Journal Pre-proof

Implications of differential Transcription Start Site selection on CML and prostate cancer cell protein expression

Arif A. Surani, Keith A. Spriggs, Christoph Ufer, Christos Polytarchou, Cristina Montiel-Duarte

PII: S2589-0042(22)01791-6

DOI: https://doi.org/10.1016/j.isci.2022.105519

Reference: ISCI 105519

To appear in: *ISCIENCE*

Received Date: 28 April 2022

Revised Date: 4 September 2022

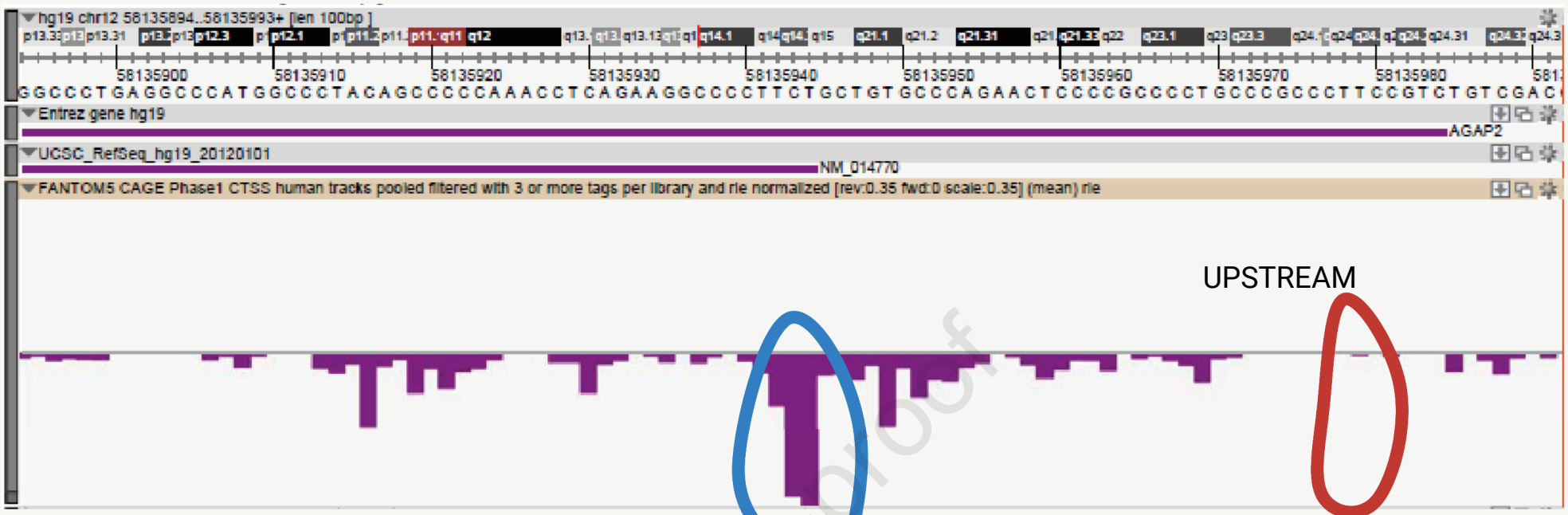
Accepted Date: 2 November 2022

Please cite this article as: Surani, A.A., Spriggs, K.A., Ufer, C., Polytarchou, C., Montiel-Duarte, C., Implications of differential Transcription Start Site selection on CML and prostate cancer cell protein expression, *ISCIENCE* (2022), doi: https://doi.org/10.1016/j.isci.2022.105519.

BROAD PROMOTER

AGAP2 (- strand) Position in chromosome 12

Transcription Start Sites (TSSs) and its frecuency



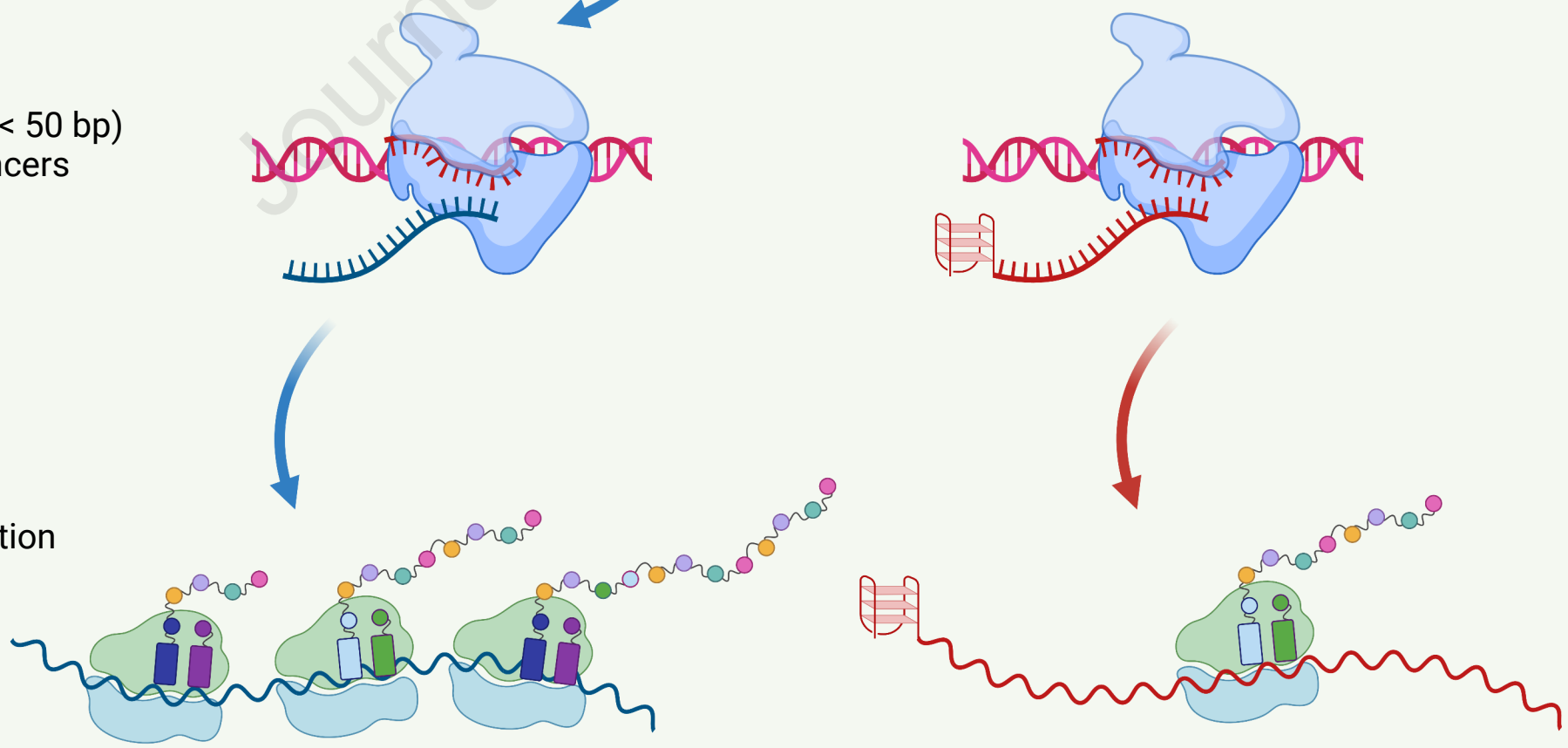
UPSTREAM

Prostate Cancer

Chronic Myeloid Leukaemia (CML)

5'UTR lenght differences (< 50 bp) generated in these cancers

Consequences in translation



Implications of differential Transcription Start Site selection on CML and prostate cancer cell protein expression

Arif A. Surani [1,2], Keith A. Spriggs [3], Christoph Ufer [4], Christos Polytarchou [1] and Cristina Montiel-Duarte [1,5,*]

[1] John van Geest Cancer Research Centre, School of Science and Technology, Nottingham Trent University, Nottingham NG11 8NS, United Kingdom.

[2] Present address: Li Ka Shining Centre, University of Cambridge, Cambridge CB2 0RE, United Kingdom.

[3] School of Pharmacy, The University of Nottingham, Nottingham NG7 2RD, United Kingdom.

[4] Formerly: Universitatsklinikum – Charite, Institut fur Biochemie CVK, Oudenarder Str. 16, 13347 Berlin, Germany.

[5] Lead contact

* Correspondence: cristina.montielduarte@ntu.ac.uk

Summary

The relevance of minor transcription start sites in broad promoters is not well understood. We have studied AGAP2 expression in prostate cancer and chronic myeloid leukaemia (CML), showing transcription is initiated from alternative transcription start sites (TSSs) within a single TSS cluster, producing cancer-type-specific *AGAP2* mRNAs with small differences in their 5' UTR length. Interestingly, in the CML cell lines where the 5' UTR is longer, AGAP2 protein levels are lower. We demonstrate that the selection of an upstream TSS involved the formation of a G quadruplex in the 5' UTR, decreasing polysome formation. After developing a bioinformatics pipeline to query data from the FANTOM project and the NCI-60 human tumour cell lines screen, we found *HK1* expression can also be regulated by the same mechanism. Overall, we present compelling data supporting TSS selection within a TSS cluster play a role on protein expression and should not be ignored.

Introduction

The transfer of information from the genome to the proteome is a coordinated multi-step process tightly controlled at the gene promoter level (influencing transcription), the mRNA level (processing and stability), and the translation level (ribosome binding and polysome association). However, the amount of mRNA in a cell does not always correlate to the amount of protein present, making RNA quantification an inexact tool to predict protein levels [1-3].

We came across discrepancies between mRNA and protein levels when studying *AGAP2* (ArfGAP with GTPase-like domain, ankyrin repeat and PH domain 2, isoform 2) promoter regulation in different cancers [4]. AGAP2 (also known as PIKE-A) is a ubiquitously expressed protein that has a role in hepatic fibrosis and cancer progression [5-7]. It is classed as a proto-oncogene involved in cell survival, apoptosis, migration, and lipid metabolism [8,9], and understanding its expression regulation would be key to modulate its functions.

One of the crucial steps in gene expression regulation is transcription initiation: a differential initiation can produce a heterogenous population of mRNA isoforms from a single gene locus. Transcription does not initiate at a single nucleotide or discrete transcription start site (TSS) within a tissue or cell culture. Instead, it is initiated across a cluster of multiple closely spaced TSSs within a promoter [10-12]. It can also be initiated from TSSs located in separate clusters (alternative promoters). In fact, alternative transcription initiation driven by multiple promoter usage has a higher contribution to tissue-dependant isoform-specific diversity than alternative splicing [13]. It is estimated that 30-50% of human genes are regulated by alternative promoters active depending on the cell type, developmental stage [14], cellular environment, or disease stages [15]. And indeed, *AGAP2* is one of these genes, presenting two alternative promoters that lead to the production of two different protein isoforms with a differential N terminus (isoform 1 and 2) that confers them unique target specificity [16].

Several attempts have been made in the last decade to precisely identify TSSs and characterise core promoter features. Notably, the FANTOM consortium and the DataBase of Transcriptional Start Sites (DBTSS) have comprehensively captured the dynamically changing landscape of TSS selection by using mRNA cap-guided deep sequencing technologies [10,17]. These databases have facilitated genome-wide analyses of promoter architecture and highlighted widespread differences in TSS selection, identifying an average of 4 robust TSS clusters per gene [10]. In addition, other studies have also found cell-specific differential distribution of TSSs within a cluster [18]. This highlights a potential relevance in gene expression regulation. After all, a differential TSS selection will change the overall length of the mRNA 5' UTR, likely altering the presence of regulatory elements such as upstream open reading frames (uORF); upstream start codons (uAUG); RNA secondary structures; and internal ribosomal entries sites [19-24].

However, although previous studies have reported the translational impact of transcripts isoforms derived from multiple closely situated promoters [19,25-29], minor TSSs (alternative TSS selection within the same cluster) have been considered as nonadaptive and the product of molecular errors [30].

We demonstrate here that *AGAP2* (isoform 2) mRNA expression is differentially initiated from alternative TSSs within the same cluster in different cancer types, directly impacting on the mRNA translational efficiency. We used 5' RACE to determine the transcription start sites in prostate cancer and CML cell lines, finding that the transcripts with a slightly longer 5' UTR contained the consensus sequence for a G quadruplex (G4), a type of secondary structure. We demonstrated the formation of the G4 using circular dichroism and an in-house

developed immunoprecipitation approach that we have termed rG4IP (RNA G4 Immunoprecipitation) [31]. We also determined that the presence of the G4 in *AGAP2* 5' UTR has a direct impact on the translation efficiency, reducing the amount of mRNA associated to polysomes. But more importantly, we hypothesised that this differential TSS selection could be a more widely used mechanism to regulate the amount of protein produced in cells. To test this, we developed a bioinformatics pipeline to interrogate data from the FANTOM project [32] and from the NCI-60 microarray (GSE32474) and NCI-60 SWATH-MS databases [3,33], finding other genes behaving in a similar manner. And we validated our findings by testing and demonstrating that *HK1* expression can also be regulated through alternative TSS selection.

Together, we present here compelling data supporting an alternative mechanism to regulate cellular protein content by controlling transcription initiation within a single TSS cluster.

Results

AGAP2 mRNA levels correlate negatively to protein levels in some cancers

Our group had previously studied the regulation of *AGAP2* promoter in prostate cancer (PC) and chronic myeloid leukaemia (CML) cell lines and noted a stronger basal promoter activity in CML cells, that resulted in higher relative *AGAP2* mRNA levels when compared to levels in PC cell lines [4]. However, when AGAP2 protein levels were analysed, we observed a significant negative correlation (Pearson's R = -0.89, $P = 0.016$) between *AGAP2* mRNA and protein in both types of cancers (Figure 1A-C). In the CML cell lines (KU812, TCC-S, and KCL-22), *AGAP2* relative mRNA expression was higher, but the protein levels were lower compared to PC cell lines (DU145, PC3 and LNCaP). In addition, the opposite occurred in PC cell lines. This mismatch between AGAP2 mRNA and protein levels was also observed in other cancer types (Figure 1D and 1E) such as hepatocarcinoma (HepG2 cells), ovarian cancer (SKOV-3 cells) and acute myeloid leukaemia (cell lines KG1 and Kasumi). However, when considering all cell lines analysed together, the negative correlation between mRNA and protein was not as strong (Pearson's R = -0.64, $P = 0.011$) as when focusing only on levels present in CML and PC cells (Figure S1A), highlighting a specific cell line-dependent regulation of AGAP2 expression.

Post-translational mechanisms can account for reduced protein levels. To rule out an enhanced protein degradation in CML cell lines by the ubiquitin-proteasome pathway, we treated CML cells with the proteasomal inhibitors MG132 and bortezomib. At the concentrations used, the inhibitors increased the levels of ubiquitinated proteins in the CML cells (Figure 1F) but did not significantly modify AGAP2 protein levels compared to untreated controls (Figure 1G).

