. PLOS ONE, 12, e0186924.

ALASKHAR ALHAMWE, B., KHALAILA, R., WOLF, J., VON BÜLOW, V., HARB, H., ALHAMDAN, F., HII, C. S., PRESCOTT, S. L., FERRANTE, A., RENZ, H., GARN, H. & POTACZEK, D. P. 2018. Histone modifications and their role in epigenetics of atopy and allergic diseases. Allergy, Asthma & Clinical Immunology, 14, 39.

ANDERSON, B. R., KARIKÓ, K. & WEISSMAN, D. 2013. Nucleofection induces transient eIF2α phosphorylation by GCN2 and PERK. Gene Ther, 20, 136-42.

ANDERSSON, R. & SANDELIN, A. 2020. Determinants of enhancer and promoter activities of regulatory elements. Nature Reviews Genetics, 21, 71-87.

ANDO, M., SAITO, Y., XU, G., BUI, N. Q., MEDETGUL-ERNAR, K., PU, M., FISCH, K., REN, S., SAKAI, A., FUKUSUMI, T., LIU, C., HAFT, S., PANG, J., MARK, A., GAYKALOVA, D. A., GUO, T., FAVOROV, A. V., YEGNASUBRAMANIAN, S., FERTIG, E. J., HA, P., TAMAYO, P., YAMASOBA, T., IDEKER, T., MESSER, K. & CALIFANO, J. A. 2019. Chromatin dysregulation and DNA methylation at transcription start sites associated with transcriptional repression in cancers. Nature communications, 10, 2188-2188.

ARAUJO, P. R., YOON, K., KO, D., SMITH, A. D., QIAO, M., SURESH, U., BURNS, S. C. & PENALVA, L. O. F. 2012. Before It Gets Started: Regulating Translation at the 5′ UTR. Comparative and Functional Genomics, 2012, 475731.

ARCE, L., YOKOYAMA, N. N. & WATERMAN, M. L. 2006. Diversity of LEF/TCF action in development and disease. Oncogene, 25, 7492-7504.

ARORA, A., DUTKIEWICZ, M., SCARIA, V., HARIHARAN, M., MAITI, S. & KURRECK, J. 2008. Inhibition of translation in living eukaryotic cells by an RNA G-quadruplex motif. RNA (New York, N.Y.), 14, 1290-1296.

ARRIBERE, J. A. & GILBERT, W. V. 2013. Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. Genome Res, 23, 977-87.

ARRICK, B. A., LEE, A. L., GRENDELL, R. L. & DERYNCK, R. 1991. Inhibition of translation of transforming growth factor-beta 3 mRNA by its 5' untranslated region. Molecular and Cellular Biology, 11, 4306-4313.

AVISSAR, Y., CHOI, J., DESAIX, J., JURUKOVSKI, V., WISE, R. & RYE, C. 2018. Biology, OpenStax.

BABENDURE, J. R., BABENDURE, J. L., DING, J. H. & TSIEN, R. Y. 2006. Control of mammalian translation by mRNA structure near caps. Rna, 12, 851-61.

BADEAUX, A. I. & SHI, Y. 2013. Emerging roles for chromatin as a signal integration and storage platform. Nature reviews. Molecular cell biology, 14, 211-224.

BAKALAR, B., HEDDI, B., SCHMITT, E., MECHULAM, Y. & PHAN, A. T. 2019. A Minimal Sequence for Left-Handed G-Quadruplex Formation. Angew Chem Int Ed Engl, 58, 2331-2335.

BALKWILL, G. D., DERECKA, K., GARNER, T. P., HODGMAN, C., FLINT, A. P. & SEARLE, M. S. 2009. Repression of translation of human estrogen receptor alpha by G-quadruplex formation. Biochemistry, 48, 11487-95.

BANCO, M. T. & FERRÉ-D'AMARÉ, A. R. 2021. The emerging structural complexity of G-quadruplex RNAs. Rna, 27, 390-402.

BANG, I. 1910. Untersuchungen über die Guanylsäre. Biochem. Z., 26, 293-311.

BARALLE, F. E. & GIUDICE, J. 2017. Alternative splicing as a regulator of development and tissue identity. Nature Reviews Molecular Cell Biology, 18, 437-451.

BARREAU, C., PAILLARD, L. & OSBORNE, H. B. 2005. AU-rich elements and associated factors: are there unifying principles? Nucleic Acids Res, 33, 7138-50.

BARRETINA, J., CAPONIGRO, G., STRANSKY, N., VENKATESAN, K., MARGOLIN, A. A., KIM, S., WILSON, C. J., LEHÁR, J., KRYUKOV, G. V., SONKIN, D., REDDY, A., LIU, M., MURRAY, L., BERGER, M. F., MONAHAN, J. E., MORAIS, P., MELTZER, J., KOREJWA, A., JANÉ-VALBUENA, J., MAPA, F. A., THIBAULT, J., BRIC-FURLONG, E., RAMAN, P., SHIPWAY, A., ENGELS, I. H., CHENG, J., YU, G. K., YU, J., ASPESI, P., DE SILVA, M., JAGTAP, K., JONES, M. D., WANG, L., HATTON, C., PALESCANDOLO, E., GUPTA, S., MAHAN, S., SOUGNEZ, C., ONOFRIO, R. C., LIEFELD, T., MACCONAILL, L., WINCKLER, W., REICH, M., LI, N., MESIROV, J. P., GABRIEL, S. B., GETZ, G., ARDLIE, K., CHAN, V., MYER, V. E., WEBER, B. L., PORTER, J., WARMUTH, M., FINAN, P., HARRIS, J. L., MEYERSON, M., GOLUB, T. R., MORRISSEY, M. P., SELLERS, W. R., SCHLEGEL, R. & GARRAWAY, L. A. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 483, 603-607.

BATUT, P., DOBIN, A., PLESSY, C., CARNINCI, P. & GINGERAS, T. R. 2013. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome Res, 23, 169-80.

BEAUDOIN, J. D. & PERREAULT, J. P. 2013. Exploring mRNA 3'-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening. Nucleic Acids Res, 41, 5898-911.

BEAUDOIN, J.-D. & PERREAULT, J.-P. 2010. 5'-UTR G-quadruplex structures acting as translational repressors. Nucleic acids research, 38, 7022-7036.

BEAUDOIN, J.-D., JODOIN, R. & PERREAULT, J.-P. 2013. In-line probing of RNA G-quadruplexes. Methods, 64, 79-87.

BEDRAT, A., LACROIX, L. & MERGNY, J.-L. 2016. Re-evaluation of G-quadruplex propensity with G4Hunter. Nucleic acids research, 44, 1746-1759.

BERKOVITS, B. D. & MAYR, C. 2015. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. Nature, 522, 363-7.

BESNARD, E., BABLED, A., LAPASSET, L., MILHAVET, O., PARRINELLO, H., DANTEC, C., MARIN, J.-M. & LEMAITRE, J.-M. 2012. Unraveling cell type–specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nature Structural & Molecular Biology, 19, 837-844.

BEZNOSKOVÁ, P., WAGNER, S., JANSEN, M. E., VON DER HAAR, T. & VALÁŠEK, L. S. 2015. Translation initiation factor eIF3 promotes programmed stop codon readthrough. Nucleic Acids Res, 43, 5099-111.

BIFFI, G., DI ANTONIO, M., TANNAHILL, D. & BALASUBRAMANIAN, S. 2014. Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells. Nature Chemistry, 6, 75-80.

BIFFI, G., TANNAHILL, D. & BALASUBRAMANIAN, S. 2012. An intramolecular G-quadruplex structure is required for binding of telomeric repeat-containing RNA to the telomeric protein TRF2. J Am Chem Soc, 134, 11974-6.

BIFFI, G., TANNAHILL, D., MCCAFFERTY, J. & BALASUBRAMANIAN, S. 2013. Quantitative visualization of DNA G-quadruplex structures in human cells. Nature Chemistry, 5, 182-186.

BJORNSTI, M. A. & HOUGHTON, P. J. 2004. Lost in translation: dysregulation of cap-dependent translation and cancer. Cancer Cell, 5, 519-23.

BLASCHKE, R. J., TÖPFER, C., MARCHINI, A., STEINBEISSER, H., JANSSEN, J. W. & RAPPOLD, G. A. 2003. Transcriptional and translational regulation of the Leri-Weill and Turner syndrome homeobox gene SHOX. J Biol Chem, 278, 47820-6.

BLIGHE, K., RANA, S. & LEWIS, M. 2020. EnhancedVolcano: Publication-ready volcano plots with 667 enhanced colouring and labeling. R package version 1.6. 0. 668 https://github. com/kevinblighe. EnhancedVolcano.

BONN, S., ZINZEN, R. P., GIRARDOT, C., GUSTAFSON, E. H., PEREZ-GONZALEZ, A., DELHOMME, N., GHAVI-HELM, Y., WILCZYŃSKI, B., RIDDELL, A. & FURLONG, E. E. 2012. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat Genet, 44, 148-56.

BONNAL, S., SCHAEFFER, C., CRÉANCIER, L., CLAMENS, S., MOINE, H., PRATS, A. C. & VAGNER, S. 2003. A single internal ribosome entry site containing a G quartet RNA structure drives fibroblast growth factor 2 gene expression at four alternative translation initiation codons. J Biol Chem, 278, 39330-6.

BOURGEOIS, C. F., MORTREUX, F. & AUBOEUF, D. 2016. The multiple functions of RNA helicases as drivers and regulators of gene expression. Nature Reviews Molecular Cell Biology, 17, 426-438.

BOUSSEMART, L., MALKA-MAHIEU, H., GIRAULT, I., ALLARD, D., HEMMINGSSON, O., TOMASIC, G., THOMAS, M., BASMADJIAN, C., RIBEIRO, N., THUAUD, F., MATEUS, C., ROUTIER, E., KAMSU-KOM, N., AGOUSSI, S., EGGERMONT, A. M., DÉSAUBRY, L., ROBERT, C. & VAGNER, S. 2014. eIF4F is a nexus of resistance to anti-BRAF and anti-MEK cancer therapies. Nature, 513, 105-9.

BUENROSTRO, J. D., WU, B., CHANG, H. Y. & GREENLEAF, W. J. 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current protocols in molecular biology, 109, 21.29.1-21.29.9.

BUGAUT, A. & BALASUBRAMANIAN, S. 2012. 5'-UTR RNA G-quadruplexes: translation regulation and targeting. Nucleic acids research, 40, 4727-4741.

BURATOWSKI, S., HAHN, S., GUARENTE, L. & SHARP, P. A. 1989. Five intermediate complexes in transcription initiation by RNA polymerase II. Cell, 56, 549-61.

BURGE, S., PARKINSON, G. N., HAZEL, P., TODD, A. K. & NEIDLE, S. 2006. Quadruplex DNA: sequence, topology and structure. Nucleic acids research, 34, 5402-5415.

BUTLER, J. E. & KADONAGA, J. T. 2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev, 16, 2583-92.

CAI, Y., WANG, J., LI, R., AYALA, G., ITTMANN, M. & LIU, M. 2009. GGAP2/PIKE-A Directly Activates Both the Akt and Nuclear Factor-κB Pathways and Promotes Prostate Cancer Progression. Cancer Research, 69, 819-827.

CAMMAS, A. & MILLEVOI, S. 2017. RNA G-quadruplexes: emerging mechanisms in disease. Nucleic acids research, 45, 1584-1595.

CAMMAS, A., DUBRAC, A., MOREL, B., LAMAA, A., TOURIOL, C., TEULADE-FICHOU, M.-P., PRATS, H. & MILLEVOI, S. 2015. Stabilization of the G-quadruplex at the VEGF IRES represses cap-independent translation. RNA Biology, 12, 320-329.

CAMPBELL, N. H. & PARKINSON, G. N. 2007. Crystallographic studies of quadruplex nucleic acids. Methods, 43, 252-63.

CARNINCI, P., KASUKAWA, T., KATAYAMA, S., GOUGH, J., FRITH, M. C., MAEDA, N., OYAMA, R., RAVASI, T., LENHARD, B., WELLS, C., KODZIUS, R., SHIMOKAWA, K., BAJIC, V. B., BRENNER, S. E., BATALOV, S., FORREST, A. R., ZAVOLAN, M., DAVIS, M. J., WILMING, L. G., AIDINIS, V., ALLEN, J. E., AMBESI-IMPIOMBATO, A., APWEILER, R., ATURALIYA, R. N., BAILEY, T. L., BANSAL, M., BAXTER, L., BEISEL, K. W., BERSANO, T., BONO, H., CHALK, A. M., CHIU, K. P., CHOUDHARY, V., CHRISTOFFELS, A., CLUTTERBUCK, D. R., CROWE, M. L., DALLA, E., DALRYMPLE, B. P., DE BONO, B., DELLA GATTA, G., DI BERNARDO, D., DOWN, T., ENGSTROM, P., FAGIOLINI, M., FAULKNER, G., FLETCHER, C. F., FUKUSHIMA, T., FURUNO, M., FUTAKI, S., GARIBOLDI, M., GEORGII-HEMMING, P., GINGERAS, T. R., GOJOBORI, T., GREEN, R. E., GUSTINCICH, S., HARBERS, M., HAYASHI, Y., HENSCH, T. K., HIROKAWA, N., HILL, D., HUMINIECKI, L., IACONO, M., IKEO, K., IWAMA, A., ISHIKAWA, T., JAKT, M., KANAPIN, A., KATOH, M., KAWASAWA, Y., KELSO, J., KITAMURA, H., KITANO, H., KOLLIAS, G., KRISHNAN, S. P., KRUGER, A., KUMMERFELD, S. K., KUROCHKIN, I. V., LAREAU, L. F., LAZAREVIC, D., LIPOVICH, L., LIU, J., LIUNI, S., MCWILLIAM, S., MADAN BABU, M., MADERA, M., MARCHIONNI, L., MATSUDA, H., MATSUZAWA, S., MIKI, H., MIGNONE, F., MIYAKE, S., MORRIS, K., MOTTAGUI-TABAR, S., MULDER, N., NAKANO, N., NAKAUCHI, H., NG, P., NILSSON, R., NISHIGUCHI, S., NISHIKAWA, S., *et al.* 2005. The transcriptional landscape of the mammalian genome. Nature Genetics, 309, 1559-63.

CARNINCI, P., SANDELIN, A., LENHARD, B., KATAYAMA, S., SHIMOKAWA, K., PONJAVIC, J., SEMPLE, C. A. M., TAYLOR, M. S., ENGSTRÖM, P. G., FRITH, M. C., FORREST, A. R. R., ALKEMA, W. B., TAN, S. L., PLESSY, C., KODZIUS, R., RAVASI, T., KASUKAWA, T., FUKUDA, S., KANAMORI-KATAYAMA, M., KITAZUME, Y., KAWAJI, H., KAI, C., NAKAMURA, M., KONNO, H., NAKANO, K., MOTTAGUI-TABAR, S., ARNER, P., CHESI, A., GUSTINCICH, S., PERSICHETTI, F., SUZUKI, H., GRIMMOND, S. M., WELLS, C. A., ORLANDO, V., WAHLESTEDT, C., LIU, E. T., HARBERS, M., KAWAI, J., BAJIC, V. B., HUME, D. A. & HAYASHIZAKI, Y. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nature Genetics, 38, 626-635.

CHAMBERS, V. S., MARSICO, G., BOUTELL, J. M., DI ANTONIO, M., SMITH, G. P. & BALASUBRAMANIAN, S. 2015. High-throughput sequencing of DNA G-quadruplex structures in the human genome. Nat Biotechnol, 33, 877-81.

CHAN, C. B., LIU, X., ENSSLIN, M. A., DILLEHAY, D. L., ORMANDY, C. J., SOHN, P., SERRA, R. & YE, K. 2010. PIKE-A is required for prolactin-mediated STAT5a activation in mammary gland development. Embo j, 29, 956-68.

CHAN, C. B., LIU, X., HE, K., QI, Q., JUNG, D. Y., KIM, J. K. & YE, K. 2011. The association of phosphoinositide 3-kinase enhancer A with hepatic insulin receptor enhances its kinase activity. EMBO Rep, 12, 847-54.

CHASSÉ, H., BOULBEN, S., COSTACHE, V., CORMIER, P. & MORALES, J. 2016. Analysis of translation using polysome profiling. Nucleic Acids Research, 45, e15-e15.

CHATTERJEE, S. & PAL, J. K. 2009. Role of 5'- and 3'-untranslated regions of mRNAs in human diseases. Biol Cell, 101, 251-62.

CHEN, X. C., CHEN, S. B., DAI, J., YUAN, J. H., OU, T. M., HUANG, Z. S. & TAN, J. H. 2018. Tracking the Dynamic Folding and Unfolding of RNA G-Quadruplexes in Live Cells. Angew Chem Int Ed Engl, 57, 4702-4706.

CHEUNG, A. C. M., SAINSBURY, S. & CRAMER, P. 2011. Structural basis of initial RNA polymerase II transcription. The EMBO journal, 30, 4755-4763.

CHOLEWA-WACLAW, J., BIRD, A., VON SCHIMMELMANN, M., SCHAEFER, A., YU, H., SONG, H., MADABHUSHI, R. & TSAI, L.-H. 2016. The Role of Epigenetic Mechanisms in the Regulation of Gene Expression in the Nervous System. The Journal of neuroscience : the official journal of the Society for Neuroscience, 36, 11427-11434.

