Implications of differential transcription start site selection on chronic myeloid leukemia and prostate cancer cell protein expression

Highlights
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 the TSS selection within a cluster can affect translation rates

Arif A. Surani, Keith A. Spriggs, Christoph Ufer, Christos Polytarchou, Cristina Montiel-Duarte
cristina.montielduarte@ntu.ac.uk

Surani et al., iScience 25, 105519 December 22, 2022 © 2022 The Author(s). https://doi.org/10.1016/j.isci.2022.105519
Implications of differential transcription start site selection on chronic myeloid leukemia and prostate cancer cell protein expression

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

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 leukemia (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 tumor 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 in 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 with the amount of protein present, making RNA quantification an inexact tool to predict protein levels. 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. AGAP2 (also known as PIKE-A) is a ubiquitously expressed protein that has a role in hepatic fibrosis and cancer progression. It is classed as a proto-oncogene involved in cell survival, apoptosis, migration, and lipid metabolism, 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 heterogeneous 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. 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. It is estimated that 30-50% of human genes are regulated by alternative promoters active depending on the cell type, developmental stage, cellular environment, or disease stages. 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.

Several attempts have been made in the last decade to precisely identify TSSs and characterize core promoter features. Notably, the FANTOM consortium and the DataBase of Transcriptional Start Sites (DBTSSs) have comprehensively captured the dynamically changing landscape of TSS selection by using mRNA cap-guided deep sequencing technologies. 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. In addition, other studies have also found cell-specific differential distribution of TSSs within a cluster. 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 (uORFs); upstream start codons (uAUG); RNA secondary structures; and internal ribosomal entries sites. However, although previous studies have reported the translational impact of transcripts isoforms derived from multiple closely situated promoters, minor TSSs (alternative TSS selection within the same cluster) have been considered as nonadaptive and the product of molecular errors.

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). 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 with polysomes. But more importantly, we hypothesized 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 and from the NCI-60 microarray (GSE32474) and NCI-60 SWATH-MS databases, 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 leukemia (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. However, when AGAP2 protein levels were analyzed, we observed a significant negative correlation (Pearson's $R = -0.89$, $p = 0.016$) between AGAP2 mRNA and protein in both types of cancers (Figures 1A–1C). 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 (Figures 1D and 1E) such as hepatocarcinoma (HepG2 cells), ovarian cancer (SKOV-3 cells), and acute myeloid leukemia (cell lines KG1 and Kasumi). However, when considering all cell lines analyzed 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 behavior 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.
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 leukemia (CML) cell lines (KU812, TCC-S, KCL-22) by RT-qPCR. The values presented were normalized 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 as a loading control. Densitometry values for the relative protein expression are represented below the blots. Differences were analyzed 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 and 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 and 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 (5 μM), TCC-S (5 μM), KCL-22 (50 μM) for 4 h) and Bortezomib (KU812 (200 nM), TCC-S (10 nM), KCL-22 (100 nM) for 6 h). 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 the 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.
Differential AGAP2 transcription start site selection within a cluster leads to slightly different 5′ untranslated regions in chronic myeloid leukemia

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-characterized role in controlling mRNA stability and localization, 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. 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 of 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, 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 at 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. 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. 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′-GGCGGGCCAGGGGCGGGG-3′) or a mutant version (5′-GAGCGAG CAGAGCGGGGG-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 was formed in vivo. However, the detection of RNA G4 structures in living cells is challenging and different approaches have variable success rates. Interestingly, a...
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 ~240 nm and a negative peak at ~260 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

---

iScience 25, 105519, December 22, 2022
Figure 3. Continued

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 it at 65°C for 15 min. The eluent is treated with DNase I and analyzed by RT-qPCR.

(C) rG4IP was performed in the TCC-S (CML) cell line and the immunoprecipitated samples are normalized by their input controls. NRAS and MMP16 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 analyzed 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 normalized by their input controls. An unspecific isotype antibody (IgG) was used as a negative control. Differences between samples were analyzed by unpaired two-tailed t-tests.

All the data shown in this figure correspond to three independent immunoprecipitations and the error bars denote SD. (*p < 0.05, **p < 0.01, ***p < 0.001).