These results suggest that the amount of protein produced (rather than the degradation) was key to the differential AGAP2 expression in these two types of cancers. However, translation is a complex mechanism with several layers of regulation. We studied the levels of the rate-limiting translation initiation factors in PC and CML cell lines, but we found no differences that could support the disparity in AGAP2 translational output (Figure S1B). Furthermore, preliminary data of *AGAP2* mRNA association to polysomes indicated a differential behaviour in CML and PC cells, with *AGAP2* mRNA associating to ribosome heavier fractions in DU145 cells (Figure S1C) and we focused on exploring this variation further.

Differential *AGAP2* Transcription Start Site selection within a cluster leads to slightly different 5' UTRs in CML

The 5' and 3' untranslated regions (UTRs) of an mRNA play a very important role in regulating translation. Whilst the 3' UTR has a well characterised role in controlling mRNA stability and localisation [34], the 5' UTR allows for ribosome binding supporting cap-dependent translation. Structures or motifs in this region can exert a post-transcriptional control in gene expression and multiple transcription initiation within a core promoter has been previously highlighted [35,36]. Therefore, we decided to focus initially on this region. Sanger sequencing of the area upstream of the start codon did not reveal any cell line-specific mutation (Figure S2A) that could support differences in AGAP2 expression. Next, we used 5' RACE to map the transcription start site (TSS) for *AGAP2* in CML and PC cells to determine if the 5' UTRs were of equal length in these cell lines. We observed that transcription initiated from the same TSS cluster, but in KU812 cells (CML) the TSS was 35 nucleotides upstream compared to DU145 cells (PC) (Figure 2A). Interestingly, *in silico* studies suggested that those extra nucleotides contained the consensus for a G quadruplex (G4) structure.

To have a better understanding of TSS selection and distribution in CML and PC, we used 5' RLM RACE and performed Sanger sequencing of the RACE products. We found a differential TSS distribution in the cell lines, with a broader distribution and upstream TSSs more frequently noted in CML (KU812) (Figure 2B). As KU812 was one the cell lines included in the FANTOM project, we were able to compare the TSSs identified in our study with the start sites detected in the FANTOM CAGE database [32], observing a highly similar distribution (Figure 2C). Interestingly, the TSS distribution for *AGAP2* in KCL-22, another CML cell line also available in this database, showed a similar widespread distribution (Figure S2B). Next, we decided to study by qPCR the expression of the *AGAP2* mRNA containing the longer 5' UTR in all the cell lines. The selection of upstream TSSs has the potential to incorporate a G4 structure in the beginning of the 5' UTR (Figure 2D). When located within the first 50 bp of the 5' UTR, those structures have been found to affect ribosome binding and influence translation rates [37]. Using primers that would detect the incorporation of the extra nucleotides that formed the G4 structure, the *AGAP2* mRNA with the longer G4-containing 5' UTR was found to be significantly more abundant in CML cell lines (Figure 2E). This confirmed a cell-specific bias in *AGAP2* TSS selection within this single TSS cluster.

Presence of a G quadruplex (G4) structure in *AGAP2* longer 5' UTR

G4 structures can modulate gene expression [38]. Given the presence of a putative G4 consensus sequence in the *AGAP2* 5' UTR, incorporated due to the selection of an upstream TSS, we evaluated the formation of the G4 structure *in vitro* and *in vivo* using circular dichroism and in-house developed immunoprecipitation technique.

To confirm the sequence found in the longer *AGAP2* 5' UTR could form a G4 *in vitro*, RNA oligos that contained the sequence under study (5'-GGGCGGGCAGGGGCGGGG-3') or a mutant version (5'-GAGCGAGCAGAGGCGGGG-3') were prepared and their circular dichroism (CD) spectrum was obtained. The results obtained were characteristic of the formation of a parallel G4 in the presence of salts (Figure 3A, left panel). When the G4 consensus was destroyed by punctual Guanine to Adenine substitutions (mutant), the characteristic peaks in the spectra were no longer observed (Figure 3A, right panel).

The next step was to demonstrate that the G4 were formed *in vivo.* However, the detection of RNA G4 structures in living cells is challenging and different approaches have variable success rates [39]. Interestingly, a structure specific G4 antibody (BG4) that selectively binds both DNA and RNA G4 was generated relatively recently [40]. Using this antibody, we developed an RNA-specific G4 immunoprecipitation technique (rG4IP) to selectively enhance the detection of cytosolic mRNAs with G4 structures (Figure 3B) [31].

Using rG4IP, we obtained an enrichment of *AGAP2* mRNA in the BG4-pulled fraction compared to the negative IgG control, detected by qPCR (Figure 3C). We also observed BG4-mediated enrichment of *NRAS* and *MMP16* mRNAs, which are known to present 5' UTR G4 structures and were used here as positive controls [41,42]. *TBP* mRNA, which lacks a G4 consensus sequence in its entire mRNA, was used as a negative control. These results highlighted the effectiveness of our rG4IP technique and demonstrated the presence of native G4 structures in *AGAP2* mRNA.

However, an analysis of *AGAP2* mRNA sequence using the psqfinder web application [43] revealed other several potential G4 consensus sequences along its entire length, apart from the one predicted in the longer 5'UTR (Figure S3). Therefore, to detect the native G4 formation specifically in the longer 5' UTR of *AGAP2* mRNA, we performed rG4IP in DU145 cells transfected with either the empty bicistronic plasmid pcDNA3 RLuc Polires FLuc [44] or

the same plasmid with the *AGAP2* longer 5' UTR cloned in front of the *Renilla* Luciferase (RLuc) gene. The results showed a significant RLuc enrichment in the cells transfected with the plasmid containing the cloned *AGAP2* 5' UTR, unequivocally demonstrating the presence of G4 structures in that region (Figure 3D).

The G4 structure in *AGAP2* longer 5' UTR influences mRNA translation negatively

The presence of G4 structures in the 5' UTR has been previously shown to decrease mRNA translational efficiency [23]. To study the influence of these structures on *AGAP2* mRNA translation, we used the same bicistronic plasmid mentioned above [44]. We generated dual-luciferase reporter constructs comprising of either the shorter 5' UTR without the G4 forming sequences (found in PC cells), the longer 5' UTR containing G quadruplex forming sequences (found in CML cells), or a mutated version of the longer 5' UTR with the G4 consensus sequence destroyed (Figure 4A). These 5' UTR variants were fused to the *Renilla* luciferase (Rluc) open reading frame (ORF) under the control of the CMV promoter. The *Firefly* luciferase (Fluc) was used as an internal control because, whilst a single mRNA is generated containing both Rluc and Fluc, its independent translation was ensured through the presence of the poliovirus IRES (Cap independent translation) sequence.

Using *in vitro* transcription and translation, we observed that the plasmid with the longer 5' UTR, containing the G4 sequence, mediated a significant decrease in the luciferase reporter activity relative to the short UTR. As this effect was reversed in the G4 mutant (Figure 4B), these results confirmed the differential role for *AGAP2* longer 5' UTR in mRNA *in vitro* translation. Transfecting these 5' UTR constructs into DU145 (PC) and KU812 (CML) cell lines demonstrated similar shifts in relative reporter activity *in vivo* (Figure 4C). Although the pattern of relative luciferase activity was found to be similar, we noted that in the CML cell line (KU812), the impact of the longer 5' UTR was less profound compared to the PC cell line. However, this could be explained by the differences in the method of transfection used. The leukaemia cell lines are notoriously difficult to transfect, and an electroporation-based technique (nucleofection) was used to achieve optimal gene transfer [45]. However, nucleofection has been shown to induce nonspecific changes in the metabolic activity of the transfected cells and to alter the phosphorylation state of the translation initiation factor eIF2α [46-48]. These non-specific effects could impact on KU812 cells response, as observed by the loss of differences in luciferase activity at later time points post-transfection and the loss of luciferase activity differences in DU145 when using nucleofection (Figure S4).

As the variations in translation efficiency found for the longer and shorter AGAP2 5' UTRs could be attributed to differences in polysome seeding and occupancy [49], we examined the polysome association profiles of *AGAP2* mRNA with the longer 5' UTR. Interestingly, we noted a decreased polyribosome association of *AGAP2* mRNA with longer 5' UTR in the CML cell lines KU812 and TCC-S (Figure 4D), implying a lower translation rate for this mRNA population with a longer 5' UTR.

Together, these results highlight the negative influence the presence of the G4 structure has on *AGAP2* mRNA translation.

***AGAP2* expression regulation is not an isolated case**

Finding that *AGAP2* expression could be regulated based in an alternative TSS selection within the same TSS cluster, raised the question of whether this was an isolated example. To examine its relevance in other genes, we performed a bioinformatics analysis to find potential G4 sequences between alternative TSS isoforms within a single cluster (Figure 5A, see also Methods). We used data from the FANTOM project [32] to select the transcripts with differential TSS usage. Then, the nucleotide sequences between alternative TSSs were

extracted and analysed for the presence of potential G4s using the pqsfinder package in R [43]. We identified 4,920 transcripts associated with 3,888 genes that contained potential G4 sequences between the two transcription start positions in the defined TSS cluster, upstream of the major TSS. And the large majority (91.9%) of these transcripts were protein-coding. In order to identify enriched pathways that could be modulated by alternatively TSS selection, we used MetaCore pathways analysis. The top three significantly enriched pathways included cytoskeleton remodelling, apoptosis and survival, and development (Figure 5B).

Next, we analysed the distribution of these genes in PC and CML cell lines and found 1,007 genes that showed differential and cancer-type-specific distribution (similar to AGAP2, differential mRNA levels in both type of cancers). To identify suitable gene targets for validation, we used the NCI-60 dataset [3] to confirm RNA expression levels (microarray data) and contrasted them to their protein (SWATH-MS) levels, compiling a reduced list of genes that had inconsistencies in RNA and protein levels whilst presenting potential G4 sequences between alternate TSSs within a cluster (Figure 5C, See Methods). From this list, we selected the *HK1* gene (hexokinase 1, NM_033496.2) as it showed very large differences in RNA *vs* protein levels in CML cell lines relative to PC cell lines.

We tested *HK1* expression in our system and, as shown in Figure 6A, the relative mRNA levels were significantly higher in two of the CML cell lines included in the study (KU812 and TCC-S) whilst their HK I relative protein levels were lower (Figure 6B). Furthermore, when analysing the abundance of the mRNA with longer 5' UTR that contained the potential G4 sequence, the levels were significantly higher in both CML cell lines (Figure 6C). We were also able to detect an enrichment of *HK1* mRNA in the BG4 fraction after performing rG4IP (Figure 6D) and confirmed that *HK1* mRNAs presented with the longer 5' UTR preferentially associated with the non-polysomal fraction (Figure 6E) although this preference was not as striking as in *AGAP2*'s case.

Overall, these results confirmed *AGAP2* is not an isolated case for protein expression regulation mediated by the presence of a G4 associated to the selection of an upstream TSS within the same cluster and we proposed this mechanism (Figure 7) as an alternative mechanism for gene expression regulation.

Discussion

Alternative transcription initiation contributes to the transcriptomic diversity of eukaryotic organisms. It produces different transcript isoforms from a single gene that qualitatively and quantitatively differ in their ability to produce proteins [15,50,51]. However, studies investigating the regulatory role of alternative transcription initiation have focused so far on the transcript isoforms derived from alternative promoters. As a result, the consequence of differential TSSs selection within a single TSS cluster is currently poorly understood. Here, we have demonstrated a differential distribution of *AGAP2* TSSs within the same TSS cluster in prostate cancer (PC) and chronic myeloid leukaemia (CML) cell lines yielding a heterogeneous population of mRNAs with small nucleotide differences in their 5' UTRs. We have highlighted that these minor changes in 5' UTR lengths can lead to the presence of regulatory elements, G quadruplexes (G4) in our case, and influence mRNA translational efficiency.

During our studies on AGAP2 role and regulation (a proto-oncogene involved in several cancer cells survival [6,7,9]), we identified a shared minimal promoter region for *AGAP2* in PC and CML cells, observing significant differences in mRNA expression levels [4]. In the current study, we have demonstrated a negative correlation between *AGAP2* mRNA and protein levels in these cancers (Figure 1). However, there are many instances where mRNA levels do not correspond with protein levels [2]. The difference here is that we have also shown a differential distribution of TSSs for *AGAP2* in PC and CML cell lines. A look at the CAGE tags representing the TSSs for *AGAP2* in the FANTOM project [52] shows a broad distribution with a dominant peak (Figure 2A). But when we analysed the TSS selection in PC and CML cell lines, we observed a single dominant peak in DU145 (PC) and a broad distribution in KU812 and KCL-22 (CML) cell lines (Figure 2 and Figure 2S), with the distinctive presence of an upstream TSS and the consequent production of a longer 5' UTR in *AGAP2* mRNA on CML cell lines (Figure 2E). Tissue-specific TSS usage within a TSS cluster has been previously described even if the consequences were unknown [36], contributing to the notion that transcription initiation is precisely regulated at promoters. But despite hints of this TSS distribution change being linked to processes such as cell cycle phases [53], a clear role for this differential selection is still missing and the concept of transcriptional 'noise' remains [30].

As the selection of an upstream TSS in CML cell lines led to the presence of a longer 5' UTR that could easily be monitored in cells, we investigated a possible differential role on translation for this isoform when compared to the 5' UTR isoform generated from the dominant peak in PC cells ('shorter 5' UTR). However, it should be noted that this longer 5' UTR isoform represented a reduced percentage of the total of 5' UTR isoforms present in CML cells, both in our hands (Figure 2B) and in the FANTOM database (Figure 2C and Figure S2B).

Changes in translation efficiency on mRNAs with differential 5' UTRs are often due to specific sequences, with longer 5' UTRs associating generally with lower translation efficiencies [26,28]. One of the features that can account for residual variance in translation rates is the presence of alternatively transcribed G quadruplexes (G4s). G4s in the 5' UTRs are generally associated with suppressed translation [41,54]. However, there are also examples of increased translation when this structure is present [55].

Different approaches have been used to detect the RNA G4 structures inside the cells, including the use of G4-stabilising ligands/ions [56,57], small molecule probes [58], RNA structural mapping [59], reverse transcription stalling [56], RNA G4 structure-protein interactions [60], ligands with fluorescence activity [61], self-biotinylation methodology [62], and a G4-structure specific antibody [40]. Most of the methodologies mentioned above used specific ligands and/or

reactive small molecules that could shift the equilibrium in the favour of G4 formation and might not be representative of actual RNA G4 conformations in living cells. Therefore, we developed the rG4IP technique to selectively enrich cytosolic RNAs with G4s, not fixing the cells and incorporating a step to degrade any trace amount of genomic DNA [31]. Using circular dichroism and rG4IP, we were able to demonstrate the formation of a G4 structure in *AGAP2* longer 5' UTR (Figure 3). And, as described for other mRNAs [41,54], this structure was responsible for a reduced protein expression (Figure 4B-C) and a reduced polysome association (Figure 4D).