CHUI, A. J., OKONDO, M. C., RAO, S. D., GAI, K., GRISWOLD, A. R., JOHNSON, D. C., BALL, D. P., TAABAZUING, C. Y., ORTH, E. L., VITTIMBERGA, B. A. & BACHOVCHIN, D. A. 2019. N-terminal degradation activates the NLRP1B inflammasome. Science (New York, N.Y.), 364, 82-85.

CIECHANOVER, A. & SCHWARTZ, A. L. 1998. The ubiquitin-proteasome pathway: The complexity and myriad functions of proteins death. Proceedings of the National Academy of Sciences, 95, 2727.

COBBOLD, L. C., WILSON, L. A., SAWICKA, K., KING, H. A., KONDRASHOV, A. V., SPRIGGS, K. A., BUSHELL, M. & WILLIS, A. E. 2010. Upregulated c-myc expression in multiple myeloma by internal ribosome entry results from increased interactions with and expression of PTB-1 and YB-1. Oncogene, 29, 2884-2891.

COCK, P. J. A., ANTAO, T., CHANG, J. T., CHAPMAN, B. A., COX, C. J., DALKE, A., FRIEDBERG, I., HAMELRYCK, T., KAUFF, F., WILCZYNSKI, B. & DE HOON, M. J. L. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics (Oxford, England), 25, 1422-1423.

COGOI, S. & XODO, L. E. 2006. G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. Nucleic Acids Research, 34, 2536-2549.

COLBEAU, A., NACHBAUR, J. & VIGNAIS, P. M. 1971. Enzymac characterization and lipid composition of rat liver subcellular membranes. Biochimica et Biophysica Acta (BBA) - Biomembranes, 249, 462-492.

COLGAN, D. F. & MANLEY, J. L. 1997. Mechanism and regulation of mRNA polyadenylation. Genes Dev, 11, 2755-66.

CONAWAY, R. C. & CONAWAY, J. W. 2011. Function and regulation of the Mediator complex. Curr Opin Genet Dev, 21, 225-30.

CORPET, F. 1988. Multiple sequence alignment with hierarchical clustering. Nucleic acids research, 16, 10881-10890.

CRAWFORD, R. A. & PAVITT, G. D. 2019. Translational regulation in response to stress in Saccharomyces cerevisiae. Yeast (Chichester, England), 36, 5-21.

CSORBA, T., QUESTA, J. I., SUN, Q. & DEAN, C. 2014. Antisense COOLAIR mediates the coordinated switching of chromatin states at FLC during vernalization. Proceedings of the National Academy of Sciences of the United States of America, 111, 16160-16165.

CULJKOVIC-KRALJACIC, B. & BORDEN, K. L. B. 2018. The Impact of Post-transcriptional Control: Better Living Through RNA Regulons. Frontiers in Genetics, 9.

CURRAN, J. A. & WEISS, B. 2016. What Is the Impact of mRNA 5' TL Heterogeneity on Translational Start Site Selection and the Mammalian Cellular Phenotype? Frontiers in genetics, 7, 156-156.

DANINO, Y. M., EVEN, D., IDESES, D. & JUVEN-GERSHON, T. 2015. The core promoter: At the heart of gene expression. Biochim Biophys Acta, 1849, 1116-31.

DAVULURI, R. V., SUZUKI, Y., SUGANO, S. & ZHANG, M. Q. 2000. CART classification of human 5' UTR sequences. Genome Res, 10, 1807-16.

DAVULURI, R. V., SUZUKI, Y., SUGANO, S., PLASS, C. & HUANG, T. H. 2008. The functional consequences of alternative promoter use in mammalian genomes. Trends Genet, 24, 167-77.

DE SOUSA ABREU, R., PENALVA, L. O., MARCOTTE, E. M. & VOGEL, C. 2009. Global signatures of protein and mRNA expression levels. Molecular BioSystems, 5, 1512-1526.

DEFORGES, J., REIS, R. S., JACQUET, P., SHEPPARD, S., GADEKAR, V. P., HART-SMITH, G., TANZER, A., HOFACKER, I. L., ISELI, C., XENARIOS, I. & POIRIER, Y. 2019. Control of Cognate Sense mRNA Translation by cis-Natural Antisense RNAs. Plant Physiology, 180, 305-322.

DEL VILLAR-GUERRA, R., TRENT, J. O. & CHAIRES, J. B. 2018. G-Quadruplex Secondary Structure Obtained from Circular Dichroism Spectroscopy. 57, 7171-7175.

DENG, W. & ROBERTS, S. G. E. 2007. TFIIB and the regulation of transcription by RNA polymerase II. Chromosoma, 116, 417-429.

DEVER, T. E., DINMAN, J. D. & GREEN, R. 2018. Translation Elongation and Recoding in Eukaryotes. Cold Spring Harbor perspectives in biology, 10, a032649.

DEVER, T. E., FENG, L., WEK, R. C., CIGAN, A. M., DONAHUE, T. F. & HINNEBUSCH, A. G. 1992. Phosphorylation of initiation factor 2 alpha by protein kinase GCN2 mediates gene-specific translational control of GCN4 in yeast. Cell, 68, 585-96.

DI ANTONIO, M., BIFFI, G., MARIANI, A., RAIBER, E.-A., RODRIGUEZ, R. & BALASUBRAMANIAN, S. 2012. Selective RNA versus DNA G-quadruplex targeting by in situ click chemistry. Angewandte Chemie (International ed. in English), 51, 11073-11078.

DICK, T. P., NUSSBAUM, A. K., DEEG, M., HEINEMEYER, W., GROLL, M., SCHIRLE, M., KEILHOLZ, W., STEVANOVIĆ, S., WOLF, D. H., HUBER, R., RAMMENSEE, H. G. & SCHILD, H. 1998. Contribution of proteasomal beta-subunits to the cleavage of peptide substrates analyzed with yeast mutants. J Biol Chem, 273, 25637-46.

DIEUDONNÉ, F.-X., O’CONNOR, P. B. F., GUBLER-JAQUIER, P., YASREBI, H., CONNE, B., NIKOLAEV, S., ANTONARAKIS, S., BARANOV, P. V. & CURRAN, J. 2015. The effect of heterogeneous Transcription Start Sites (TSS) on the translatome: implications for the mammalian cellular phenotype. BMC Genomics, 16, 986.

DING, L., MYCHALECKYJ, J. C. & HEGDE, A. N. 2007. Full length cloning and expression analysis of splice variants of regulator of G-protein signaling RGS4 in human and murine brain. Gene, 401, 46-60.

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29, 15-21.

DOUSH, Y. 2015. AGAP2 expression and regulation in the CML and prostate cancer cells. Doctoral, Nottingham Trent University.

DOUSH, Y., SURANI, A. A., NAVARRO-CORCUERA, A., MCARDLE, S., BILLETT, E. E. & MONTIEL-DUARTE, C. 2019. SP1 and RARα regulate AGAP2 expression in cancer. Scientific Reports, 9, 390.

DRYGIN, D., SIDDIQUI-JAIN, A., O'BRIEN, S., SCHWAEBE, M., LIN, A., BLIESATH, J., HO, C. B., PROFFITT, C., TRENT, K., WHITTEN, J. P., LIM, J. K., VON HOFF, D., ANDERES, K. & RICE, W. G. 2009. Anticancer activity of CX-3543: a direct inhibitor of rRNA biogenesis. Cancer Res, 69, 7653-61.

DUNCAN, C. D. S. & MATA, J. 2017. Effects of cycloheximide on the interpretation of ribosome profiling experiments in Schizosaccharomyces pombe. Scientific Reports, 7, 10331.

DUNHAM, I., KUNDAJE, A., ALDRED, S. F., COLLINS, P. J., DAVIS, C. A., DOYLE, F., EPSTEIN, C. B., FRIETZE, S., HARROW, J., KAUL, R., KHATUN, J., LAJOIE, B. R., LANDT, S. G., LEE, B.-K., PAULI, F., ROSENBLOOM, K. R., SABO, P., SAFI, A., SANYAL, A., SHORESH, N., SIMON, J. M., SONG, L., TRINKLEIN, N. D., ALTSHULER, R. C., BIRNEY, E., BROWN, J. B., CHENG, C., DJEBALI, S., DONG, X., DUNHAM, I., ERNST, J., FUREY, T. S., GERSTEIN, M., GIARDINE, B., GREVEN, M., HARDISON, R. C., HARRIS, R. S., HERRERO, J., HOFFMAN, M. M., IYER, S., KELLIS, M., KHATUN, J., KHERADPOUR, P., KUNDAJE, A., LASSMANN, T., LI, Q., LIN, X., MARINOV, G. K., MERKEL, A., MORTAZAVI, A., PARKER, S. C. J., REDDY, T. E., ROZOWSKY, J., SCHLESINGER, F., THURMAN, R. E., WANG, J., WARD, L. D., WHITFIELD, T. W., WILDER, S. P., WU, W., XI, H. S., YIP, K. Y., ZHUANG, J., BERNSTEIN, B. E., BIRNEY, E., DUNHAM, I., GREEN, E. D., GUNTER, C., SNYDER, M., PAZIN, M. J., LOWDON, R. F., DILLON, L. A. L., ADAMS, L. B., KELLY, C. J., ZHANG, J., WEXLER, J. R., GREEN, E. D., GOOD, P. J., FEINGOLD, E. A., BERNSTEIN, B. E., BIRNEY, E., CRAWFORD, G. E., DEKKER, J., ELNITSKI, L., FARNHAM, P. J., GERSTEIN, M., GIDDINGS, M. C., GINGERAS, T. R., GREEN, E. D., GUIGÓ, R., HARDISON, R. C., HUBBARD, T. J., KELLIS, M., KENT, W. J., LIEB, J. D., MARGULIES, E. H., MYERS, R. M., SNYDER, M., STAMATOYANNOPOULOS, J. A., TENENBAUM, S. A., *et al.* 2012. An integrated encyclopedia of DNA elements in the human genome. Nature, 489, 57-74.

HE, K., JANG, S. W., JOSHI, J., YOO, M. H. & YE, K. 2011. Akt-phosphorylated PIKE-A inhibits UNC5B-induced apoptosis in cancer cell lines in a p53-dependent manner. Mol Biol Cell, 22, 1943-54.

EDDY, J. & MAIZELS, N. 2006. Gene function correlates with potential for G4 DNA formation in the human genome. Nucleic acids research, 34, 3887-3896.

EDFORS, F., DANIELSSON, F., HALLSTRÖM, B. M., KÄLL, L., LUNDBERG, E., PONTÉN, F., FORSSTRÖM, B. & UHLÉN, M. 2016. Gene-specific correlation of RNA and protein levels in human cells and tissues. Molecular systems biology, 12, 883-883.

EINARSON, O. J. & SEN, D. 2017. Self-biotinylation of DNA G-quadruplexes via intrinsic peroxidase activity. Nucleic Acids Res, 45, 9813-9822.

ELKAHLOUN, A. G., KRIZMAN, D. B., WANG, Z., HOFMANN, T. A., ROE, B. & MELTZER, P. S. 1997. Transcript mapping in a 46-kb sequenced region at the core of 12q13.3 amplification in human cancers. Genomics, 42, 295-301.

ESTÈVE, P. O., CHIN, H. G., BENNER, J., FEEHERY, G. R., SAMARANAYAKE, M., HORWITZ, G. A., JACOBSEN, S. E. & PRADHAN, S. 2009. Regulation of DNMT1 stability through SET7-mediated lysine methylation in mammalian cells. Proc Natl Acad Sci U S A, 106, 5076-81.

FAN, Y. & YOU, G. 2020. Proteasome Inhibitors Bortezomib and Carfilzomib Stimulate the Transport Activity of Human Organic Anion Transporter 1. Molecular Pharmacology, 97, 384-391.

FAY, M. M., LYONS, S. M. & IVANOV, P. 2017. RNA G-Quadruplexes in Biology: Principles and Molecular Mechanisms. J Mol Biol, 429, 2127-2147.

FAYE, M. D., GRABER, T. E., LIU, P., THAKOR, N., BAIRD, S. D., DURIE, D. & HOLCIK, M. 2013. Nucleotide composition of cellular internal ribosome entry sites defines dependence on NF45 and predicts a posttranscriptional mitotic regulon. Molecular and cellular biology, 33, 307-318.

FENG, Y., ZHANG, F., LOKEY, L. K., CHASTAIN, J. L., LAKKIS, L., EBERHART, D. & WARREN, S. T. 1995. Translational suppression by trinucleotide repeat expansion at FMR1. Science, 268, 731-4.

FLEMING, A. M., DING, Y. & BURROWS, C. J. 2017. Oxidative DNA damage is epigenetic by regulating gene transcription via base excision repair. Proceedings of the National Academy of Sciences, 114, 2604-2609.

FORMAN, S. L., FETTINGER, J. C., PIERACCINI, S., GOTTARELLI, G. & DAVIS, J. T. 2000. Toward artificial ion channels: A lipophilic G-quadruplex. Journal of the American Chemical Society, 122, 4060-4067.

FORREST, A. R., KAWAJI, H., REHLI, M., BAILLIE, J. K., DE HOON, M. J., HABERLE, V., LASSMANN, T., KULAKOVSKIY, I. V., LIZIO, M., ITOH, M., ANDERSSON, R., MUNGALL, C. J., MEEHAN, T. F., SCHMEIER, S., BERTIN, N., JØRGENSEN, M., DIMONT, E., ARNER, E., SCHMIDL, C., SCHAEFER, U., MEDVEDEVA, Y. A., PLESSY, C., VITEZIC, M., SEVERIN, J., SEMPLE, C., ISHIZU, Y., YOUNG, R. S., FRANCESCATTO, M., ALAM, I., ALBANESE, D., ALTSCHULER, G. M., ARAKAWA, T., ARCHER, J. A., ARNER, P., BABINA, M., RENNIE, S., BALWIERZ, P. J., BECKHOUSE, A. G., PRADHAN-BHATT, S., BLAKE, J. A., BLUMENTHAL, A., BODEGA, B., BONETTI, A., BRIGGS, J., BROMBACHER, F., BURROUGHS, A. M., CALIFANO, A., CANNISTRACI, C. V., CARBAJO, D., CHEN, Y., CHIERICI, M., CIANI, Y., CLEVERS, H. C., DALLA, E., DAVIS, C. A., DETMAR, M., DIEHL, A. D., DOHI, T., DRABLØS, F., EDGE, A. S., EDINGER, M., EKWALL, K., ENDOH, M., ENOMOTO, H., FAGIOLINI, M., FAIRBAIRN, L., FANG, H., FARACH-CARSON, M. C., FAULKNER, G. J., FAVOROV, A. V., FISHER, M. E., FRITH, M. C., FUJITA, R., FUKUDA, S., FURLANELLO, C., FURINO, M., FURUSAWA, J., GEIJTENBEEK, T. B., GIBSON, A. P., GINGERAS, T., GOLDOWITZ, D., GOUGH, J., GUHL, S., GULER, R., GUSTINCICH, S., HA, T. J., HAMAGUCHI, M., HARA, M., HARBERS, M., HARSHBARGER, J., HASEGAWA, A., HASEGAWA, Y., HASHIMOTO, T., HERLYN, M., HITCHENS, K. J., HO SUI, S. J., HOFMANN, O. M., HOOF, I., HORI, F., HUMINIECKI, L., *et al.* 2014. A promoter-level mammalian expression atlas. Nature, 507, 462-70.

FOUQUEREL, E., PARIKH, D. & OPRESKO, P. 2016. DNA damage processing at telomeres: The ends justify the means. DNA Repair, 44, 159-168.

FRASER, D. J., PHILLIPS, A. O., ZHANG, X., VAN ROEYEN, C. R., MUEHLENBERG, P., EN-NIA, A. & MERTENS, P. R. 2008. Y-box protein-1 controls transforming growth factor-&#x3b2;1 translation in proximal tubular cells. Kidney International, 73, 724-732.

FRIEDMAN, R. C., FARH, K. K.-H., BURGE, C. B. & BARTEL, D. P. 2009. Most mammalian mRNAs are conserved targets of microRNAs. Genome research, 19, 92-105.

FRIEDRICH, M., VAXEVANIS, C. K., BIEHL, K., MUELLER, A. & SELIGER, B. 2020. Targeting the coding sequence: opposing roles in regulating classical and non-classical MHC class I molecules by miR-16 and miR-744. Journal for ImmunoTherapy of Cancer, 8, e000396.

FROHMAN, M. A. 1993. Rapid amplification of complementary DNA ends for generation of full-length complementary DNAs: thermal RACE. Methods Enzymol, 218, 340-56.

FROHMAN, M. A., DUSH, M. K. & MARTIN, G. R. 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc Natl Acad Sci U S A, 85, 8998-9002.

GARANT, J.-M., PERREAULT, J.-P. & SCOTT, M. S. 2017. Motif independent identification of potential RNA G-quadruplexes by G4RNA screener. Bioinformatics (Oxford, England), 33, 3532-3537.

GELLERT, M., LIPSETT, M. N. & DAVIES, D. R. 1962. Helix formation by guanylic acid. Proceedings of the National Academy of Sciences of the United States of America, 48, 2013-2018.

GERSHENZON, N. I. & IOSHIKHES, I. P. 2005. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. Bioinformatics, 21, 1295-300.

