Codon Usage and Splicing Jointly Influence mRNA Localization

Citation for published version:
Mordstein, C, Savisaar, R, Young, RS, Bazile, J, Talmane, L, Luft, J, Liss, M, Taylor, MS, Hurst, LD & Kudla, G 2020, 'Codon Usage and Splicing Jointly Influence mRNA Localization', *Cell Systems*, vol. 10, no. 4, pp. 351-362.e8. https://doi.org/10.1016/j.cels.2020.03.001

Digital Object Identifier (DOI):
10.1016/j.cels.2020.03.001

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Cell Systems

Codon usage and splicing jointly influence mRNA localization

Christine Mordstein[1,2], Rosina Savisaar[2,3], Robert S Young[1,4], Jeanne Bazile[1], Lana Talmane[1], Juliet Luft[1], Michael Liss[5], Martin S Taylor[1], Laurence D Hurst[2], Grzegorz Kudla[1*]

[1]MRC Human Genetics Unit, Institute for Genetics and Molecular Medicine, The University of Edinburgh, Edinburgh, UK

[2]Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, UK

[3]Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal

[4]Centre for Global Health Research, Usher Institute, The University of Edinburgh, Edinburgh, UK

[5]Thermo Fisher Scientific, GENEART GmbH, Regensburg, Germany

*Lead Contact (Corresponding Author): Grzegorz Kudla (gkudla@gmail.com)

Abstract

In the human genome, most genes undergo splicing and patterns of codon usage are splicing-dependent: guanine and cytosine (GC) content is highest within single-exon genes and within first exons of multi-exon genes. However, the effects of codon usage on gene expression are typically characterized in unspliced model genes. Here, we measured the effects of splicing on expression in a panel of synonymous reporter genes that varied in nucleotide composition. We found that high GC content increased protein yield, mRNA yield, cytoplasmic mRNA localization and translation of unspliced reporters. Splicing did not affect the expression of GC-rich variants. However, splicing promoted the expression of AT-rich variants by increasing their steady-state protein and mRNA levels, in part through promoting cytoplasmic localization of mRNA. We propose that splicing promotes the nuclear export of AU-rich mRNAs and that codon- and splicing-dependent effects on expression are under evolutionary pressure in the human genome.

Introduction

Mammalian genomes are characterised by large regional variation in base composition (Bernardi, 1993). Regions with a high density of G and C nucleotides (GC-rich regions) are in an open, transcriptionally active state, are gene-dense, and replicate early. In contrast, AT-rich regions are enriched with heterochromatin, contain large gene deserts and replicate late (Arhondakis et al., 2011; Lander et al., 2001; Vinogradov, 2003). The mechanisms that give rise to this compositional heterogeneity have been under debate for years and many researchers believe that the pattern originates from the process of GC-biased gene conversion (Duret and Galtier, 2009), though other neutral and selective mechanisms have been proposed as well (Eyre-Walker, 1991; Galtier et al., 2018; Plotkin and Kudla, 2011; Sharp and Li, 1987b).

The sequence composition of mammalian genes correlates with the GC-content of their genomic location. Thus, introns and exons of genes located in GC-rich parts of the genome are themselves GC-rich. This can potentially influence gene expression in multiple ways: nucleotide composition affects the physical properties of DNA, the thermodynamic stability of RNA folding, the propensity of RNA to interact with other RNAs and proteins, the codon adaptation of mRNA to tRNA pools, and the propensity for RNA modifications, such as m6A (Dominissini et al., 2012) and ac4C (Arango et al., 2018). However, studies of the effects of nucleotide composition on gene expression in human cells have led to opposing conclusions. On the one hand, heterologous expression experiments typically report large positive effects of increased GC content on protein production in a wide variety of transgenes, including fluorescent reporter genes, human cDNAs, and virus genes (Bauer et al., 2010; Kosovac et al., 2011; Kotsopoulou et al., 2000; Kudla et al., 2006; Zolotukhin et al., 1996). As a result, increasing the GC content of transgenes has become a common strategy in coding sequence optimization for heterologous expression in human cells (Fath et al., 2011). On the other hand, genome-wide analyses of endogenous genes typically show little or no correlation of GC content with expression (Duan et al., 2013; Lercher et al., 2003; Rudolph et al., 2016; Semon et al., 2005).

We hypothesized that the conflicting results in heterologous and endogenous gene expression studies might be explained by RNA splicing. Most transgenes used in heterologous expression systems have no introns, whereas 97% of genes in the human genome contain one or more introns. Splicing is known to influence gene expression at multiple stages, including nuclear RNP assembly, RNA export, and translation. If splicing selectively increased the expression of AT-rich genes, it could account for the lack of correlation of GC content and gene expression in previous genome-wide studies. We therefore compared spliced and unspliced genes with respect to their (1) genomic codon usage, (2) expression levels of reporter genes in transient and stable transfection experiments and (3) global expression patterns in human transcriptome studies. We show that splicing increases the expression of AT-rich genes, but not GC-rich genes, in part through effects on cytoplasmic RNA enrichment.

Results

Codon usage of human protein-coding genes depends on RNA splicing

We first analysed the relationship between the nucleotide composition of human genes and splicing. GC4 content (guanine and cytosine content at 4-fold degenerate sites of codons) correlates negatively with the number of exons in humans (Figure 1A; Spearman's $\rho = -0.27$; $p < 2.2\times10^{-16}$; see also (Carels and Bernardi, 2000; Ressayre et al., 2015; Savisaar and Hurst, 2016)). In addition, GC4 content is highest in 5'-proximal exons (Figure 1B; Spearman's $\rho = -0.18$; $p < 2.2\times10^{-16}$), and first exons have a higher GC4 content than second exons ($p < 2.2\times10^{-16}$, one-tailed Wilcoxon test). Although these patterns could result from proximity to GC-rich transcription start sites (TSSs) (Zhang et al., 2004), we found that first exons have significantly higher GC4 content than second exons even when controlling for the distance from the TSS (Figure 1C). This suggests that splicing contributes to the observed enrichment of G and C nucleotides in the 5'-proximal exons in humans. Interestingly, there is little association between exon counts and GC content among human lncRNAs (Figure S1).

To understand the causal links between splicing and nucleotide composition, we studied the compositional patterns of retrogenes. Retrotransposition provides a natural evolutionary experiment of what happens when a previously spliced gene suddenly loses its introns. We first analysed a set of 49 parent-retrogene pairs for which both the parent and the retrocopy ORFs have been retained in human and mouse. We found that the retrocopies had a significantly higher GC4 content than their parents (median $GC4_{retrocopy}$ - $GC4_{parent}$ = 11.5%; $p = 2.1\times10^{-4}$ from one-tailed Wilcoxon test; Figure 1D). It thus appears that after retrotransposition, newly integrated intronless genes come under selective pressure for increased GC content. In a comparison of 31 parent-retrogene pairs retained between human and macaque, the median GC4 difference is not significant (0.09%; p = 0.13, Wilcoxon test), but this may be explained by duplication events in macaques being more recent (dS ~ 0.08) than in mouse (dS ~ 0.56) (Gradnigo et al., 2016; Ponting and Goodstadt, 2009) so that changes in GC composition might not have had time to accumulate. As a control, we analysed retrocopies classified as pseudogenes (Figure S1D) and found their GC4 content to be significantly lower compared to their parental genes (-2.9%; $p < 2.2\times10^{-16}$, Wilcoxon test). Furthermore, the genomic neighbourhood of functional retrocopies and pseudogenes had significantly lower GC content than the neighbourhood of their respective parental genes (Figure S1E), suggesting that increased GC content is not intrinsically connected with retrotransposition, but is required for maintaining long-term functionality of retrogenes. Taken together, these results support a splicing-dependent mechanism shaping conserved patterns of nucleotide composition across functional protein-coding genes.

GC-content is a strong predictor of expression of unspliced reporter genes

The above analyses show a connection between splicing and genomic GC content of endogenous human genes. To test whether splicing differentially affects the expression of genes depending on their GC content, we designed 22 synonymous variants of GFP that span a broad range of GC3 content (GC content at the third positions of codons) (Mittal et al., 2018) (Figure S2). The collection encompasses most of the variation in GC3 content found among human genes. All variants were independently designed by randomly drawing each codon from an appropriate probability distribution, to ensure uniform GC content and statistical independence between sequences. We cloned these variants into two mammalian expression vectors: an intronless vector with a CMV promoter (pCM3) and a version of the same vector with a synthetic intron located in the 5' UTR (pCM4). The GC content profiles of the 5' UTRs were similar in both vectors (Figure S2E,F), and the intron was spliced efficiently in all variants tested, independently of the coding sequence GC content (Figure S3A). The vectors also encoded a far-red fluorescent protein, mKate2, which we used to normalize GFP protein abundance (normalization reduced measurement noise, but similar results were obtained with and without normalization). Transient transfections of HeLa cells with three independent preparations of each plasmid showed reproducible expression with a large dynamic range: synonymous variants differed in GFP protein production 46-fold. Consistent with previous studies, GFP fluorescence was strongly correlated with GC3 content in unspliced genes (Figure 2A). Introduction of an intron into the 5' UTR increased the expression of most, but not all variants. Typically, GC-poor variants experienced a large increase of expression in the presence of an intron, whereas GC-rich variants were unaffected or experienced a moderate increase (Figure 2B,C).

We obtained similar results in stably transfected HEK293 and HeLa cells (Figure S3B,C) and when expressing an independently designed collection of 25 synonymous variants of mKate2 in HeLa cells (Figure 2D-F). A Fisher's exact test revealed that the expression of GC-poor variants was more likely to be increased by splicing, compared to GC-rich variants (GC3<60% vs GC3>60%, p=0.02, N=47, GFP and mKate variants combined). These experiments show that many AT-rich genetic variants are expressed inefficiently in human cells, but low expression can be partially rescued by splicing. Notably, the average GC content of the human genome is 41% (Li, 2011). In our experiments, genes with GC content at or below 41% are expressed extremely inefficiently, unless they contain an intron (Figure 2A,B). This may provide a strong selective pressure for maintaining introns in human genes.

To establish which stages of expression are responsible for these observations, we first measured mRNA abundance of GFP variants in transiently transfected HeLa cells by quantitative RT-PCR (qRT-PCR). High GC content may introduce unwanted bias in PCR, so to allow fair comparison of all variants irrespective of their GC content, PCR primers were placed in the untranslated regions, whose sequence did not vary. Similar to protein levels, mRNA abundance varied widely between synonymous variants of GFP. GC-poor variants experienced a large increase of expression in the presence of an intron, whereas GC-rich variants were less affected (Figure 2G-I). The range of variation in mRNA abundance was much smaller in constructs with an intron than without intron (Figure 2I), indicating that splicing compensates the effects of GC content on expression.

We then asked if changes in mRNA abundance arose at transcriptional or post-transcriptional levels. As a proxy for transcriptional efficiency, we measured the abundance of intronic RNA for GFP variants expressed from the intron-containing plasmid. Coding sequence GC content did not correlate with intronic RNA abundance (Figure 2J), suggesting that transcription of the 5' UTR intron does not depend on GC content of the coding sequence. We further performed metabolic labelling of nascent RNA using 4-thiouridine (4sU) in cell lines stably expressing GC-poor and GC-rich GFP variants, expressed both with and without 5’ UTR intron, followed by nascent RNA purification and qRT-PCR (Figure S3D,E). We did not observe any systematic variation in nascent GFP RNA levels that could be explained by either GC content or splicing. Conversely, high GC content was associated with stabilization in unspliced and spliced constructs (Figure 2K). Taken together, these experiments show that high GC content enhances gene expression at a post-transcriptional level, and that the effect of GC content on expression is modulated by splicing.