Next, we used a bioinformatics approach to detect other genes with protein levels negatively associated to the presence of a G4 and the selection of an upstream TSS within the same cluster. We identified a list of potential target genes implicated in key cellular pathways (Figure 5). However, it is likely that our approach might have missed many other targets as, for example, the SWATH-MS data for protein levels only included proteins common to all the cell lines in the NCI-60 database and could lead to the underestimation of targets. Still, we were able to validate this association for *HK1* expression (Figure 6), a key protein in glucose metabolism and implicated in neurodevelopmental abnormalities [63].

Our data supports TSS selection within a TSS cluster as a mechanism to modulate protein levels. And as the longer 5' UTR isoforms with the G4 are present when the levels of mRNA are higher, it would be interesting to explore their role as a mRNA reservoir ready to be translated under specific signals. Further research into TSS selection is also necessary as we know it can be influenced by different factors such as the type of promoter [35], methylation patterns [64], chromatin remodelling and histone modifications [65], but a detailed understanding would open new possibilities for protein expression manipulation.

In conclusion, when comparing CML and prostate cancer cell lines, our study has highlighted the relevance of TSS selection within a TSS cluster as a regulatory mechanism involving the differential formation of a G4 structure in the longer 5' UTR isoforms, altering mRNA translation efficiency and associating with lower protein expression levels.

Limitation of the study

The negative correlation between *AGAP2* mRNA and protein observed in Figure 1 cannot be fully explained by the reduced protein expression obtained by the presence of the longer 5' UTR, as this is not a major isoform in these cells. Therefore, the other 5' UTRs generated in CML cells will likely be contributors to this reduced protein output and it would be worth studying them for the presence of specific motives/structures. In particular uORFs, as there are several predicted functional uORFs between the start codon and the main TSS peak [66] that could become operative in the shorter 5' UTR isoforms.

Acknowledgments

The authors would like to thank Dr Yegor Doush for his initial contribution to the study of *AGAP2* mRNA and protein levels; and Prof Ellen E. Billet and Dr Stephanie McArdle for constructive comments during the research carried out. The authors acknowledge Nottingham Trent University for the facilities and Vice Chancellor's Bursary Scholarship received by A.A.S. to undertake this work. The authors also acknowledge that the Graphical Abstract was created with BioRender.com.

Author Contributions

Conceptualisation: C.M.D; Methodology: A.A.S and C.M.D.; Software: A.A.S.; Investigation: A.A.S., K.A.S., C.U. and C.M.D.; Writing – Original Draft: A.A.S and C.M.D.; Writing – Review and Editing: A.A.S., K.A.S., C.P. and C.M.D.; Supervision: K.A.S., C.P. and C.M.D.; Funding Acquisition: C.P. and C.M.D.

Declaration of Interests

The authors declare no competing interests.

Inclusion and diversity

We support inclusive, diverse, and equitable conduct of research.

Figure titles and legends

Figure 1. *AGAP2* mRNA and protein levels discrepancies. **(A)** *AGAP2* mRNA basal levels were measured in prostate cancer (PC) cell lines (DU145, PC3, LNCaP) and chronic myeloid leukaemia (CML) cell lines (KU812, TCC-S, KCL-22) by RT-qPCR. The values presented were normalised against the levels of the housekeeping gene *HPRT* and shown relative to the prostate cancer 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. **(B)** Representative image of AGAP2 protein levels detected by immunoblotting in CML and PC cell lines. β-Actin was used a loading control. Densitometry values for the relative protein expression are represented below the blots. Differences were analysed using Kruskal-Wallis [H (5) = 14.71, P =0.012] followed by uncorrected Dunn's test. **(C)** Strong negative correlation between AGAP2 mRNA (x-axis) and protein levels (y-axis) in PC and CML cell lines (Pearson's R= -0.89, p=0.016). The data presented is relative to DU145 (PC cell line). **(D, E)** *AGAP2* relative mRNA levels **(D)** and protein **(E)** in different cancer cell lines, assessed as described in **(A)** and **(B)**. Statistical analyses for mRNA levels in **(D)** were carried out by one-way ANOVA [F (9, 20) = 41.30, $P < 0.001$)] with post-hoc Sidak's multiple comparison tests. **(F,G)** Western blot analysis for the accumulation of ubiquitinated proteins (as positive control for the proteasomal inhibitors) and AGAP2 levels in CML cell lines treated with proteasomal inhibitors: MG132 [KU812 (5μM), TCC-S (5μM), KCL-22 (50μM) for 4 hours] and Bortezomib [KU812 (200nM), TCC-S (10nM), KCL-22 (100nM) for 6 hours]. β-Actin levels were used as a loading control. All data shown in the graphs in this figure are the mean ± SD from three independent experiments (performed in triplicate in the case of qPCRs); (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

Figure 2. Alternative TSS usage for *AGAP2* in PC and CML cell lines leads to differential 5' UTR isoforms. (A) Image derived from the ZENBU browser (http://fantom.gsc.riken.jp/zenbu/) showing the main TSSs in *AGAP2* and its frequency (height of peaks). 5' RACE was performed according to manufacturer's instructions using adenines for the tailing reaction. KU812, a CML cell line, presents with an upstream TSS that produces an *AGAP2* mRNA with a slightly longer (35 bp) 5' UTR than the one found in DU145, a PC cell line. In those extra nucleotides there is a repetition of guanine residues that fits the pattern predicted for the formation of G quadruplexes. **(B)** Comparison of the frequency of alternative TSSs used in DU145 (PC) and KU812 (CML) cell lines, mapped by 5' RLM-RACE. The relative frequencies (in percentages) are shown as bars placed at the nucleotide position upstream from the start codon (n = 10). **(C)** TSSs for AGAP2 in KU812 obtained by 5' RLM-RACE is plotted against the TSSs noted by the FANTOM CAGE database. **(D)** Cartoon representing *AGAP2* core promoter and its TSSs. The differential TSS selection creates slight differences in the length of the 5' UTR, with upstream/earlier TSSs resulting in longer 5' UTRs. The selection of an earlier TSS in CML KU812 produces an mRNA that encodes extra nucleotides in the 5' UTR containing the consensus for a G quadruplex structure. **(E)** Relative levels of the longer *AGAP2* 5' UTR containing the G quadruplex consensus sequence in PC and CML cell lines. The data represents the mean ± SD of three independent experiments. 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. (*$P < 0.05$; ***$P < 0.001$).

Figure 3. Presence of a G quadruplex structure in the longer *AGAP2* 5' UTR. (A) RNA oligos containing either the sequence corresponding to the G quadruplex consensus found

in the longer *AGAP2* 5' UTR or a mutated version were folded in the presence of 100 mM NaCl, 100 mM KCl, or no salts and its CD spectra represented here. The characteristic pattern of a parallel G quadruplex (G4) structure was noted, exhibiting a positive peak at ~260 nm and a negative peak at ~240 nm in the presence of salts (left). This pattern was lost in the mutant RNA oligo where key guanosines were changed to adenosines (right). **(B)** Overview of the RNA G quadruplex immunoprecipitation (rG4IP) technique [31]. Cells were treated with 25 µg/mL digitonin and the extracted cytoplasmic fraction precleared and incubated overnight with a structure-specific G4 antibody (BG4) bound to protein G magnetic beads. After incubation, the beads were washed, and the bound RNA eluted by unfolding the G4 by heating at 65 °C for 15 minutes. The eluent is treated with DNase I and analysed by RT-qPCR. **(C)** rG4IP was performed in the TCC-S (CML) cell line and the immunoprecipitated samples are 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 was used as a negative control as it lacks G4 consensus sequences in its entire mRNA. Differences between samples were analysed with unpaired t test. **(D)** rG4IP was performed in DU145 cells transfected with either an empty vector (with no 5' UTR) or the same vector containing *AGAP2* longer 5' UTR in front of the *Renilla* luciferase gene. The levels of *Renilla* mRNA in the immunoprecipitated samples were normalised by their input controls. An unspecific isotype antibody (IgG) was used as a negative control. Differences between samples were analysed by unpaired two-tailed t-tests. All the data shown in this figure correspond to three independent immunoprecipitations and the error bars denote standard deviation. (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

Figure 4. The G quadruplex (G4) structure in *AGAP2* longer 5' UTR influences mRNA translation negatively. (A) Schematic representation of the fragments cloned into the bicistronic luciferase reporter (pcDNA3 RLUC POLIRES FLUC) plasmid [44]. The *AGAP2* 5' UTR fragments (shorter 5' UTR, longer 5' UTR with G4 consensus, and longer 5' UTR with G4 consensus mutated) were inserted at the unique *Nhe*I restriction site proximal to the *Renilla* luciferase (Rluc) ORF. The Rluc is driven by cap-dependent mRNA translation through the cloned 5′ UTR. The *Firefly* luciferase cistron was used as an internal control for normalisation. **(B)** Relative luciferase activity of the *AGAP2* 5' UTR constructs measured using an *in vitro* transcription and translation system. The graph represents the mean of 4 independent experiments and data are expressed as the Rluc/Fluc ratio relative to the activity of the shorter 5' UTR. Differences were analysed using a Kruskal-Wallis [H (2) = 47.13, P =<0.001] followed by Mann–Whitney U test (***$P < 0.001$). **(C)** Relative luciferase activity after transfecting the different constructs in DU145 and KU812 cells. The luciferase activity was analysed 48 hours after transfection in DU145 and 6 hours after transfection in KU812. The graph represents the mean of three independent experiments performed in duplicate and expressed as relative Rluc/Fluc ratios. Differences between samples were analysed with a Kruskal Wallis test followed by the Mann–Whitney U test, ***$P < 0.001$. The bars represent the mean ± standard deviation. **(D)** Lysates for polysome profiling were prepared from KU812 and TCC-S cells and fractionated through a sucrose gradient. The profiles were monitored by measuring the absorbance at 254 nm ($A_{254\ nm}$). A representative polysome profile from a KU812 extraction is shown on the left. The relative distribution of the mRNA for *AGAP2* longer 5' UTR isoform (concentrated in the non-polysomal fractions) is shown on the right. The abundance of the RNA detected per fraction is presented as the percentage of the total RNA. mRNA levels were normalised to exogenous spike-in luciferase control mRNA. The graph represents the mean ± SEM of two independent experiments.

Figure 5. Identification of potential targets regulated in a similar manner to AGAP2. (A) Workflow diagram used to identify genes whose expression could be regulated by alternative selection of a TSSs, involving the presence of a G quadruplex (G4). First, the FANTOM database was used to identify all the transcripts that contained a G quadruplex (G4) consensus sequence between alternative TSSs within their defined TSS cluster. The G4 consensus sequences were identified using the pqsfinder package in R [67]. The FANTOM database was also used to detect differential expression in PC (DU145, PC3) and CML (KU812, K562, KCL-22) cell lines for those genes that would encode a G4 consensus between alternative TSSs. Microarray and SWATH-MS data from the NCI-60 database were integrated to characterise genes that demonstrated discrepancies between their mRNA and protein levels (high mRNA and low protein) within those genes showing a differential 5' UTRs with G4 sequences. **(B)** Metacore pathway enrichment analysis of mRNAs with alternative 5' UTRs that contain G4 consensus sequences. 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) **(C)** Venn diagram illustrating the overlapping genes in the FANTOM and NCI-60 databases showing differential 5' UTRs with G4 consensus, with differentially expressed mRNA (≥ 1 log FC), and either no statistically significant differences in protein levels or significantly lower proteins levels in CML cell lines (left) or PC cell lines (right). The differential expression and TSS distribution were computed by linear modelling followed by empirical Bayes statistics.

Figure 6. *HK1* expression is also regulated by an alternative TSSs and mediation of a G4 structure. (A) *HK1* mRNA basal levels were detected in prostate cancer (PC) cell lines (DU145, PC3 and LNCaP) and chronic myeloid leukaemia (CML) cell lines (KU812, TCC-S and KCL-22) by RT-qPCR. *HK1* expression was normalised against levels for the housekeeping gene *HPRT* and it is shown relative to levels in the PC cell line 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 (***$P < 0.001$). The error bars denote standard deviation. **(B)** HK I protein levels were detected by western blot. The graph below shows overall densitometry values relative to those in DU145 cell line. **(C)** Relative expression levels for the *HK1* isoform with the longer 5' UTR containing a G4 consensus sequence, detected by RT-qPCR in PC and CML cell lines. The bars represent the mean ± SD of three independent experiments. 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 (*$P < 0.05$; **$P < 0.01$). **(D)** rG4IP followed by *HK1* detection by RT-qPCR. Expression levels were normalised by input control, and the data presented correspond to three independent immunoprecipitations. The error bars denote standard deviation. Differences between samples were analysed by unpaired two-tailed t-tests, P-values shown. **(E)** Polysomal fractionation: relative abundance of the *HK1* mRNA with the longer 5' UTR in polysome fractions in TCC-S cells (left) and KU812 cells (right).

Figure 7. Model for an alternative regulatory mechanism. (A) The selection of an earlier (upstream) TSS within a TSS cluster results in a slightly longer 5' UTR that contains a G quadruplex (G4) structure. This G4 structure decreases the translational efficiency of the mRNA possibly by impeding ribosome scanning, decreasing the formation of polysomes, and resulting in a reduced translational output. **(B)** The selection of a downstream TSS yields a shorter 5' UTR without the G4 sequence. This mRNA isoform prominently associates with ribosomes, forming polysomes and increasing translation efficiency.

STAR Methods

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Dr Cristina Montiel-Duarte (cristina.montielduarte@ntu.ac.uk).

Materials Availability

Plasmids generated in this study will be deposited to Addgene.

Data and code availability

- All data reported in this paper will be shared by the lead contact upon request. However, this paper also analyses existing, publicly available data. The accession numbers for these datasets are listed in the key resources table.
- All original code is available in this paper's supplemental information (scripts S1-S3).
- Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines and culture conditions

DU145 (RRID:CVCL_0105), HEPG2 (RRID:CVCL_0027) , HuH7(RRID:CVCL_0336), and U-2 OS (RRID:CVCL_0042) were cultivated in DMEM GlutaMAX supplemented with 10% FBS. LNCaP (ATCC), KU812 (RRID:CVCL_0379), TCC-S [69], KCL-22 (ATCC), KASUMI-1 (RRID:CVCL_0589), and RAJI (RRID:CVCL_0511) were cultured in RPMI supplemented with 2 mM L-Glutamine and 10% FBS. PC3 (RRID:CVCL_0035) was grown in DMEM/F12 containing 2 mM L-Glutamine and 10% FBS. MCF-7 (RRID:CVCL_0031) was cultured in DMEM GlutaMAX supplemented with 10% FBS and 0.01 mg/mL human recombinant insulin. KG1 (RRID:CVCL_0374) was cultivated in Iscove's Modified Dulbecco's Medium, 2mM Glutamine, and 20% FBS. All cell lines were maintained at 37°C in a 5% CO_2 incubator and tested negative for mycoplasma contamination.