GHOSH, G. & VAN DUYNE, G. D. 1996. Pieces of the puzzle: assembling the preinitiation complex of Pol II. Structure, 4, 891-895.

GOMEZ, D., GUÉDIN, A., MERGNY, J. L., SALLES, B., RIOU, J. F., TEULADE-FICHOU, M. P. & CALSOU, P. 2010. A G-quadruplex structure within the 5'-UTR of TRF2 mRNA represses translation in human cells. Nucleic Acids Res, 38, 7187-98.

GOOSSENS, S., JANSSENS, B., VANPOUCKE, G., DE RYCKE, R., VAN HENGEL, J. & VAN ROY, F. 2007. Truncated isoform of mouse alphaT-catenin is testis-restricted in expression and function. Faseb j, 21, 647-55.

GOSS, D. J. & THEIL, E. C. 2011. Iron responsive mRNAs: a family of Fe2+ sensitive riboregulators. Accounts of chemical research, 44, 1320-1328.

GOWDA, M., LI, H., ALESSI, J., CHEN, F., PRATT, R. & WANG, G.-L. 2006. Robust analysis of 5'-transcript ends (5'-RATE): a novel technique for transcriptome analysis and genome annotation. Nucleic acids research, 34, e126-e126.

GREENBAUM, D., COLANGELO, C., WILLIAMS, K. & GERSTEIN, M. 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biology, 4, 117.

GU, Z., EILS, R. & SCHLESNER, M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics, 32, 2847-2849.

GUO, J. U. & BARTEL, D. P. 2016. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. Science, 353, aaf5371.

GUO, T., LUNA, A., RAJAPAKSE, V. N., KOH, C. C., WU, Z., LIU, W., SUN, Y., GAO, H., MENDEN, M. P., XU, C., CALZONE, L., MARTIGNETTI, L., AUWERX, C., BULJAN, M., BANAEI-ESFAHANI, A., ORI, A., ISKAR, M., GILLET, L., BI, R., ZHANG, J., ZHANG, H., YU, C., ZHONG, Q., VARMA, S., SCHMITT, U., QIU, P., ZHANG, Q., ZHU, Y., WILD, P. J., GARNETT, M. J., BORK, P., BECK, M., LIU, K., SAEZ-RODRIGUEZ, J., ELLOUMI, F., REINHOLD, W. C., SANDER, C., POMMIER, Y. & AEBERSOLD, R. 2019. Quantitative Proteome Landscape of the NCI-60 Cancer Cell Lines. iScience, 21, 664-680.

GUO, Y., XIAO, P., LEI, S., DENG, F., XIAO, G. G., LIU, Y., CHEN, X., LI, L., WU, S., CHEN, Y., JIANG, H., TAN, L., XIE, J., ZHU, X., LIANG, S. & DENG, H. 2008. How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes. Acta Biochim Biophys Sin (Shanghai), 40, 426-36.

GUPTA, K., SARI-AK, D., HAFFKE, M., TROWITZSCH, S. & BERGER, I. 2016. Zooming in on Transcription Preinitiation. Journal of molecular biology, 428, 2581-2591.

HAIMOV, O., SINVANI, H. & DIKSTEIN, R. 2015. Cap-dependent, scanning-free translation initiation mechanisms. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 1849, 1313-1318.

HALBEISEN, R. E., GALGANO, A., SCHERRER, T. & GERBER, A. P. 2007. Post-transcriptional gene regulation: From genome-wide studies to principles. Cellular and Molecular Life Sciences, 65, 798.

HAMPSEY, M. 1998. Molecular genetics of the RNA polymerase II general transcriptional machinery. Microbiol Mol Biol Rev, 62, 465-503.

HAN, M., GU, Y., LU, P., LI, J., CAO, H., LI, X., QIAN, X., YU, C., YANG, Y., YANG, X., HAN, N., DOU, D., HU, J. & DONG, H. 2020. Exosome-mediated lncRNA AFAP1-AS1 promotes trastuzumab resistance through binding with AUF1 and activating ERBB2 translation. Molecular cancer, 19, 26.

HAN, W. D., YU, L., LOU, F. D., WANG, Q. S., ZHAO, Y., SHI, Z. J. & JIN, H. J. 2001. [The Application of RACE Technique to Clone the Full-Length cDNA of A Novel Leukemia Associated Gene LRP16]. Zhongguo Shi Yan Xue Ye Xue Za Zhi, 9, 18-21.

HÄNSEL-HERTSCH, R., SPIEGEL, J., MARSICO, G., TANNAHILL, D. & BALASUBRAMANIAN, S. 2018. Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. Nature protocols, 13, 551-564.

HARDIN, C. C., WATSON, T., CORREGAN, M. & BAILEY, C. 1992. Cation-dependent transition between the quadruplex and Watson-Crick hairpin forms of d(CGCG3GCG). Biochemistry, 31, 833-41.

HASHIMOTO, S., SUZUKI, Y., KASAI, Y., MOROHOSHI, K., YAMADA, T., SESE, J., MORISHITA, S., SUGANO, S. & MATSUSHIMA, K. 2004. 5'-end SAGE for the analysis of transcriptional start sites. Nat Biotechnol, 22, 1146-9.

HE, K., JANG, S.-W., JOSHI, J., YOO, M.-H. & YE, K. 2011. Akt-phosphorylated PIKE-A inhibits UNC5B-induced apoptosis in cancer cell lines in a p53-dependent manner. Molecular Biology of the Cell, 22, 1943-1954.

HEBERLE, H., MEIRELLES, G. V., DA SILVA, F. R., TELLES, G. P. & MINGHIM, R. 2015. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics, 16, 169.

HEINEMEYER, W., FISCHER, M., KRIMMER, T., STACHON, U. & WOLF, D. H. 1997. The active sites of the eukaryotic 20 S proteasome and their involvement in subunit precursor processing. J Biol Chem, 272, 25200-9.

HENDERSON, A., WU, Y., HUANG, Y. C., CHAVEZ, E. A., PLATT, J., JOHNSON, F. B., BROSH, R. M., JR, SEN, D. & LANSDORP, P. M. 2013. Detection of G-quadruplex DNA in mammalian cells. Nucleic Acids Research, 42, 860-869.

HENGST, L. & REED, S. I. 1996. Translational control of p27Kip1 accumulation during the cell cycle. Science, 271, 1861-4.

HENTZE, M. W., CASTELLO, A., SCHWARZL, T. & PREISS, T. 2018. A brave new world of RNA-binding proteins. Nature Reviews Molecular Cell Biology, 19, 327-341.

HERDY, B., MAYER, C., VARSHNEY, D., MARSICO, G., MURAT, P., TAYLOR, C., D'SANTOS, C., TANNAHILL, D. & BALASUBRAMANIAN, S. 2018. Analysis of NRAS RNA G-quadruplex binding proteins reveals DDX3X as a novel interactor of cellular G-quadruplex containing transcripts. Nucleic Acids Res, 46, 11592-11604.

HERSHEY, J. W., MATHEWS, M. & SONENBERG, N. 2000. Translational control of gene expression, Cold Spring Harbor Laboratory Press.

HERSHKO, A. & CIECHANOVER, A. 1998. The ubiquitin system. Annu Rev Biochem, 67, 425-79.

HIGGINS, S. J. & HAMES, B. D. 1999. Protein expression: a practical approach, Oxford University Press, USA.

HINNEBUSCH, A. G. & LORSCH, J. R. 2012. The mechanism of eukaryotic translation initiation: new insights and challenges. Cold Spring Harbor perspectives in biology, 4, a011544.

HINNEBUSCH, A. G., IVANOV, I. P. & SONENBERG, N. 2016. Translational control by 5'-untranslated regions of eukaryotic mRNAs. Science (New York, N.Y.), 352, 1413-1416.

HOLLERER, I., BARKER, J. C., JORGENSEN, V., TRESENRIDER, A., DUGAST-DARZACQ, C., CHAN, L. Y., DARZACQ, X., TJIAN, R., ÜNAL, E. & BRAR, G. A. 2019. Evidence for an Integrated Gene Repression Mechanism Based on mRNA Isoform Toggling in Human Cells. G3: Genes|Genomes|Genetics, 9, 1045-1053.

HON, J., MARTÍNEK, T., ZENDULKA, J. & LEXA, M. 2017. pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. Bioinformatics, 33, 3373-3379.

HUNTZINGER, E. & IZAURRALDE, E. 2011. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nature Reviews Genetics, 12, 99-110.

HUPPERT, J. L. & BALASUBRAMANIAN, S. 2005. Prevalence of quadruplexes in the human genome. Nucleic Acids Research, 33, 2908-2916.

HUPPERT, J. L., BUGAUT, A., KUMARI, S. & BALASUBRAMANIAN, S. 2008. G-quadruplexes: the beginning and end of UTRs. Nucleic acids research, 36, 6260-6268.

HWANG, A., MCKENNA, W. G. & MUSCHEL, R. J. 1998. Cell cycle-dependent usage of transcriptional start sites. A novel mechanism for regulation of cyclin B1. J Biol Chem, 273, 31505-9.

IAKOVA, P., WANG, G.-L., TIMCHENKO, L., MICHALAK, M., PEREIRA-SMITH, O. M., SMITH, J. R. & TIMCHENKO, N. A. 2004. Competition of CUGBP1 and calreticulin for the regulation of p21 translation determines cell fate. The EMBO Journal, 23, 406-417.

INFANTE, J. J., LAW, G. L. & YOUNG, E. T. 2012. Analysis of Nucleosome Positioning Using a Nucleosome-Scanning Assay. In: MORSE, R. H. (ed.) Chromatin Remodeling: Methods and Protocols. Totowa, NJ: Humana Press.

INGOLIA, N. T., LAREAU, L. F. & WEISSMAN, J. S. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell, 147, 789-802.

IPSARO, J. J. & JOSHUA-TOR, L. 2015. From guide to target: molecular insights into eukaryotic RNA-interference machinery. Nature Structural & Molecular Biology, 22, 20-28.

ISHIGURO, A., KIMURA, N., WATANABE, Y., WATANABE, S. & ISHIHAMA, A. 2016. TDP-43 binds and transports G-quadruplex-containing mRNAs into neurites for local translation. Genes Cells, 21, 466-81.

JACKSON, R. J., HELLEN, C. U. T. & PESTOVA, T. V. 2012. Termination and post-termination events in eukaryotic translation. In: MARINTCHEV, A. (ed.) Advances in Protein Chemistry and Structural Biology. Academic Press.

JANSEN, R.-P. 2001. mRNA localization: message on the move. Nature Reviews Molecular Cell Biology, 2, 247-256.

JAVAHERY, R., KHACHI, A., LO, K., ZENZIE-GREGORY, B. & SMALE, S. T. 1994. DNA sequence requirements for transcriptional initiator activity in mammalian cells. Molecular and cellular biology, 14, 116-127.

JAYARAJ, G. G., PANDEY, S., SCARIA, V. & MAITI, S. 2012. Potential G-quadruplexes in the human long non-coding transcriptome. RNA Biol, 9, 81-6.

JENKINS, R. H., BENNAGI, R., MARTIN, J., PHILLIPS, A. O., REDMAN, J. E. & FRASER, D. J. 2010. A Conserved Stem Loop Motif in the 5'Untranslated Region Regulates Transforming Growth Factor-β1 Translation. PLOS ONE, 5, e12283.

JIANG, C. & PUGH, B. F. 2009. Nucleosome positioning and gene regulation: advances through genomics. Nature reviews. Genetics, 10, 161-172.

JIANG, G., ZHANG, S., YAZDANPARAST, A., LI, M., PAWAR, A. V., LIU, Y., INAVOLU, S. M. & CHENG, L. 2016. Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. BMC Genomics, 17, 525.

JIANG, Y., LIU, M., SPENCER, C. A. & PRICE, D. H. 2004. Involvement of Transcription Termination Factor 2 in Mitotic Repression of Transcription Elongation. Molecular Cell, 14, 375-386.

JIMENO, S., CAMARILLO, R., MEJÍAS-NAVARRO, F., FERNÁNDEZ-ÁVILA, M. J., SORIA-BRETONES, I., PRADOS-CARVAJAL, R. & HUERTAS, P. 2018. The Helicase PIF1 Facilitates Resection over Sequences Prone to Forming G4 Structures. Cell Reports, 24, 3262-3273.e4.

JOACHIMI, A., BENZ, A. & HARTIG, J. S. 2009. A comparison of DNA and RNA quadruplex structures and stabilities. Bioorg Med Chem, 17, 6811-5.

JOHNSON, J. E., CAO, K., RYVKIN, P., WANG, L.-S. & JOHNSON, F. B. 2010. Altered gene expression in the Werner and Bloom syndromes is associated with sequences having G-quadruplex forming potential. Nucleic acids research, 38, 1114-1122.

JUNG, H.-Y., FATTET, L., TSAI, J. H., KAJIMOTO, T., CHANG, Q., NEWTON, A. C. & YANG, J. 2019. Apical–basal polarity inhibits epithelial–mesenchymal transition and tumour metastasis by PAR-complex-mediated SNAI1 degradation. Nature Cell Biology, 21, 359-371.

JUVEN-GERSHON, T. & KADONAGA, J. T. 2010. Regulation of gene expression via the core promoter and the basal transcriptional machinery. Dev Biol, 339, 225-9.

KADONAGA, J. T. 2012. Perspectives on the RNA polymerase II core promoter. Wiley Interdiscip Rev Dev Biol, 1, 40-51.

KAGOHARA, L. T., STEIN-O'BRIEN, G. L., KELLEY, D., FLAM, E., WICK, H. C., DANILOVA, L. V., EASWARAN, H., FAVOROV, A. V., QIAN, J., GAYKALOVA, D. A. & FERTIG, E. J. 2017. Epigenetic regulation of gene expression in cancer: techniques, resources and analysis. Briefings in Functional Genomics, 17, 49-63.

KAHN, R. A., BRUFORD, E., INOUE, H., LOGSDON, J. M., JR., NIE, Z., PREMONT, R. T., RANDAZZO, P. A., SATAKE, M., THEIBERT, A. B., ZAPP, M. L. & CASSEL, D. 2008. Consensus nomenclature for the human ArfGAP domain-containing proteins. Journal of Cell Biology, 182, 1039-1044.

KAMURA, T., KATSUDA, Y., KITAMURA, Y. & IHARA, T. 2020. G-quadruplexes in mRNA: A key structure for biological function. Biochemical and Biophysical Research Communications, 526, 261-266.

KANAMORI-KATAYAMA, M., ITOH, M., KAWAJI, H., LASSMANN, T., KATAYAMA, S., KOJIMA, M., BERTIN, N., KAIHO, A., NINOMIYA, N., DAUB, C. O., CARNINCI, P., FORREST, A. R. & HAYASHIZAKI, Y. 2011. Unamplified cap analysis of gene expression on a single-molecule sequencer. Genome Res, 21, 1150-9.

KARLSSON, K., LÖNNERBERG, P. & LINNARSSON, S. 2017. Alternative TSSs are co-regulated in single cells in the mouse brain. 13, 930.

KAUFMANN, J. & SMALE, S. T. 1994. Direct recognition of initiator elements by a component of the transcription factor IID complex. Genes Dev, 8, 821-9.

KAWAJI, H., FRITH, M. C., KATAYAMA, S., SANDELIN, A., KAI, C., KAWAI, J., CARNINCI, P. & HAYASHIZAKI, Y. 2006. Dynamic usage of transcription start sites within core promoters. Genome Biol, 7, R118.

KAWAJI, H., LIZIO, M., ITOH, M., KANAMORI-KATAYAMA, M., KAIHO, A., NISHIYORI-SUEKI, H., SHIN, J. W., KOJIMA-ISHIYAMA, M., KAWANO, M., MURATA, M., NINOMIYA-FUKUDA, N., ISHIKAWA-KATO, S., NAGAO-SATO, S., NOMA, S., HAYASHIZAKI, Y., FORREST, A. R. R., CARNINCI, P. & CONSORTIUM, F. 2014. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. Genome research, 24, 708-717.

KAWAMATA, T. & TOMARI, Y. 2010. Making RISC. Trends in Biochemical Sciences, 35, 368-376.

KEENE, J. D. & TENENBAUM, S. A. 2002. Eukaryotic mRNPs may represent posttranscriptional operons. Mol Cell, 9, 1161-7.

KEJNOVSKÁ, I., RENČIUK, D., PALACKÝ, J. & VORLÍČKOVÁ, M. 2019. CD Study of the G-Quadruplex Conformation. Methods Mol Biol, 2035, 25-44.

KHAREL, P., BECKER, G., TSVETKOV, V. & IVANOV, P. 2020. Properties and biological impact of RNA G-quadruplexes: from order to turmoil and back. Nucleic Acids Res, 48, 12534-12555.

KIKIN, O., D'ANTONIO, L. & BAGGA, P. S. 2006. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic acids research, 34, W676-W682.

KIM, W. & KYUNG LEE, E. 2012. Post-transcriptional regulation in metabolic diseases. RNA biology, 9, 772-780.

KIM, T. K. & EBERWINE, J. H. 2010. Mammalian cell transfection: the present and the future. Analytical and bioanalytical chemistry, 397, 3173-3178.