High GC content at the 5’ end correlates with efficient expression

To further explore the sequence determinants of expression, we assembled a pool of 217 synonymous variants of GFP that included the 22 variants studied above, 137 variants from our earlier study (Kudla et al., 2009), and 58 additional variants. We cloned the collection into plasmids with and without a 5' UTR intron. We then established pools of HeLa Flp-In T-REx cells that stably express these constructs from a single genomic locus under a doxycycline-inducible promoter and measured the protein levels of all variants by Flow-Seq (Kosuri et al., 2013). We also performed Flow-Seq in HEK293 cells using the intronless constructs only. In Flow-Seq, a pool of cells is sorted by FACS into bins of increasing fluorescence and the distribution of variants in each bin is probed by amplicon sequencing to quantify protein abundance (Figure 3A). All variants could be quantified with good technical and biological reproducibility, and high correlation was found between Flow-Seq and spectrofluorometric measurement of individual constructs (Figure S4). Most variants showed the expected unimodal distribution across fluorescence bins, but some variants showed bimodal distributions, possibly indicative of gene silencing in a fraction of cells.

All Flow-Seq experiments showed substantial variation of expression between synonymous variants of GFP (Figure 3B). GFP protein levels in HeLa cells (with intron), HeLa cells (without intron), and HEK293 cells (without intron) were all correlated with each other, but the moderate degree of correlation (r=0.51 HEK293 (without intron) vs HeLa (without intron); r=0.36 Hela (with intron) vs HeLa (without intron)) suggests that the effects of codon usage on expression are modulated by splicing and by cell line identity - in agreement with prior observations of tissue-specific codon usage (Burow et al., 2018; Gingold et al., 2014; Plotkin et al., 2004; Rudolph et al., 2016). Flow-Seq confirms the positive correlation of synonymous site GC-content with expression of unspliced variants, whereas no significant correlation was found among intron-containing variants (Figure 3C). In contrast to results reported by us and others in bacteria and yeast (Cambray et al., 2018; Goodman et al., 2013; Kudla et al., 2009; Shah et al., 2013), but consistently with the positive correlation between GC content and expression, strong mRNA folding near the beginning of the coding sequence correlated with increased expression (Spearman's ρ = 0.27 in HeLa cells; ρ = 0.4 in HEK293 cells). Expression was positively correlated with CpG content and codon adaptation index (CAI), and negatively correlated with the estimated density of AU-rich elements (ARE) or cryptic splice sites (see STAR methods for definitions of all sequence features tested). Because of the strong correlation between GC content, CpG content, CAI and mRNA folding energy, a multiple regression analysis could not resolve which of these properties was causally related to expression.

Some of the variants analysed by Flow-Seq featured large regional variation in GC content (Figure S5A) and we asked whether the localization of low-GC and high-GC regions within the coding sequence influences expression. We found that the GC3 content in the first half of the coding sequence (nt 1-360), but not in the second half (nt 361-720), was positively correlated with expression of intronless GFP variants in the HeLa and HEK293 cells (Figure 3D). The GC3 content in either half of the gene showed no correlation with expression in the intron-containing constructs.

To further test whether GC content at the 5' end of genes has a particularly important effect on expression, we constructed in-frame fusions between GC-rich and GC-poor variants of GFP and mKate2 genes and quantified their protein and mRNA abundance in transient transfection experiments. RNA and protein yields showed a dependence on the GC content profile: GC-poor mKate2 showed nearly undetectable expression on its own, or when fused to the 5' end of GC-rich GFP, but it was efficiently expressed when fused to the 3' end of GC-rich GFP (Figure 3E, left panels). Similarly, expression of GC-poor GFP was significantly enhanced when it was fused to the 3' end of GC-rich mKate2 (Figure 3E, right panels). By contrast, pairs of GC-rich variants were efficiently expressed when fused in either orientation. N-terminal fusion of GC-rich GFP had a slightly larger positive effect on expression compared to GC-rich mKate, perhaps because of differences in codon usage or protein folding. Taken together, these experiments confirm that GC content near the 5' end of the coding sequence has a large effect on expression.

Introns within the coding sequence enhance GC-poor gene expression

While the experiments described above utilised an intron placed in the 5’ UTR, it should be noted that most introns within human genes are found within the CDS. To examine the relationship between intron location and gene expression changes relating to codon usage, we modified two GFP variants by moving their introns from the 5' UTR into the coding sequence (Figure 3F). We chose variants that were AT-rich (GC3=0.38 and 0.37), poorly expressed (HeLa Flow-Seq scores 3.71 and 4.4.) and experienced a large increase in expression when expressed with a 5’ UTR intron (HeLa Flow-seq scores 6.18 and 5.98). Transient transfections confirmed the positive effect of a 5’ UTR intron on expression of both variants (Figure 3F, first 2 bars in each plot). When the intron was placed within the coding sequence, expression was also increased compared to the intronless counterparts, suggesting that the positive effects of splicing on expression are not inherently linked to the intron position. For one of the variants, the inclusion of both 5’ UTR and CDS introns led to a further increase in expression. This is consistent with our genome-wide observation that codon usage is linked to number of introns. Taken together, these results support a splicing-dependent effect of codon usage on gene expression.

High GC content leads to cytoplasmic enrichment of mRNA and higher ribosome association

We then used the pooled HeLa cell lines to analyse the effects of GC content on mRNA localization. We separated the cells into nuclear and cytoplasmic fractions, isolated RNA and performed amplicon sequencing of each fraction to analyse mRNA localization of each GFP variant. Analysis of fractions showed the expected enrichment of the lncRNA MALAT1 in the nucleus, and of tRNA in the cytoplasm, confirming the quality of fractionations (Figure 4A). For each GFP variant, we calculated the relative cytoplasmic concentration of its mRNA (RCC) as the ratio of cytoplasmic read counts to the sum of reads from both fractions (RCC = c_cyto / (c_cyto+c_nuc); Figure 4B). A value of 0 therefore indicates 100% nuclear retention, whereas a value of 1 indicates 100% cytoplasmic localization. In the absence of splicing, RCC scores ranged from 0.09 to 0.64 and RCC correlated significantly with GC content ($r=0.51$, $p=3.85\times10^{-13}$, Figure 4C). In the presence of a 5’ UTR intron, we observed a significant increase in RCC score for GFP variants with low GC content, but no increase in RCC for GC-rich variants (Figure 4D). GC3 content at the beginning of the coding sequence was significantly correlated with RCC in the absence of splicing ($r=0.5$, $p=2.0\times10^{-11}$), but not in the presence of splicing ($r<0.01$, $p=0.48$; Figure S5B). Thus, high GC content at the 5' end of genes increases gene expression in part through facilitating the cytoplasmic localization of mRNA.

To assess whether GC content also affects translational dynamics, we performed polysome profiling on HEK293 GFP pool cells using sucrose gradient fractionation (Figure 5A). qRT-PCR analysis of RNA extracted from all collected fractions showed a broad distribution of GFP across fractions, with enrichment within polysome-associated fractions. In order to determine distribution patterns of individual GFP variants, RNA from several fractions was pooled (as indicated in Figure 5B) and subjected to high-throughput sequencing. The resulting read distribution indicates that GC-rich variants are associated with denser polysomal fractions (ribosome density, Figure 5C, left panel; $R^2=0.55$, $p < 2.2\times10^{-16}$) and are more likely to be translated (ribosome association, Figure 5C, right panel; $R^2=0.28$, $p<9.03\times10^{-15}$), compared to GC-poor variants. This suggests that enhanced translational dynamics also contribute to more efficient expression of GC-rich genes.

The expression fate of endogenous RNA depends on splicing, nucleotide composition, and cell type

To test whether splicing- and position-dependent effects of codon usage can be observed among human genes, we turned to genome-wide measurements of expression at endogenous human loci and related these measurements to codon usage and splicing. Although the correlations between GC content and expression depended on the experimental measure and type of cells under study, we find that GC4 content usually has a more positive effect on gene expression in unspliced genes relative to spliced ones (Figure 6, Table S1). In particular, unspliced mRNAs show a more positive/less negative correlation of GC4 with transcription initiation (GRO-cap data); cytoplasmic stability (exosome mutant); RNA (whole cell RNA-seq); cytoplasmic enrichment (cell fractionation), translation rate (ribosome profiling vs whole cell RNA-seq); and protein amount (mass-spec). These analyses suggest that GC4 content has an effect on the RNA abundance of intronless mRNA molecules, which is carried through to the protein expression. Taken together, these genome-wide analyses support our observation of a splicing-dependent relationship between codon usage and expression in human cells.

Discussion

We have shown that the effects of GC content on gene expression in human cells are splicing-dependent (the effect is larger in unspliced genes compared to spliced genes) and position-dependent (the effect is larger at the 5' end of genes than at the 3' end). In addition, human genes show striking patterns of codon usage, which differ between spliced and unspliced genes and between first and subsequent exons. Our results have implications for the understanding of the evolution of human genes and the functional consequences of synonymous codon usage.

Mechanisms of splicing- and position-dependent effects of codon usage

Specific patterns of codon usage have previously been found at the 5' ends of genes in bacteria, yeast and other species (Gu et al., 2010; Kudla et al., 2009; Tuller et al., 2010). In bacteria and yeast, strong mRNA folding near the start codon prevents ribosome binding and reduces translation efficiency, resulting in selection against strongly folded 5' mRNA regions (Kudla et al., 2009; Shah et al., 2013). In addition a "ramp" of rare codons has been observed near the 5' end of RNAs in multiple species, with a possible role in preventing a wasteful accumulation of ribosomes on mRNAs (Tuller et al., 2010) or reducing the strength of mRNA folding (Bentele et al., 2013). These phenomena cannot explain our results in human, because both the folding energy and codon ramp models predict low GC content near the start codon, whereas we observe high GC content within first exons of human protein-coding genes (Figure 1B). Furthermore, our experiments show that high GC content near the start codon increases expression, whereas the folding energy and codon ramp models would predict low expression.

We propose instead that splicing- and position-dependent effects of GC content are explained by early post-transcriptional events in the lifetime of an mRNA. Using matched reporter gene libraries, we show that most, but not all, variants show an increase in expression when spliced. Splicing typically increases the expression of AT-rich variants, but it does not further increase the expression of GC-rich transcripts, which suggests that splicing and high GC content influence expression through at least one common mechanism. Splicing increases transcription (Kwek et al., 2002), prevents nuclear degradation (Nott et al., 2003), facilitates nuclear-cytoplasmic mRNA export through the Aly/REF-TREX pathway (Muller-McNicoll et al., 2016), and stimulates translation (Nott et al., 2004). High GC content might increase RNA polymerase processivity (Bauer et al., 2010; Zhou et al., 2016); AT-rich genes are more likely to contain cryptic polyadenylation sites (consensus sequence: AAUAAA) (Higgs et al., 1983; Zhou et al., 2018) or destabilizing AU-Rich Elements (AREs); and AU-rich mRNAs may be preferentially localized in P-bodies (Courel et al., 2019) or in the nucleus (this study). GC-rich sequence elements of endogenous unspliced genes were previously shown to route transcripts into the splicing-independent ALREX nuclear export pathway, allowing efficient cytoplasmic accumulation (Palazzo et al., 2007). In agreement with this, low expression caused by inhibitory sequence features (such as low GC-content) can be rescued by extending the mRNA at the 5'end with a GC-rich sequence (Figure 3E). This may act as a compensatory mechanism when gene expression cannot rely on the positive regulatory effects of splicing (Palazzo and Akef, 2012). In contrast, it was recently shown that binding of HNRNPK to the GC-rich SIRLOIN motif leads to nuclear enrichment of lncRNAs (and also some mRNAs) (Lubelsky and Ulitsky, 2018). Our genomic analyses of lncRNA sequences do not show the same splicing-dependent compositional patterns as observed in mRNAs and it is therefore likely that antagonistic pathways act simultaneously in shaping the RNA expression landscape. Thus, we propose that the genomic patterns and their consequences on gene expression reported here are general features of protein-coding genes.