METHOD DETAILS

Protein extraction and western blot analysis

For total protein extraction and western blot analysis, cells were lysed in ice-cold RIPA buffer [50 mM Tris-Cl (pH 8.0), 1% NP-40, 1% sodium deoxycholate, 0.1% SDS, 150 mM NaCl] supplemented with protease inhibitors (Roche). The lysate was 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 amount of protein was quantified using the Pierce BCA Protein Assay Kit (ThermoFisher). Typically, 50 µg of total protein in Laemmli buffer [2% SDS, 10% glycerol, 50 mM Tris-HCl (pH 6.8), bromophenol blue 0.02%, 1% β-mercaptoethanol] was heated to 95°C for 5 minutes and separated by SDS-PAGE and subsequently transferred to Amersham Protran 0.2 µm nitrocellulose membrane (GE Healthcare). The membrane was

blocked with 5% non-fat dry milk in TBST [20 mM Tris-HCl (pH 7.6), 150 mM NaCl, 0.1% Tween20] and probed with indicated primary antibodies overnight at 4°C. Membranes were then washed with TBST three times for 10 min at room temperature followed by incubation for 1 hour with the appropriate secondary antibody. The membrane was washed again three times and signals were detected using ECL Western Blot Substrate (BioRad) and the luminescent image analyser LAS-4000 (Fujifilm).

RNA extraction and Real-time Quantitative PCR

Total RNA from cell lines were isolated by ReliaPrep RNA Miniprep System (Promega) according to the manufacturer's protocol. 2 µg of the total RNA was reverse transcribed using M-MLV Reverse Transcriptase (Promega) with Random hexamers (Promega). The quantitative real-time PCR (qPCR) was performed in triplicate using GoTaq qPCR SYBR master mix (Promega) on the Rotor-Gene Q real-time PCR cycler (Qiagen). The expression levels of *AGAP2* and *HK1* were normalized to the house-keeping gene (HPRT). Primer sequences are presented in Table S1. The relative gene expression levels were calculated using the comparative Ct method ($2^{-\Delta\Delta Ct}$) [73]. For amplification of *AGAP2* 5' UTR isoforms, a nested PCR with outer and inner forward and reverse primers were used (Table S1). The first-round PCR products were diluted 50-fold as the template for the second-round of qPCR.

5' RNA ligase-mediated rapid amplification of cDNA ends

The 5' RNA ligase-mediated rapid amplification of cDNA ends (5' RLM-RACE) [GeneRacer kit (ThermoFisher)] was performed following the manufacturer's instructions. Briefly, 3 µg of total RNA were treated with calf intestinal phosphatase and tobacco acid pyrophosphatase 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 primers (Table S1). The cDNA was amplified using the adaptor and gene-specific nested primers (Table S1) and purified by 2% agarose gel electrophoresis. The purified product was cloned for sequencing using TOPO TA Cloning Kit (ThermoFisher) and, at least, ten independent clones were sequenced for each cell line by Sanger Sequencing (Source Bioscience).

Plasmid Constructs, Transient Transfection, and Dual luciferase Reporter assay

The plasmid (pcDNA3 RLUC POLIRES FLUC) was a gift from Nahum Sonenberg (Addgene pcDNA3; RRID:Addgene_45642). The 5' UTR isoforms (shorter, longer and mutated longer) were designed and purchased from GeneScript (Hong Kong) (Table S2). The 5' UTR isoforms and the plasmid were digested with *Nhe*I and the products were separated in a 1% agarose gel, purified with Wizard SV kit (Promega), and treated with alkaline phosphatase (Promega). 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 with 100 µg/mL Ampicillin (Sigma-Aldrich). Positive clones were chosen, purified, and confirmed by Sanger sequencing (Source Bioscience).

Reporter constructs were transfected into the PC cell line (DU145) using JetPRIME transfection reagent (Polyplus) according to the manufacturer's protocol. Cells were seeded at a density of 2.5 x 10^5 cells/ well in 6-well plates for 24 hours before transfection and collected 48 hours after being transfected. For transient transfection of the CML cell line (KU812), 1 µg of each reporter plasmid was electroporated into 2 X 10^6 cells using Amaxa Cell Line Nucleofector solution V (Lonza), program X-001, and collected after 6 hours.

After the indicated time points, cells were lysed with Passive Lysis Buffer (Promega) and their luciferase activities were measured using the Dual-Luciferase Reporter Assay System (Promega) on a CLARIOstar microplate reader (BMG Labtech). The *Firefly* luciferase activity was used as an internal normalising control.

In vitro Transcription and Translation assay

The plasmid constructs were transcribed *in vitro* by T7 RNA polymerase using mMESSAGE mMACHINE Transcription Kit (Thermo Scientific) following the manufacturer's guidelines and precipitated using Lithium chloride (ThermoFisher). The resulting RNAs were translated *in vitro* using Flexi Rabbit Reticulocyte Lysate Translation System (Promega). The RNA was translated for 90 min at 30°C and the luciferase activity of the translation products was analysed using Dual-Luciferase Reporter Assay, as described above.

Circular Dichroism Spectroscopy

The Circular dichroism (CD) experiments were performed on 5 µM of RNA oligos (See KRT) 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. The data presented are an average of six spectral scans with baseline buffer correction.

Polysome fractionation

Prior to harvesting, 25 x 10^6 cells were treated with 100 µg/mL cycloheximide (CHX) and incubated at 37°C and 5% CO_2 for 10 min. The cells were washed and resuspended in 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 incubated for 10 minutes on ice followed by centrifugation at 12,000 x g for 10 minutes at 4°C to pellet the nuclei and debris. The RNA supernatant was layered on the top of a 10-50% sucrose 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. 500 µL of each sucrose fractions were collected and RNA was isolated using TRIzol extraction. Briefly, each fraction was resuspended in TRIzol (ThermoFisher) and chloroform (Sigma-Aldrich), mixed vigorously and centrifuged at 13,000 x g for 15 min at 4°C to separate into 3 layers. The RNA in the top aqueous layer was precipitated using ethanol and 3M sodium acetate (pH 5.2) and spiked with 500 ng of *in vitro* transcribed *Renilla* luciferase RNA control. The RNA was further cleaned and concentrated using ReliaPrep RNA Miniprep Systems (Promega). The samples were reverse transcribed and amplified using qPCR, as mentioned above.

RNA G quadruplex immunoprecipitation (rG4IP) [31]

Briefly, rG4IP was performed with a structure-specific G quadruplex (BG4) antibody (Absolute Antibody). TCC-S cells (15 X 10^6) were collected, washed with ice-cold PBS, and resuspended in ice-cold lysis buffer (150mM KCL, 50mM HEPES, 25µg/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 min 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. Precleared lysates were incubated overnight with 3 µg of BG4 antibody bound to Protein G magnetic beads (Biorad). After incubation, the beads were magnetised, washed thrice with lysis buffer, and eluted by incubating 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.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistics

All statistical analysis was performed using GraphPad Prism software (version 8). 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. The data were considered to be statistically significant if $P < 0.05$ and this is indicated in the figure legends by asterisks (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

Mapping G quadruplex consensus sequences between alternative TSSs

The normalised 5' CAGE tag density (Tags Per Million - TPM) data for all the available human samples in the FANTOM database was downloaded from the ZENBU genome browser [32]. The CAGE tag starting sites mapped to a +/- 50 base pair region around the annotated transcription start sites for each gene transcripts were selected. The TSS annotations were downloaded from Ensembl GRCh38.p13 (release 98) [74].The tags with a normalised density less than 2 TPM were removed to select robust CAGE tag starting sites. The overlapping CAGE tags around the annotated TSS, as defined above, were considered as part of a single cluster. The sequence between CAGE tags with the highest TPM and furthermost upstream tag within the same cluster were extracted for all the transcripts using Bio. Entrez module in Biopython [75] (Script S1). The sequences were then analysed for the presence of G quadruplex consensus using pqsfinder package in R [67] (Script S2). For analysing the differential distribution of G quadruplex forming TSSs in PC and CML cell lines, the 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 proportions of G quadruplex forming TSSs were estimated by dividing the numbers of tags within a 21 bp subregion upstream of the G quadruplex starting position and the total tags in the selected TSS cluster. 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 [67,76]. The differential distribution was computed by linear modelling and empirical Bayes approach using the Limma package in R (Script S3 for data wrangling and analysis).

Identification of genes with a discrepancy in mRNA and Protein levels:

To identify gene showing discrepancies in mRNA and protein level as noted for AGAP2, the NCI-60 microarray data (GSE64674) and NCI-60 SWATH-MS database were used [33,68]. The differentially expressed RNAs in PC (DU145, PC3) and CML (K562) were analysed using GEO2R (NCBI) (Script S3). The significant differences in protein mass spectral intensity values were evaluated using Limma (Script S3). The genes with differential RNA expression 2-fold or greater (RNA logFC ≥ 1) and no statistically significant differences in protein levels and/or significantly lower proteins levels were considered to have a discrepancy in mRNA and protein levels. The genes with discrepancies were analysed for the presence of

alternative 5’ UTRs with G quadruplex consensus sequences to identify targets for validation.

Pathway Analysis

Functional pathway maps of genes with alternatively transcribed G quadruplex structure was created using Metacore (Clarivate Analytics). The ranked hypergeometric test was used to determine enriched pathways and processes.

References

1. Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. FEBS Lett *24,* 3966-3973. DOI: 10.1016/j.febslet.2009.10.036.

2. Ghaemmaghami, S., Huh, W., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *6959,* 737-741. DOI: 10.1038/nature02046.

3. Guo, Y., Xiao, P., Lei, S., Deng, F., Xiao, G.G., Liu, Y., Chen, X., Li, L., Wu, S., Chen, Y. *et al*. (2008). How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes. Acta Biochim Biophys Sin (Shanghai) *5,* 426-436. DOI: 10.1111/j.1745-7270.2008.00418.x.

4. Doush, Y., Surani, A.A., Navarro-Corcuera, A., McArdle, S., Billett, E.E., and Montiel-Duarte, C. (2019). SP1 and RARα regulate AGAP2 expression in cancer. Sci Rep *1,* 390. DOI: 10.1038/s41598-018-36888-x.

5. Navarro-Corcuera, A., López-Zabalza, M.J., Martínez-Irujo, J.J., Álvarez-Sola, G., Ávila, M.A., Iraburu, M.J., Ansorena, E., and Montiel-Duarte, C. (2019). Role of AGAP2 in the profibrogenic effects induced by TGFβ in LX-2 hepatic stellate cells. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research *4,* 673-685. DOI: 10.1016/j.bbamcr.2019.01.008.

6. Cai, Y., Wang, J., Li, R., Ayala, G., Ittmann, M., and Liu, M. (2009). GGAP2/PIKE-a directly activates both the Akt and nuclear factor-kappaB pathways and promotes prostate cancer progression. Cancer Res *3,* 819-827. DOI: 10.1158/0008-5472.CAN-08-2537.

7. Ahn, J., Hu, Y., Kroll, T.G., Allard, P., and Ye, K. (2004). PIKE-A is amplified in human cancers and prevents apoptosis by up-regulating Akt. Proc Natl Acad Sci U S A *18,* 6993-6998. DOI: 10.1073/pnas.0400921101.

8. Tse, M.C.L., Liu, X., Yang, S., Ye, K., and Chan, C.B. (2013). Fyn regulates adipogenesis by promoting PIKE-A/STAT5a interaction. Mol Cell Biol *9,* 1797-1808. DOI: 10.1128/MCB.01410-12.

9. Liu, X., Hu, Y., Hao, C., Rempel, S.A., and Ye, K. (2007). PIKE-A is a proto-oncogene promoting cell growth, transformation and invasion. Oncogene *34,* 4918-4927. DOI: 10.1038/sj.onc.1210290.

10. Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, Michiel J. L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., Itoh, M. *et al*. (2014). A promoter-level mammalian expression atlas. Nature *7493,* 462-470. DOI: 10.1038/nature13182.

11. Karlsson, K., Lönnerberg, P., and Linnarsson, S. (2017). Alternative TSSs are co-regulated in single cells in the mouse brain. Mol Syst Biol *5,* 930. DOI: 10.15252/msb.20167374.

12. Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Morishita, S. *et al*. (2001). Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep *5,* 388-393. DOI: 10.1093/embo-reports/kve085.

13. Reyes, A., and Huber, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. Nucleic Acids Res *2,* 582-592. DOI: 10.1093/nar/gkx1165.

14. Danks, G.B., Navratilova, P., Lenhard, B., and Thompson, E.M. (2018). Distinct core promoter codes drive transcription initiation at key developmental transitions in a marine chordate. BMC Genomics *1,* 164. DOI: 10.1186/s12864-018-4504-5.

15. Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C., and Huang, T.H.-. (2008). The functional consequences of alternative promoter use in mammalian genomes. Trends Genet *4,* 167-177. DOI: 10.1016/j.tig.2008.01.008.

16. Ahn, J., Rong, R., Kroll, T.G., Van Meir, E.G., Snyder, S.H., and Ye, K. (2004). PIKE (phosphatidylinositol 3-kinase enhancer)-A GTPase stimulates Akt activity and mediates cellular invasion. J Biol Chem *16,* 16441-16451. DOI: 10.1074/jbc.M312175200.

17. Yamashita, R., Wakaguri, H., Sugano, S., Suzuki, Y., and Nakai, K. (2010). DBTSS provides a tissue specific dynamic view of Transcription Start Sites. Nucleic Acids Res *Database issue,* 98. DOI: 10.1093/nar/gkp1017.

18. Ohmiya, H., Vitezic, M., Frith, M.C., Itoh, M., Carninci, P., Forrest, A.R.R., Hayashizaki, Y., and Lassmann, T. (2014). RECLU: a pipeline to discover reproducible transcriptional start sites and their alternative regulation using capped analysis of gene expression (CAGE). BMC Genomics 269. DOI: 10.1186/1471-2164-15-269.

19. Dieudonné, F., O'Connor, P.B.F., Gubler-Jaquier, P., Yasrebi, H., Conne, B., Nikolaev, S., Antonarakis, S., Baranov, P.V., and Curran, J. (2015). The effect of heterogeneous Transcription Start Sites (TSS) on the translatome: implications for the mammalian cellular phenotype. BMC Genomics 986. DOI: 10.1186/s12864-015-2179-8.