KIMURA, K., WAKAMATSU, A., SUZUKI, Y., OTA, T., NISHIKAWA, T., YAMASHITA, R., YAMAMOTO, J.-I., SEKINE, M., TSURITANI, K., WAKAGURI, H., ISHII, S., SUGIYAMA, T., SAITO, K., ISONO, Y., IRIE, R., KUSHIDA, N., YONEYAMA, T., OTSUKA, R., KANDA, K., YOKOI, T., KONDO, H., WAGATSUMA, M., MURAKAWA, K., ISHIDA, S., ISHIBASHI, T., TAKAHASHI-FUJII, A., TANASE, T., NAGAI, K., KIKUCHI, H., NAKAI, K., ISOGAI, T. & SUGANO, S. 2006. Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. Genome research, 16, 55-65.

KISSELEV, A. F., CALLARD, A. & GOLDBERG, A. L. 2006. Importance of the different proteolytic sites of the proteasome and the efficacy of inhibitors varies with the protein substrate. J Biol Chem, 281, 8582-90.

KISSELEV, A. F., GARCIA-CALVO, M., OVERKLEEFT, H. S., PETERSON, E., PENNINGTON, M. W., PLOEGH, H. L., THORNBERRY, N. A. & GOLDBERG, A. L. 2003. The caspase-like sites of proteasomes, their substrate specificity, new inhibitors and substrates, and allosteric interactions with the trypsin-like sites. J Biol Chem, 278, 35869-77.

KODZIUS, R., KOJIMA, M., NISHIYORI, H., NAKAMURA, M., FUKUDA, S., TAGAMI, M., SASAKI, D., IMAMURA, K., KAI, C., HARBERS, M., HAYASHIZAKI, Y. & CARNINCI, P. 2006. CAGE: cap analysis of gene expression. Nat Methods, 3, 211-22.

KOMAR, A. A., MAZUMDER, B. & MERRICK, W. C. 2012. A new framework for understanding IRES-mediated translation. Gene, 502, 75-86.

KONTOYIANNIS, D., PASPARAKIS, M., PIZARRO, T. T., COMINELLI, F. & KOLLIAS, G. 1999. Impaired On/Off Regulation of TNF Biosynthesis in Mice Lacking TNF AU-Rich Elements: Implications for Joint and Gut-Associated Immunopathologies. Immunity, 10, 387-398.

KOROMILAS, A. E., LAZARIS-KARATZAS, A. & SONENBERG, N. 1992. mRNAs containing extensive secondary structure in their 5' non-coding region translate efficiently in cells overexpressing initiation factor eIF-4E. Embo j, 11, 4153-8.

KOZAK, M. 1986. Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. Proc Natl Acad Sci U S A, 83, 2850-4.

KOZAK, M. 1989. Circumstances and mechanisms of inhibition of translation by secondary structure in eucaryotic mRNAs. Mol Cell Biol, 9, 5134-42.

KOZAK, M. 2001. New ways of initiating translation in eukaryotes? Mol Cell Biol, 21, 1899-907.

KRIVEGA, I. & DEAN, A. 2012. Enhancer and promoter interactions-long distance calls. Curr Opin Genet Dev, 22, 79-85.

KRUSE, K. B., BRODSKY, J. L. & MCCRACKEN, A. A. 2006. Autophagy: an ER protein quality control process. Autophagy, 2, 135-7.

KUERSTEN, S. & GOODWIN, E. B. 2003. The power of the 3' UTR: translational control and development. Nat Rev Genet, 4, 626-37.

KUERSTEN, S., RADEK, A., VOGEL, C. & PENALVA, L. O. 2013. Translation regulation gets its 'omics' moment. Wiley Interdiscip Rev RNA, 4, 617-30.

KUGEL, J. F. & GOODRICH, J. A. 2017. Finding the start site: redefining the human initiator element. Genes & development, 31, 1-2.

KUMARI, S., BUGAUT, A. & BALASUBRAMANIAN, S. 2008. Position and stability are determining factors for translation repression by an RNA G-quadruplex-forming sequence within the 5' UTR of the NRAS proto-oncogene. Biochemistry, 47, 12664-12669.

KUMARI, S., BUGAUT, A., HUPPERT, J. L. & BALASUBRAMANIAN, S. 2007. An RNA G-quadruplex in the 5′ UTR of the NRAS proto-oncogene modulates translation. Nature Chemical Biology, 3, 218-221.

KWOK, C. K. & BALASUBRAMANIAN, S. 2015. Targeted Detection of G-Quadruplexes in Cellular RNAs. Angew Chem Int Ed Engl, 54, 6751-4.

KWOK, C. K., MARSICO, G., SAHAKYAN, A. B., CHAMBERS, V. S. & BALASUBRAMANIAN, S. 2016a. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. Nature Methods, 13, 841-844.

KWOK, C. K., SAHAKYAN, A. B. & BALASUBRAMANIAN, S. 2016b. Structural Analysis using SHALiPE to Reveal RNA G-Quadruplex Formation in Human Precursor MicroRNA. Angew Chem Int Ed Engl, 55, 8958-61.

LABUDOVÁ, D., HON, J. & LEXA, M. 2019. pqsfinder web: G-quadruplex prediction using optimized pqsfinder algorithm. Bioinformatics, 36, 2584-2586.

LACERDA, R., MENEZES, J. & ROMÃO, L. 2017. More than just scanning: the importance of cap-independent mRNA translation initiation for cellular stress response and cancer. Cell Mol Life Sci, 74, 1659-1680.

LACKNER, D. H. & BÄHLER, J. 2008. Chapter 5 Translational Control of Gene Expression: From Transcripts to Transcriptomes. International Review of Cell and Molecular Biology. Academic Press.

LAGUERRE, A., HUKEZALIE, K., WINCKLER, P., KATRANJI, F., CHANTELOUP, G., PIRROTTA, M., PERRIER-CORNET, J. M., WONG, J. M. & MONCHAUD, D. 2015. Visualization of RNA-Quadruplexes in Live Cells. J Am Chem Soc, 137, 8521-5.

LAMMICH, S., KAMP, F., WAGNER, J., NUSCHER, B., ZILOW, S., LUDWIG, A. K., WILLEM, M. & HAASS, C. 2011. Translational repression of the disintegrin and metalloprotease ADAM10 by a stable G-quadruplex secondary structure in its 5'-untranslated region. J Biol Chem, 286, 45063-72.

LANDRY, J. R., MAGER, D. L. & WILHELM, B. T. 2003. Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet, 19, 640-8.

LAT, P. K., LIU, K., KUMAR, D. N., WONG, K. K. L., VERHEYEN, E. M. & SEN, D. 2020. High specificity and tight spatial restriction of self-biotinylation by DNA and RNA G-Quadruplexes complexed in vitro and in vivo with Heme. Nucleic acids research, 48, 5254-5267.

LECKER, S. H., GOLDBERG, A. L. & MITCH, W. E. 2006. Protein Degradation by the Ubiquitin–Proteasome Pathway in Normal and Disease States. Journal of the American Society of Nephrology, 17, 1807-1819.

LEE, D. S. M., GHANEM, L. R. & BARASH, Y. 2020. Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. Nature Communications, 11, 527.

LEE, TONG I. & YOUNG, RICHARD A. 2013. Transcriptional Regulation and Its Misregulation in Disease. Cell, 152, 1237-1251.

LEENEN, F. A., VERNOCCHI, S., HUNEWALD, O. E., SCHMITZ, S., MOLITOR, A. M., MULLER, C. P. & TURNER, J. D. 2016. Where does transcription start? 5'-RACE adapted to next-generation sequencing. Nucleic Acids Res, 44, 2628-45.

LENHARD, B., SANDELIN, A. & CARNINCI, P. 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet, 13, 233-45.

LEPPEK, K., DAS, R. & BARNA, M. 2018. Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. Nature reviews. Molecular cell biology, 19, 158-174.

LEXA, M., STEFLOVA, P., MARTINEK, T., VORLICKOVA, M., VYSKOT, B. & KEJNOVSKY, E. 2014. Guanine quadruplexes are formed by specific regions of human transposable elements. BMC Genomics, 15, 1032.

LI, B., CAREY, M. & WORKMAN, J. L. 2007. The Role of Chromatin during Transcription. Cell, 128, 707-719.

LI, H., BAI, L., LI, H., LI, X., KANG, Y., ZHANG, N., SUN, J. & SHAO, Z. 2019. Selective translational usage of TSS and core promoters revealed by translatome sequencing. BMC Genomics, 20, 282.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2078-9.

LIANOGLOU, S., GARG, V., YANG, J. L., LESLIE, C. S. & MAYR, C. 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. Genes Dev, 27, 2380-96.

LIAO, B., HU, Y. & BREWER, G. 2007. Competitive binding of AUF1 and TIAR to MYC mRNA controls its translation. Nat Struct Mol Biol, 14, 511-8.

LINDSTEIN, T., JUNE, C., LEDBETTER, J., STELLA, G. & THOMPSON, C. 1989. Regulation of lymphokine messenger RNA stability by a surface-mediated T cell activation pathway. Science, 244, 339-343.

LIU, R., TIAN, B., GEARING, M., HUNTER, S., YE, K. & MAO, Z. 2008. Cdk5-mediated regulation of the PIKE-A-Akt pathway and glioblastoma cell invasion. Proc Natl Acad Sci U S A, 105, 7570-5.

LIU, X., HU, Y., HAO, C., REMPEL, S. A. & YE, K. 2007. PIKE-A is a proto-oncogene promoting cell growth, transformation and invasion. Oncogene, 26, 4918-27.

LIVAK, K. J. & SCHMITTGEN, T. D. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods, 25, 402-8.

LIZIO, M., HARSHBARGER, J., SHIMOJI, H., SEVERIN, J., KASUKAWA, T., SAHIN, S., ABUGESSAISA, I., FUKUDA, S., HORI, F., ISHIKAWA-KATO, S., MUNGALL, C. J., ARNER, E., BAILLIE, J. K., BERTIN, N., BONO, H., DE HOON, M., DIEHL, A. D., DIMONT, E., FREEMAN, T. C., FUJIEDA, K., HIDE, W., KALIYAPERUMAL, R., KATAYAMA, T., LASSMANN, T., MEEHAN, T. F., NISHIKATA, K., ONO, H., REHLI, M., SANDELIN, A., SCHULTES, E. A., 'T HOEN, P. A. C., TATUM, Z., THOMPSON, M., TOYODA, T., WRIGHT, D. W., DAUB, C. O., ITOH, M., CARNINCI, P., HAYASHIZAKI, Y., FORREST, A. R. R., KAWAJI, H. & THE, F. C. 2015. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biology, 16, 22.

LOPES, J., PIAZZA, A., BERMEJO, R., KRIEGSMAN, B., COLOSIO, A., TEULADE-FICHOU, M.-P., FOIANI, M. & NICOLAS, A. 2011. G-quadruplex-induced instability during leading-strand replication. The EMBO Journal, 30, 4033-4046.

LORENZ, R., BERNHART, S. H., QIN, J., HÖNER ZU SIEDERDISSEN, C., TANZER, A., AMMAN, F., HOFACKER, I. L. & STADLER, P. F. 2013. 2D meets 4G: G-quadruplexes in RNA secondary structure prediction. IEEE/ACM Trans Comput Biol Bioinform, 10, 832-44.

LOTHROP, A. P., TORRES, M. P. & FUCHS, S. M. 2013. Deciphering post-translational modification codes. FEBS letters, 587, 1247-1257.

LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15, 550.

LUO, R., BAI, C., YANG, L., ZHENG, Z., SU, G., GAO, G., WEI, Z., ZUO, Y. & LI, G. 2018. DNA methylation subpatterns at distinct regulatory regions in human early embryos. Open Biology, 8, 180131.

MAIER, T., GÜELL, M. & SERRANO, L. 2009. Correlation of mRNA and protein in complex biological samples. FEBS Letters, 583, 3966-3973.

MALIK, S. & ROEDER, R. G. 2010. The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. Nat Rev Genet, 11, 761-72.

MALTBY, C. J., SCHOFIELD, J. P. R., HOUGHTON, S. D., O'KELLY, I., VARGAS-CABALLERO, M., DEINHARDT, K. & COLDWELL, M. J. 2020. A 5' UTR GGN repeat controls localisation and translation of a potassium leak channel mRNA through G-quadruplex formation. Nucleic Acids Research, 48, 9822-9839.

MAO, S.-Q., GHANBARIAN, A. T., SPIEGEL, J., MARTÍNEZ CUESTA, S., BERALDI, D., DI ANTONIO, M., MARSICO, G., HÄNSEL-HERTSCH, R., TANNAHILL, D. & BALASUBRAMANIAN, S. 2018. DNA G-quadruplex structures mold the DNA methylome. Nature Structural & Molecular Biology, 25, 951-957.

MARCEL, V., TRAN, P. L., SAGNE, C., MARTEL-PLANCHE, G., VASLIN, L., TEULADE-FICHOU, M. P., HALL, J., MERGNY, J. L., HAINAUT, P. & VAN DYCK, E. 2011. G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms. Carcinogenesis, 32, 271-8.

MARTINEAU, Y., LE BEC, C., MONBRUN, L., ALLO, V., CHIU, I. M., DANOS, O., MOINE, H., PRATS, H. & PRATS, A. C. 2004. Internal ribosome entry site structural motifs conserved among mammalian fibroblast growth factor 1 alternatively spliced mRNAs. Molecular and Cellular Biology, 24, 7622-7635.

MARUYAMA, K. & SUGANO, S. 1994. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene, 138, 171-4.

MASUZAWA, T. & OYOSHI, T. 2020. Roles of the RGG Domain and RNA Recognition Motif of Nucleolin in G-Quadruplex Stabilization. ACS Omega, 5, 5202-5208.

MATERA, A. G. & WANG, Z. 2014. A day in the life of the spliceosome. Nature reviews. Molecular cell biology, 15, 108-121.

MATHAD, R. I. & YANG, D. 2011. G-quadruplex structures and G-quadruplex-interactive compounds. Methods Mol Biol, 735, 77-96.

MATSUI, M., YACHIE, N., OKADA, Y., SAITO, R. & TOMITA, M. 2007. Bioinformatic analysis of post-transcriptional regulation by uORF in human and mouse. FEBS Letters, 581, 4184-4188.

MAYR, C. 2016. Evolution and Biological Roles of Alternative 3'UTRs. Trends in cell biology, 26, 227-237.

MELLO DE QUEIROZ, F., SÁNCHEZ, A., AGARWAL, J. R., STÜHMER, W. & PARDO, L. A. 2012. Nucleofection induces non-specific changes in the metabolic activity of transfected cells. Molecular biology reports, 39, 2187-2194.

MERGNY, J. L., LACROIX, L., TEULADE-FICHOU, M. P., HOUNSOU, C., GUITTAT, L., HOARAU, M., ARIMONDO, P. B., VIGNERON, J. P., LEHN, J. M., RIOU, J. F., GARESTIER, T. & HÉLÈNE, C. 2001. Telomerase inhibitors based on quadruplex ligands selected by a fluorescence assay. Proceedings of the National Academy of Sciences of the United States of America, 98, 3062-3067.

MERKHOFER, E. C., HU, P. & JOHNSON, T. L. 2014. Introduction to cotranscriptional RNA splicing. Methods in molecular biology (Clifton, N.J.), 1126, 83-96.

MESTRE-FOS, S., PENEV, P. I., SUTTAPITUGSAKUL, S., HU, M., ITO, C., PETROV, A. S., WARTELL, R. M., WU, R. & WILLIAMS, L. D. 2019. G-Quadruplexes in Human Ribosomal RNA. J Mol Biol, 431, 1940-1955.

MIGLANI, G. S. A. 2014. Gene expression, Oxford, U. K., Alpha Science International Ltd.

MINGUEZ, P., LETUNIC, I., PARCA, L., GARCIA-ALONSO, L., DOPAZO, J., HUERTA-CEPAS, J. & BORK, P. 2014. PTMcode v2: a resource for functional associations of post-translational modifications within and between proteins. Nucleic Acids Research, 43, D494-D502.

MINGUEZ, P., PARCA, L., DIELLA, F., MENDE, D. R., KUMAR, R., HELMER-CITTERICH, M., GAVIN, A. C., VAN NOORT, V. & BORK, P. 2012. Deciphering a global network of functionally associated post-translational modifications. Mol Syst Biol, 8, 599.

MITCHELL, S. A., SPRIGGS, K. A., COLDWELL, M. J., JACKSON, R. J. & WILLIS, A. E. 2003. The Apaf-1 internal ribosome entry segment attains the correct structural conformation for function via interactions with PTB and unr. Mol Cell, 11, 757-71.

MOORE, L. D., LE, T. & FAN, G. 2013. DNA methylation and its basic function. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology, 38, 23-38.

MOORE, M. J. B., SCHULTES, C. M., CUESTA, J., CUENCA, F., GUNARATNAM, M., TANIOUS, F. A., WILSON, W. D. & NEIDLE, S. 2006. Trisubstituted Acridines as G-quadruplex Telomere Targeting Agents. Effects of Extensions of the 3,6- and 9-Side Chains on Quadruplex Binding, Telomerase Activity, and Cell Proliferation. Journal of Medicinal Chemistry, 49, 582-599.

MORRIS, D. R. & GEBALLE, A. P. 2000. Upstream open reading frames as regulators of mRNA translation. Mol Cell Biol, 20, 8635-42.

MORRIS, M. J. & BASU, S. 2009. An unusually stable G-quadruplex within the 5'-UTR of the MT3 matrix metalloproteinase mRNA represses translation in eukaryotic cells. Biochemistry, 48, 5313-9.