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 the AGAP2 mRNA sequence using the psqfinder web application33 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 FLuc44 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.51 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 leukemia 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 non-specific changes in the metabolic activity of the transfected cells and to alter the phosphorylation state of the translation
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 NheI 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 normalization.

(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 +/- standard deviation and data are expressed as the Rluc/Fluc ratio relative to the activity of the shorter 5' UTR. Differences were analyzed 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 analyzed 48 h after transfection in DU145 and 6 h 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 analyzed 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 (A254 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 normalized 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 the 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. The FANTOM database was also used to detect...
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, 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 on 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 S5A, see also STAR methods). We used data from the FANTOM project to select the transcripts with differential TSS usage. Then, the nucleotide sequences between alternative TSSs were extracted and analyzed for the presence of potential G4s using the pqfinder package in R. 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 remodeling, apoptosis and survival, and development (Figure 5B).

Next, we analyzed 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 types of cancers). To identify suitable gene targets for validation, we used the NCI-60 dataset 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 SC, See STAR 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 HK1 relative protein levels were lower (Figure 6B). Furthermore, when analyzing 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.
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 leukemia (CML) cell lines (KU812, TCC-S, and KCL-22) by RT-qPCR. HK1 expression was normalized 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 analyzed 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) HK1 protein levels were detected by western blot. The graph later in discussion shows overall densitometry values relative to those in the 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 analyzed 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 normalized by input control, and the data presented correspond to three independent immunoprecipitations. The error bars denote the standard deviation. Differences between samples were analyzed 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).
Overall, these results confirmed AGAP2 is not an isolated case for protein expression regulation mediated by the presence of a G4 associated with 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 different in their ability to produce proteins.\(^{15,51,52}\) 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 leukemia (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 cell survival\(^{35,37}\)), we identified a shared minimal promoter region for AGAP2 in PC and CML cells, observing significant differences in mRNA expression levels.\(^3\) 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.\(^5\) 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\(^{53}\) shows a broad distribution with a dominant peak (Figure 2A). But when we analyzed 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 (Figures 2 and S2), 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,\(^54\) a clear role for this differential selection is still missing and the concept of transcriptional “noise” remains.\(^70\)

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 in 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 (Figures 2C and S2B).

Changes in translation efficiency on mRNAs with differential 5' UTRs are often due to specific sequences, with longer 5' UTRs associated generally with lower translation efficiencies.\(^76,28\) 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.\(^41,55\) However, there are also examples of increased translation when this structure is present.\(^56\)

Different approaches have been used to detect the RNA G4 structures inside the cells, including the use of G4-stabilising ligands/ions,\(^52,58\) small molecule probes,\(^57\) RNA structural mapping,\(^60\) reverse transcription stalling,\(^57\) RNA G4 structure-protein interactions,\(^51\) ligands with fluorescence activity,\(^62\) self-biotinylation methodology,\(^63\) 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 favor 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.\(^51\) 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,\(^41,55\) this structure was responsible for a reduced protein expression (Figures 4B and 4C) and a reduced polysome association (Figure 4D).

Next, we used a bioinformatics approach to detect other genes with protein levels negatively associated with 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.\(^64\)

Our data support 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,\(^36\) methylation patterns,\(^65\) chromatin remodeling, and histone modifications,\(^66\) 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\(^67\) that could become operative in the shorter 5' UTR isoforms.

**STAR★METHODS**

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead contact
  - Materials availability
  - Data and code availability
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
  - Cell lines and culture conditions
- **METHOD DETAILS**
  - Protein extraction and western blot analysis
  - RNA extraction and real-time quantitative PCR
  - 5' RNA ligase-mediated rapid amplification of cDNA ends
  - Plasmid constructs, transient transfection, and dual luciferase reporter assay
  - In vitro transcription and translation assay
  - Circular dichroism spectroscopy
  - Polysome fractionation
  - RNA G quadruplex immunoprecipitation (rG4IP)\(^31\)
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
  - Statistics
  - Mapping G quadruplex consensus sequences between alternative TSSs
  - Identification of genes with a discrepancy in mRNA and protein levels
  - Pathway analysis

**SUPPLEMENTAL INFORMATION**

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105519.