Recent studies highlight patterns of codon usage as major determinants of RNA stability in yeast (Presnyak et al., 2015), zebrafish (Mishima and Tomari, 2016) and other species (Bazzini et al., 2016). The usage of less common, 'non-optimal' codons within transcripts was shown to control poly-A tail length and RNA half-life in a translation-dependent manner through the coupled activity of different CCR4-NOT nucleases (Radhakrishnan et al., 2016; Webster et al., 2018). Consistent with these findings, we observed that CAI is positively correlated with mRNA expression levels in human cells. However, it remains to be seen whether the correlation of CAI with mRNA expression depends on translation. Because of the strong correlation between GC content and CAI, it is difficult to disentangle independent contributions of these variables. Additionally, we find that the correlation between GC content (or CAI) and expression is position- and splicing-dependent, whereas no evidence for such context-dependence has been reported for the CCR4-NOT-mediated mechanism.

Other instances in which the effects of codon usage are context-dependent have been described. Most notably, tRNA populations and transcriptome codon usage patterns were shown to differ between mammalian tissues (Dittmar et al., 2006; Gingold et al., 2014; Plotkin et al., 2004; Rudolph et al., 2016). Intriguingly, genes preferentially expressed in proliferating cells and tissue-specific genes tend to be AT-rich, whereas genes expressed in differentiated cell types and housekeeping genes are more GC-rich (Gingold et al., 2014; Vinogradov, 2003). Although these differences have been interpreted in terms of the match between codon usage and cellular tRNA pools, it is plausible that translation-independent mechanisms contribute to context-dependent effects of codon usage. Accordingly, in Drosophila, codon optimality determines mRNA stability in whole cell embryos, but not in the nervous system, independent of tRNA abundance (Burow et al., 2018). Recently, it was shown that Zinc-finger Antiviral Protein (ZAP) selectively recognises high CpG-containing viral transcripts as a mechanism to distinguish self from non-self (Takata et al., 2017). We speculate that similar regulatory proteins and mechanisms exist for cellular expressed genes. The cell lines used in the present study, HeLa and HEK293, are both rapidly proliferating and experimental results are correlated (r=0.36, Flow-Seq data), but divergent expression of some GFP variants was also observed. Similarly, the effect size of GC content on the expression of endogenously expressed genes varies with cell type. It would be interesting to compare the expression of our variants in other cell types to further address the question of tissue-specific codon usage and adaptation to tRNA pools.

Implications for the evolution of protein-coding genes

The fact that long, multi-exon genes are often found in GC-poor regions of the genome might result from regional mutation bias, but an alternative explanation is possible: GC-poor genes may be under selective pressure to retain their introns, and intronless genes may experience selective pressure to increase their GC content. These alternative explanations are supported by multiple observations: Firstly, endogenous intronless genes are on average more GC-rich than intron-containing genes. Secondly, the GC content of functional (but not non-functional) retrogenes is higher compared to their respective intron-containing parental genes, which cannot be explained by a systematic integration bias. Thirdly, in genome-wide analysis, correlations between GC-content and expression are generally more positive (or less negative) for unspliced compared to spliced genes. Taken together, this suggests that for the long-term success of an unspliced gene (i.e. stable conservation of expression and functionality) an increase in GC content is essential. By contrast, splicing allows genes to remain functional even when mutation bias or other mechanisms lead to a decrease of their GC content.

Acknowledgments

We thank Elisabeth Freyer from the IGMM FACS facility for help with cell sorting; Andrew Jackson, Nick Gilbert and Aleksandra Helwak for gifts of cell lines and plasmids; James Brindle for technical assistance; members of Kudla and Hurst groups for discussions; Edinburgh Genomics (University of Edinburgh) and the Imperial BRC Genomics facility for next-generation sequencing; and the IGMM technical support facility for help with media preparation and sequencing. This work was supported by the Wellcome Trust (Fellowships 097383 and 207507 to GK), the European Research Council (Advanced grant ERC-2014-ADG 669207 to LDH), the Medical Research Council (Grants MC_UU_00007/11 to MST. and MC_UU_00007/12 to GK and PhD studentship to CM), and Thermofisher (Cross Collaboration Grant to ML and GK).

Author Contributions

CM and GK conceived the work and designed experiments. CM and JB performed experiments. ML provided reagents and analysis tools. CM, RS, RSY, LT, JL and GK analysed data. ML, MST and LDH provided expertise and feedback. CM and GK wrote the paper.

Declaration of Interests

The authors declare no competing interests.

Figure 1. Splicing- and position-dependent patterns of nucleotide composition in human genes.

(A) GC4 distribution of human protein-coding genes, grouped by number of exons per gene. The Y axis indicates the proportion of genes within a given range of GC4.

(B) Mean GC4 content in protein-coding exons, grouped by exon position (rank) and by number of exons per gene.

(C) Mean GC4 for individual codons within exons of rank 1 (black dots) or rank 2 (white dots) downstream of the transcription start site (TSS).

(D) GC4 distribution of functional retrogenes (dark grey) and their corresponding parental genes (light grey) conserved between mouse and human

($p=2.1\times10^{-4}$, from one-tailed Wilcoxon signed rank test, n=49). See also Figure S1.

Figure 2. The effect of GC content on gene expression depends on splicing.

(A-B) Protein levels of 22 GFP variants when transiently expressed as unspliced (A) or spliced (B) constructs in HeLa cells and quantified by spectrofluorometry. Each data point represents the mean of 9 replicates, +/- SEM. GFP Relative Fluorescence Units (RFU) are defined as (GFP fluorescence - background GFP fluorescence)/(mKate fluorescence - background mKate fluorescence), where background fluorescence was measured in mock-transfected cells.

(C) Correlation of protein levels between unspliced and spliced variants of GFP (n=22, $R^2=0.69$, $p=9.0\times10^{-7}$). The dashed line indicates x=y.

(D-E) Protein levels of 23 mKate2 variants in the absence (D) or presence (E) of splicing. Each data point represents the mean of 9 replicates, +/- SEM. mKate RFU are defined as (mKate fluorescence - background mKate fluorescence), where background fluorescence was measured in mock-transfected cells.

(F) Correlation of protein levels between unspliced and spliced variants of mKate2 (n=23, $R^2=0.29$, $p=2.8\times10^{-4}$).

(G-H) mRNA levels of 10 GFP variants when transiently expressed as unspliced (G) or spliced (H) constructs in HeLa cells and quantified by qRT-PCR. Data points represent the mean of 3 replicates, +/- SEM, calculated as (GFP RNA)/(NeoR RNA).

(I) Comparison of mRNA expression from spliced and unspliced GFP variants (n=10, $R^2=0.49$, p=0.014).

(J) Intronic RNA levels of GFP variants measured by qRT-PCR, calculated as (GFP intronic RNA)/(NeoR RNA).

(K) RNA stability time course of 6 GFP variants expressed from stably transfected HEK293 Flp-in cells after blocking transcription with 500 nM triptolide. Variants were expressed as unspliced and spliced constructs. Results represent the averages of 2 independent experiments. RNA stability of c-myc (n=12) and GAPDH (n=6) are shown as unstable and stable RNA controls. See also Figures S2 and S3.

Figure 3. Splicing- and position-dependent effects of codon usage on protein production.

(A) Schematic outline of Flow-Seq experimental workflow. Stable HeLa and HEK293 cell pools expressing 217 GFP variants were established using a multiplex Flp-In integration approach, followed by FACS sorting, sequencing and calculation of a fluorescence score for each variant (see Figure S4).

(B) Heatmap representation of Flow-Seq results. Rows represent normalised read distributions of individual GFP variants across 8 fluorescence bins (columns). The average difference between lowest and highest fluorescence bins is around 100-fold. Data shown represents the average of 3 Flow-Seq measurements for HeLa cells, the average of 2 Flow-Seq experiments for HeLa with intron and 1 experiment for HEK293 cells.

(C) Pearson correlation matrix of experimental measurements obtained by Flow-Seq and sequence covariates. The colour of squares indicates the correlation coefficient; crosses indicate non-significant correlations ($p>0.05$).

(D) Correlations between Flow-Seq measurements and GC3 content of 1st (nt 1-360) and 2nd (nt 361 - 720) halves of GFP sequences.

(E) Protein and mRNA measurements of translational fusion constructs between GC-poor (30% GC3, Kpoor) and GC-rich (85% GC3, Krich) variants of mKate2 with a GC-rich (97% GC3, Grich) or GC-poor (33%, Gpoor) variants of GFP. Data represents the mean of 3 replicates, +/- SEM. GFP protein RFU, mKate protein RFU and RNA AU were defined as in Figure 2.

(F) Protein fluorescence measurements of 2 GC-poor GFP variants (GFP_154; GC3=0.38 and GFP_403; GC3=0.37) expressed either as unspliced constructs, or with an intron placed within the 5' UTR, the CDS or both. Data represents the mean of 3 replicates, +/- SEM. All intron-containing constructs differ significantly from their intronless counterparts ($p<0.05$, t-test). GFP protein RFU were defined as (GFP fluorescence - background GFP fluorescence). See also Figures S4 and S5.

Figure 4. High GC content increases cytoplasmic localisation of mRNA.

(A) Stable HeLa pools expressing 217 GFP variants +/- intron were fractionated into nuclear and cytoplasmic portions before RNA extraction. Specific markers of subcellular compartments were quantified by qRT-PCR before amplicon-library preparation.

(B) Relative cytoplasmic concentration (RCC) of unspliced and spliced GFP variants. Data represents the mean of 2 replicates. ***p=2×10^{-6}.

(C) Correlation between GC3 content and RCC for unspliced and spliced GFP RNA. Data points represent the means of 2 replicates.

(D) Correlation between RCC scores of unspliced and spliced GFP (R^2=0.1, p=2.6×10^{-5}). See also Figure S5.

Figure 5. High GC content leads to increased ribosome association.

(A) (Left) A stable pool of HEK293 cells expressing 217 unspliced GFP variants was subjected to polysome profiling using sucrose gradient centrifugation. (Right, from top to bottom) UV absorbance profile, GFP mRNA abundance, GAPDH mRNA abundance, ethidium bromide staining of gradient fractions. GFP and GAPDH mRNA were quantified by qRT-PCR.

(B) RNA from collected fractions was combined into 4 pools (as indicated by coloured boxes) before amplicon library preparation for high-throughput sequencing: unbound ribonucleoprotein complexes (red), monosomes (yellow), light polysomes (light green) and heavy polysomes (dark green). Resulting read distributions (in %) for GFP variants are represented as heatmap.

(C) Correlation plot between mean ribosome density (left panel) and ribosome association (right panel) of GFP variants and their corresponding GC3 content. Triangles indicate outliers (Ribosome association values 24.89 (GC3=0.84) and 24.80 (GC3=0.90)). The ribosome density and ribosome association measures were calculated as described in the methods section.

Figure 6. Splicing-dependent codon usage shapes global gene expression.

Effects of GC4 content on the expression of unspliced (y-axis) and spliced (x-axis) endogenous human genes, both on RNA and protein level. Each point corresponds to the regression coefficient of an individual experiment (cell line and/or biological replicate). Error bars indicate the standard error of these regression coefficients. Surrounding ellipses indicate the 95% confidence interval for 1,000 bootstraps of underlying data (see Methods, Figure S6 and Table S1). The diagonal indicates x=y. See also Figure S6 and Table S1.

STAR Methods

Lead contact and materials availability

Further information and requests for resources and reagents should be directed to, and will be fulfilled by, Grzegorz Kudla (gkudla@gmail.com). Plasmids generated in this study will be distributed by Grzegorz Kudla.

Experimental model and subject details

HeLa Flp-in T-Rex cells were obtained from the Andrew Jackson group, HEK293 Flp-in T-Rex cells were sourced from ThermoFisher, and HeLa cells were from ATCC.