MORRIS, M. J., NEGISHI, Y., PAZSINT, C., SCHONHOFT, J. D. & BASU, S. 2010. An RNA G-quadruplex is essential for cap-independent translation initiation in human VEGF IRES. J Am Chem Soc, 132, 17831-9.

MOURIAUX, F., ZANIOLO, K., BERGERON, M. A., WEIDMANN, C., DE LA FOUCHARDIÈRE, A., FOURNIER, F., DROIT, A., MORCOS, M. W., LANDREVILLE, S. & GUÉRIN, S. L. 2016. Effects of Long-term Serial Passaging on the Characteristics and Properties of Cell Lines Derived From Uveal Melanoma Primary Tumors. Invest Ophthalmol Vis Sci, 57, 5288-5301.

MURAT, P., MARSICO, G., HERDY, B., GHANBARIAN, A., PORTELLA, G. & BALASUBRAMANIAN, S. 2018. RNA G-quadruplexes at upstream open reading frames cause DHX36- and DHX9-dependent translation of human mRNAs. Genome Biology, 19, 229.

MURAT, P., ZHONG, J., LEKIEFFRE, L., COWIESON, N. P., CLANCY, J. L., PREISS, T., BALASUBRAMANIAN, S., KHANNA, R. & TELLAM, J. 2014. G-quadruplexes regulate Epstein-Barr virus-encoded nuclear antigen 1 mRNA translation. Nat Chem Biol, 10, 358-64.

NAGASE, T., SEKI, N., ISHIKAWA, K., TANAKA, A. & NOMURA, N. 1996. Prediction of the coding sequences of unidentified human genes. V. The coding sequences of 40 new genes (KIAA0161-KIAA0200) deduced by analysis of cDNA clones from human cell line KG-1 (supplement). DNA Res, 3, 43-53.

NAKAYAMA, J., RICE, J. C., STRAHL, B. D., ALLIS, C. D. & GREWAL, S. I. 2001. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. Science, 292, 110-3.

NAVARRO-CORCUERA, A., ANSORENA, E., MONTIEL-DUARTE, C. & IRABURU, M. J. 2020. AGAP2: Modulating TGFβ1-Signaling in the Regulation of Liver Fibrosis. Int J Mol Sci, 21.

NAVARRO-CORCUERA, A., LÓPEZ-ZABALZA, M. J., MARTÍNEZ-IRUJO, J. J., ÁLVAREZ-SOLA, G., ÁVILA, M. A., IRABURU, M. J., ANSORENA, E. & MONTIEL-DUARTE, C. 2019. Role of AGAP2 in the profibrogenic effects induced by TGFβ in LX-2 hepatic stellate cells. Biochim Biophys Acta Mol Cell Res, 1866, 673-685.

NEININGER, K., MARSCHALL, T. & HELMS, V. 2019. SNP and indel frequencies at transcription start sites and at canonical and alternative translation initiation sites in the human genome. PLOS ONE, 14, e0214816.

NGUYEN, G. H., TANG, W., ROBLES, A. I., BEYER, R. P., GRAY, L. T., WELSH, J. A., SCHETTER, A. J., KUMAMOTO, K., WANG, X. W., HICKSON, I. D., MAIZELS, N., MONNAT, R. J. & HARRIS, C. C. 2014. Regulation of gene expression by the BLM helicase correlates with the presence of G-quadruplex DNA motifs. Proceedings of the National Academy of Sciences, 111, 9905-9910.

NIE, Z., FEI, J., PREMONT, R. T. & RANDAZZO, P. A. 2005. The Arf GAPs AGAP1 and AGAP2 distinguish between the adaptor protein complexes AP-1 and AP-3. J Cell Sci, 118, 3555-66.

NOELS, H., SOMERS, R., LIU, H., YE, H., DU, M. Q., DE WOLF-PEETERS, C., MARYNEN, P. & BAENS, M. 2009. Auto-ubiquitination-induced degradation of MALT1-API2 prevents BCL10 destabilization in t(11;18)(q21;q21)-positive MALT lymphoma. PLoS One, 4, e4822.

NUSSBAUM, A. K., DICK, T. P., KEILHOLZ, W., SCHIRLE, M., STEVANOVIĆ, S., DIETZ, K., HEINEMEYER, W., GROLL, M., WOLF, D. H., HUBER, R., RAMMENSEE, H. G. & SCHILD, H. 1998. Cleavage motifs of the yeast 20S proteasome beta subunits deduced from digests of enolase 1. Proc Natl Acad Sci U S A, 95, 12504-9.

O'BRIEN, J., HAYDER, H., ZAYED, Y. & PENG, C. 2018. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Front Endocrinol (Lausanne), 9, 402.

OHMIYA, H., VITEZIC, M., FRITH, M. C., ITOH, M., CARNINCI, P., FORREST, A. R. R., HAYASHIZAKI, Y., LASSMANN, T. & AND THE, F. C. 2014. RECLU: a pipeline to discover reproducible transcriptional start sites and their alternative regulation using capped analysis of gene expression (CAGE). BMC Genomics, 15, 269.

OKUR, V., CHO, M. T., VAN WIJK, R., VAN OIRSCHOT, B., PICKER, J., COURY, S. A., GRANGE, D., MANWARING, L., KRANTZ, I., MURARESKU, C. C., HULICK, P. J., MAY, H., PIERCE, E., PLACE, E., BUJAKOWSKA, K., TELEGRAFI, A., DOUGLAS, G., MONAGHAN, K. G., BEGTRUP, A., WILSON, A., RETTERER, K., ANYANE-YEBOA, K. & CHUNG, W. K. 2019. De novo variants in HK1 associated with neurodevelopmental abnormalities and visual impairment. European journal of human genetics : EJHG, 27, 1081-1089.

O'LEARY, N. A., WRIGHT, M. W., BRISTER, J. R., CIUFO, S., HADDAD, D., MCVEIGH, R., RAJPUT, B., ROBBERTSE, B., SMITH-WHITE, B., AKO-ADJEI, D., ASTASHYN, A., BADRETDIN, A., BAO, Y., BLINKOVA, O., BROVER, V., CHETVERNIN, V., CHOI, J., COX, E., ERMOLAEVA, O., FARRELL, C. M., GOLDFARB, T., GUPTA, T., HAFT, D., HATCHER, E., HLAVINA, W., JOARDAR, V. S., KODALI, V. K., LI, W., MAGLOTT, D., MASTERSON, P., MCGARVEY, K. M., MURPHY, M. R., O'NEILL, K., PUJAR, S., RANGWALA, S. H., RAUSCH, D., RIDDICK, L. D., SCHOCH, C., SHKEDA, A., STORZ, S. S., SUN, H., THIBAUD-NISSEN, F., TOLSTOY, I., TULLY, R. E., VATSAN, A. R., WALLIN, C., WEBB, D., WU, W., LANDRUM, M. J., KIMCHI, A., TATUSOVA, T., DICUCCIO, M., KITTS, P., MURPHY, T. D. & PRUITT, K. D. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res, 44, D733-45.

OYAMA, M., ITAGAKI, C., HATA, H., SUZUKI, Y., IZUMI, T., NATSUME, T., ISOBE, T. & SUGANO, S. 2004. Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. Genome Res, 14, 2048-52.

PALAVECINO, C. E., CARRASCO-VÉLIZ, N., QUEST, A. F. G., GARRIDO, M. P. & VALENZUELA-VALDERRAMA, M. 2020. The 5’ untranslated region of the anti-apoptotic protein Survivin contains an inhibitory upstream AUG codon. Biochemical and Biophysical Research Communications, 526, 898-905.

PARDEE, T. S., BANGUR, C. S. & PONTICELLI, A. S. 1998. The N-terminal region of yeast TFIIB contains two adjacent functional domains involved in stable RNA polymerase II binding and transcription start site selection. J Biol Chem, 273, 17859-64.

PARKES, G. M. & NIRANJAN, M. 2019. Uncovering extensive post-translation regulation during human cell cycle progression by integrative multi-’omics analysis. BMC Bioinformatics, 20, 536.

PELLETIER, J. & SONENBERG, N. 1985. Insertion mutagenesis to increase secondary structure within the 5' noncoding region of a eukaryotic mRNA reduces translational efficiency. Cell, 40, 515-26.

PICKERING, B. M. & WILLIS, A. E. 2005. The implications of structured 5' untranslated regions on translation and disease. Semin Cell Dev Biol, 16, 39-47.

PICKERING, B. M., MITCHELL, S. A., SPRIGGS, K. A., STONELEY, M. & WILLIS, A. E. 2004. Bag-1 internal ribosome entry segment activity is promoted by structural changes mediated by poly(rC) binding protein 1 and recruitment of polypyrimidine tract binding protein 1. Mol Cell Biol, 24, 5595-605.

PISAREVA, V. P., PISAREV, A. V., KOMAR, A. A., HELLEN, C. U. & PESTOVA, T. V. 2008. Translation initiation on mammalian mRNAs with structured 5'UTRs requires DExH-box protein DHX29. Cell, 135, 1237-50.

PITCHER, D. S., DE MATTOS-SHIPLEY, K., TZORTZIS, K., AUNER, H. W., KARADIMITRIS, A. & KLEIJNEN, M. F. 2015. Bortezomib Amplifies Effect on Intracellular Proteasomes by Changing Proteasome Structure. EBioMedicine, 2, 642-648.

PLOTKIN, J. B. & KUDLA, G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet, 12, 32-42.

PLOTKIN, J. B. 2010. Transcriptional regulation is only half the story. Molecular systems biology, 6, 406-406.

POULIN, F., GINGRAS, A. C., OLSEN, H., CHEVALIER, S. & SONENBERG, N. 1998. 4E-BP3, a new member of the eukaryotic initiation factor 4E-binding protein family. J Biol Chem, 273, 14002-7.

POZNER, A., GOLDENBERG, D., NEGREANU, V., LE, S. Y., ELROY-STEIN, O., LEVANON, D. & GRONER, Y. 2000. Transcription-coupled translation control of AML1/RUNX1 is mediated by cap- and internal ribosome entry site-dependent mechanisms. Mol Cell Biol, 20, 2297-307.

POZNER, A., LOTEM, J., XIAO, C., GOLDENBERG, D., BRENNER, O., NEGREANU, V., LEVANON, D. & GRONER, Y. 2007. Developmentally regulated promoter-switch transcriptionally controls Runx1 function during embryonic hematopoiesis. BMC Dev Biol, 7, 84.

PRIMERANO, B., TASSONE, F., HAGERMAN, R. J., HAGERMAN, P., AMALDI, F. & BAGNI, C. 2002. Reduced FMR1 mRNA translation efficiency in fragile X patients with premutations. Rna, 8, 1482-8.

PRINGLE, E. S., MCCORMICK, C. & CHENG, Z. 2019. Polysome Profiling Analysis of mRNA and Associated Proteins Engaged in Translation. Current Protocols in Molecular Biology, 125, e79.

PROUD, C. G. 2005. eIF2 and the control of cell physiology. Semin Cell Dev Biol, 16, 3-12.

PUIG LOMBARDI, E. & LONDOÑO-VALLEJO, A. 2020. A guide to computational methods for G-quadruplex prediction. Nucleic acids research, 48, 1-15.

QI, Q., KANG, S. S., ZHANG, S., PHAM, C., FU, H., BRAT, D. J. & YE, K. 2017. Co-amplification of phosphoinositide 3-kinase enhancer A and cyclin-dependent kinase 4 triggers glioblastoma progression. Oncogene, 36, 4562-4572.

QIN, Z., STOILOV, P., ZHANG, X. & XING, Y. 2018. SEASTAR: systematic evaluation of alternative transcription start sites in RNA. Nucleic acids research, 46, e45-e45.

QIU, C., JIN, H., VVEDENSKAYA, I., LLENAS, J. A., ZHAO, T., MALIK, I., VISBISKY, A. M., SCHWARTZ, S. L., CUI, P., ČABART, P., HAN, K. H., LAI, W. K. M., METZ, R. P., JOHNSON, C. D., SZE, S.-H., PUGH, B. F., NICKELS, B. E. & KAPLAN, C. D. 2020. Universal promoter scanning by Pol II during transcription initiation in Saccharomyces cerevisiae. Genome Biology, 21, 132.

QUELLE, D. E., ZINDY, F., ASHMUN, R. A. & SHERR, C. J. 1995. Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. Cell, 83, 993-1000.

QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26, 841-2.

RACHWAL, P. A. & FOX, K. R. 2007. Quadruplex melting. Methods, 43, 291-301.

RADMAN-LIVAJA, M. & RANDO, O. J. 2010. Nucleosome positioning: how is it established, and why does it matter? Developmental biology, 339, 258-266.

RAJAPAKSE, V. N., LUNA, A., YAMADE, M., LOMAN, L., VARMA, S., SUNSHINE, M., IORIO, F., SOUSA, F. G., ELLOUMI, F., ALADJEM, M. I., THOMAS, A., SANDER, C., KOHN, K. W., BENES, C. H., GARNETT, M., REINHOLD, W. C. & POMMIER, Y. 2018. CellMinerCDB for Integrative Cross-Database Genomics and Pharmacogenomics Analyses of Cancer Cell Lines. iScience, 10, 247-264.

RANISH, J. A., YUDKOVSKY, N. & HAHN, S. 1999. Intermediates in formation and activity of the RNA polymerase II preinitiation complex: holoenzyme recruitment and a postrecruitment role for the TATA box and TFIIB. Genes & development, 13, 49-63.

REINHOLD, W. C., SUNSHINE, M., VARMA, S., DOROSHOW, J. H. & POMMIER, Y. 2015. Using CellMiner 1.6 for Systems Pharmacology and Genomic Analysis of the NCI-60. Clinical cancer research : an official journal of the American Association for Cancer Research, 21, 3841-3852.

REYES, A. & HUBER, W. 2018. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. Nucleic acids research, 46, 582-592.

RIBEIRO DE ALMEIDA, C., DHIR, S., DHIR, A., MOGHADDAM, A. E., SATTENTAU, Q., MEINHART, A. & PROUDFOOT, N. J. 2018. RNA Helicase DDX1 Converts RNA G-Quadruplex Structures into R-Loops to Promote IgH Class Switch Recombination. Mol Cell, 70, 650-662.e8.

RIGAULT, C., LE BORGNE, F. & DEMARQUOY, J. 2006. Genomic structure, alternative maturation and tissue expression of the human BBOX1 gene. Biochim Biophys Acta, 1761, 1469-81.

RILEY, A., JORDAN, L. E. & HOLCIK, M. 2010. Distinct 5′ UTRs regulate XIAP expression under normal growth conditions and during cellular stress. Nucleic Acids Research, 38, 4665-4674.

RITCHIE, M. E., PHIPSON, B., WU, D., HU, Y., LAW, C. W., SHI, W. & SMYTH, G. K. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research, 43, e47-e47.

ROBERT, F. & PELLETIER, J. 2013. Perturbations of RNA helicases in cancer. Wiley Interdiscip Rev RNA, 4, 333-49.

ROCK, K. L., GRAMM, C., ROTHSTEIN, L., CLARK, K., STEIN, R., DICK, L., HWANG, D. & GOLDBERG, A. L. 1994. Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules. Cell, 78, 761-71.

ROJAS-DURAN, M. F. & GILBERT, W. V. 2012. Alternative transcription start site selection leads to large differences in translation activity in yeast. RNA (New York, N.Y.), 18, 2299-2305.

ROULEAU, S. G., GARANT, J.-M., BOLDUC, F., BISAILLON, M. & PERREAULT, J.-P. 2018. G-Quadruplexes influence pri-microRNA processing. RNA Biology, 15, 198-206.

ROULEAU, S., GLOUZON, J. S., BRUMWELL, A., BISAILLON, M. & PERREAULT, J. P. 2017. 3' UTR G-quadruplexes regulate miRNA binding. Rna, 23, 1172-1179.

ROUX, P. P. & TOPISIROVIC, I. 2018. Signaling Pathways Involved in the Regulation of mRNA Translation. Mol Cell Biol, 38.

SAHAKYAN, A. B., CHAMBERS, V. S., MARSICO, G., SANTNER, T., DI ANTONIO, M. & BALASUBRAMANIAN, S. 2017. Machine learning model for sequence-driven DNA G-quadruplex formation. Scientific reports, 7, 14535-14535.

SALVATI, E., LEONETTI, C., RIZZO, A., SCARSELLA, M., MOTTOLESE, M., GALATI, R., SPERDUTI, I., STEVENS, M. F., D'INCALCI, M., BLASCO, M., CHIORINO, G., BAUWENS, S., HORARD, B., GILSON, E., STOPPACCIARO, A., ZUPI, G. & BIROCCIO, A. 2007. Telomere damage induced by the G-quadruplex ligand RHPS4 has an antitumor effect. J Clin Invest, 117, 3236-47.

SAXONOV, S., BERG, P. & BRUTLAG, D. L. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences, 103, 1412-1417.

SCHAFFITZEL, C., BERGER, I., POSTBERG, J., HANES, J., LIPPS, H. J. & PLÜCKTHUN, A. 2001. In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei. Proceedings of the National Academy of Sciences, 98, 8572-8577.