**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**

Conceptualization: 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. Pharmacology 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.
REFERENCES

1. Maier, T., Guell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. FEBS Lett. 583, 3966–3973. https://doi.org/10.1016/j.febslet.2009.10.036.

2. Ghemmaghami, S., Huh, W.K., Bower, K., Howson, P.W., Belle, A., Dephore, N., O’Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature 425, 737–741. https://doi.org/10.1038/nature2046.

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. 40, 426–436. https://doi.org/10.1111/j.1745-7270.2008.00418.x.

4. Doush, Y., Surani, A.A., Navarro-Corcuera, A., McArtredle, S., Billett, E.E., and Montiel-Duarte, C. (2019). SP1 and RARα regulate AGAP2 expression in cancer. Sci. Rep. 9, 390. https://doi.org/10.1038/s41598-019-43888-x.

5. Navarro-Corcuera, A., López-Zabalza, M.J., Martínez-Injujo, J.J., Álvarez-Sola, G., Ávila, M.A., Iraburu, M.J., Ansorena, E., and 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. Biochim. Biophys. Acta. Mol. Cell. Res. 1866, 673–685. https://doi.org/10.1016/j.bbamcr.2019.01.008.

6. Cai, Y., Wang, J., Li, R., Ayala, G., Ittmann, M., and Liu, M. (2009). GGA2/PIKE-a directly activates both the Akt and nuclear factor-kappaB pathways and promotes prostate cancer progression. Cancer Res. 69, 819–827. https://doi.org/10.1158/0008-5472.CAN-08-2537.

7. Ahn, J.Y., 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. USA 101, 6993–6998. https://doi.org/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. 33, 1797–1808. https://doi.org/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 26, 4918–4927. https://doi.org/10.1038/sj.onc.1210290.

10. FANTOM Consortium and the RIKEN PMI and CLST DGT, Forrest, A.R.R., Kawaji, H., Rehli, M., Bailie, J.K., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., et al. (2014). A promoter-level mammalian expression atlas. Nature 507, 462–470. https://doi.org/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. 13, 930. https://doi.org/10.15252/msb.2016734.

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. 2, 388–393. https://doi.org/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. 46, 582–592. https://doi.org/10.1093/nar/gko1165.

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 Genom. 19, 164. https://doi.org/10.1186/s12864-018-4504-5.

15. Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C., and Huang, T.H.M. (2008). The functional consequences of alternative promoter usage in mammalian genomes. Trends Genet. 24, 167–177. https://doi.org/10.1016/j.tig.2008.01.008.

16. Ahn, J.Y., 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 stress and aging. Cell Rep. 4, 2737–2747.e5. https://doi.org/10.1016/j.celrep.2019.04.114.

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. 38, 98–104. https://doi.org/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.; FANTOM Consortium (2014). RECLUB: a pipeline to discover reproducible transcriptional start sites and their alternative regulation using capped analysis of gene expression (CAGE). BMC Genom. 15, 269. https://doi.org/10.1186/1471-2164-15-269.

19. Dieudonné, F.X., O’Connor, P.B.F., Gubler-Jaquier, P., Yaribi, H., Conne, B., Nikolaev, S., Antonarakis, S., Baranov, P.V., and Curran, J. (2015). The effect of heterogeneous Transcription Start Sites (TSS) on the translational implications for the mammalian cellular phenotype. BMC Genom. 16, 986. https://doi.org/10.1186/s12864-015-2179-8.

20. Hollerer, I., Barker, J.C., Jorgensen, V., Treierinder, A., Dugas-Darzacq, C., Chan, L.Y., Darzacq, X., Tjian, R., Unal, E., and Brar, G.A. (2019). Evidence for an integrated gene repression mechanism based on mRNA isoform toggling in human cells. G3 9, 1045–1053. https://doi.org/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. 16, 39–47. https://doi.org/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. 20, 2291–2307. https://doi.org/10.1128/MCB.20.7.2291-2307.2000.

23. Bugaut, A., and Balaubrariansian, S. (2012). 5′-UTR RNA G-quadruplexes: translation regulation and targeting. Nucleic Acids Res. 40, 4727–4741. https://doi.org/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 Len-Weil and Turner syndrome homeobox gene SHOX. J. Biol. Chem. 278, 47820–47826. https://doi.org/10.1074/jbc.M312175200.