20. Hollerer, I., Barker, J.C., Jorgensen, V., Tresenrider, A., Dugast-Darzacq, C., Chan, L.Y., Darzacq, X., Tjian, R., Ünal, E., and Brar, G.A. (2019). Evidence for an Integrated Gene Repression Mechanism Based on mRNA Isoform Toggling in Human Cells. G3 (Bethesda) *4,* 1045-1053. DOI: 10.1534/g3.118.200802.

21. Pickering, B.M., and Willis, A.E. (2005). The implications of structured 5' untranslated regions on translation and disease. Semin Cell Dev Biol *1,* 39-47. DOI: 10.1016/j.semcdb.2004.11.006.

22. Pozner, A., Goldenberg, D., Negreanu, V., Le, S.Y., Elroy-Stein, O., Levanon, D., and Groner, Y. (2000). Transcription-coupled translation control of AML1/RUNX1 is mediated by cap- and internal ribosome entry site-dependent mechanisms. Mol Cell Biol *7,* 2297-2307. DOI: 10.1128/MCB.20.7.2297-2307.2000.

23. Bugaut, A., and Balasubramanian, S. (2012). 5'-UTR RNA G-quadruplexes: translation regulation and targeting. Nucleic Acids Res *11,* 4727-4741. DOI: 10.1093/nar/gks068.

24. Blaschke, R.J., Töpfer, C., Marchini, A., Steinbeisser, H., Janssen, J.W.G., and Rappold, G.A. (2003). Transcriptional and translational regulation of the Leri-Weill and Turner syndrome homeobox gene SHOX. J Biol Chem *48,* 47820-47826. DOI: 10.1074/jbc.M306685200.

25. Zeitz, M.J., Calhoun, P.J., James, C.C., Taetzsch, T., George, K.K., Robel, S., Valdez, G., and Smyth, J.W. (2019). Dynamic UTR Usage Regulates Alternative Translation to Modulate Gap Junction Formation during Stress and Aging. Cell Rep *9,* 2737-2747.e5. DOI: 10.1016/j.celrep.2019.04.114.

26. Wang, X., Hou, J., Quedenau, C., and Chen, W. (2016). Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. Mol Syst Biol *7,* 875. DOI: 10.15252/msb.20166941.

27. Li, H., Bai, L., Li, H., Li, X., Kang, Y., Zhang, N., Sun, J., and Shao, Z. (2019). Selective translational usage of TSS and core promoters revealed by translatome sequencing. BMC Genomics *1,* 282. DOI: 10.1186/s12864-019-5650-0.

28. Rojas-Duran, M.F., and Gilbert, W.V. (2012). Alternative transcription start site selection leads to large differences in translation activity in yeast. RNA *12,* 2299-2305. DOI: 10.1261/rna.035865.112.

29. Arribere, J.A., and Gilbert, W.V. (2013). Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. Genome Res *6,* 977-987. DOI: 10.1101/gr.150342.112.

30. Xu, C., Park, J., and Zhang, J. (2019). Evidence that alternative transcriptional initiation is largely nonadaptive. PLoS Biol *3,* e3000197. DOI: 10.1371/journal.pbio.3000197.

31. Surani, A.A., and Montiel-Duarte, C. (2022). Native RNA G quadruplex immunoprecipitation (rG4IP) from mammalian cells. STAR Protoc *2,* 101372. DOI: 10.1016/j.xpro.2022.101372.

32. Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S. *et al*. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol 22. DOI: 10.1186/s13059-014-0560-6.

33. Reinhold, W.C., Sunshine, M., Varma, S., Doroshow, J.H., and Pommier, Y. (2015). Using CellMiner 1.6 for Systems Pharmacology and Genomic Analysis of the NCI-60. Clin Cancer Res *17,* 3841-3852. DOI: 10.1158/1078-0432.CCR-15-0335.

34. Mayr, C. (2019). What Are 3' UTRs Doing? Cold Spring Harb Perspect Biol *10,* DOI: 10.1101/cshperspect.a034728.

35. 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. *et al*. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet *6,* 626-635. DOI: 10.1038/ng1789.

36. Kawaji, H., Frith, M.C., Katayama, S., Sandelin, A., Kai, C., Kawai, J., Carninci, P., and Hayashizaki, Y. (2006). Dynamic usage of transcription start sites within core promoters. Genome Biol *12,* R118. DOI: 10.1186/gb-2006-7-12-r118.

37. Kumari, S., Bugaut, A., and 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 *48,* 12664-12669. DOI: 10.1021/bi8010797.

38. Holder, I.T., and Hartig, J.S. (2014). A matter of location: influence of G-quadruplexes on Escherichia coli gene expression. Chem Biol *11,* 1511-1521. DOI: 10.1016/j.chembiol.2014.09.014.

39. Kharel, P., Becker, G., Tsvetkov, V., and Ivanov, P. (2020). Properties and biological impact of RNA G-quadruplexes: from order to turmoil and back. Nucleic Acids Res *22,* 12534-12555. DOI: 10.1093/nar/gkaa1126.

40. Biffi, G., Tannahill, D., McCafferty, J., and Balasubramanian, S. (2013). Quantitative visualization of DNA G-quadruplex structures in human cells. Nat Chem *3,* 182-186. DOI: 10.1038/nchem.1548.

41. Kumari, S., Bugaut, A., Huppert, J.L., and Balasubramanian, S. (2007). An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation. Nat Chem Biol *4,* 218-221. DOI: 10.1038/nchembio864.

42. Morris, M.J., and Basu, S. (2009). An unusually stable G-quadruplex within the 5'-UTR of the MT3 matrix metalloproteinase mRNA represses translation in eukaryotic cells. Biochemistry *23,* 5313-5319. DOI: 10.1021/bi900498z.

43. Labudová, D., Hon, J., and Lexa, M. (2020). pqsfinder web: G-quadruplex prediction using optimized pqsfinder algorithm. Bioinformatics *8,* 2584-2586. DOI: 10.1093/bioinformatics/btz928.

44. Poulin, F., Gingras, A.C., Olsen, H., Chevalier, S., and Sonenberg, N. (1998). 4E-BP3, a new member of the eukaryotic initiation factor 4E-binding protein family. J Biol Chem *22,* 14002-14007. DOI: 10.1074/jbc.273.22.14002.

45. Schakowski, F., Buttgereit, P., Mazur, M., Märten, A., Schöttker, B., Gorschlüter, M., and Schmidt-Wolf, I.G. (2004). Novel non-viral method for transfection of primary leukemia cells and cell lines. Genet Vaccines Ther *1,* 1. DOI: 10.1186/1479-0556-2-1.

46. Anderson, B.R., Karikó, K., and Weissman, D. (2013). Nucleofection induces transient eIF2α phosphorylation by GCN2 and PERK. Gene Ther *2,* 136-142. DOI: 10.1038/gt.2012.5.

47. Mello de Queiroz, F., Sánchez, A., Agarwal, J.R., Stühmer, W., and Pardo, L.A. (2012). Nucleofection induces non-specific changes in the metabolic activity of transfected cells. Mol Biol Rep *3,* 2187-2194. DOI: 10.1007/s11033-011-0967-z.

48. Zhang, M., Ma, Z., Selliah, N., Weiss, G., Genin, A., Finkel, T.H., and Cron, R.Q. (2014). The impact of Nucleofection® on the activation state of primary human CD4 T cells. J Immunol Methods 123-131. DOI: 10.1016/j.jim.2014.05.014.

49. Van Der Kelen, K., Beyaert, R., Inzé, D., and De Veylder, L. (2009). Translational control of eukaryotic gene expression. Crit Rev Biochem Mol Biol *4,* 143-168. DOI: 10.1080/10409230902882090.

50. Landry, J., Mager, D.L., and Wilhelm, B.T. (2003). Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet *11,* 640-648. DOI: 10.1016/j.tig.2003.09.014.

51. Dik, E., Naamati, A., Asraf, H., Lehming, N., and Pines, O. (2016). Human Fumarate Hydratase Is Dual Localized by an Alternative Transcription Initiation Mechanism. Traffic *7,* 720-732. DOI: 10.1111/tra.12397.

52. Lizio, M., Abugessaisa, I., Noguchi, S., Kondo, A., Hasegawa, A., Hon, C.C., de Hoon, M., Severin, J., Oki, S., Hayashizaki, Y. *et al*. (2019). Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. Nucleic Acids Research *D1,* D752-D758. DOI: 10.1093/nar/gky1099.

53. Wragg, J.W., Roos, L., Vucenovic, D., Cvetesic, N., Lenhard, B., and Müller, F. (2020). Embryonic tissue differentiation is characterized by transitions in cell cycle dynamic-associated core promoter regulation. Nucleic Acids Res *15,* 8374-8392. DOI: 10.1093/nar/gkaa563.

54. Lee, D.S.M., Ghanem, L.R., and Barash, Y. (2020). Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. Nat Commun *1,* 527. DOI: 10.1038/s41467-020-14404-y.

55. Agarwala, P., Pandey, S., Mapa, K., and Maiti, S. (2013). The G-quadruplex augments translation in the 5' untranslated region of transforming growth factor β2. Biochemistry *9,* 1528-1538. DOI: 10.1021/bi301365g.

56. Kwok, C.K., and Balasubramanian, S. (2015). Targeted Detection of G-Quadruplexes in Cellular RNAs. Angew Chem Int Ed Engl *23,* 6751-6754. DOI: 10.1002/anie.201500891.

57. Biffi, G., Di Antonio, M., Tannahill, D., and Balasubramanian, S. (2014). Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells. Nat Chem *1,* 75-80. DOI: 10.1038/nchem.1805.

58. Yang, S.Y., Lejault, P., Chevrier, S., Boidot, R., Robertson, A.G., Wong, J.M.Y., and Monchaud, D. (2018). Transcriptome-wide identification of transient RNA G-quadruplexes in human cells. Nat Commun *1,* 4730. DOI: 10.1038/s41467-018-07224-8.

59. Guo, J.U., and Bartel, D.P. (2016). RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. Science *6306,* DOI: 10.1126/science.aaf5371.

60. Herdy, B., Mayer, C., Varshney, D., Marsico, G., Murat, P., Taylor, C., D'Santos, C., Tannahill, D., and 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 *21,* 11592-11604. DOI: 10.1093/nar/gky861.

61. Chen, X., Chen, S., Dai, J., Yuan, J., Ou, T., Huang, Z., and Tan, J. (2018). Tracking the Dynamic Folding and Unfolding of RNA G-Quadruplexes in Live Cells. Angew Chem Int Ed Engl *17,* 4702-4706. DOI: 10.1002/anie.201801999.

62. Einarson, O.J., and Sen, D. (2017). Self-biotinylation of DNA G-quadruplexes via intrinsic peroxidase activity. Nucleic Acids Res *17,* 9813-9822. DOI: 10.1093/nar/gkx765.

63. 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. *et al*. (2019). De novo variants in HK1 associated with neurodevelopmental abnormalities and visual impairment. Eur J Hum Genet *7,* 1081-1089. DOI: 10.1038/s41431-019-0366-9.

64. Wilkinson, E.J., Woodworth, A.M., Parker, M., Phillips, J.L., Malley, R.C., Dickinson, J.L., and Holloway, A.F. (2020). Epigenetic regulation of the ITGB4 gene in prostate cancer. Exp Cell Res *2,* 112055. DOI: 10.1016/j.yexcr.2020.112055.

65. Niederacher, G., Klopf, E., and Schüller, C. (2011). Interplay of dynamic transcription and chromatin remodeling: lessons from yeast. Int J Mol Sci *8,* 4758-4769. DOI: 10.3390/ijms12084758.

66. McGillivray, P., Ault, R., Pawashe, M., Kitchen, R., Balasubramanian, S., and Gerstein, M. (2018). A comprehensive catalog of predicted functional upstream open reading frames in humans. Nucleic Acids Res *7,* 3326-3338. DOI: 10.1093/nar/gky188.

67. Hon, J., Martínek, T., Zendulka, J., and Lexa, M. (2017). pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. Bioinformatics *21,* 3373-3379. DOI: 10.1093/bioinformatics/btx413.

68. Guo, T., Luna, A., Rajapakse, V.N., Koh, C.C., Wu, Z., Liu, W., Sun, Y., Gao, H., Menden, M.P., Xu, C. *et al*. (2019). Quantitative Proteome Landscape of the NCI-60 Cancer Cell Lines. iScience 664-680. DOI: 10.1016/j.isci.2019.10.059.

69. Van, P.N.T., Xinh, P.T., Kano, Y., Tokunaga, K., and 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 *1,* 25-33. DOI: 10.1111/j.1749-0774.2005.tb00054.x.

70. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *12,* 550. DOI: 10.1186/s13059-014-0550-8.

71. Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res *22,* 10881-10890. DOI: 10.1093/nar/16.22.10881.

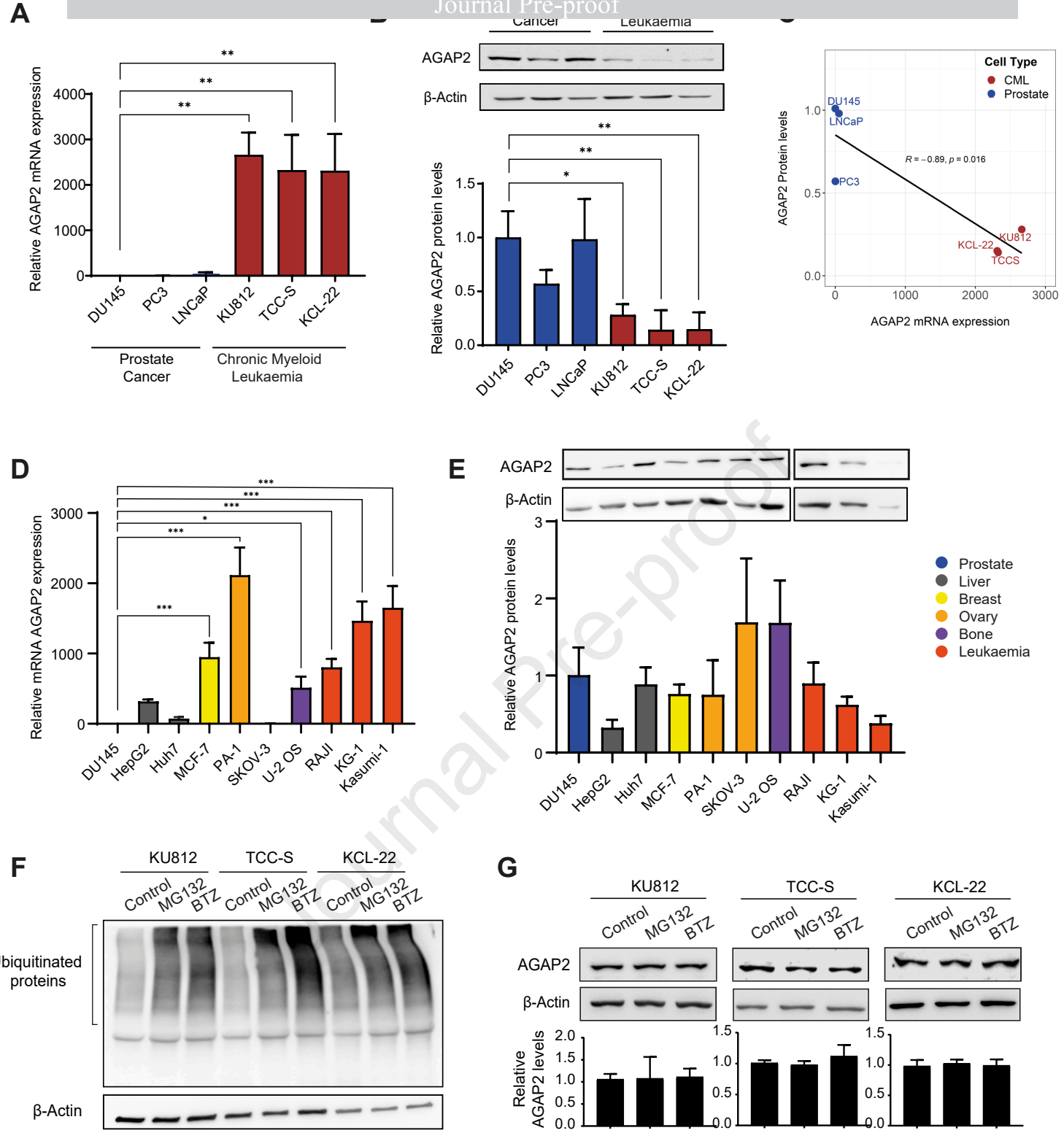
72. Heberle, H., Meirelles, G.V., da Silva, F.R., Telles, G.P., and Minghim, R. (2015). InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics 169. DOI: 10.1186/s12859-015-0611-3.