SCHAKOWSKI, F., BUTTGEREIT, P., MAZUR, M., MÄRTEN, A., SCHÖTTKER, B., GORSCHLÜTER, M. & SCHMIDT-WOLF, I. G. H. 2004. Novel non-viral method for transfection of primary leukemia cells and cell lines. Genetic vaccines and therapy, 2, 1-1.

SCHENA, M. 1989. The evolutionary conservation of eukaryotic gene transcription. Experientia, 45, 972-983.

SCHNEIDER-POETSCH, T., JU, J., EYLER, D. E., DANG, Y., BHAT, S., MERRICK, W. C., GREEN, R., SHEN, B. & LIU, J. O. 2010. Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. Nature Chemical Biology, 6, 209-217.

SCHULZ, I. 1990. Permeabilizing cells: Some methods and applications for the study of intracellular processes. Methods in Enzymology. Academic Press.

SEN, D. & GILBERT, W. 1988. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. Nature, 334, 364-366.

SENNEPIN, A. D., CHARPENTIER, S., NORMAND, T., SARRÉ, C., LEGRAND, A. & MOLLET, L. M. 2009. Multiple reprobing of Western blots after inactivation of peroxidase activity by its substrate, hydrogen peroxide. Analytical Biochemistry, 393, 129-131.

SEVERIN, J., LIZIO, M., HARSHBARGER, J., KAWAJI, H., DAUB, C. O., HAYASHIZAKI, Y., BERTIN, N., FORREST, A. R. R. & THE, F. C. 2014. Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. Nature Biotechnology, 32, 217-219.

SHAHID, R., BUGAUT, A. & BALASUBRAMANIAN, S. 2010. The BCL-2 5' untranslated region contains an RNA G-quadruplex-forming motif that modulates protein expression. Biochemistry, 49, 8300-6.

SHEN, S., PARK, J. W., LU, Z.-X., LIN, L., HENRY, M. D., WU, Y. N., ZHOU, Q. & XING, Y. 2014. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proceedings of the National Academy of Sciences, 111, E5593-E5601.

SHIH, J.-W. & LEE, Y.-H. W. 2014. Human DExD/H RNA helicases: Emerging roles in stress survival regulation. Clinica Chimica Acta, 436, 45-58.

SHIMOTOHNO, K., KODAMA, Y., HASHIMOTO, J. & MIURA, K.-I. 1977. Importance of 5'-terminal blocking structure to stabilize mRNA in eukaryotic protein synthesis. Proceedings of the National Academy of Sciences, 74, 2734-2738.

SHIVALINGAM, A., IZQUIERDO, M. A., MAROIS, A. L., VYŠNIAUSKAS, A., SUHLING, K., KUIMOVA, M. K. & VILAR, R. 2015. The interactions between a small molecule and G-quadruplexes are visualized by fluorescence lifetime imaging microscopy. Nature Communications, 6.

SIDDIQUI-JAIN, A., GRAND, C. L., BEARSS, D. J. & HURLEY, L. H. 2002. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. Proceedings of the National Academy of Sciences of the United States of America, 99, 11593-11598.

SIMON, J. A. & KINGSTON, R. E. 2009. Mechanisms of polycomb gene silencing: knowns and unknowns. Nat Rev Mol Cell Biol, 10, 697-708.

SIMONSSON, T. 2001. G-quadruplex DNA structures--variations on a theme. Biol Chem, 382, 621-8.

SIMS, R. J., 3RD, BELOTSERKOVSKAYA, R. & REINBERG, D. 2004. Elongation by RNA polymerase II: the short and long of it. Genes Dev, 18, 2437-68.

SKOURTI-STATHAKI, K. & PROUDFOOT, N. J. 2014. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. Genes Dev, 28, 1384-96.

SMALE, S. T. & BALTIMORE, D. 1989. The "initiator" as a transcription control element. Cell, 57, 103-113.

SMALE, S. T. & KADONAGA, J. T. 2003. The RNA polymerase II core promoter. Annu Rev Biochem, 72, 449-79.

SMALE, S. T. 2001. Core promoters: active contributors to combinatorial gene regulation. Genes Dev, 15, 2503-8.

SMITH, C. L. & PETERSON, C. L. 2005. ATP-dependent chromatin remodeling. Curr Top Dev Biol, 65, 115-48.

SMITH, F. W. & FEIGON, J. 1992. Quadruplex structure of Oxytricha telomeric DNA oligonucleotides. Nature, 356, 164-168.

SMITH, J., SEN, S., WEEKS, R. J., ECCLES, M. R. & CHATTERJEE, A. 2020. Promoter DNA Hypermethylation and Paradoxical Gene Activation. Trends in Cancer, 6, 392-406.

SMITH, P. K., KROHN, R. I., HERMANSON, G. T., MALLIA, A. K., GARTNER, F. H., PROVENZANO, M. D., FUJIMOTO, E. K., GOEKE, N. M., OLSON, B. J. & KLENK, D. C. 1985. Measurement of protein using bicinchoninic acid. Anal Biochem, 150, 76-85.

SOBCZAK, K. & KRZYZOSIAK, W. J. 2002. Structural determinants of BRCA1 translational regulation. J Biol Chem, 277, 17349-58.

SONENBERG, N. & HINNEBUSCH, A. G. 2009. Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. Cell, 136, 731-745.

SONG, P., YANG, F., JIN, H. & WANG, X. 2021. The regulation of protein translation and its implications for cancer. Signal Transduction and Targeted Therapy, 6, 68.

SPRIGGS, K. A., BUSHELL, M. & WILLIS, A. E. 2010. Translational regulation of gene expression during conditions of cell stress. Mol Cell, 40, 228-37.

SPRIGGS, K. A., COBBOLD, L. C., RIDLEY, S. H., COLDWELL, M., BOTTLEY, A., BUSHELL, M., WILLIS, A. E. & SIDDLE, K. 2009. The human insulin receptor mRNA contains a functional internal ribosome entry segment. Nucleic Acids Research, 37, 5881-5893.

SPRIGGS, K. A., STONELEY, M., BUSHELL, M. & WILLIS, A. E. 2008. Re-programming of translation following cell stress allows IRES-mediated translation to predominate. Biol Cell, 100, 27-38.

STAMM, S., BEN-ARI, S., RAFALSKA, I., TANG, Y., ZHANG, Z., TOIBER, D., THANARAJ, T. A. & SOREQ, H. 2005. Function of alternative splicing. Gene, 344, 1-20.

STOECKLIN, G., GROSS, B., MING, X. F. & MORONI, C. 2003. A novel mechanism of tumor suppression by destabilizing AU-rich growth factor mRNA. Oncogene, 22, 3554-61.

SUBRAMANIAN, M., RAGE, F., TABET, R., FLATTER, E., MANDEL, J. L. & MOINE, H. 2011. G-quadruplex RNA structure as a signal for neurite mRNA targeting. EMBO Rep, 12, 697-704.

SUN, D. & HURLEY, L. H. 2010. Biochemical techniques for the characterization of G-quadruplex structures: EMSA, DMS footprinting, and DNA polymerase stop assay. Methods in molecular biology (Clifton, N.J.), 608, 65-79.

SUN, M., NIE, F., WANG, Y., ZHANG, Z., HOU, J., HE, D., XIE, M., XU, L., DE, W., WANG, Z. & WANG, J. 2016. LncRNA HOXA11-AS Promotes Proliferation and Invasion of Gastric Cancer by Scaffolding the Chromatin Modification Factors PRC2, LSD1, and DNMT1. Cancer Res, 76, 6299-6310.

SUNDQUIST, W. I. & KLUG, A. 1989. Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. Nature, 342, 825-829.

SURANI, A. A., COLOMBO, S., BARLOW, G., FOULDS, G. & & MONTIEL-DUARTE, C. 2021. Optimising cell synchronisation using nocodazole or double thymidine blocks, Springer (In press).

SUZUKI, A., KAWANO, S., MITSUYAMA, T., SUYAMA, M., KANAI, Y., SHIRAHIGE, K., SASAKI, H., TOKUNAGA, K., TSUCHIHARA, K., SUGANO, S., NAKAI, K. & SUZUKI, Y. 2018. DBTSS/DBKERO for integrated analysis of transcriptional regulation. Nucleic Acids Res, 46, D229-d238.

SUZUKI, Y., TAIRA, H., TSUNODA, T., MIZUSHIMA-SUGANO, J., SESE, J., HATA, H., OTA, T., ISOGAI, T., TANAKA, T., MORISHITA, S., OKUBO, K., SAKAKI, Y., NAKAMURA, Y., SUYAMA, A. & SUGANO, S. 2001. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO reports, 2, 388-393.

SUZUKI, Y., YAMASHITA, R., NAKAI, K. & SUGANO, S. 2002. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. Nucleic acids research, 30, 328-331.

SWINDELL, W. R., REMMER, H. A., SARKAR, M. K., XING, X., BARNES, D. H., WOLTERINK, L., VOORHEES, J. J., NAIR, R. P., JOHNSTON, A., ELDER, J. T. & GUDJONSSON, J. E. 2015. Proteogenomic analysis of psoriasis reveals discordant and concordant changes in mRNA and protein abundance. Genome medicine, 7, 86-86.

SZAMECZ, B., RUTKAI, E., CUCHALOVÁ, L., MUNZAROVÁ, V., HERRMANNOVÁ, A., NIELSEN, K. H., BURELA, L., HINNEBUSCH, A. G. & VALÁSEK, L. 2008. eIF3a cooperates with sequences 5' of uORF1 to promote resumption of scanning by post-termination ribosomes for reinitiation on GCN4 mRNA. Genes Dev, 22, 2414-25.

TAKAHASHI, H., KATO, S., MURATA, M. & CARNINCI, P. 2012a. CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. Methods in molecular biology (Clifton, N.J.), 786, 181-200.

TAKAHASHI, H., LASSMANN, T., MURATA, M. & CARNINCI, P. 2012b. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nature protocols, 7, 542-561.

TANG, C.-F. & SHAFER, R. H. 2006. Engineering the Quadruplex Fold:  Nucleoside Conformation Determines Both Folding Topology and Molecularity in Guanine Quadruplexes. Journal of the American Chemical Society, 128, 5966-5973.

TANG, X., FENG, Y. & YE, K. 2007. Src-family tyrosine kinase fyn phosphorylates phosphatidylinositol 3-kinase enhancer-activating Akt, preventing its apoptotic cleavage and promoting cell survival. Cell Death Differ, 14, 368-77.

THANDAPANI, P., SONG, J., GANDIN, V., CAI, Y., ROULEAU, S. G., GARANT, J.-M., BOISVERT, F.-M., YU, Z., PERREAULT, J.-P., TOPISIROVIC, I. & RICHARD, S. 2015. Aven recognition of RNA G-quadruplexes regulates translation of the mixed lineage leukemia protooncogenes. eLife, 4, e06234.

THOMAS, M. C. & CHIANG, C. M. 2006. The general transcription machinery and general cofactors. Crit Rev Biochem Mol Biol, 41, 105-78.

THORVALDSDÓTTIR, H., ROBINSON, J. T. & MESIROV, J. P. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics, 14, 178-192.

TILLETT, D., BURNS, B. P. & NEILAN, B. A. 2000. Optimized rapid amplification of cDNA ends (RACE) for mapping bacterial mRNA transcripts. Biotechniques, 28, 448, 450, 452-3, 456.

TODD, A. K., JOHNSTON, M. & NEIDLE, S. 2005. Highly prevalent putative quadruplex sequence motifs in human DNA. Nucleic Acids Research, 33, 2901-2907.

TOLSMA, THOMAS O. & HANSEN, JEFFREY C. 2019. Post-translational modifications and chromatin dynamics. Essays in Biochemistry, 63, 89-96.

TOMEK, W. & WOLLENHAUPT, K. 2012. The “closed loop model” in controlling mRNA translation during development. Animal Reproduction Science, 134, 2-8.

TSE, M. C. L., LIU, X., YANG, S., YE, K. & CHAN, C. B. 2013. Fyn Regulates Adipogenesis by Promoting PIKE-A/STAT5a Interaction. Molecular and Cellular Biology, 33, 1797-1808.

TSUCHIHARA, K., SUZUKI, Y., WAKAGURI, H., IRIE, T., TANIMOTO, K., HASHIMOTO, S., MATSUSHIMA, K., MIZUSHIMA-SUGANO, J., YAMASHITA, R., NAKAI, K., BENTLEY, D., ESUMI, H. & SUGANO, S. 2009. Massive transcriptional start site analysis of human genes in hypoxia cells. Nucleic Acids Res, 37, 2249-63.

TULLER, T., CARMI, A., VESTSIGIAN, K., NAVON, S., DORFAN, Y., ZABORSKE, J., PAN, T., DAHAN, O., FURMAN, I. & PILPEL, Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell, 141, 344-54.

TUROWSKI, T. W. & TOLLERVEY, D. 2020. Extended ncRNAs Interfere with Promoter Nucleosome Dynamics. Trends in Genetics, 36, 637-639.

VAN HERCK, J. L., DE MEYER, G. R. Y., MARTINET, W., BULT, H., VRINTS, C. J. & HERMAN, A. G. 2009. Proteasome inhibitor bortezomib promotes a rupture-prone plaque phenotype in ApoE-deficient mice. Basic Research in Cardiology, 105, 39.

VAN, P. N., XINH, P. T., KANO, Y., TOKUNAGA, K. & SATO, Y. 2005. Establishment and characterization of A novel Philadelphia-chromosome positive chronic myeloid leukemia cell line, TCC-S, expressing P210 and P190 BCR/ABL transcripts but missing normal ABL gene. Hum Cell, 18, 25-33.

VANNIER, J.-B., PAVICIC-KALTENBRUNNER, V., PETALCORIN, MARK I. R., DING, H. & BOULTON, SIMON J. 2012. RTEL1 Dismantles T Loops and Counteracts Telomeric G4-DNA to Maintain Telomere Integrity. Cell, 149, 795-806.

VARIZHUK, A., ISCHENKO, D., TSVETKOV, V., NOVIKOV, R., KULEMIN, N., KALUZHNY, D., VLASENOK, M., NAUMOV, V., SMIRNOV, I. & POZMOGOVA, G. 2017. The expanding repertoire of G4 DNA structures. Biochimie, 135, 54-62.

VENNE, A. S., KOLLIPARA, L. & ZAHEDI, R. P. 2014. The next level of complexity: crosstalk of posttranslational modifications. Proteomics, 14, 513-24.

VLASOVA, I. A., TAHOE, N. M., FAN, D., LARSSON, O., RATTENBACHER, B., STERNJOHN, J. R., VASDEWANI, J., KARYPIS, G., REILLY, C. S., BITTERMAN, P. B. & BOHJANEN, P. R. 2008. Conserved GU-rich elements mediate mRNA decay by binding to CUG-binding protein 1. Mol Cell, 29, 263-70.

VO NGOC, L., CASSIDY, C. J., HUANG, C. Y., DUTTKE, S. H. C. & KADONAGA, J. T. 2017. The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. Genes & development, 31, 6-11.

VOGEL, C. & MARCOTTE, E. M. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet, 13, 227-32.

VOGEL, C., ABREU RDE, S., KO, D., LE, S. Y., SHAPIRO, B. A., BURNS, S. C., SANDHU, D., BOUTZ, D. R., MARCOTTE, E. M. & PENALVA, L. O. 2010. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol Syst Biol, 6, 400.

VOLLOCH, V., SCHWEITZER, B. & RITS, S. 1994. Ligation-mediated amplification of RNA from murine erythroid cells reveals a novel class of beta globin mRNA with an extended 5'-untranslated region. Nucleic acids research, 22, 2507-2511.

VU, T. H. & HOFFMAN, A. R. 1994. Promoter-specific imprinting of the human insulin-like growth factor-II gene. Nature, 371, 714-7.

VVEDENSKAYA, I. O., VAHEDIAN-MOVAHED, H., ZHANG, Y., TAYLOR, D. M., EBRIGHT, R. H. & NICKELS, B. E. 2016. Interactions between RNA polymerase and the core recognition element are a determinant of transcription start site selection. Proceedings of the National Academy of Sciences, 113, E2899-E2905.

WANG, X., HOU, J., QUEDENAU, C. & CHEN, W. 2016. Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. Mol Syst Biol, 12, 875.

WANG, Z. & LIU, J.-P. 2017. Effects of the central potassium ions on the G-quadruplex and stabilizer binding. Journal of Molecular Graphics and Modelling, 72, 168-177.

WANROOIJ, P. H., UHLER, J. P., SIMONSSON, T., FALKENBERG, M. & GUSTAFSSON, C. M. 2010. G-quadruplex structures in RNA stimulate mitochondrial transcription termination and primer formation. Proc Natl Acad Sci U S A, 107, 16072-7.

WARNER, J. R., KNOPF, P. M. & RICH, A. 1963. A multiple ribosomal structure in protein synthesis. Proc Natl Acad Sci U S A, 49, 122-9.

WEINGARTEN-GABBAY, S., ELIAS-KIRMA, S., NIR, R., GRITSENKO, A. A., STERN-GINOSSAR, N., YAKHINI, Z., WEINBERGER, A. & SEGAL, E. 2016. Systematic discovery of cap-independent translation sequences in human and viral genomes. Science, 351, aad4939.

WEINSTEIN, L. S., XIE, T., ZHANG, Q. H. & CHEN, M. 2007. Studies of the regulation and function of the Gs alpha gene Gnas using gene targeting technology. Pharmacol Ther, 115, 271-91.