25. Zeita, 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. 27, 2737–2747.e5. https://doi.org/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. 12, 875. https://doi.org/10.1525/msb.20166941.
27. Li, H., Bai, L., Li, H., Xi, X., Kang, Y., Zhang, N., Sun, J., and Shao, Z. (2019). Selective translational usage of TSS and core promoters revealed by translome sequencing. BMC Genom. 20, 282. https://doi.org/10.1186/s12864-019-6560-0.

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 18, 2299-2305. https://doi.org/10.1261/rna.253865.112.

29. Arribere, J.A., and Gilbert, W.V. (2013). Roles of location: influence of G-quadruplexes on expression. Chem. Biol. 20, 1077-1089. https://doi.org/10.1016/j.chembiol.2013.06.003.

30. Xu, C., Park, J.K., and Zhang, J. (2019). A matter of location: influence of G-quadruplexes on Escherichia coli gene expression. Chem. Biol. 26, 1511-1523. https://doi.org/10.1016/j.chembiol.2019.09.014.

31. 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. 48, 12534–12555. https://doi.org/10.1093/nar/gkaa1126.

32. Biffi, G., Tannahill, D., McCafferty, J., and Balasubramanian, S. (2013). Quantitative visualization of DNA G-quadruplex structures in human cells. Nat. Chem. 5, 182–186. https://doi.org/10.1038/nchem.1548.

33. Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Porjovic, J., Semple, C.A.M., Taylor, M.S., Engstro¨m, P.G., Frith, M.C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. 38, 626–635. https://doi.org/10.1038/ng1789.

34. Kawaji, H., Frith, M.C., Katayama, S., Sandelin, A., Kai, C., Kawai, J., Carninci, P., and Hayashizaki, Y. (2006). Dynamic usage of alternative promoters in mammalian genomes. Trends Genet. 19, 640-648. https://doi.org/10.1016/j.tig.2003.09.014.

35. Holder, J.T., and Hartig, J.S. (2014). A matter of location: influence of G-quadruplexes on Escherichia coli gene expression. Chem. Biol. 21, 2584–2586. https://doi.org/10.1016/j.chembiol.2014.09.014.

36. Xu, C., Park, J.K., and Zhang, J. (2019). A matter of location: influence of G-quadruplexes on expression. Chem. Biol. 20, 1077-1089. https://doi.org/10.1016/j.chembiol.2013.06.003.

37. Cawaj, 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. 7, R118. https://doi.org/10.1186/gb-2006-7-12-r118.

38. Carninci, P., Sandelin, A., Lenhard, B.,Katayama, S., Shimokawa, K., Porjovic, J., Semple, C.A.M., Taylor, M.S., Engstro¨m, P.G., Frith, M.C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. 38, 626–635. https://doi.org/10.1038/ng1789.

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. 48, 12534–12555. https://doi.org/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. 5, 182–186. https://doi.org/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 regulates translation in eukaryotic cells. Biochemistry 46, 5313–5319. https://doi.org/10.1021/bi060949z.

42. Labudová, D., Hon, J., and Lexa, M. (2020). Pqsfinder web: G-quadruplex prediction using optimized pqsfinder algorithm. Bioinformatics 36, 2584–2586. https://doi.org/10.1093/bioinformatics/btaa323.

43. Poulin, F., Gingras, A.C., Olsen, H., Chevalier, S., and Sonenberg, N. (1998). 4E-BP3, a new translation initiation factor family. J. Biol. Chem. 273, 14002–14007. https://doi.org/10.1074/jbc. 273.22.14002.

44. Schakowski, F., Buttgert, P., Mazur, M., Marten, A., Schöttker, B., Gorschütz, M., and Schmidt-Wolf, I.G. (2004). Novel non-viral method for transfection of primary leukemia cells and cell lines. Genet. Vaccines Ther. 2, 1 https://doi.org/10.1186/1479-0556-2-1.

45. Anderson, B.R., Kankö, K., and Weissman, D. (2013). Nucleofection induces transient elfP2 phosphorylation by GCN2 and PERK. Gene Ther. 20, 136–142. https://doi.org/10.1038/gt.2012.5.

46. 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. 39, 2187–2194. https://doi.org/10.1007/s11033-011-0667-2.