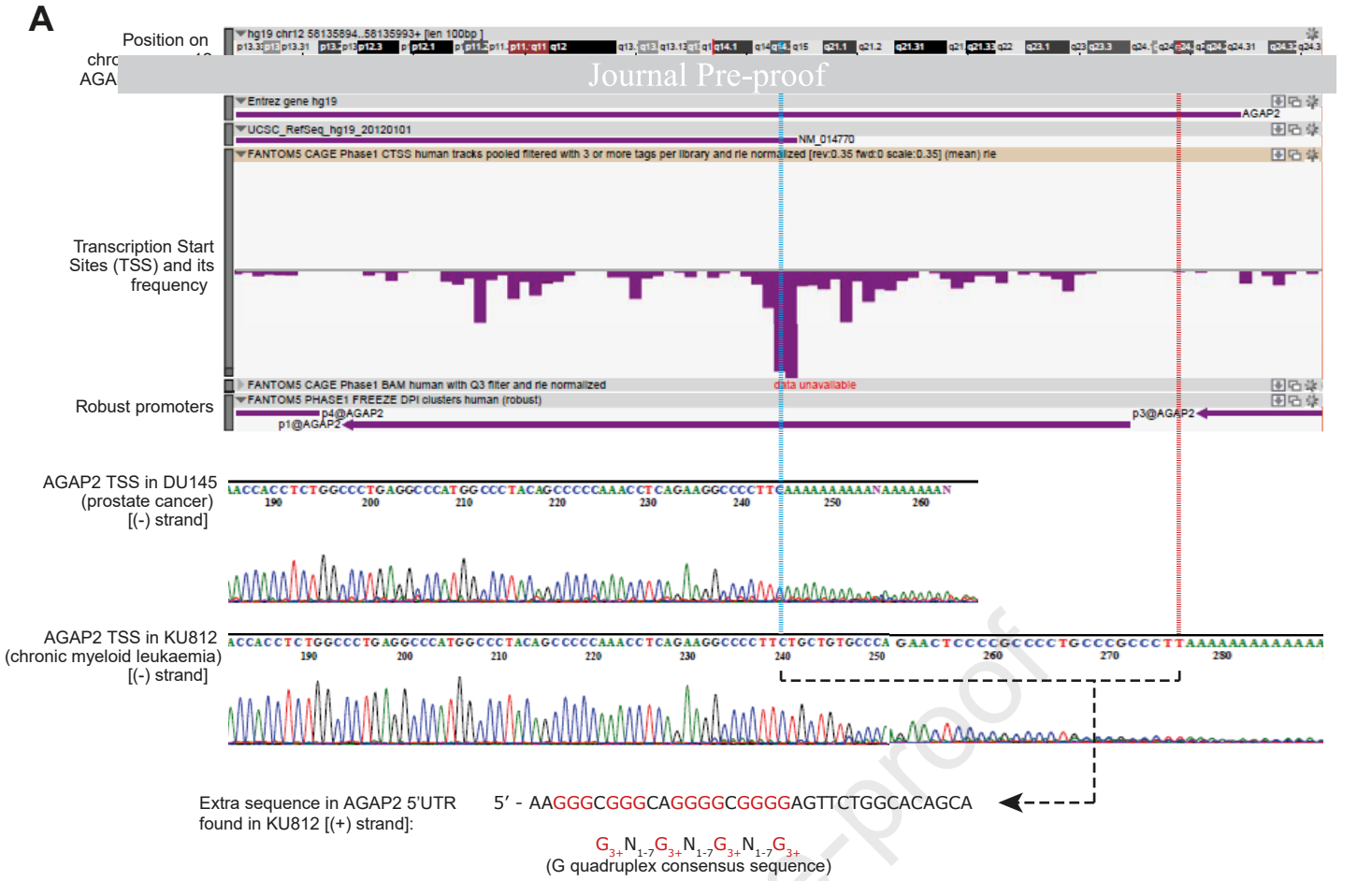
73. Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods *4,* 402-408. DOI: 10.1006/meth.2001.1262.

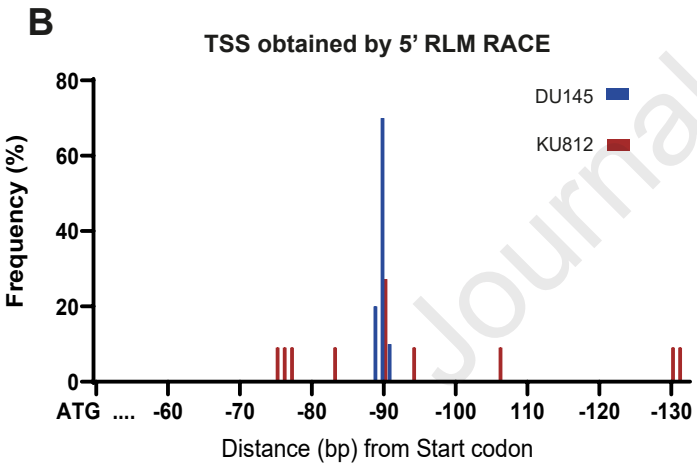
74. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R. *et al*. (2020). Ensembl 2020. Nucleic Acids Res *D1,* D682-D688. DOI: 10.1093/nar/gkz966.

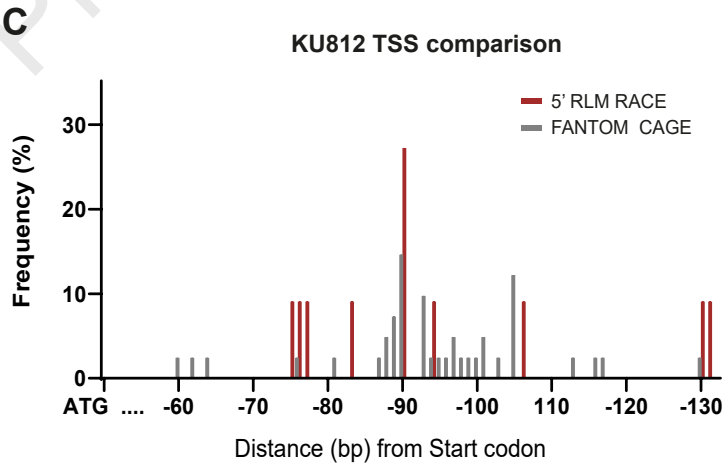
75. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, Michiel J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *11,* 1422-1423. DOI: 10.1093/bioinformatics/btp163.

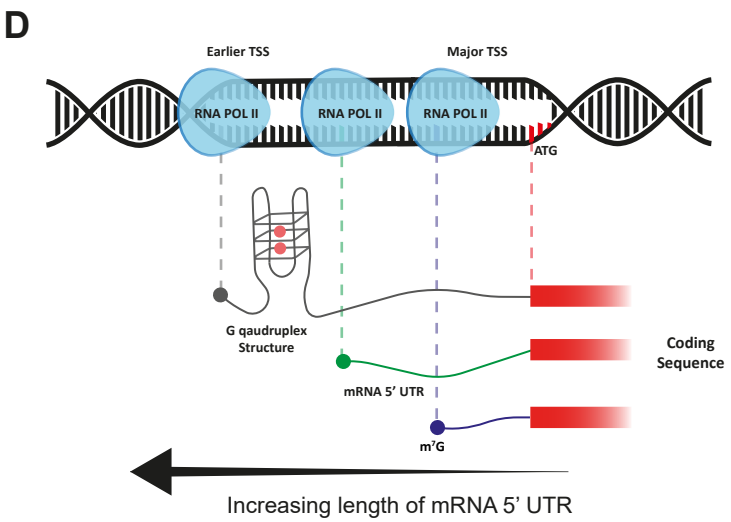
76. Takahashi, H., Kato, S., Murata, M., and Carninci, P. (2012). CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. Methods Mol Biol 181-200. DOI: 10.1007/978-1-61779-292-2_11.