Genes and plasmids

The library of 217 synonymous GFP variants used here consists of 138 variants from an earlier study (Kudla et al., 2009), 59 new variants assembled using the PCR-based method described in (Kudla et al., 2009), and 22 variants that were designed *in silico* and ordered as synthetic gene fragments (gBlocks) from Integrated DNA Technologies (IDT) (Mittal et al., 2018). Each of the 22 variants was designed by setting a target GC3 content (between 25 and 95%) and randomly replacing each codon with one of its synonymous codons, such that the expected GC3 content at each codon position corresponded to the target GC3 content. For example, to design a GFP variant with GC3 content of 25%, each glycine codon was replaced with one of the four synonymous glycine codons with the following probabilities: GGA, 37.5%; GGC, 12.5%, GGG, 12.5%; GGT, 37.5%. We also generated 23 mKate2 sequences using an analogous procedure and ordered the variants as gBlocks from IDT. All the genes were cloned into the Gateway Entry vector pGK3 (Kudla et al., 2009).

Construction of transient expression vectors

Plasmids used in transient transfection experiments are based on pCI-neo (Promega), a CMV-driven mammalian expression vector that contains a chimeric intron upstream of the multiple cloning site (MCS) within the 5’ UTR. This intron consists of the 5’ splice donor site from the first intron of the human beta-globin gene and the branch and 3’ splice acceptor site from the intron of immunoglobulin gene heavy chain variable region (see pCI-neo vector technical bulletin, Promega). This vector was adapted to be compatible with Gateway recombination cloning by inserting the Gateway-destination cassette, RfA, using the unique EcoRV and SmaI restriction sites present within the MCS of pCI-neo, generating pCM2. This plasmid was then further modified by removing the intron contained within the 5’ UTR by site-directed deletion mutagenesis using Phusion-Taq (ThermoScientific) and primers ‘pCI_del_F’ and ‘pCI_del_R’ (see Table S2 for list of all primers used), generating plasmid pCM1.

To be able to normalise spectrophotometric measurements from single GFP transfection experiments, pCM1 and pCM2 were further modified to contain a separate expression cassette driving the expression of a second fluorescent reporter gene, mKate2. The mKate2 gene cassette from pmKate2-N (Evrogen) was inserted via Gibson assembly cloning: First, the entire mKate2 expression cassette was amplified using primers ‘mKate2_gibs_F’ and ‘mKate2_gibs_R’ which add overhangs homologous to the pCM insertion site. Next, pCM1 and pCM2 were linearised by PCR using primers ‘pCI_gib_F’ and ‘pCI_gib_R’. All PCR products were purified using the Qiagen PCR purification kit and fragments with homologous sites recombined using the Gibson assembly cloning kit (NEB) according to manufacturer’s instructions (NEB). Successful integration was validated by Sanger sequencing. This generated plasmids pCM3 (-intron, +mKate2) and pCM4 (+intron, +mKate2).

Transient plasmid transfections for spectrofluorometric measurements

Plasmids for transient expression of fluorescent genes were transfected into HeLa cells grown in 96-well plates. Per plasmid construct, 3 replicates were tested by reverse transfection. Enough transfection mix for 4 wells was prepared by diluting 280ng plasmid DNA in 40ul OptiMem (Gibco). 1ul Lipofectamine2000 (Invitrogen; 0.25ul per well) was diluted in 40ul OptiMem and incubated for

5min at room temperature. Both plasmid and Lipofectamine2000 dilutions were then mixed (80ul total volume) and further incubated for 20-30min. 20ul of transfection complex was then pipetted into each of 3 wells before adding 200ul of HeLa cell suspension (45,000 cells/ml; 9,000 cells/well) in phenol red-free DMEM (Biochrom, F0475). Media was exchanged 3-4h post-transfection to reduce toxicity. Cells were then grown for a further 24h or 48h at 37C, 5% CO2.

After incubation, cells were lysed by removing media and adding 200ul of cell lysis buffer (25mM Tris, pH 7.4, 150mM NaCl, 1% Triton X-100, 1mM EDTA, pH 8). Fluorescence readings were obtained using a Tecan Infinite M200pro multimode plate reader. The plate was first incubated under gentle shaking for 15min followed by fluorescence measurements using the following settings: Ex486nm/Em 515nm for GFP and Ex588nm/Em633nm for mKate2; reading mode: bottom; number of reads: 10 per well; gain: optimal.

For data analysis, measurements of untransfected cells were subtracted as background from all other wells. For comparability of different plates within a set of experiments, the same 3 genes were transfected on every plate to account for technical variability. In the screen of individual GFP variants (see Figure 2), GFP measurements were divided by mKate2 measurements from same wells to reduce noise caused by well-to-well variation in transfection efficiency, but similar results were obtained without normalisation.

Transient transfections and RNA extraction for qRT-PCR analysis

HeLa cells were reverse transfected in 12-well plates using 800ng plasmid DNA and 2ul Lipofectamine 2000 (Invitrogen). DNA and Lipofectamine 2000 were diluted in 100ul OptiMEM (Gibco) each, incubated for 5min, mixed and further incubated for 20min. The transfection complex was then added to each well before adding 10^5 HeLa cells. Cells were incubated for 24h at 37C, 5% CO2 before harvesting. Cells were then harvested by adding 1ml Trizol reagent (Life technologies). RNA was extracted according to manufacturer's instructions. Resulting RNA was further treated with DNAse I using the Turbo DNase kit (Ambion) to remove any residual plasmid and genomic DNA.

RT-PCR analysis

cDNA for qRT-PCR analysis was prepared using SuperScript III Reverse Transcriptase (Life technologies) according to the manufacturer's recommendations with 500ng total RNA as template and 500ng random hexamers (Promega). All qRT-PCRs were carried out on a Roche LightCycler 480 using Roche LightCycler480 SYBR Green I Master Mix and 0.3uM gene-specific primers. Samples were analysed in triplicate as 20ul reactions, using 2ul of diluted cDNA. Cycling settings: DNA was first denatured for 5min at 95°C before entering a cycle (50-60x) of denaturing for 10sec at 95°C, annealing for 7sec at 55-60°C (depending on primers used), extension for 10sec at 72°C and data acquisition. DNA was then gradually heated up by 2.20 °C/s from 65 to 95°C for 5sec each and data continuously collected (Melting curve analysis). Data was evaluated using the comparative Ct method (Livak and Schmittgen, 2001). RNA measurements from transient transfection experiments were normalised to the abundance of neomycin resistance marker (NeoR) RNA, which is expressed from the same plasmid, to control for differences in transfection efficiency (primers 'Neo_F' and 'Neo_R'). PCRs performed on cDNA from stable Flp-in T-Rex cell lines to measure splicing efficiency were performed on an Eppendcorf Mastercycler nexus X2 in 20ul reaction volumes, using Accuprime Pfx (ThermoFisher) according to manufacturer's instructions, using 0.3uM primers (intron-independent: pc5_5UTR_F & pc5_3UTR_R1; intron specific: pc5_INT_F & pc5_3UTR_R2).

Subcellular fractionation

This protocol is based on the cellular fractionation protocol published by (Gagnon et al., 2014) but includes a further clean-up step using a sucrose cushion as described by (Zaghlool et al., 2013) and a second lysis step as described by (Wang et al., 2006). Cell lysis and nuclear integrity was monitored throughout by light microscopy following Trypan blue staining (Sigma). Cells were grown in 10cm plates for 24h to about 90% confluency. Cells were then washed with PBS and trypsinised briefly using 1ml of 1xTrypsin/EDTA. After stopping the reaction with 5ml DMEM, cells were transferred into 15ml falcon tubes and collected by spinning at 100g for 5min. Resulting cell pellets were resuspended in 500ul ice-cold PBS, transferred into 1.5ml reaction tubes and spun at 500g for 5min, 4°C.

The supernatant was discarded and cells resuspended in 250ul HLB (10mM Tris (pH 7.5), 10mM NaCl, 3mM MgCl2, 0.5% (v/v) NP40, 10% (v/v) Glycerol, 0.32M sucrose) containing 10% RNase inhibitors (RNasin Plus, Life Technologies) by gently vortexing. Samples were then incubated on ice for 10min. After incubation, samples were vortexed gently, spun at 1000g for 3min, 4°C, and supernatants and pellets were processed separately as indicated in a) and b) below.

a) Cytoplasmic extract:

The supernatant was carefully layered over 250ul of a 1.6M sucrose cushion and spun at 21,000g for 5min. The supernatant was then transferred into a fresh 1.5ml tube and 1ml Trizol was added and mixed by vortexing.

b) Nuclear extract:

The pellets were washed 3 times with HLB containing RNase inhibitors by gently pipetting up and down 10 times followed by a spin at 300g for 2min. After the 3rd wash, nuclei were resuspended in 250ul HLB and 25ul (10%) of detergent mix (3.3% (wt/wt) sodium deoxycholate/6.6% (vol/vol) Tween 40) dropwise added while vortexing slowly (600rpm). Nuclei were then incubated for 5min on ice before spinning at 500g for 2min. The supernatant was discarded and pellets resuspended in 1ml Trizol (Ambion) by vortexing. 10ul 0.5M EDTA are added to each nuclear sample in Trizol and tubes heated to 65°C for 10min to disrupt very strong Protein-RNA and DNA-RNA interactions. Tubes were then left to reach room temperature and RNA was extracted following the manufacturer's instructions.

Transcription inhibition assay

HeLa T-Rex Flp-in cell lines were grown to 80-90% confluency in 6 well for 24h before treatment with 500nM Triptolide (Sigma). Cells were harvested at indicated time points and RNA extracted using the Qiagen RNeasy kit (Qiagen, 74104). Control cells were treated with an equal volume of DMSO (drug carrier). To assess transcript levels, qRT-PCR was performed as described above using primers 'pc5_3UTR_F' and 'pc5_3UTR_R1'. GFP levels were normalised to levels of 7SK, a RNA polymerase III-transcribed non-coding RNA, whose expression levels are not affected by Triptolide treatment. Relative transcript levels of c-Myc are shown as an example of a relatively unstable transcript, while levels of Gapdh are shown as a stable transcript. Transcript half-lives ($t_{1/2}$) were calculated by first fitting an exponential decay curve, $y(x) = a \times e^{kx}$, through the data points to obtain the decay constant k. The half-life is then calculated as $t_{1/2} = \ln(2)\,/k$.

Generation of stable Flp-in cell lines

We adopted a multiplex-Gateway integration method to create a pool of 217 GFP plasmids which are compatible with the T-Rex Flp-in system (Invitrogen) for creating stable, doxycycline-inducible cell lines, in which each variant is expressed from the same genomic locus, allowing direct comparison of expression levels.

pcDNA5/FRT/TO/DEST (Aleksandra Helwak, University of Edinburgh) contains the Gateway-compatible attB destination cassette to allow the subcloning of genes from any Gateway-entry vectors. This plasmid was further modified to contain the same 5' UTR intron sequence as in pCM4 used in transient expression experiments using Gibson Assembly (NEB): the intronic sequence was amplified from pCM4 by PCR using primers 'Gib_intr_F' and 'Gib_intr_R' using Q5 High-Fidelity Polymerase (NEB). The primers added 15nt overhangs which are homologous to the ends of pcDNA5/FRT/TO/DEST when linearised with AflII. The Gibson assembly reaction was performed as per manufacturer's instructions (NEB), generating pcDNA5/FRT/TO/DEST/INT.