47. 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 408, 123–131. https://doi.org/10.1016/j.jim.2014.05.014.

48. Van Der Kelen, K., Beyaert, R., Inze´ , D., and Monnens, L. (2003). Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet. 19, 640-648. https://doi.org/10.1016/j.tig.2003.09.014.

49. Lee, D.S.M., Ghanem, L.R., and Brash, Y. (2020). Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. Nat. Commun. 11, 527. https://doi.org/10.1038/s41467-020-14404-y.

50. 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 52, 1528–1538. https://doi.org/10.1021/ bi301365g.

51. Kowk, C.K., and Balasubramanian, S. (2015). Targeted detection of G-quadruplexes in cellular RNAs. Angew. Chem. Int. Ed. Engl. 54, 6751–6754. https://doi.org/10.1002/anie.201500891.

52. 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. 6, 75–80. https://doi.org/10.1038/ncchem.1805.

53. Yang, S.Y., Lejault, P., Chevrier, S., Boidot, R., Robertson, A.G., Wong, J.M.Y., and Moncouch, D. (2018). Transcriptome-wide identification of transient RNA G-quadruplexes in human cells. Nat. Chem. 10, 4730. https://doi.org/10.1038/s41467-018-07224-8.

54. Guo, J.U., and Bartel, D.P. (2016). RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. Science 353, aaa5371. https://doi.org/10.1126/science.aaa5371.

55. 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 for potential quadruplex-forming sequences in R. Bioinformatics 33, 3373-3379. https://doi. org/10.1093/bioinformatics/btx413.
of cellular G-quadruplex containing transcripts. Nucleic Acids Res. 46, 11592–11604. https://doi.org/10.1093/nar/gky861.

62. Chen, X.C., Chen, S.B., Dai, J., Yuan, J.H., Ou, T.M., Huang, Z.S., and 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. https://doi.org/10.1002/anie.201801999.

63. Einarson, O.J., and Sen, D. (2017). Self-biotinylation of DNA G-quadruplexes via intrinsic peroxidase activity. Nucleic Acids Res. 45, 9813–9822. https://doi.org/10.1093/nar/gkx765.

64. 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. 27, 1081–1089. https://doi.org/10.1038/s41431-019-0366-9.

65. 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. 392, 112055. https://doi.org/10.1016/j.yexcr.2020.112055.

66. Niederacher, G., Klopf, E., and Schüller, C. (2011). Interplay of dynamic transcription and chromatin remodeling: lessons from yeast. Int. J. Mol. Sci. 12, 4758–4769. https://doi.org/10.3390/ijms12064758.

67. 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. 46, 3326–3338. https://doi.org/10.1093/nar/gky188.

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 21, 664–680. https://doi.org/10.1016/j.isci.2019.05.099.

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 18, 25–33. https://doi.org/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. 15, 550. https://doi.org/10.1186/s13059-014-0550-8.

71. Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. 16, 10881–10890. https://doi.org/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 Bioinf. 16, 169. https://doi.org/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 25, 402–408. https://doi.org/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. 48, D682–D688. https://doi.org/10.1093/nar/gkaa266.

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., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

76. Takahashi, H., Kato, S., Murata, M., and Carinci, P. (2012). CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. Methods Mol. Biol. 786, 181–200. https://doi.org/10.1007/978-1-61779-292-2_11.
## STAR★METHODS

### 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#PR9333; 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 |
| MGl32, 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 |

(Continued on next page)
| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| 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 |
| 1 kb DNA Ladder     | Promega | Cat#G5711 |
| 100 bp DNA Ladder   | Promega | Cat#G2101 |
| Precision Plus Protein™ Dual Color Standards | Biorad | Cat#1610374 |

**Critical commercial assays**

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| 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 Super-Script III | ThermoFisher | Cat#L150201 |
| RT and TOPO TA Cloning for 5’ RLM-RACE | 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/sRNA transfection reagent | Polyplus | Cat#114-01 |
| DNeasy Blood & Tissue Kit | QIAGEN | Cat#69504 |