WELDON, C., BEHM-ANSMANT, I., HURLEY, L. H., BURLEY, G. A., BRANLANT, C., EPERON, I. C. & DOMINGUEZ, C. 2017. Identification of G-quadruplexes in long functional RNAs using 7-deazaguanine RNA. Nature Chemical Biology, 13, 18-20.

WELDON, C., DACANAY, J. G., GOKHALE, V., BODDUPALLY, P. V. L., BEHM-ANSMANT, I., BURLEY, G. A., BRANLANT, C., HURLEY, L. H., DOMINGUEZ, C. & EPERON, I. C. 2018. Specific G-quadruplex ligands modulate the alternative splicing of Bcl-X. Nucleic Acids Res, 46, 886-896.

WELDON, C., EPERON, I. C. & DOMINGUEZ, C. 2016. Do we know whether potential G-quadruplexes actually form in long functional RNA molecules? Biochemical Society Transactions, 44, 1761-1768.

WILKIE, G. S., DICKSON, K. S. & GRAY, N. K. 2003. Regulation of mRNA translation by 5'- and 3'-UTR-binding factors. Trends Biochem Sci, 28, 182-8.

WOLFE, A. L., SINGH, K., ZHONG, Y., DREWE, P., RAJASEKHAR, V. K., SANGHVI, V. R., MAVRAKIS, K. J., JIANG, M., RODERICK, J. E., VAN DER MEULEN, J., SCHATZ, J. H., RODRIGO, C. M., ZHAO, C., RONDOU, P., DE STANCHINA, E., TERUYA-FELDSTEIN, J., KELLIHER, M. A., SPELEMAN, F., PORCO JR, J. A., PELLETIER, J., RÄTSCH, G. & WENDEL, H. G. 2014. RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. Nature, 513, 65-70.

WU, Y., ZHAO, Y., MA, X., ZHU, Y., PATEL, J. & NIE, Z. 2013. The Arf GAP AGAP2 interacts with β-arrestin2 and regulates β(2)-adrenergic receptor recycling and ERK activation. The Biochemical journal, 452, 411-421.

XIA, C., MA, W., STAFFORD, L. J., LIU, C., GONG, L., MARTIN, J. F. & LIU, M. 2003. GGAPs, a new family of bifunctional GTP-binding and GTPase-activating proteins. Mol Cell Biol, 23, 2476-88.

XIA, Z., DONEHOWER, L. A., COOPER, T. A., NEILSON, J. R., WHEELER, D. A., WAGNER, E. J. & LI, W. 2014. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. Nat Commun, 5, 5274.

XIAO, S., ZHANG, J. Y., ZHENG, K. W., HAO, Y. H. & TAN, Z. 2013. Bioinformatic analysis reveals an evolutional selection for DNA:RNA hybrid G-quadruplex structures as putative transcription regulatory elements in warm-blooded animals. Nucleic Acids Res, 41, 10379-90.

XU, C., PARK, J. K. & ZHANG, J. 2019. Evidence that alternative transcriptional initiation is largely nonadaptive. PLoS Biol., 17, e3000197.

XU, J., LU, Y., LIU, Q., XIA, A., ZHAO, J., XU, X., SUN, Q., QI, F. & SUN, B. 2020. Long noncoding RNA GMAN promotes hepatocellular carcinoma progression by interacting with eIF4B. Cancer Lett, 473, 1-12.

XU, L., CHEN, Y., SONG, Q., XU, D., WANG, Y. & MA, D. 2009. PDCD5 interacts with Tip60 and functions as a cooperator in acetyltransferase activity and DNA damage-induced apoptosis. Neoplasia, 11, 345-54.

XU, N., CHEN, C. Y. & SHYU, A. B. 2001. Versatile role for hnRNP D isoforms in the differential regulation of cytoplasmic mRNA turnover. Molecular and cellular biology, 21, 6960-6971.

XU, T. P., MA, P., WANG, W. Y., SHUAI, Y., WANG, Y. F., YU, T., XIA, R. & SHU, Y. Q. 2019c. KLF5 and MYC modulated LINC00346 contributes to gastric cancer progression through acting as a competing endogeous RNA and indicates poor outcome. Cell Death Differ, 26, 2179-2193.

XU, W., SAN LUCAS, A., WANG, Z. & LIU, Y. 2014. Identifying microRNA targets in different gene regions. BMC Bioinformatics, 15, S4.

XU, Y., GUO, W., LI, P., ZHANG, Y., ZHAO, M., FAN, Z., ZHAO, Z. & YAN, J. 2016. Long-Range Chromosome Interactions Mediated by Cohesin Shape Circadian Gene Expression. PLOS Genetics, 12, e1005992.

YAMASHITA, R., SUGANO, S., SUZUKI, Y. & NAKAI, K. 2012. DBTSS: DataBase of Transcriptional Start Sites progress report in 2012. Nucleic Acids Research, 40, D150-D154.

YAMASHITA, R., WAKAGURI, H., SUGANO, S., SUZUKI, Y. & NAKAI, K. 2010. DBTSS provides a tissue specific dynamic view of Transcription Start Sites. Nucleic Acids Res, 38, D98-104.

YAN, J., LEI, J., CHEN, L., DENG, H., DONG, D., JIN, T., LIU, X., YUAN, R., QIU, Y., GE, J., PENG, X. & SHAO, J. 2018. Human Leukocyte Antigen F Locus Adjacent Transcript 10 Overexpression Disturbs WISP1 Protein and mRNA Expression to Promote Hepatocellular Carcinoma Progression. Hepatology, 68, 2268-2284.

YANG, C., BOLOTIN, E., JIANG, T., SLADEK, F. M. & MARTINEZ, E. 2007. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene, 389, 52-65.

YANG, D. 2019. G-Quadruplex DNA and RNA. Methods Mol Biol, 2035, 1-24.

YANG, S. Y., LEJAULT, P., CHEVRIER, S., BOIDOT, R., ROBERTSON, A. G., WONG, J. M. Y. & MONCHAUD, D. 2018. Transcriptome-wide identification of transient RNA G-quadruplexes in human cells. Nature Communications, 9, 4730.

YANG, X., CHEEMA, J., ZHANG, Y., DENG, H., DUNCAN, S., UMAR, M. I., ZHAO, J., LIU, Q., CAO, X., KWOK, C. K. & DING, Y. 2020. RNA G-quadruplex structures exist and function in vivo in plants. Genome Biology, 21, 226.

YANG, X., PANG, J., SHEN, N., YAN, F., WU, L.-C., AL-KALI, A., LITZOW, M. R., PENG, Y., LEE, R. J. & LIU, S. 2016. Liposomal bortezomib is active against chronic myeloid leukemia by disrupting the Sp1-BCR/ABL axis. Oncotarget, 7, 36382-36394.

YATES, A. D., ACHUTHAN, P., AKANNI, W., ALLEN, J., ALLEN, J., ALVAREZ-JARRETA, J., AMODE, M. R., ARMEAN, I. M., AZOV, A. G., BENNETT, R., BHAI, J., BILLIS, K., BODDU, S., MARUGÁN, J. C., CUMMINS, C., DAVIDSON, C., DODIYA, K., FATIMA, R., GALL, A., GIRON, C. G., GIL, L., GREGO, T., HAGGERTY, L., HASKELL, E., HOURLIER, T., IZUOGU, O. G., JANACEK, S. H., JUETTEMANN, T., KAY, M., LAVIDAS, I., LE, T., LEMOS, D., MARTINEZ, J. G., MAUREL, T., MCDOWALL, M., MCMAHON, A., MOHANAN, S., MOORE, B., NUHN, M., OHEH, D. N., PARKER, A., PARTON, A., PATRICIO, M., SAKTHIVEL, M. P., ABDUL SALAM, A. I., SCHMITT, B. M., SCHUILENBURG, H., SHEPPARD, D., SYCHEVA, M., SZUBA, M., TAYLOR, K., THORMANN, A., THREADGOLD, G., VULLO, A., WALTS, B., WINTERBOTTOM, A., ZADISSA, A., CHAKIACHVILI, M., FLINT, B., FRANKISH, A., HUNT, S. E., IISLEY, G., KOSTADIMA, M., LANGRIDGE, N., LOVELAND, J. E., MARTIN, F. J., MORALES, J., MUDGE, J. M., MUFFATO, M., PERRY, E., RUFFIER, M., TREVANION, S. J., CUNNINGHAM, F., HOWE, K. L., ZERBINO, D. R. & FLICEK, P. 2019. Ensembl 2020. Nucleic Acids Research, 48, D682-D688.

YE, K., HURT, K. J., WU, F. Y., FANG, M., LUO, H. R., HONG, J. J., BLACKSHAW, S., FERRIS, C. D. & SNYDER, S. H. 2000. Pike. A nuclear gtpase that enhances PI3kinase activity and is regulated by protein 4.1N. Cell, 103, 919-30.

YE, K. & SNYDER, S. H. 2004. PIKE GTPase: a novel mediator of phosphoinositide signaling. Journal of Cell Science, 117, 155-161.

YUAN, F., HANKEY, W., WAGNER, E. J., LI, W. & WANG, Q. 2021. Alternative polyadenylation of mRNA and its role in cancer. Genes & Diseases, 8, 61-72.

ZEITZ, M. J., CALHOUN, P. J., JAMES, C. C., TAETZSCH, T., GEORGE, K. K., ROBEL, S., VALDEZ, G. & SMYTH, J. W. 2019. Dynamic UTR Usage Regulates Alternative Translation to Modulate Gap Junction Formation during Stress and Aging. Cell Rep, 27, 2737-2747.e5.

ZHANG, M., MA, Z., SELLIAH, N., WEISS, G., GENIN, A., FINKEL, T. H. & CRON, R. Q. 2014. The impact of Nucleofection® on the activation state of primary human CD4 T cells. Journal of immunological methods, 408, 123-131.

ZHANG, P., DIMONT, E., HA, T., SWANSON, D. J., HIDE, W. & GOLDOWITZ, D. 2017. Relatively frequent switching of transcription start sites during cerebellar development. BMC Genomics, 18, 461.

ZHANG, S., SHENG, H., ZHANG, X., QI, Q., CHAN, C. B., LI, L., SHAN, C. & YE, K. 2019. Cellular energy stress induces AMPK-mediated regulation of glioblastoma cell proliferation by PIKE-A phosphorylation. 10, 222.

ZHANG, S., SUN, H., WANG, L., LIU, Y., CHEN, H., LI, Q., GUAN, A., LIU, M. & TANG, Y. 2018. Real-time monitoring of DNA G-quadruplexes in living cells with a small-molecule fluorescent probe. Nucleic Acids Research, 46, 7522-7532.

ZHENG, N. & SHABEK, N. 2017. Ubiquitin Ligases: Structure, Function, and Regulation. Annu Rev Biochem, 86, 129-157.

ZHOU, T., WEEMS, M. & WILKE, C. O. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. Mol Biol Evol, 26, 1571-80.

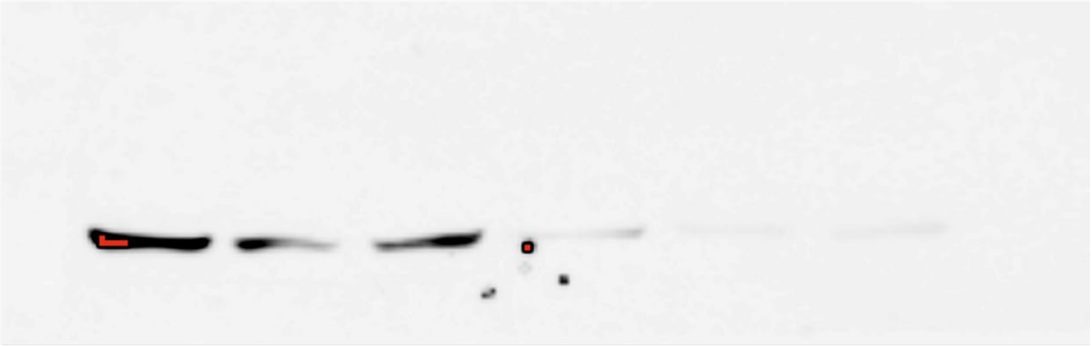
ZHU, Y., WU, Y., KIM, J. I., WANG, Z., DAAKA, Y. & NIE, Z. 2009. Arf GTPase-activating protein AGAP2 regulates focal adhesion kinase activity and focal adhesion remodeling. J Biol Chem, 284, 13489-96.

# Chapter 8:

# Appendix

## 8.1 Appendix 1:

**(A)**



**(B)**

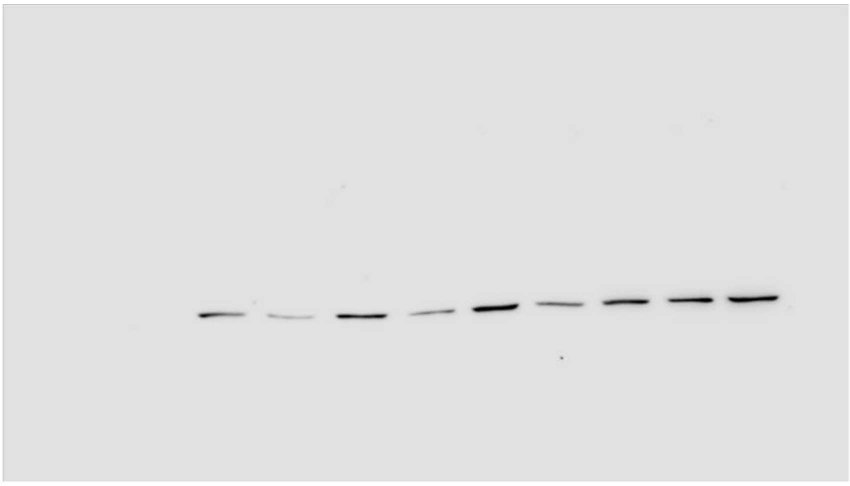
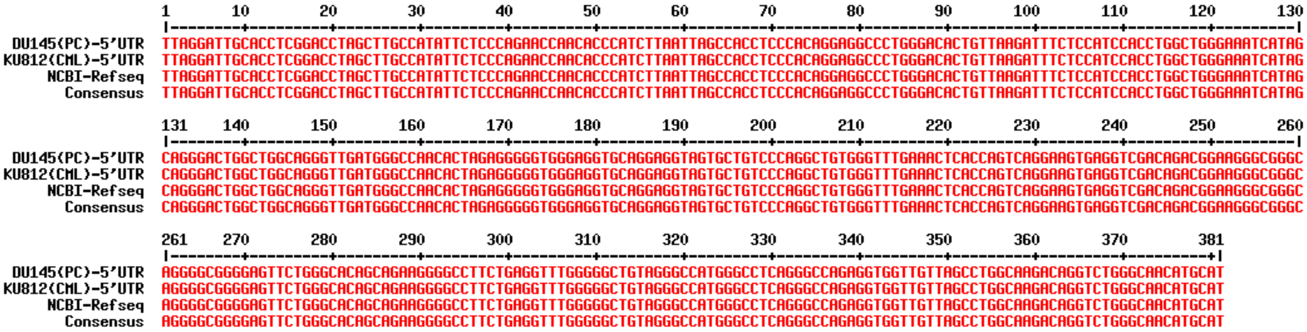


**Figure 8.1: Full immunoblot for Figure 3.5.** Representative full immunoblot for AGAP2 protein detection in PC and CML cell lines **(A)** and in different cell lines (DU145, HepG2, MCF7, PA-1, SKOV3, and U-2OS, and RAJI **(B).**

## 8.2 Appendix 2:



```
                 1       10        20        30        40        50        60        70        80        90       100       110       120       130
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
 DU145(PC)-5’UTR TTAGGATTGCACCTCGGACCTAGCTTGCCATATTCTCCCAGAACCAACACCCATCTTAATTAGCCACCTCCCACAGGAGGCCCTGGGACACTGTTAAGATTTCTCCATCCACCTGGCTGGGAAATCATAG
KU812(CML)-5’UTR TTAGGATTGCACCTCGGACCTAGCTTGCCATATTCTCCCAGAACCAACACCCATCTTAATTAGCCACCTCCCACAGGAGGCCCTGGGACACTGTTAAGATTTCTCCATCCACCTGGCTGGGAAATCATAG
     NCBI-Refseq TTAGGATTGCACCTCGGACCTAGCTTGCCATATTCTCCCAGAACCAACACCCATCTTAATTAGCCACCTCCCACAGGAGGCCCTGGGACACTGTTAAGATTTCTCCATCCACCTGGCTGGGAAATCATAG
       Consensus TTAGGATTGCACCTCGGACCTAGCTTGCCATATTCTCCCAGAACCAACACCCATCTTAATTAGCCACCTCCCACAGGAGGCCCTGGGACACTGTTAAGATTTCTCCATCCACCTGGCTGGGAAATCATAG

                 131    140       150       160       170       180       190       200       210       220       230       240       250       260
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
 DU145(PC)-5’UTR CAGGGACTGGCTGGCAGGGTTGATGGGCCAACACTAGAGGGGGTGGGAGGTGCAGGAGGTAGTGCTGTCCCAGGCTGTGGGTTTGAAACTCACCAGTCAGGAAGTGAGGTCGACAGACGGAAGGGCGGGC
KU812(CML)-5’UTR CAGGGACTGGCTGGCAGGGTTGATGGGCCAACACTAGAGGGGGTGGGAGGTGCAGGAGGTAGTGCTGTCCCAGGCTGTGGGTTTGAAACTCACCAGTCAGGAAGTGAGGTCGACAGACGGAAGGGCGGGC
     NCBI-Refseq CAGGGACTGGCTGGCAGGGTTGATGGGCCAACACTAGAGGGGGTGGGAGGTGCAGGAGGTAGTGCTGTCCCAGGCTGTGGGTTTGAAACTCACCAGTCAGGAAGTGAGGTCGACAGACGGAAGGGCGGGC
       Consensus CAGGGACTGGCTGGCAGGGTTGATGGGCCAACACTAGAGGGGGTGGGAGGTGCAGGAGGTAGTGCTGTCCCAGGCTGTGGGTTTGAAACTCACCAGTCAGGAAGTGAGGTCGACAGACGGAAGGGCGGGC

                 261    270       280       290       300       310       320       330       340       350       360       370        381
                 |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+|
 DU145(PC)-5’UTR AGGGGCGGGGAGTTCTGGGCACAGCAGAAGGGGCCTTCTGAGGTTTGGGGGCTGTAGGGCCATGGGCCTCAGGGCCAGAGGTGGTTGTTAGCCTGGCAAGACAGGTCTGGGCAACATGCAT
KU812(CML)-5’UTR AGGGGCGGGGAGTTCTGGGCACAGCAGAAGGGGCCTTCTGAGGTTTGGGGGCTGTAGGGCCATGGGCCTCAGGGCCAGAGGTGGTTGTTAGCCTGGCAAGACAGGTCTGGGCAACATGCAT
     NCBI-Refseq AGGGGCGGGGAGTTCTGGGCACAGCAGAAGGGGCCTTCTGAGGTTTGGGGGCTGTAGGGCCATGGGCCTCAGGGCCAGAGGTGGTTGTTAGCCTGGCAAGACAGGTCTGGGCAACATGCAT
       Consensus AGGGGCGGGGAGTTCTGGGCACAGCAGAAGGGGCCTTCTGAGGTTTGGGGGCTGTAGGGCCATGGGCCTCAGGGCCAGAGGTGGTTGTTAGCCTGGCAAGACAGGTCTGGGCAACATGCAT
```

**Figure 8.2: Alignment of *AGAP2* 5’ UTR region between cancer cell lines (DU145 and KU812) and NCBI RefSeq database** Multalin alignment of the amplified sequence in the *AGAP2* TSS region in PC and CML cell lines and NCBI RefSeq database.