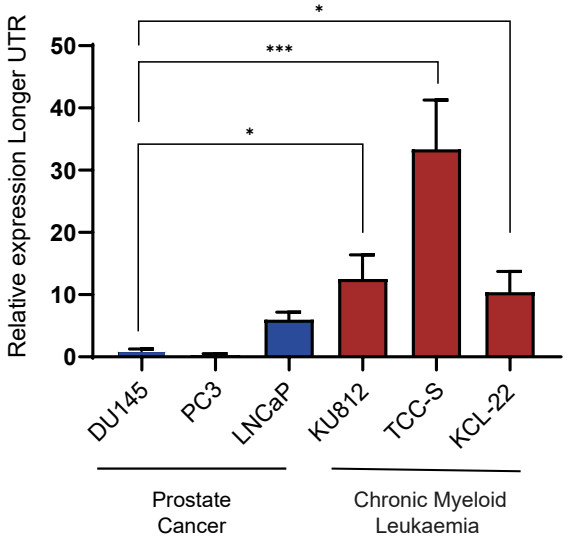


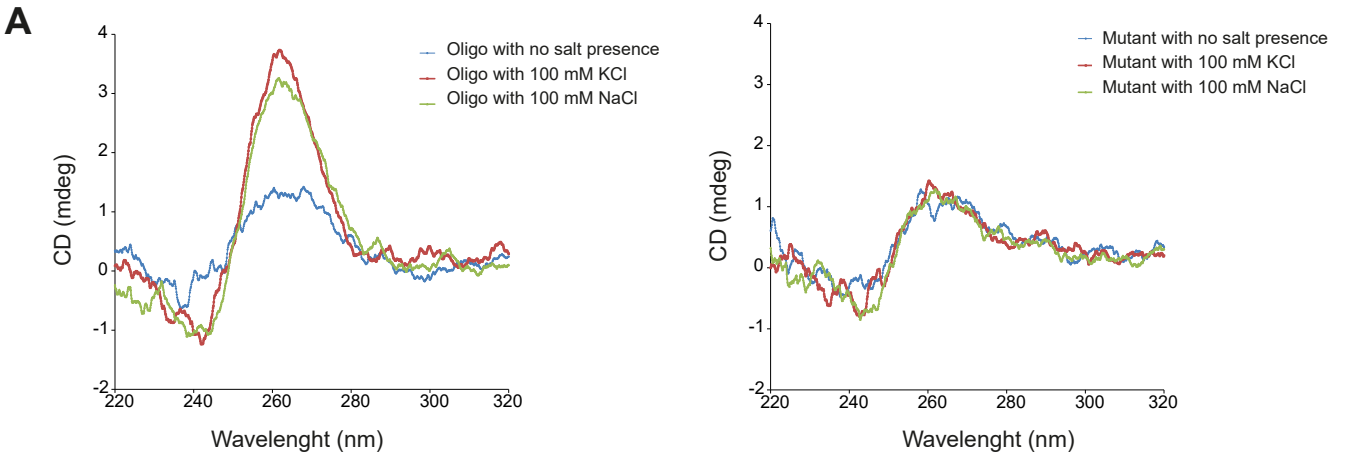


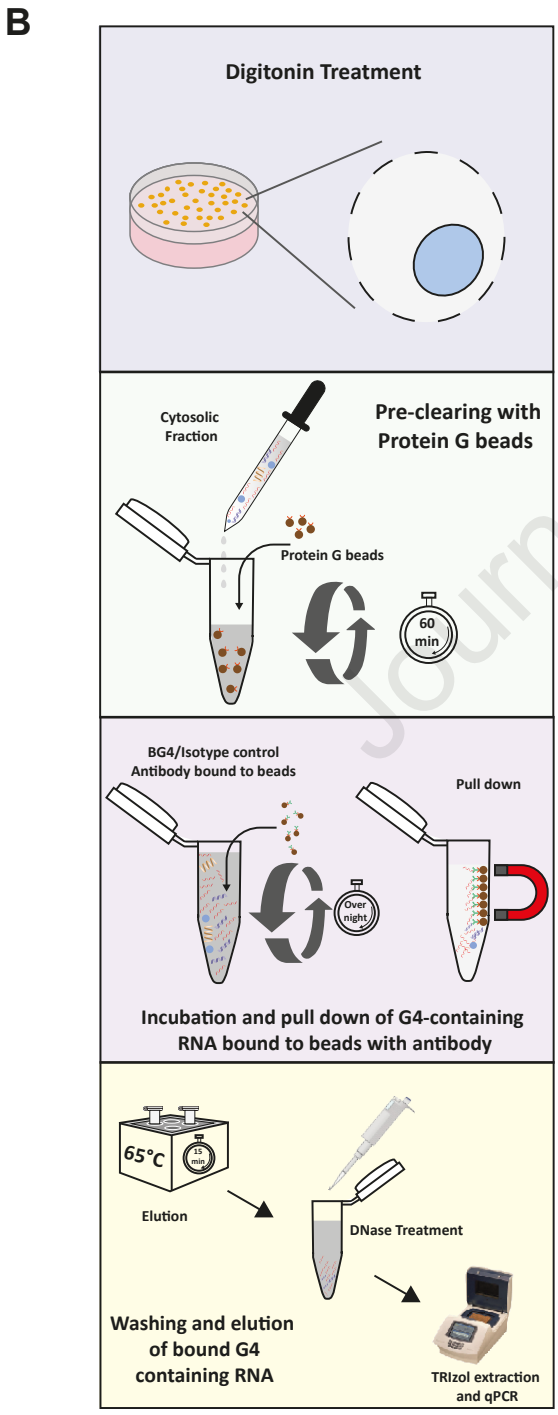


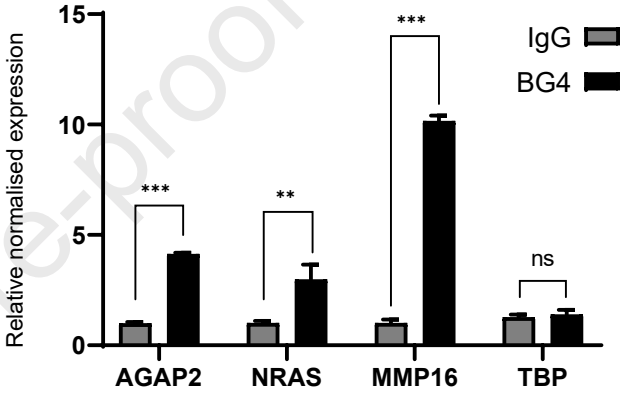


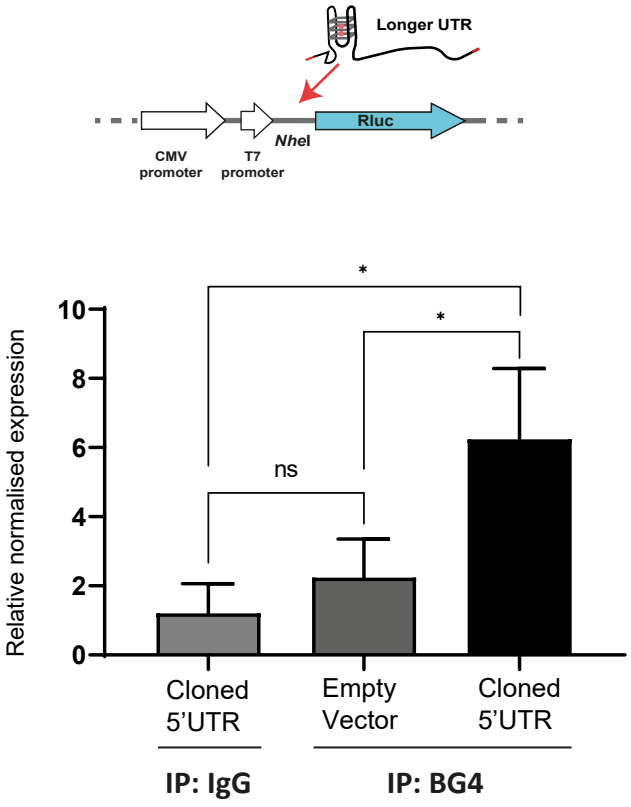


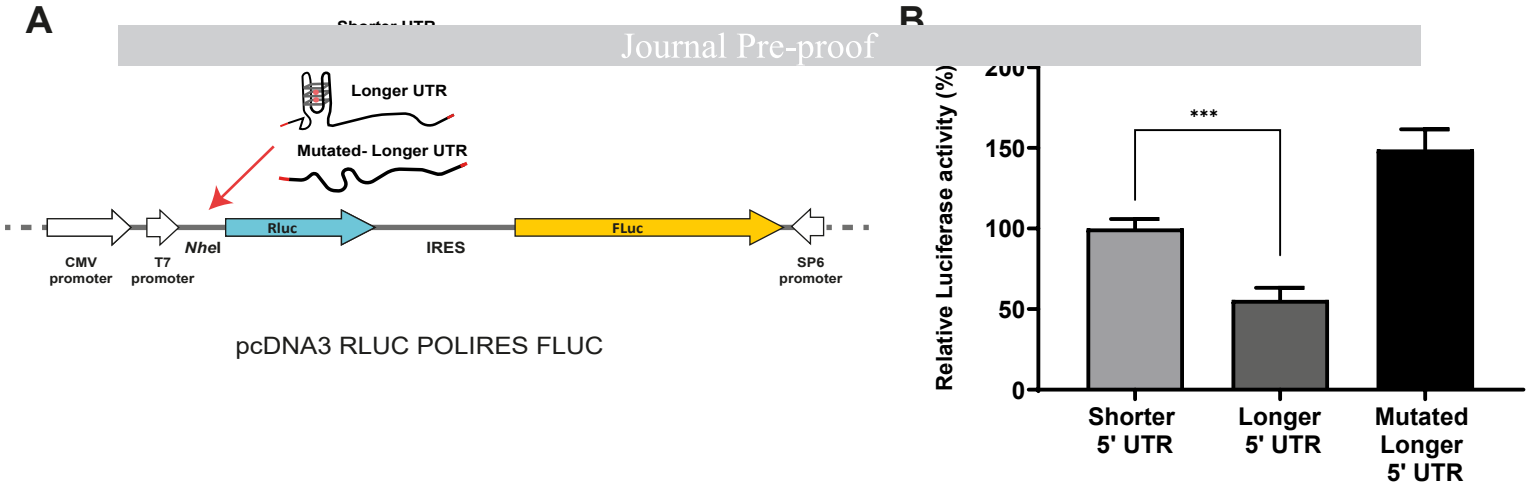


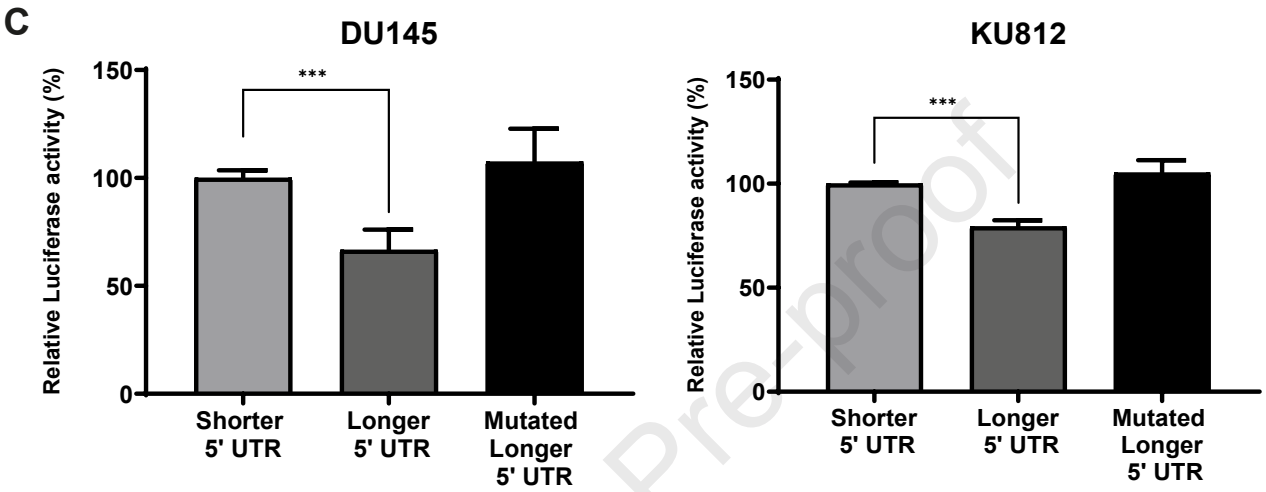


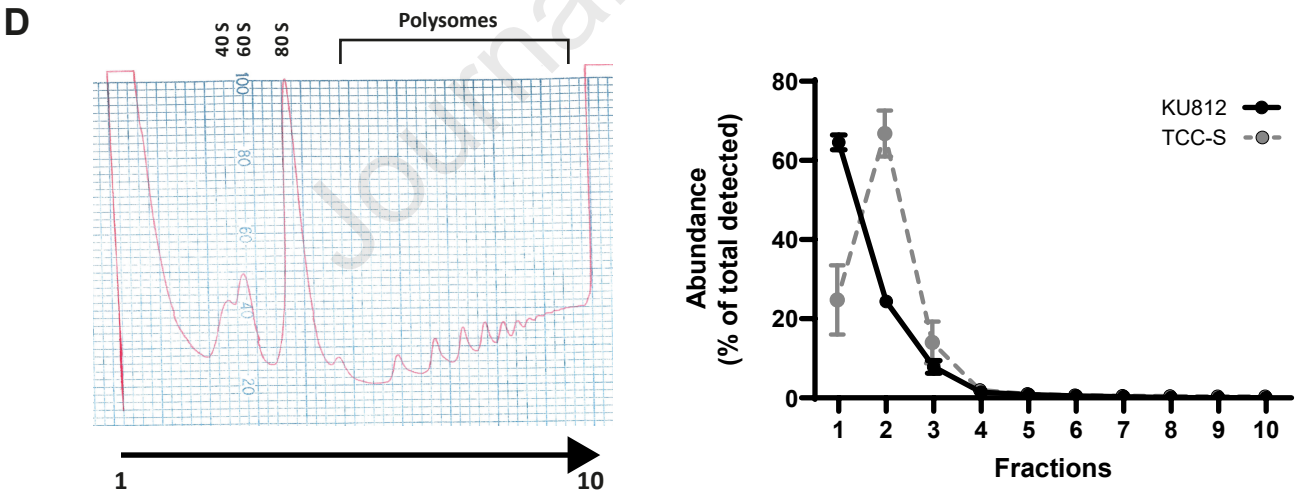












A

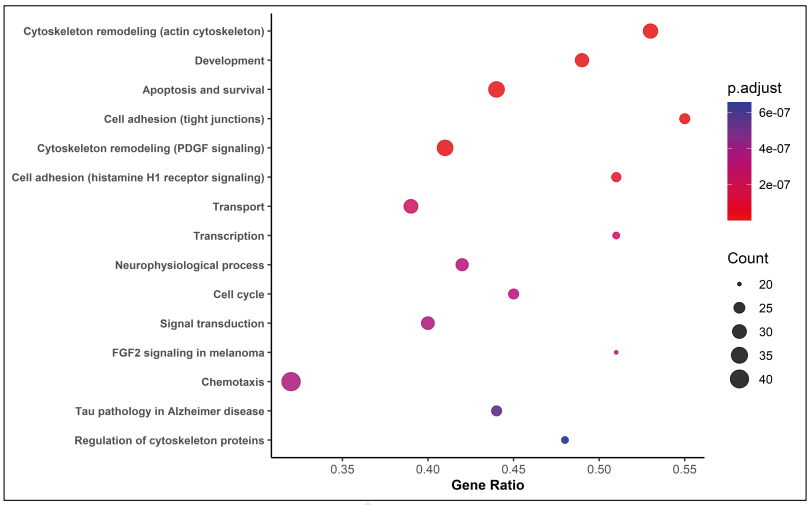
FANTOM Database
Transcription Start Site

Sequences extracted between alternative TSSs within a core promoter

Identification of G4 sequences using pqsfinder

List of genes with G4 sequences between alternative TSSs

Differential transcript counts of those genes in PC and CML cell lines (FANTOM database)



SWATH-MS

Proteins with similar or lower levels in NCI-60 dataset

Microarray

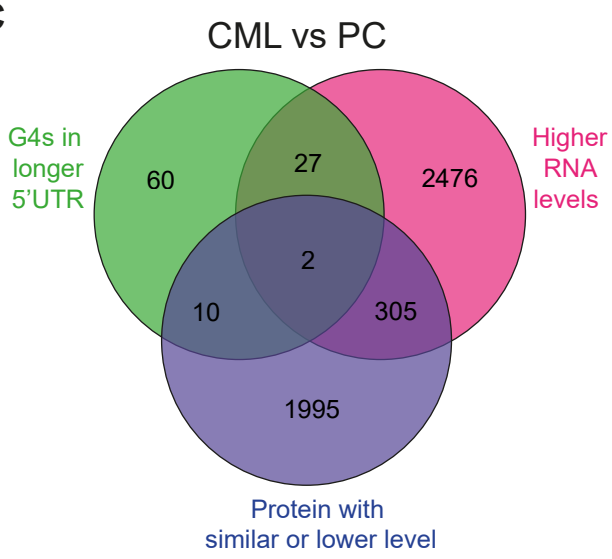
Differentially expressed genes in NCI-60 dataset

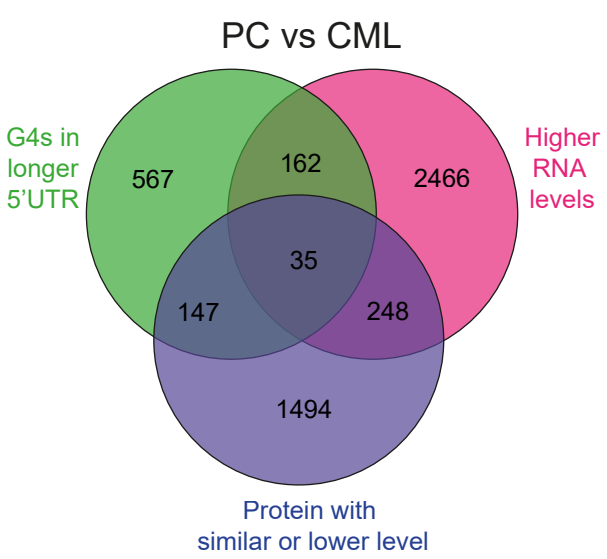
Genes with inconsistencies in mRNA and protein expression levels and G4 sequences in alternative 5' UTR isoforms