**Deposited data**

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| FANTOM5 database for TSS profiles | Lizio et al., 32 | https://fantom.gsc.riken.jp/5/datafiles/latest/ |
| NCI-60 microarray dataset | Reinhold et al., 33 | https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32474; GEO: GSE32474 |
| NCI-60 SWATH-MS dataset | Guo et al., 68 | PRIDE: PXD003539 |

**Experimental models: Cell lines**

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| KU812 (Human chronic myelogenous leukaemia) | ATCC | Cat#CRL-209; RRID:CVCL_0379 |
| TCC-S (Human myelogenous leukaemia) | Van et al., 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#HTB-8065; RRID:CVCL_0027 |
| HuH7 (Human liver cancer) | JCRB | Cat#JCRB0403; RRID:CVCL_0336 |

(Continued on next page)
### REAGENT or RESOURCE

| REAGENT or RESOURCE                                      | SOURCE       | IDENTIFIER                  |
|----------------------------------------------------------|--------------|-----------------------------|
| 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
- Random Primers: Promega, Cat#C1181
- RNA oligo (CD spectroscopy): GGGCGGGCGAGGGCGGGCCGGGG
- Mutant RNA oligo (CD spectroscopy): GAGCGAGCACAGGGCGAG

### Recombinant DNA

- pcDNA3 RLUC POLIRES FLUC: Addgene; Poulin et al., Cat#45642, RRID:Addgene_45642
- pcDNA3 RLUC POLIRES FLUC-G1 (AGAP longer 5' UTR): This Study
- pcDNA3 RLUC POLIRES FLUC-G2 (AGAP shorter 5' UTR): This Study
- pcDNA3 RLUC POLIRES FLUC-G3 (AGAP mutated longer 5' UTR): This Study

### 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: Love et al., http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html
- Pqsfinder: Hon et al., http://bioconductor.org/packages/r/analysis/bioc/html/pqsfinder.html
- Python programming language: Version 3.6.8 https://www.python.org/
- Multiple sequence alignment: Corpet et al., http://multalin.toulouse.inra.fr/multalin/
- InteractiVenn for Venn diagram: Heberle et al., 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

(Continued on next page)
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 (Data 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, 2 mM Glutamine, and 20% FBS. All cell lines were maintained at 37°C in a 5% CO2 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 min and sonicated with ice-cooling for 3 x 5 s pulses at a frequency of 5 microns using a Soniprep 150 plus (MSEs) 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 min 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% Tween 20] 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 h 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^(-ΔΔCt)).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; RRD-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 NheI 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 × 10^5 cells/well in 6-well plates for 24 h before transfection and collected 48 h after being transfected. For transient transfection of the CML cell line (KU812), 1 μg of each reporter plasmid was electroporated into 2 × 10^6 cells using Amaxa Cell Line Nucleofector solution V (Lonza), program X-001, and collected after 6 h.

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 analyzed 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 key resources table) 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 and 320 nm at a scan speed of 50 nm/min with a response time of 2 s. 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% CO₂ 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 min on ice followed by centrifugation at 12,000 x g for 10 min 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 min at 4°C. The gradients were then fractionated from top to bottom while measuring absorbance at 254 nm. 500 μL of each sucrose fraction 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)
Briefly, rG4IP was performed with a structure-specific G quadruplex (BG4) antibody (Absolute Antibody). TCC-S cells (15 x 10⁶) 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 min 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 10⁶ DU145 cells were seeded in a 100 mm dish, transfected using JetPRIME transfection reagent (Polyplus), trypsinised after 48 h, 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 min to release the bound nucleic acids. The eluent was treated with 2 U of RNase-free DNase I (ThermoFisher) for 15 min 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. 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). 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 furthermore upstream tag within the same cluster were extracted for all the transcripts using Bio. Entrez module in Biopython (Data S1). The sequences were then analyzed for the presence of G quadruplex consensus using pqsfinder package in R (Data 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.\textsuperscript{50,76} The differential distribution was computed by linear modeling and empirical Bayes approach using the Limma package in R (Data 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.\textsuperscript{33,68} The differentially expressed RNAs in PC (DU145, PC3) and CML (K562) were analyzed using GEO2R (NCBI) (Data S3). The significant differences in protein mass spectral intensity values were evaluated using Limma (Data S3). The genes with differential RNA expression 2-fold or greater (RNA logFC $\geq$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 analyzed 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.