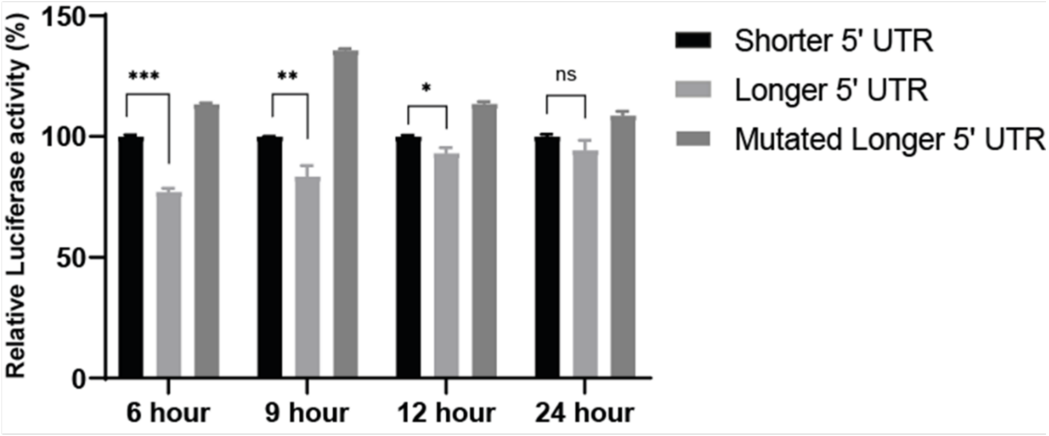
## 8.3 Appendix 3:

**Effect of electroporation on offsetting the impact of G4 in the longer 5' UTR:**

The leukaemia cell lines are notoriously difficult to transfect, and electroporation-based techniques (nucleofection) are usually used to achieve optimal gene transfer (Schakowski *et al.,* 2004). However, nucleofection has been shown to induce non-specific changes in the metabolic activity of the transfected cells and alters the phosphorylation state of certain translation initiation factors (Anderson *et al.,* 2013, Mello de Queiroz *et al.,* 2012, Zhang *et al.,* 2014). The stressors that disrupt cellular homeostasis were also shown to activate DExD/H RNA helicases and initiate a multifaceted translational regulation response (Reviewed in Shih and Lee, 2014). These non-specific effects could be responsible to offset the impact of G4 in the longer 5' UTR, as observed by the loss of differences in the relative luciferase levels at later time points after transfection and when reporter plasmids were transfected in DU145 using nucleofection (*Figure 8.3*).

As indicated in *Figure 8.3A*, the significant differences between shorter and longer 5' UTR decreased by increasing the incubation time after electroporation in the KU812 cell line. It could be due to the deranged stress response with overexpression of certain helicases that unfold secondary structures, including G4, resulting in uninhibited translation. Further studies are required to understand the folding/unfolding dynamics of G4 structures during the cellular stress response. We also noted a loss of significant differences when the reporter constructs were transfected into the DU145 cell line using electroporation (*Figure 8.3B*). These reporter constructs when transfected using a chemical-based method showed significant difference (*Figure 8.3B*) which were no longer present when electroporation was used as a method of transfection. Taken together, these results elucidated a prominent role of electroporation in mediating the observed translational differences which are less likely due to the cell-specific factors. Other transfection methods, for example, transduction of lentiviral vector could be used to avoid biases introduced by electroporation.

(A)



(B)

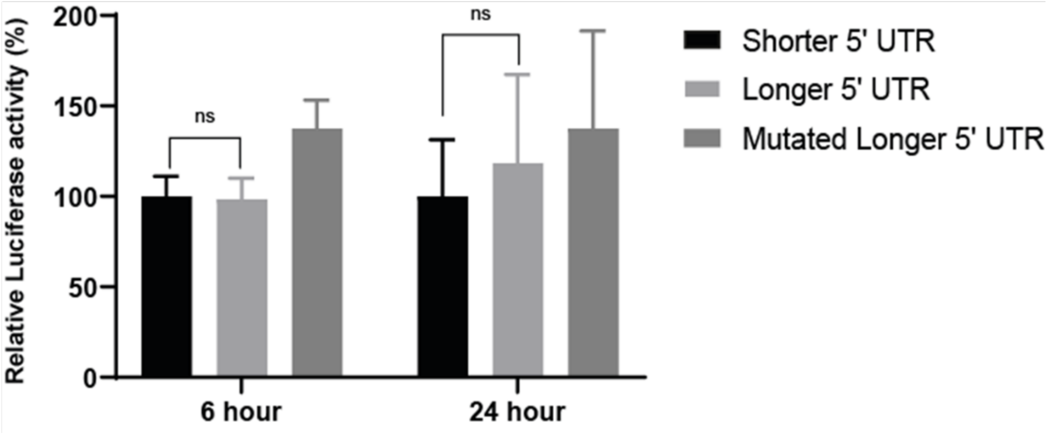


**Figure 8.3: Electroporation and relative luciferase activity in KU812 and DU145. (A)** KU812 cell line was transiently transfected with reporter constructs using electroporation with a nucleofector device (program X-001). After the indicated time points (x-axis), the cells were collected for the dual-luciferase reporter assay. The graph represents the mean of 2 independent experiments (n=2) and expressed as relative Rluc/Fluc ratio. The differences between shorter and longer 5’ UTR isoforms were analysed with Mann–Whitney U test, *P*-values shown. **(B)** DU145 cell line was transfected with reporter constructs using electroporation as above and cells were collected for the reporter assay at 6-hour and 24-hour. The graph represents the mean of 2 independent experiments (n=2) and expressed as relative Rluc/Fluc ratio. (**P* < 0.05; ***P* < 0.01; ****P* < 0.001; ns: not significant). The firefly luciferase was used as an internal control.

## 8.4 Appendix 4:

We also analysed two other genes namely *PON2* and *CAV-1* from the PC cell line group as part of the validation studies. These genes are not part of the gene list generated by the current bioinformatics analysis. They were identified by an earlier bioinformatics analysis when linear modelling and empirical Bayes was not applied to identify significantly different G4 TSS proportions. The mRNA and protein expression analysis of PON2 and CAV-1 are shown in *Figure 8.4*. As evident from the figure, both of these genes have higher relative mRNA expression and also an associated increase in the relative protein expression. Both for *PON2* and *CAV-1*, higher relative mRNA corresponds to higher relative protein abundance. Since these genes did not display a relative inconsistency in mRNA and protein expression, and they did not appear as significant in the newer analysis either, these were not selected for validation studies.

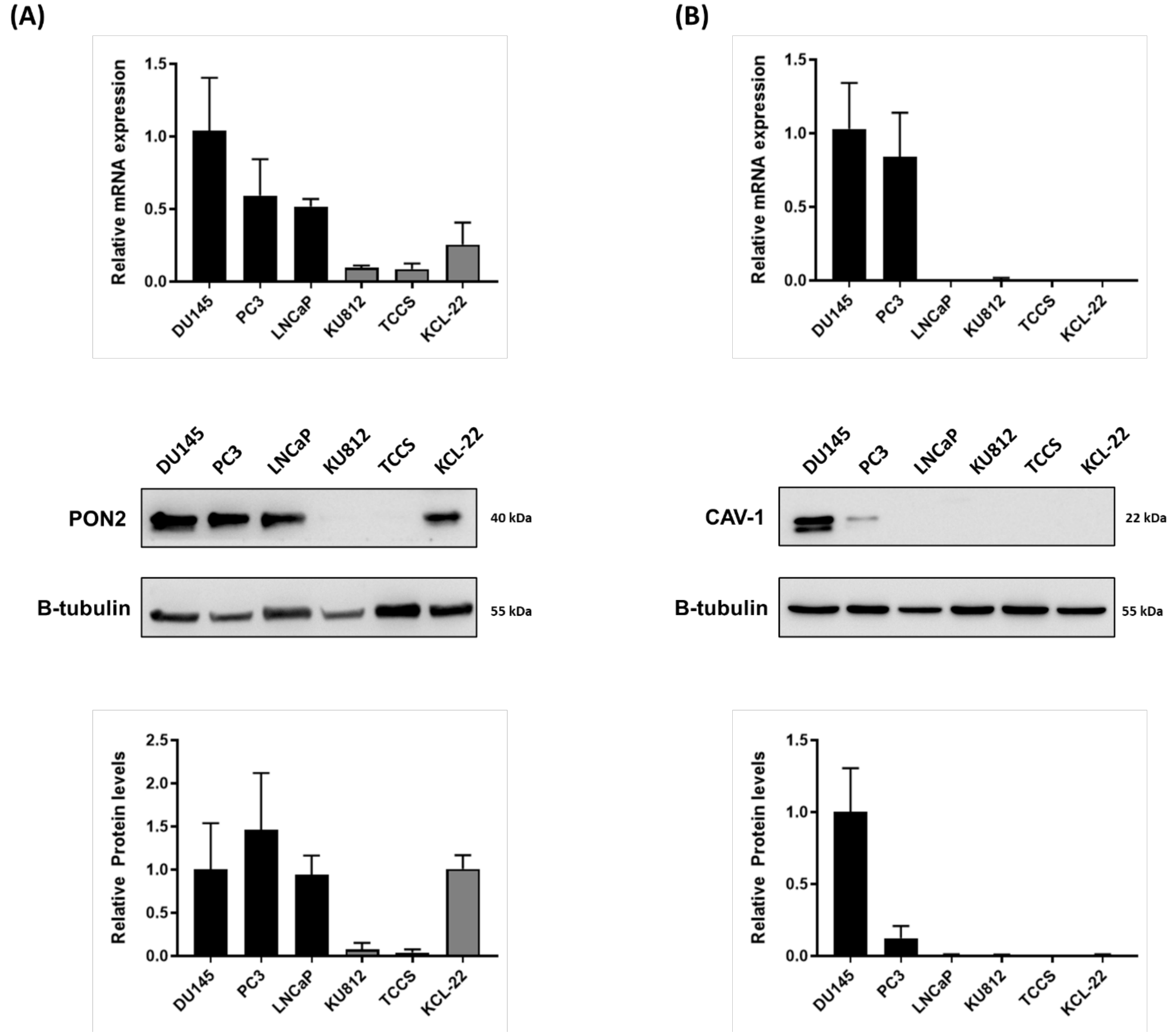


**Figure 8.4: RNA and protein expression profiles of PON2 and CAV-1.** The RNA basal levels for PON2 **(A)** and CAV-1 **(B)** were measured in PC cell lines (DU-145, PC3, LNCaP) and CML cell lines (KU812, TCCS, KCL-22) by qRT-PCR. The values presented are normalised against the levels of the housekeeping gene *HPRT* and shown relative to the PC cell line (DU145). The protein levels examined using western blotting are shown in the middle. The proteins were detected by resolving 50 µg of protein using 10% SDS-PAGE followed by immunoblotting with isoform-

specific antibody. β-tubulin was used as a loading control. Densitometry values for the relative protein expression are presented below the blots and displayed relative to DU145. All the data shown are the mean ± SD of three independent experiments (n=3), error bars represent S.D.

## 8.5 Appendix 5

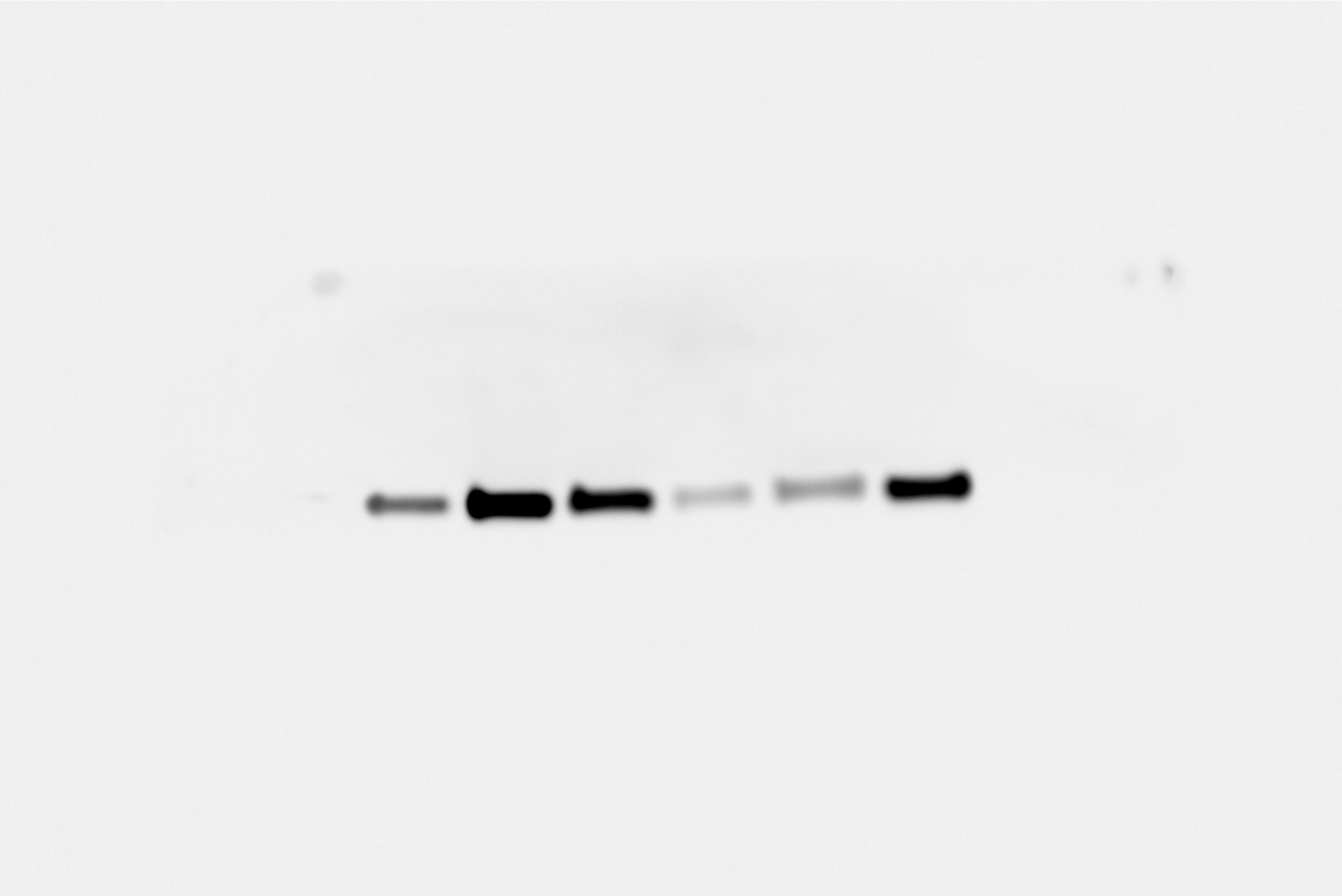


**Figure 8.5: Full immunoblot for Figure 5.8B.** Representative full immunoblot for HK1 protein detection in PC and CML cell lines.