217 individual GFP variants stored in Gateway-entry vector pGK3 were mixed with a concentration of 0.06ng of each GFP variant. For each pcDNA5 destination vector, a separate Gateway LR reaction was set-up in a total volume of 45ul using 500ng destination vector, 5ul LR Clonase enzyme mix, 38ul of the mixed 217 pGK3-GFP plasmids and TE (pH 8). The reactions were incubated at 25C overnight followed by Proteinase K digest (5ul, LR Clonase kit) for 10min at 37C. The total 50ul reaction mix was transformed into 2.5ml highly competent DH5alpha in a 15ml Falcon tube by heat-shocking cells for 2min 30s at 42C, followed by cooling on ice for 3min, before adding 10ml SOC medium and incubating while shaking for 1h at 37C. After incubation, cells were spun down at 3000g for 3min and resulting bacterial pellets resuspended in 1ml fresh SOC. 10x100ul were plated onto L-Ampicillin agar plates and incubated overnight at

37C resulting in >800 colonies per plate. Bacterial colonies were scraped off the plates and collected in a falcon tube. Plasmid DNA was extracted using a Qiagen Midiprep kit according to the manufacturer's instructions, resulting in two plasmid pools: pCDNA5/GFPpool and pcDNA5/INT/GFPpool. Both pools were subjected to high-throughput sequencing to confirm the presence of different GFP variants.

HeLa T-Rex Flp-in cells (gifted by the Andrew Jackson lab, The University of Edinburgh) and HEK293 T-Rex Flp-in (Thermo Scientific) were grown to 80% confluency in 6 well plates. For GFP plasmid pool transfections, pCDNA5/GFPpool or pCDNA5/INT/GFPpool were mixed in a 9:1 ratio with the Flp-recombinase expression plasmid pOG44 (Invitrogen) to give 2ug in total (1.8ug pOG44 + 0.2ug pCDNA5) and diluted in OptiMEM (Gibco) to 100ul. Transfections were performed with 9ul Lipofectamine2000 (Invitrogen) and 91ul OptiMEM per well by incubating 5min at room temperature before mixing with plasmid DNA and a further 15min incubation. The transfection mix was then added dropwise to the cells. Media were replaced with conditioned media 4h post-transfection. Cells were incubated for further 48h before chemical selection to select for successful gene integration using 10ng/ul Blasticidin S (ThermoFisher) and 400mg/ml (HeLa T-Rex Flp-in) or 100mg/ml (HEK293 T-Rex Flp-in) Hygromycin B (Life Technologies). Successful selection was determined by monitoring cell death in untransfected cells. Chemically resistant cells represent pools of cell lines expressing different GFP variants from the same genomic locus. High-throughput sequencing of the GFP integration site within each generated cell line pool confirmed the successful integration of all variants.

HeLa T-Rex Flp-in and HEK293 T-Rex Flp-in cell lines expressing individual intron-containing and intronless GFP variants were generated using the same protocol.

Flow-Seq: FACS sorting and genomic DNA extraction

80x15cm cell culture plates of HeLa T-Rex Flp-in GFP pool cells and 40x15cm cell culture plates of HEK293 T-Rex Flp-in GFP pool cells were induced with 1ug/ml Doxycyline (Sigma, D9891) in phenol red-free DMEM (Biochrom, F0475) supplemented with 10% FCS (Sigma, F-7524) and 2mM L-Glutamine. After 24h or 48h, cells were harvested by gentle trypsinisation and cells were sorted into 8 fluorescence bins using a BD FACS Aria II cell sorter. To define the range of GFP positive signal, cells without stable GFP expression were used as negative control. 80% of HeLa and 90% HEK293 GFP pool cells fell into the GFP-positive range. Each fluorescence bin was chosen to comprise roughly 10% of the GFP-positive population. The bin spacing was kept the same for the sorting of HeLa cell pools expressing unspliced and spliced GFP variants to allow direct comparisons of the fluorescence profiles of individual variants.

About 10^7 cells per bin were collected in Polypropylene collection tubes (Falcon) coated with 1% BSA/PBS, cushioned with 200ul 20%FBS/PBS. Cell suspensions were decanted into 15ml tubes and cells collected by spinning 5min at 500g. The supernatant was transferred into fresh 15ml tubes and precipitated using 2 volumes of 100% EtOH/0.1 volume Sodium Acetate (pH 5.3) and 10ul Glycoblue (Ambion). Tubes were shaken vigorously for 10s before incubating at -20C for 15min, followed by spinning at 3000g for 20min. Resulting pellets were air-dried, resuspended in 1ml digest buffer (100mM Tris pH 8.5, 5mM EDTA, 0.2% SDS, 200mM NaCl) and then combined with the respective cell pellet. 10ul RNAse A (Qiagen, 70U) was added and samples gently rotated at 37C. After 1h, 1ul/ml Proteinase K (20mg/ml, Roche) was added to the samples before rotating a further 2h at 55C. Genomic DNA was purified 3 times by using 1 volume Phenol:Chloroform:Isoamyl alcohol (PCI, 25:24:1, Sigma). After each addition of PCI, samples were shaken vigorously for 10s before spinning at 3000g for 20min (first extraction) or 5min (all following). The resulting bottom layers including the interphase were removed before each PCI addition. After the last PCI extraction, the upper layer was transferred into a fresh 15ml tube and 1 extraction performed using 1 volume chloroform:isoamyl alcohol (CI,24:1, Sigma). After a 5min spin at 3000g, the upper layer was transferred into a fresh 15ml tube and DNA precipitated using EtOH/Sodium Acetate as before. After a 5min incubation on ice, DNA was collected by spinning for 30min at 3000g. The resulting DNA pellets were washed 2 times with 75% EtOH before air-drying and resuspending in 200ul Tris-EDTA (10mM). The quality of the extracted genomic DNA was assessed on a 0.8% Agarose/TBE gel.

Polysome profiling

HEK293 Flp-in GFP pool cell lines were grown to 90% confluency on 15cm dishes. Cells were treated for 20min with 100ug/ul Cycloheximide before harvesting cells by removing media, washing with 2x ice-cold PBS followed by scraping cells into 1ml PBS and transferring into 1.5ml tubes. Cells were pelleted at 7000rpm, 4°C for 1min and resulting cell pellet carefully resuspended by pipetting up and down in 250ul RSB (10x RSB: 200mM Tris (pH 7.5), 1M KCl, 100mM MgCl2) containing 1/40 RNasin (40U/ul, Promega), until no clumps were visible. 250ul of polysome extraction buffer was then added (1ml 10x RSB + 50ul NP-40 (Sigma) + 9ml H2O + 1 complete mini EDTA-free protease inhibitor pill (Roche)) and lysate passed 5x through a 25G needle avoiding bubble formation. The lysate was then incubated on ice for 10min before spinning 10min at 10,000g, 4°C. The supernatant was then transferred into a fresh 1.5ml tube and the RNA concentration estimated by measuring the OD at 260nm. Sucrose gradients (10–45%) containing 20 mM Tris, pH 7.5, 10 mM MgCl2, and 100 mM KCl were made using the BioComp gradient master. 100ug of Lysate were loaded on sucrose gradients and spun at 41,000rpm for 2.5h in a Sorvall centrifuge with a SW41Ti rotor. Following centrifugation, gradients were fractionated using a BioComp gradient station model 153 (BioComp 23 Instruments, New Brunswick, Canada) by measuring cytosolic RNA at 254 nm and collecting 18 fractions.

RNA from all fractions was precipitated using 1 volume of 100% EtOH and 1ul Glycoblue (Ambion), before extracting RNA using the Trizol method (Life Technologies). Equal volumes of RNA of each fraction was run on a 1.3% Agarose/TBE gel to assess the quality of fractionation and RNA integrity. Additionally, equal volumes of RNA of each fraction were used in cDNA synthesis using SuperScript III (ThermoFisher) and 2uM gene-specific primers for GFP ('pcDNA5-UTR_R') and GAPDH ('GAPDH_R') followed by qRT-PCR analysis. For high-throughput sequencing, total RNA from collected fractions was combined in equal volumes into 4 pools (as indicated in Figure 5B; free ribonucleoprotein (RNP) complexes, monosomes, light polysomes (2-4) and heavy polysomes (5+)) before amplicon library preparation (as described below).

High-throughput library preparation and sequencing

Sequencing libraries were generated by PCR using primers specific for GFP amplification (Table S2) which carry the required adaptor sequences for paired-end MiSeq sequencing, as well as 6nt indices for library multiplexing. Between 6-10ug of total genomic DNA were used in multiple PCR reactions (200ng per 50ul reaction). All PCRs were performed using Accuprime Pfx (NEB) according to manufacturer's recommendations using 0.4ul Accuprime Pfx Polymerase and 0.3uM of each primer ('PE_PCR_left' and 'S_indexX_right_PEPCR'). The cycling conditions were as follows: Initial denaturation at 95C for 2min, followed by 30 cycles of denaturation at 95C for 15sec, annealing at 51C for 30sec, extension at 68C for 1min. The final extension was performed at 68C for 2min. After PCR, all reactions of the same template were pooled and 1/3 of the reaction purified using the Qiagen PCR purification kit according to the manufacturer's instructions. DNA was eluted in 50ul H2O. Library size selection was performed using the Invitrogen E-gel system (Clonewell gels, 0.8% agarose) followed by Qiagen MinElute PCR purification. Correct fragment sizes were confirmed and quantified using the Agilent Bioanalyzer 2100 system.

For library preparation of RNA samples, 500ng RNA was first converted into cDNA using 2nmol GFP-specific primers ('S_indexX_right_PEPCR') using SuperScript III (Life technologies) according to manufacturer's protocol, using 50C as extension temperature. Resulting cDNA was then treated with 1ul RNaseH (NEB) for 20min at 37C, followed by heat inactivation at 65C for 5min. Samples were diluted 1:2.5 before using 2ul as template in PCR for library preparation. A minimum of 8x50ul PCR reactions were set up and pooled for each sample before PCR purification, followed by E-gel purification as described above.

High-throughput sequencing was conducted by Edinburgh Genomics (The University of Edinburgh) and Imperial BRC Genomics facility (Imperial College London) using the Illumina MiSeq platform (2x300nt paired-end reads).

4sU labelling and separation of nascent RNA

GFP expression was induced for 24h using 1ug/ml Doxycyline (Sigma, D9891) at 80% confluency in 15cm cell culture dishes. To label nascent RNA, 4sU (Sigma, T4509) was added to the media to a final concentration of 500 uM. Cells were then further incubated at 37C, 5%CO2 for 20min. After incubation, cells were harvested using 5ml Trizol reagent and RNA extracted following manufacturer's instructions using 1ml Chloroform and Phase Lock Gel Heavy tubes (15ml, Eppendorf). Resulting RNA pellet was resuspended in 100ul RNAse-free water, followed by a DNAse digest step using the TURBO DNA-free kit (Ambion) following manufacturer's instructions.

Biotin labelling reactions were set up as following: 100ug RNA + 2ul Biotin-HPDP (1mg/ml in DMF; Pierce, 21341) + 1ul 10x Biotinylation buffer (100mM Tris pH 7.4, 10mM EDTA) + H2O to 1ml. Reactions were then incubated for 1.5h at RT with rotation. Unincorporated biotin-HPDP was removed by 2 x chloroform extraction (1 volume) using Phase lock tubes (2ml, Eppendorf). The upper phase was then transferred to a DNA lobind tube (Eppendorf, 0030108051) and RNA precipitated using 1/10 reaction volume 5M NaCl and an equal reaction volume of 100% Isopropanol. Resulting RNA pellet was washed with 70% Ethanol before resuspending biotinylated RNA in 100ul RNAse-free water.

Streptavidin pull-down reactions were set up using 100ul biototinylated RNA (up to 100ug RNA) + 100ul Streptavidin beads (Miltenyi, 130074101) and reaction incubated for 15min at RT with gentle shaking. Streptavidin beads were then isolated using uMACS columns (Miltenyi, 130074101) attached to a magnetic stand. Columns were equilibrated with Washing buffer (WB; 100mM Tris pH 7.5, 10mM EDTA, 1M NaCl, 0.1% Tween20) before adding Streptavidin reaction mixtures to the column. Columns were then washed 3 times with WB heated to 65C, followed by 3 times with WB at RT. RNA was then eluted using 100ul freshly prepared 100mM DTT, followed by purification using the Qiagen RNeasy Minelute kit (Qiagen, 74204). RNA was eluted in 20ul RNAse-free water and concentration determined using the Qubit RNA HS assay kit (Life technologies, Q32852). cDNA synthesis was performed using equal amounts of RNA across all samples using SuperScript III and qRT-PCRs performed as described in section 'RT-PCR analysis' using primers specific for the 3' UTR (‘pc5_3UTR_F’ + ‘pc5_3UTR_R1’) and intronic sequence (‘pCI-premRNA_F’ + ‘pCI-premRNA-R’).