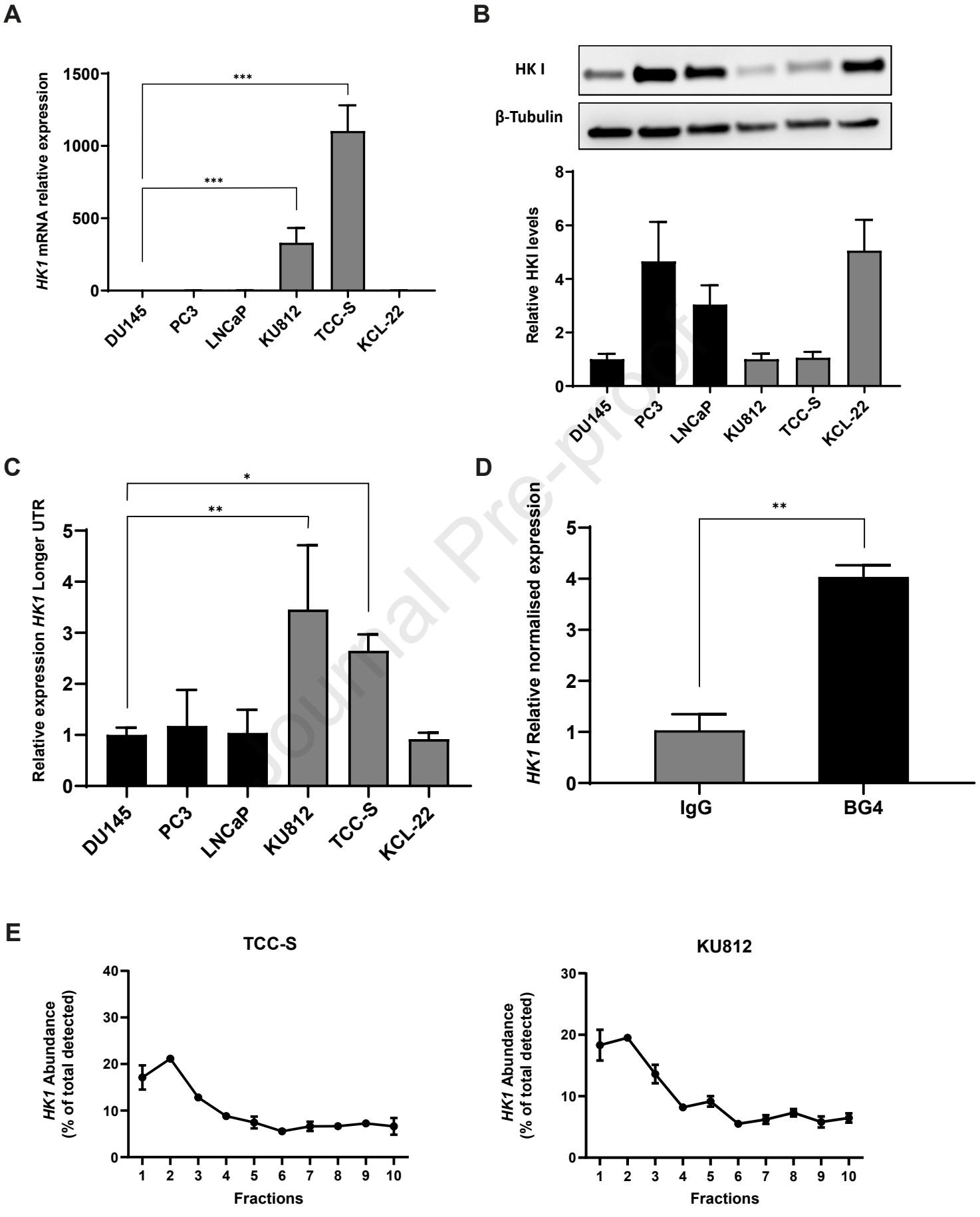
Targets in CML cell lines

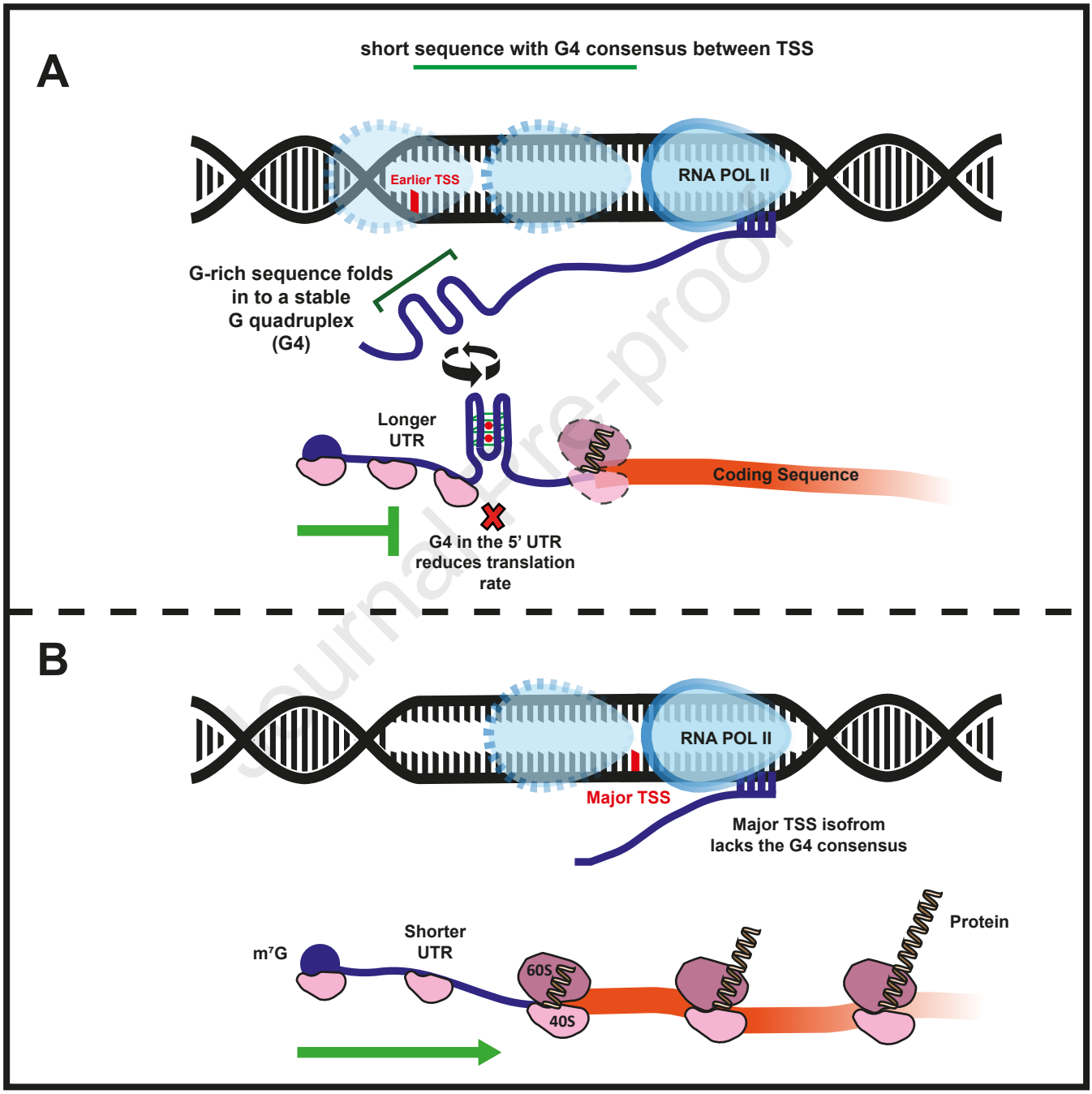
Targets in PC cell lines

C






A
HK1 mRNA relative expression
1500
1000
500
0
***
***
DU145
PC3
LNCaP
KU812
TCC-S
KCL-22
B
HK I
β-Tubulin
Relative HKI levels
8
6
4
2
0
DU145
PC3
LNCaP
KU812
TCC-S
KCL-22
C
Relative expression HK1 Longer UTR
*
**
5
4
3
2
1
0
DU145
PC3
LNCaP
KU812
TCC-S
KCL-22
D
HK1 Relative normalised expression
**
5
4
3
2
1
0
IgG
BG4
E
TCC-S
HK1 Abundance
(% of total detected)
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10
Fractions
KU812
HK1 Abundance
(% of total detected)
30
20
10
0
1 2 3 4 5 6 7 8 9 10
Fractions



Differential TSS distribution in CML and prostate cancer cells for *AGAP2* and *HK1*

Less than 50 bp differences in the 5'UTR isoforms generated in these cells

Longer mRNA 5'UTR isoform contains a G4 structure and reduced translation rates

Evidence here supports TSS selection within a cluster can affect translation rates

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Goat polyclonal anti-AGAP2	Sigma-Aldrich	Cat#SAB2501250; RRID:AB_10620617
Mouse monoclonal anti-HK I	Santa Cruz Biotechnology	Cat#sc-46695; RRID:AB_627721
Mouse monoclonal anti-DNA/RNA G-quadruplex (clone BG4)	Absolute Antibody	Cat#Ab00174-1.1
Mouse IgG Isotype Control antibody	ThermoFisher	Cat#31903; RRID:AB_10959891
Rabbit polyclonal anti-Ubiquitin	Cell Signaling Technology	Cat#3933; RRID:AB_2180538
Mouse monoclonal anti-β-Actin	Sigma-Aldrich	Cat#A2228; RRID:AB_476697
Mouse monoclonal anti-β-Tubulin	Sigma-Aldrich	Cat#T8328; RRID:AB_1844090
Rabbit monoclonal anti-eIF4A (clone C32B4)	Cell Signaling Technology	Cat#2013; RRID:AB_2097363
Rabbit Polyclonal anti-eIF4A1	Cell Signaling Technology	Cat#2490; RRID:AB_823487
Rabbit Polyclonal anti-eIF4B	Cell Signaling Technology	Cat#3592; RRID:AB_2293388
Rabbit monoclonal anti-eIF4E (clone C46H6)	Cell Signaling Technology	Cat#2067; RRID:AB_2097675
Rabbit monoclonal anti-eIF4G (clone C45A4)	Cell Signaling Technology	Cat#2469; RRID:AB_2096028
Rabbit monoclonal anti-eIF4H (clone D85F2))	Cell Signaling Technology	Cat#3469; RRID:AB_2096038
Anti-rabbit IgG, HRP-linked Antibody	Cell Signaling Technology	Cat#7074; RRID:AB_2099233
Anti-mouse IgG, HRP-linked Antibody	Cell Signaling Technology	Cat#7076; RRID:AB_330924
Anti-goat IgG, HRP-linked Antibody	Sigma-Aldrich	Cat#A4174; RRID:AB_258138
Bacterial and Virus Strains		
DH5α	Thermo-Fisher	Cat#18265017
One Shot TOP10 Chemically Competent E. coli	Thermo-Fisher	Cat#C404010
Chemicals, Peptides, and Recombinant Proteins		
Potassium chloride	Sigma-Aldrich	Cat#P9333; CAS:7447-40-7
HEPES	Sigma-Aldrich	Cat# H3375; CAS:7365-45-9
NP-40	Sigma-Aldrich	Cat# I3021; CAS:9002-93-1
Digitonin	Abcam	ab141501; CAS:11024-24-1
Absolute Ethanol for molecular biology	Fischer Scientific	Cat#10644795
2-Propanol for molecular biology	Sigma-Aldrich	Cat#278475
Nuclease free water	Promega	Cat#P1193

MG132, proteasome inhibitor	Sigma-Aldrich	Cat#474790; CAS:133407-82-6
Bortezomib, proteasome inhibitor	Santa Cruz	Cat#sc-217785; CAS:179324-69-7
Cyclohexamide	Santa Cruz	Cat#sc-3508; 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
NheI restriction endonuclease	Promega	Cat#R6501
XhoI restriction endonuclease	Promega	Cat#R6161
Taq DNA polymerase	Promega	Cat#M7841
Alkaline Calf Intestinal Phosphatase	Promega	Cat# M1821
T4 DNA Ligase	Promega	Cat#M1801
TRIzol Reagent	ThermoFisher	Cat#15596026
LiCl Precipitation Solution	ThermoFisher	Cat#AM9480
Chloroform	Sigma-Aldrich	Cat#C2432
30% Acrylamide	Severn Biotech	Cat#20-2100-10
1kb DNA Ladder	Promega	Cat#G5711
100bp DNA Ladder	Promega	Cat#G2101
Precision Plus Protein™ Dual Color Standards	Biorad	Cat#1610374
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
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
Deposited Data		
FANTOM5 database for TSS profiles	[32]	https://fantom.gsc.riken.jp/5/datafiles/latest/
NCI-60 microarray dataset	[33]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32474; GEO: GSE32474
NCI-60 SWATH-MS dataset	[68]	PRIDE: PXD003539
Experimental Models: Cell Lines		

KU812 (Human chronic myelogenous leukaemia)	ATCC	Cat#CRL-209; RRID:CVCL_0379
TCC-S (Human myelogenous leukaemia)	[69]	N/A
KCL-22 (Human myelogenous leukaemia)	ATCC	N/A
DU145 (Human prostate cancer)	ATCC	Cat#HTB-81; RRID:CVCL_0105
PC3 (Human prostate adenocarcinoma)	ATCC	Cat#CRL-1435; RRID:CVCL_0035
LNCaP (Human prostate cancer)	ATCC	N/A
HepG2 (hepatocellular carcinoma)	ATCC	Cat#HB-8065; RRID:CVCL_0027
HuH7 (Human liver cancer)	JCRB	Cat#JCRB0403; RRID:CVCL_0336
MCF-7 (Human breast adenocarcinoma)	ATCC	Cat#HTB-22; RRID:CVCL_0031
PA-1 (Human ovary teratocarcinoma)	ATCC	Cat#CRL-1572; RRID:CVCL_0479
SK-OV-3 (Human ovary adenocarcinoma)	ATCC	Cat#HTB-77; RRID:CVCL_0532
U-2 OS (Human osteosarcoma)	ECACC	Cat#92022711; RRID:CVCL_0042
RAJI (Human Burkitt's lymphoma)	ATCC	Cat#CCL-86; RRID:CVCL_0511
KG1 (Human acute myelogenous leukemia)	ATCC	Cat#CRL-8031; RRID:CVCL_0374
KASUMI-1 (Human acute myeloblastic leukemia)	ATCC	Cat#CRL-2724; RRID:CVCL_0589
Oligonucleotides		
A full list of DNA oligos is available in Table S1	N/A	N/A
Random Primers	Promega	Cat#C1181
RNA oligo (CD spectroscopy) GGGCGGGCAGGGGCGGGG	This Study	N/A
Mutant RNA oligo (CD spectroscopy) GAGCGAGCAGAGGCGAGG	This Study	N/A
Recombinant DNA		
pcDNA3 RLUC POLIRES FLUC	Addgene [44]	Cat#45642; RRID:Addgene_45642
pcDNA3 RLUC POLIRES FLUC-G1 (AGAP longer 5' UTR)	This Study	N/A
pcDNA3 RLUC POLIRES FLUC-G2 (AGAP shorter 5' UTR)	This Study	N/A
pcDNA3 RLUC POLIRES FLUC-G3 (AGAP mutated longer 5' UTR)	This Study	N/A
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)	Broad Institute	http://software.broadinstitute.org/software/igv/
BEDTOOLS v2.28	N/A	https://bedtools.readthedocs.io/en/latest/index.html
Rstudio	Rstudio team	https://www.rstudio.com/
DESeq2	[70]	http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html.
pqsfinder	[67]	http://bioconductor.org/packages/release/bioc/html/pqsfinder.html
Python programming language	Version 3.6.8	https://www.python.org/
Multiple sequence alignment	[71]	http://multalin.toulouse.inra.fr/multalin/
InteractiVenn for Venn diagram	[72]	http://www.interactivenn.net/
Other		
Nitrocellulose membrane	GE Healthcare	Cat#10600006
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