Quantification and Statistical analysis

Analysis of GFP pool experiments

Raw sequencing files (database accession number PRJNA596086) were demultiplexed by 6nt indices by the respective sequencing facility. To remove the plasmid sequence, the second reads from paired-end sequencing were trimmed using flexbar (-as ATGTGCAGGGCCGCGAATTCTTA -ao 4 -m 15 -u 30). Reads were then mapped to the GFP library using bowtie2 (-X 750) and filtered using samtools (-f 99).

For Flow-seq data, only variants with a minimum of 1000 reads across all 8 sequencing bins were used for further analysis. For each GFP variant, the number of reads in each bin (n(i)) was multiplied by the respective bin index (i) before taking the sum and dividing by the total number of reads across all bins:

Fluorescence (variant) = $\sum_{i=1}^{8} i * n(i) / \sum_{i=1}^{8} n(i)$

For cell fractionation experiments, only data with a minimum of 1000 reads across both cytoplasmic and nuclear fractions was used to calculate the relative cytoplasmic concentration (‘RCC’) for each variant: $RCC = \frac{n(cyto)}{n(cyto)+n(nuc)}$

For polysome profiling, only variants with a minimum of 1000 reads across all 4 sequencing bins were used for further analysis. To estimate ribosome density, for each GFP variant, the number of reads in each bin (n(i)) was multiplied by the respective bin index i (free RNA, i=1; monosomes, i=2; light polysomes, i=3; heavy polysomes, i=4) before taking the sum and dividing by the total sum of reads across all fractions:

Ribosome density(variant) = $\sum_{i=1}^{4} i * n(i) / \sum_{i=1}^{4} n(i)$

Ribosome association for each variant was calculated as the sum of reads (n) in light polysomes, heavy polysomes and monosomal fractions, divided by the sum of reads found in the free RNP fraction:

Ribosome association(variant) =

$$\frac{(\ n(monosomes) + n(light\ polysomes) + n(heavy\ polysomes))}{n(free\ RNPs)}$$

Definition of calculated sequence features

GC3: GC content in the third position of codons

CpG: number of CpG dinucleotides

dG: The minimum free energy of predicted mRNA secondary structure around the start codon was calculated using the hybrid-ss-min program version 3.8 (default settings: NA = RNA, t = 37, [Na+] = 1, [Mg++] = 0, maxloop = 30, prefilter = 2/2) in the 42-nt window (-4 to 38) as in (Kudla et al., 2009).

CAI: Codon Adaptation Index (*H. sapiens*) (Sharp and Li, 1987a) was calculated using a reference list of highly expressed human genes collected from the EMBL-EBI expression atlas https://www.ebi.ac.uk/gxa.

tAI: tRNA adaptation index (dos Reis et al., 2004)

ARE: top score of ATTTA motif match in each sequence.

AT-stretch: number of times motif (AT){9} was identified in each sequence.

GC-stretch: number of times motif (GC){9} was identified in each sequence.

Poly_A: number of times the position-specific scoring matrix ((47,3,0,50)(18,6,9,67)(53,12,12,23)(59,6,0,35)(70,6,6,18)) was identified in each sequence.

SD_cryptic: number of times RSGTNNHT motif was identified in each sequence.

SD_PSSM: number of times the position-specific scoring matrix ((60,13,13,14)(9,3,80,7)(0,0,100,0)(0,0,0,100)(53,3,42,3)(71,8,12,9)(7,6,81,6)(1 6,17,21,46)) was identified in each sequence.

FIMO (http://meme-suite.org) was calculated to identify and count sequence motifs. Open-source packages available for R were used for generating correlation matrices (corrplot), heatmaps (ggplot2), boxplots (graphics/ggplot2), The GC3 of all human coding sequences (assembly: GRCg38_hg38; only CDS exons) was calculated using R package ‘seqinr’.

Analysis of GC content variation in the human genome

The GRCh38 sequence of the human genome, as well as the corresponding gene annotations (Ensembl release 85), was retrieved from the Ensembl FTP site (Zerbino et al., 2018). The full coding sequences (CDSs) of protein-coding genes were extracted, filtered for quality and clustered into putative paralogous families (see (Savisaar and Hurst, 2016) for full details). For all analyses, a random member was picked from each putative paralogous cluster. In addition, only one transcript isoform (the longest) was considered from each gene. Note that exon rank was always counted from the first exon of the gene, even if it was not coding. In Figure 1A, density was calculated using the ggplot2 geom_density() function. For Figure 1C, GC4 was averaged across all sites that were at the same nucleotide distance to the TSS and within an exon of the same rank. For the functional retrocopies analysis, the parent-retrocopy genes derived in (Parmley et al., 2007) were used. Pseudogenic retrocopies were retrieved from RetrogeneDB (Rosikiewicz et al., 2017). Retrocopy annotations were filtered to only leave human genes with a one-to-one ortholog in *Macaca mulatta*. Next, only ortholog pairs where both the human and the macaque copy were annotated as not having an intact reading frame and where the human copy was annotated as *KNOWN_PSEUDOGENE* were retained. For the analyses reported in Figure S1, the functional retrocopies were also retrieved from RetrogeneDB, as we could not access genomic locations for the (Parmley et al., 2007) set. The functional retrogenes were retrieved similarly to pseudogenes, except that both the human and the macaque copy were required to have an intact open reading frame and the human copy could not be annotated as *KNOWN_PSEUDOGENE*.

Python 3.4.2. was used for data processing and R 3.1.2 was used for statistics and plotting (R Development Core Team, 2005).

Computation methods for analysis of endogenous gene expression

Data Collection

See also Table S1 for summary of datasets used.

1. GC4 content was calculated for each protein-coding transcript annotated in GENCODE version 19 as the GC content of the third codon position across all fourfold-degenerate codons (CT*, GT*, TC*, CC*, AC*, GC*, GA*, CC*, GC*). The core promoter of each transcript is further defined as -300 bp/+100 bp around the annotated TSS.

2. The level of transcription initiation was quantified in K562 and Gm12878 cells as the number of GRO-cap reads from the same strand which overlap the core promoter.

3. Nuclear stability was assessed using CAGE data obtained in triplicate from Egfp, Mtr4 and Rrp40 knockdowns (GSE62047; (Andersson et al., 2014)). Similarly to the approach used for the GRO-cap data, we calculated the RPKM across core promoters for each library separately. The baseMean expression for each treatment was quantified using DESeq2, where promoters with no reads across any replicate were first removed from each comparison. Nuclear stability was then assessed as the fold-change between the Egfp and Mtr4 knockdown and cytoplasmic stability by the estimated fold-change between the Mtr4 and Rrp40 knockdowns.

4. The level of the mature mRNA was quantified using RNA-seq libraries from whole cell samples (prepared as described elsewhere for HEK293 cells and downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq for Gm12878, HepG2, HeLa, Huvec and K562 cells). Reads were pseudoaligned against GENCODE transcript models using Kallisto, set with 100 bootstraps. All other parameters were left at their default. Transcript expressions were extracted as the estimated TPM (tags per million) values.

5. The level of the mature mRNA in the nuclear and cytoplasmic fractions was quantified using Kallisto as previously. As transcript stability was similar in both fractions (linear regression coefficient 0.97, $p < 2.2\times10^{-16}$), nuclear export was determined as the fraction TPM from these two compartments which was present in the nuclear fraction.

6. Ribosome-sequencing data from HEK293 (GSE94460) and HeLa (GSE79664) cells were used to quantify the level of mRNA translation in these two cells. Both of these measures were determined at the gene level, and so these observations were applied to all GENCODE transcripts

annotated to these associated genes. These data were normalised to the mean mRNA expression in the relevant cell types (from step 4).

7. Protein expression was assessed using mass-spectrometry data (Geiger et al., 2012) (Supp. Table 2) as the mean LFQ intensity across three replicates for each uniprot-annotated gene in each cell line for which data were available. Only data from genes where the UniProt ID is uniquely linked to a single transcript were considered in the analyses presented here.
8. Protein stability was calculated as the level of the mature protein in HEK293 and HeLa cells (step 7) relative to the mean rate of mRNA translation in these cells (step 6).

Regression modelling

A pseudocount of 0.0001 was added to each measurement of gene expression and, excluding the nuclear export data, these values were then log2-transformed to generate a normal distribution of expression for subsequent analysis. Transcripts with an expression value of 0 were removed from downstream analysis and the resulting distributions used for regression analysis are displayed in Figure S6. Transcripts were separated into unspliced and spliced, where splicing was defined as containing more than one exon in the GENCODE transcript model. Expression measurements were then linearly regressed against the GC4 content separately for each class of transcript and the coefficients along with their associated standard errors. These data were then bootstrapped by sampling with replacement and recalculating the regression coefficients for spliced and unspliced transcripts. The 95% confidence interval of these coefficients (discounting the standard error in these estimations) obtained by 1,000 samplings of this type was used to draw the ellipses shown in Figure 6**.**

Data and Software availability

Raw sequencing files have been deposited in SRA and can be accessed under database accession number PRJNA596086.

Reference list

Andersson, R., Refsing Andersen, P., Valen, E., Core, L.J., Bornholdt, J., Boyd, M., Heick Jensen, T., and Sandelin, A. (2014). Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nature communications *5*, 5336.

Arango, D., Sturgill, D., Alhusaini, N., Dillman, A.A., Sweet, T.J., Hanson, G., Hosogane, M., Sinclair, W.R., Nanan, K.K., Mandler, M.D.*, et al.* (2018). Acetylation of Cytidine in mRNA Promotes Translation Efficiency. Cell *175*, 1872-1886 e1824.

Arhondakis, S., Auletta, F., and Bernardi, G. (2011). Isochores and the regulation of gene expression in the human genome. Genome Biol Evol *3*, 1080-1089.

Bauer, A.P., Leikam, D., Krinner, S., Notka, F., Ludwig, C., Langst, G., and Wagner, R. (2010). The impact of intragenic CpG content on gene expression. Nucleic Acids Res *38*, 3891-3908.

Bazzini, A.A., Del Viso, F., Moreno-Mateos, M.A., Johnstone, T.G., Vejnar, C.E., Qin, Y., Yao, J., Khokha, M.K., and Giraldez, A.J. (2016). Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. EMBO J *35*, 2087-2103.

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., and Bluthgen, N. (2013). Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol *9*, 675.

Bernardi, G. (1993). The vertebrate genome: isochores and evolution. Mol Biol Evol *10*, 186-204.

Burow, D.A., Martin, S., Quail, J.F., Alhusaini, N., Coller, J., and Cleary, M.D. (2018). Attenuated Codon Optimality Contributes to Neural-Specific mRNA Decay in Drosophila. Cell reports *24*, 1704-1712.

Cambray, G., Guimaraes, J.C., and Arkin, A.P. (2018). Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. Nat Biotechnol *36*, 1005-1015.

Carels, N., and Bernardi, G. (2000). Two classes of genes in plants. Genetics *154*, 1819-1825.

Courel, M., Clement, Y., Bossevain, C., Foretek, D., Vidal Cruchez, O., Yi, Z., Benard, M., Benassy, M.N., Kress, M., Vindry, C.*, et al.* (2019). GC content shapes mRNA storage and decay in human cells. eLife *8*.

Dittmar, K.A., Goodenbour, J.M., and Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. PLoS Genet *2*, e221.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M.*, et al.* (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature *485*, 201-206.

dos Reis, M., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res *32*, 5036-5044.

Duan, J., Shi, J., Ge, X., Dolken, L., Moy, W., He, D., Shi, S., Sanders, A.R., Ross, J., and Gejman, P.V. (2013). Genome-wide survey of interindividual differences of RNA stability in human lymphoblastoid cell lines. Scientific reports *3*, 1318.

Duret, L., and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet *10*, 285-311.

Eyre-Walker, A.C. (1991). An analysis of codon usage in mammals: selection or mutation bias? J Mol Evol *33*, 442-449.

Fath, S., Bauer, A.P., Liss, M., Spriestersbach, A., Maertens, B., Hahn, P., Ludwig, C., Schafer, F., Graf, M., and Wagner, R. (2011). Multiparameter RNA and codon optimization: a standardized tool to assess and enhance autologous mammalian gene expression. PLoS One *6*, e17596.

Gagnon, K.T., Li, L., Janowski, B.A., and Corey, D.R. (2014). Analysis of nuclear RNA interference in human cells by subcellular fractionation and Argonaute loading. Nat Protoc *9*, 2045-2060.

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glemin, S., Bierne, N., and Duret, L. (2018). Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. Mol Biol Evol *35*, 1092-1103.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics *11*, M111 014050.

Gingold, H., Tehler, D., Christoffersen, N.R., Nielsen, M.M., Asmar, F., Kooistra, S.M., Christophersen, N.S., Christensen, L.L., Borre, M., Sorensen, K.D.*, et al.* (2014). A dual program for translation regulation in cellular proliferation and differentiation. Cell *158*, 1281-1292.

Goodman, D.B., Church, G.M., and Kosuri, S. (2013). Causes and effects of N-terminal codon bias in bacterial genes. Science *342*, 475-479.

Gradnigo, J.S., Majumdar, A., Norgren, R.B., Jr., and Moriyama, E.N. (2016). Advantages of an Improved Rhesus Macaque Genome for Evolutionary Analyses. PLoS One *11*, e0167376.

Gu, W., Zhou, T., and Wilke, C.O. (2010). A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol *6*, e1000664.

Higgs, D.R., Goodbourn, S.E., Lamb, J., Clegg, J.B., Weatherall, D.J., and Proudfoot, N.J. (1983). Alpha-thalassaemia caused by a polyadenylation signal mutation. Nature *306*, 398-400.

Kosovac, D., Wild, J., Ludwig, C., Meissner, S., Bauer, A.P., and Wagner, R. (2011). Minimal doses of a sequence-optimized transgene mediate high-level and long-term EPO expression in vivo: challenging CpG-free gene design. Gene Ther *18*, 189-198.

Kosuri, S., Goodman, D.B., Cambray, G., Mutalik, V.K., Gao, Y., Arkin, A.P., Endy, D., and Church, G.M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proc Natl Acad Sci U S A *110*, 14024-14029.

Kotsopoulou, E., Kim, V.N., Kingsman, A.J., Kingsman, S.M., and Mitrophanous, K.A. (2000). A Rev-independent human immunodeficiency virus type 1 (HIV-1)-based vector that exploits a codon-optimized HIV-1 gag-pol gene. J Virol *74*, 4839-4852.

Kudla, G., Lipinski, L., Caffin, F., Helwak, A., and Zylicz, M. (2006). High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol *4*, e180.

Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. Science *324*, 255-258.

Kwek, K.Y., Murphy, S., Furger, A., Thomas, B., O'Gorman, W., Kimura, H., Proudfoot, N.J., and Akoulitchev, A. (2002). U1 snRNA associates with TFIIH and regulates transcriptional initiation. Nat Struct Biol *9*, 800-805.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lercher, M.J., Urrutia, A.O., Pavlicek, A., and Hurst, L.D. (2003). A unification of mosaic structures in the human genome. Hum Mol Genet *12*, 2411-2415.

Li, W. (2011). On parameters of the human genome. J Theor Biol *288*, 92-104.

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.

Lubelsky, Y., and Ulitsky, I. (2018). Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. Nature *555*, 107-111.

Mishima, Y., and Tomari, Y. (2016). Codon Usage and 3' UTR Length Determine Maternal mRNA Stability in Zebrafish. Mol Cell *61*, 874-885.

Mittal, P., Brindle, J., Stephen, J., Plotkin, J.B., and Kudla, G. (2018). Codon usage influences fitness through RNA toxicity. Proc Natl Acad Sci U S A *115*, 8639-8644.

Muller-McNicoll, M., Botti, V., de Jesus Domingues, A.M., Brandl, H., Schwich, O.D., Steiner, M.C., Curk, T., Poser, I., Zarnack, K., and Neugebauer, K.M. (2016). SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. Genes Dev *30*, 553-566.

Nott, A., Le Hir, H., and Moore, M.J. (2004). Splicing enhances translation in mammalian cells: an additional function of the exon junction complex. Genes Dev *18*, 210-222.

Nott, A., Meislin, S.H., and Moore, M.J. (2003). A quantitative analysis of intron effects on mammalian gene expression. RNA *9*, 607-617.

Palazzo, A.F., and Akef, A. (2012). Nuclear export as a key arbiter of "mRNA identity" in eukaryotes. Biochim Biophys Acta *1819*, 566-577.

Palazzo, A.F., Springer, M., Shibata, Y., Lee, C.S., Dias, A.P., and Rapoport, T.A. (2007). The signal sequence coding region promotes nuclear export of mRNA. PLoS Biol *5*, e322.

Parmley, J.L., Urrutia, A.O., Potrzebowski, L., Kaessmann, H., and Hurst, L.D. (2007). Splicing and the evolution of proteins in mammals. PLoS biology *5*, e14.

Plotkin, J.B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet *12*, 32-42.

Plotkin, J.B., Robins, H., and Levine, A.J. (2004). Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci U S A *101*, 12588-12591.

Ponting, C.P., and Goodstadt, L. (2009). Separating derived from ancestral features of mouse and human genomes. Biochem Soc Trans *37*, 734-739.

Presnyak, V., Alhusaini, N., Chen, Y.H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R.*, et al.* (2015). Codon optimality is a major determinant of mRNA stability. Cell *160*, 1111-1124.

R Development Core Team (2005). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).

Radhakrishnan, A., Chen, Y.H., Martin, S., Alhusaini, N., Green, R., and Coller, J. (2016). The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. Cell *167*, 122-132 e129.

Ressayre, A., Glemin, S., Montalent, P., Serre-Giardi, L., Dillmann, C., and Joets, J. (2015). Introns Structure Patterns of Variation in Nucleotide Composition in Arabidopsis thaliana and Rice Protein-Coding Genes. Genome Biol Evol *7*, 2913-2928.

Rosikiewicz, W., Kabza, M., Kosinski, J.G., Ciomborowska-Basheer, J., Kubiak, M.R., and Makalowska, I. (2017). RetrogeneDB-a database of plant and animal retrocopies. Database (Oxford) *2017*.

Rudolph, K.L., Schmitt, B.M., Villar, D., White, R.J., Marioni, J.C., Kutter, C., and Odom, D.T. (2016). Codon-Driven Translational Efficiency Is Stable across Diverse Mammalian Cell States. PLoS Genet *12*, e1006024.

Savisaar, R., and Hurst, L.D. (2016). Purifying Selection on Exonic Splice Enhancers in Intronless Genes. Mol Biol Evol *33*, 1396-1418.

Semon, M., Mouchiroud, D., and Duret, L. (2005). Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. Hum Mol Genet *14*, 421-427.

Shah, P., Ding, Y., Niemczyk, M., Kudla, G., and Plotkin, J.B. (2013). Rate-limiting steps in yeast protein translation. Cell *153*, 1589-1601.

Sharp, P.M., and Li, W.H. (1987a). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res *15*, 1281-1295.

Sharp, P.M., and Li, W.H. (1987b). The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol *4*, 222-230.

Takata, M.A., Goncalves-Carneiro, D., Zang, T.M., Soll, S.J., York, A., Blanco-Melo, D., and Bieniasz, P.D. (2017). CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature *550*, 124-127.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell *141*, 344-354.

Vinogradov, A.E. (2003). Isochores and tissue-specificity. Nucleic Acids Res *31*, 5212-5220.

Wang, Y., Zhu, W., and Levy, D.E. (2006). Nuclear and cytoplasmic mRNA quantification by SYBR green based real-time RT-PCR. Methods *39*, 356-362.

Webster, M.W., Chen, Y.H., Stowell, J.A.W., Alhusaini, N., Sweet, T., Graveley, B.R., Coller, J., and Passmore, L.A. (2018). mRNA Deadenylation Is Coupled to Translation Rates by the Differential Activities of Ccr4-Not Nucleases. Mol Cell *70*, 1089-1100 e1088.

Zaghlool, A., Ameur, A., Nyberg, L., Halvardson, J., Grabherr, M., Cavelier, L., and Feuk, L. (2013). Efficient cellular fractionation improves RNA sequencing analysis of mature and nascent transcripts from human tissues. BMC Biotechnol *13*, 99.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Giron, C.G.*, et al.* (2018). Ensembl 2018. Nucleic Acids Res *46*, D754-D761.

Zhang, L., Kasif, S., Cantor, C.R., and Broude, N.E. (2004). GC/AT-content spikes as genomic punctuation marks. Proceedings of the National Academy of Sciences *101*, 16855-16860.

Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C.H., Fu, J., Chen, S., and Liu, Y. (2016). Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proc Natl Acad Sci U S A *113*, E6117-E6125.

Zhou, Z., Dang, Y., Zhou, M., Yuan, H., and Liu, Y. (2018). Codon usage biases co-evolve with transcription termination machinery to suppress premature cleavage and polyadenylation. eLife *7*.

Zolotukhin, S., Potter, M., Hauswirth, W.W., Guy, J., and Muzyczka, N. (1996). A "humanized" green fluorescent protein cDNA adapted for high-level expression in mammalian cells. J Virol *70*, 4646-4654.

KEY RESOURCES TABLE

The table highlights the genetically modified organisms and strains, cell lines, reagents, software, and source data **essential** to reproduce results presented in the manuscript. Depending on the nature of the study, this may include standard laboratory materials (i.e., food chow for metabolism studies), but the Table is **not** meant to be comprehensive list of all materials and resources used (e.g., essential chemicals such as SDS, sucrose, or standard culture media don't need to be listed in the Table). **Items in the Table must also be reported in the Method Details section within the context of their use.** The number of **primers and RNA sequences** that may be listed in the Table is restricted to no more than ten each. If there are more than ten primers or RNA sequences to report, please provide this information as a supplementary document and reference this file (e.g., See Table S1 for XX) in the Key Resources Table.

Please note that ALL references cited in the Key Resources Table must be included in the References list. Please report the information as follows:

- **REAGENT or RESOURCE:** Provide full descriptive name of the item so that it can be identified and linked with its description in the manuscript (e.g., provide version number for software, host source for antibody, strain name). In the Experimental Models section, please include all models used in the paper and describe each line/strain as: model organism: name used for strain/line in paper: genotype. (i.e., Mouse: $\text{OXTR}^{fl/fl}$: B6.129(SJL)-$\text{Oxtr}^{tm1.1Wsy/J}$). In the Biological Samples section, please list all samples obtained from commercial sources or biological repositories. Please note that software mentioned in the Methods Details or Data and Software Availability section needs to be also included in the table. See the sample Table at the end of this document for examples of how to report reagents.

- **SOURCE:** Report the company, manufacturer, or individual that provided the item or where the item can obtained (e.g., stock center or repository). For materials distributed by Addgene, please cite the article describing the plasmid and include "Addgene" as part of the identifier. If an item is from another lab, please include the name of the principal investigator and a citation if it has been previously published. If the material is being reported for the first time in the current paper, please indicate as "this paper." For software, please provide the company name if it is commercially available or cite the paper in which it has been initially described.

- **IDENTIFIER:** Include catalog numbers (entered in the column as "Cat#" followed by the number, e.g., Cat#3879S). Where available, please include unique entities such as RRIDs, Model Organism Database numbers, accession numbers, and PDB or CAS IDs. For antibodies, if applicable and available, please also include the lot number or clone identity. For software or data resources, please include the URL where the resource can be downloaded. Please ensure accuracy of the identifiers, as they are essential for generation of hyperlinks to external sources when available. Please see the Elsevier list of Data Repositories with automated bidirectional linking for details. When listing more than one identifier for the same item, use semicolons to separate them (e.g. Cat#3879S; RRID: AB_2255011). If an identifier is not available, please enter "N/A" in the column.
  - ***A NOTE ABOUT RRIDs:*** We highly recommend using RRIDs as the identifier (in particular for antibodies and organisms, but also for software tools and databases). For more details on how to obtain or generate an RRID for existing or newly generated resources, please visit the RII or search for RRIDs.

Please use the empty table that follows to organize the information in the sections defined by the subheading, skipping sections not relevant to your study. Please do not add subheadings. To add a row, place the cursor at the end of the row above where you would like to add the row, just outside the right border of the table. Then press the ENTER key to add the row. Please delete empty rows. Each entry must be on a separate row; do not list multiple items in a single table cell. Please see the sample table at the end of this document for examples of how reagents should be cited.

TABLE FOR AUTHOR TO COMPLETE

Please upload the completed table as a separate document. ***Please do not add subheadings to the Key Resources Table.*** *If you wish to make an entry that does not fall into one of the subheadings below, please contact your handling editor. (**NOTE:** For authors publishing in Current Biology, please note that references within the KRT should be in numbered style, rather than Harvard.)*

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
DH5alpha	Life Technologies	18265017
One Shot ccdB Survival 2 T1R Competent Cells	ThermoFisher	A10460
Biological Samples		
Chemicals, Peptides, and Recombinant Proteins		
EcoRV	NEB	R0195
SmaI	NEB	R0141
LR Clonase II mix	Invitrogen	11791100
EcoRI	NEB	R0101
BamHI	NEB	R0136
T4 DNA Ligase	NEB	M0202
Glycoblue	Invitrogen	AM9516
Phusion Taq Polymerase	Thermo Scientific	F530S
Accuprime Pfx Polymerase	ThermoFisher	12344024
RNeasy purification kit	Qiagen	74104
Trizol reagent	Invitrogen	15596026
Turbo DNA-free kit	Invitrogen	AM1907
RNAse-free DNAse kit	Qiagen	79254
Opti-MEM reduced serum medium	Gibco	31985062
Phenol red-free DMEM	Biochrom	F0475
Random hexamers	Promega	C1181
SuperScript III Reverse Transcriptase	Invitrogen	18080044
Lightcycler480 SYBR Green I Master Mix	Roche	04707516001
Trypan blue	Sigma-Aldrich	T8154
Trypsin solution	Sigma-Aldrich	T4174
RNasin plus	Promega	N2611
Proteinase K	Roche	3115836001
Blasticidin S	Gibco	R21001
Hygromycin B	Gibco	10687010
Doxycyline	Sigma-Aldrich	D9891
RNase A	Qiagen	19101
Phenol:Chlorofom:Isoamyl alcohol	Sigma-Aldrich	P2069

Cycloheximide		
4-Thiouridine	Sigma-Aldrich	T4509
dCTP, [α-32P]- 3000Ci/mmol	Perkin Elmer	NEG013H250UC
Biotin-HPDP	Pierce	21341
Dimethylformamide	Pierce	20673
Triptolide	Sigma-Aldrich	T3652
Lipofectamine 2000	Invitrogen	11668019
Critical Commercial Assays		
Gibson Assembly Cloning Kit	NEB	E5510S
Qiaquick PCR purification kit	Qiagen	28104
MinElute PCR purification kit	Qiagen	28004
µMACS Streptavidin Kit	Miltenyi Biotec	130-074-101
DMEM	LifeTechnologies	41965039
Trypsin EDTA solution	Sigma	T4174
Deposited Data		
Sequencing data	SRA	PRJNA596086
Experimental Models: Cell Lines		
HEK293 T-REx Flp-in	ThermoFisher	R78007
HeLa T-REx Flp-in	Andrew Jackson Lab, MRC Human Genetics Unit, Edinburgh, UK.	N/A
Experimental Models: Organisms/Strains		
Oligonucleotides		
MiSeq library and sequencing primers	This paper, Sigma	Table S1
Cloning primers	This paper, Sigma	Table S1
(q)RT-PCR primers	This paper, Sigma	Table S1
Recombinant DNA		
pGK3 (Gateway entry vector)	Kudla et al., 2009	N/A

GFP variants	Kudla et al., 2009, Mittal et al., 2018	N/A
mKate2 variants	This paper	N/A
pCI-neo	Promega	E1841
pBluescript-RfA	Grzegorz Kudla, MRC Human Genetics Unit, Edinburgh, UK.	N/A
pmKate2-N	Evrogen	FP182
pcDNA5/FRT/TO/DEST	David Tollervey Lab, University of Edinburgh,Edinburgh, UK.	N/A
pOG44 (Flp-recombinase vector)	ThermoFisher	V600520
Software and Algorithms		
Python		Version 3.4.2
R		Version 3.1.2
FIMO	http://meme-suite.org	
Other		
Infinite M200 Pro plate reader	Tecan	N/A

TABLE WITH EXAMPLES FOR AUTHOR REFERENCE

REAGENT or RESOURCE	**SOURCE**	**IDENTIFIER**
Antibodies		
Rabbit monoclonal anti-Snail	Cell Signaling Technology	Cat#3879S; RRID: AB_2255011
Mouse monoclonal anti-Tubulin (clone DM1A)	Sigma-Aldrich	Cat#T9026; RRID: AB_477593
Rabbit polyclonal anti-BMAL1	This paper	N/A
Bacterial and Virus Strains		

pAAV-hSyn-DIO-hM3D(Gq)-mCherry	Krashes et al., 2011	Addgene AAV5; 44361-AAV5
AAV5-EF1a-DIO-hChR2(H134R)-EYFP	Hope Center Viral Vectors Core	N/A
Cowpox virus Brighton Red	BEI Resources	NR-88
Zika-SMGC-1, GENBANK: KX266255	Isolated from patient (Wang et al., 2016)	N/A
Staphylococcus aureus	ATCC	ATCC 29213
Streptococcus pyogenes: M1 serotype strain: strain SF370; M1 GAS	ATCC	ATCC 700294
Biological Samples		
Healthy adult BA9 brain tissue	University of Maryland Brain & Tissue Bank; http://medschool.umaryland.edu/btbank/	Cat#UMB1455
Human hippocampal brain blocks	New York Brain Bank	http://nybb.hs.columbia.edu/
Patient-derived xenografts (PDX)	Children's Oncology Group Cell Culture and Xenograft Repository	http://cogcell.org/
Chemicals, Peptides, and Recombinant Proteins		
MK-2206 AKT inhibitor	Selleck Chemicals	S1078; CAS: 1032350-13-2
SB-505124	Sigma-Aldrich	S4696; CAS: 694433-59-5 (free base)
Picrotoxin	Sigma-Aldrich	P1675; CAS: 124-87-8
Human TGF-β	R&D	240-B; GenPept: P01137
Activated S6K1	Millipore	Cat#14-486
GST-BMAL1	Novus	Cat#H00000406-P01
Critical Commercial Assays		
EasyTag EXPRESS 35S Protein Labeling Kit	Perkin-Elmer	NEG772014MC
CaspaseGlo 3/7	Promega	G8090
TruSeq ChIP Sample Prep Kit	Illumina	IP-202-1012
Deposited Data		
Raw and analyzed data	This paper	GEO: GSE63473
B-RAF RBD (apo) structure	This paper	PDB: 5J17
Human reference genome NCBI build 37, GRCh37	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
Nanog STILT inference	This paper; Mendeley Data	http://dx.doi.org/10.17632/wx6s4mj7s8.2
Affinity-based mass spectrometry performed with 57 genes	This paper; and Mendeley Data	Table S8; http://dx.doi.org/10.17632/5hvpvspw82.1
Experimental Models: Cell Lines		
Hamster: CHO cells	ATCC	CRL-11268

D. melanogaster: Cell line S2: S2-DRSC	Laboratory of Norbert Perrimon	FlyBase: FBtc0000181
Human: Passage 40 H9 ES cells	MSKCC stem cell core facility	N/A
Human: HUES 8 hESC line (NIH approval number NIHhESC-09-0021)	HSCI iPS Core	hES Cell Line: HUES-8
Experimental Models: Organisms/Strains		
C. elegans: Strain BC4011: srl-1(s2500) II; dpy-18(e364) III; unc-46(e177)rol-3(s1040) V.	Caenorhabditis Genetics Center	WB Strain: BC4011; WormBase: WBVar00241916
D. melanogaster: RNAi of Sxl: y[1] sc[*] v[1]; P{TRiP.HMS00609}attP2	Bloomington Drosophila Stock Center	BDSC:34393; FlyBase: FBtp0064874
S. cerevisiae: Strain background: W303	ATCC	ATTC: 208353
Mouse: R6/2: B6CBA-Tg(HDexon1)62Gpb/3J	The Jackson Laboratory	JAX: 006494
Mouse: OXTRfl/fl: B6.129(SJL)-Oxtr$^{tm1.1Wsy}$/J	The Jackson Laboratory	RRID: IMSR_JAX:008471
Zebrafish: Tg(Shha:GFP)t10: t10Tg	Neumann and Nuesslein-Volhard, 2000	ZFIN: ZDB-GENO-060207-1
Arabidopsis: 35S::PIF4-YFP, BZR1-CFP	Wang et al., 2012	N/A
Arabidopsis: JYB1021.2: pS24(AT5G58010)::cS24:GFP(-G):NOS #1	NASC	NASC ID: N70450
Oligonucleotides		
siRNA targeting sequence: PIP5K I alpha #1: ACACAGUACUCAGUUGAUA	This paper	N/A
Primers for XX, see Table SX	This paper	N/A
Primer: GFP/YFP/CFP Forward: GCACGACTTCTTCAAGTCCGCCATGCC	This paper	N/A
Morpholino: MO-pax2a GGTCTGCTTTGCAGTGAATATCCAT	Gene Tools	ZFIN: ZDB-MRPHLNO-061106-5
ACTB (hs01060665_g1)	Life Technologies	Cat#4331182
RNA sequence: hnRNPA1_ligand: UAGGGACUUAGGGUUCUCUCUAGGGACUUAGGGUUCUCUCUAGGGA	This paper	N/A
Recombinant DNA		
pLVX-Tight-Puro (TetOn)	Clonetech	Cat#632162
Plasmid: GFP-Nito	This paper	N/A
cDNA GH111110	Drosophila Genomics Resource Center	DGRC:5666; FlyBase:FBcl0130415
AAV2/1-hsyn-GCaMP6- WPRE	Chen et al., 2013	N/A
Mouse raptor: pLKO mouse shRNA 1 raptor	Thoreen et al., 2009	Addgene Plasmid #21339
Software and Algorithms		
ImageJ	Schneider et al., 2012	https://imagej.nih.gov/ij/

Bowtie2	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Samtools	Li et al., 2009	http://samtools.sourceforge.net/
Weighted Maximal Information Component Analysis v0.9	Rau et al., 2013	https://github.com/ChristophRau/wMICA
ICS algorithm	This paper; Mendeley Data	http://dx.doi.org/10.17632/5hvpvspw82.1
Other		
Sequence data, analyses, and resources related to the ultra-deep sequencing of the AML31 tumor, relapse, and matched normal.	This paper	http://aml31.genome.wustl.edu
Resource website for the AML31 publication	This paper	https://github.com/chrisamiller/aml31SuppSite

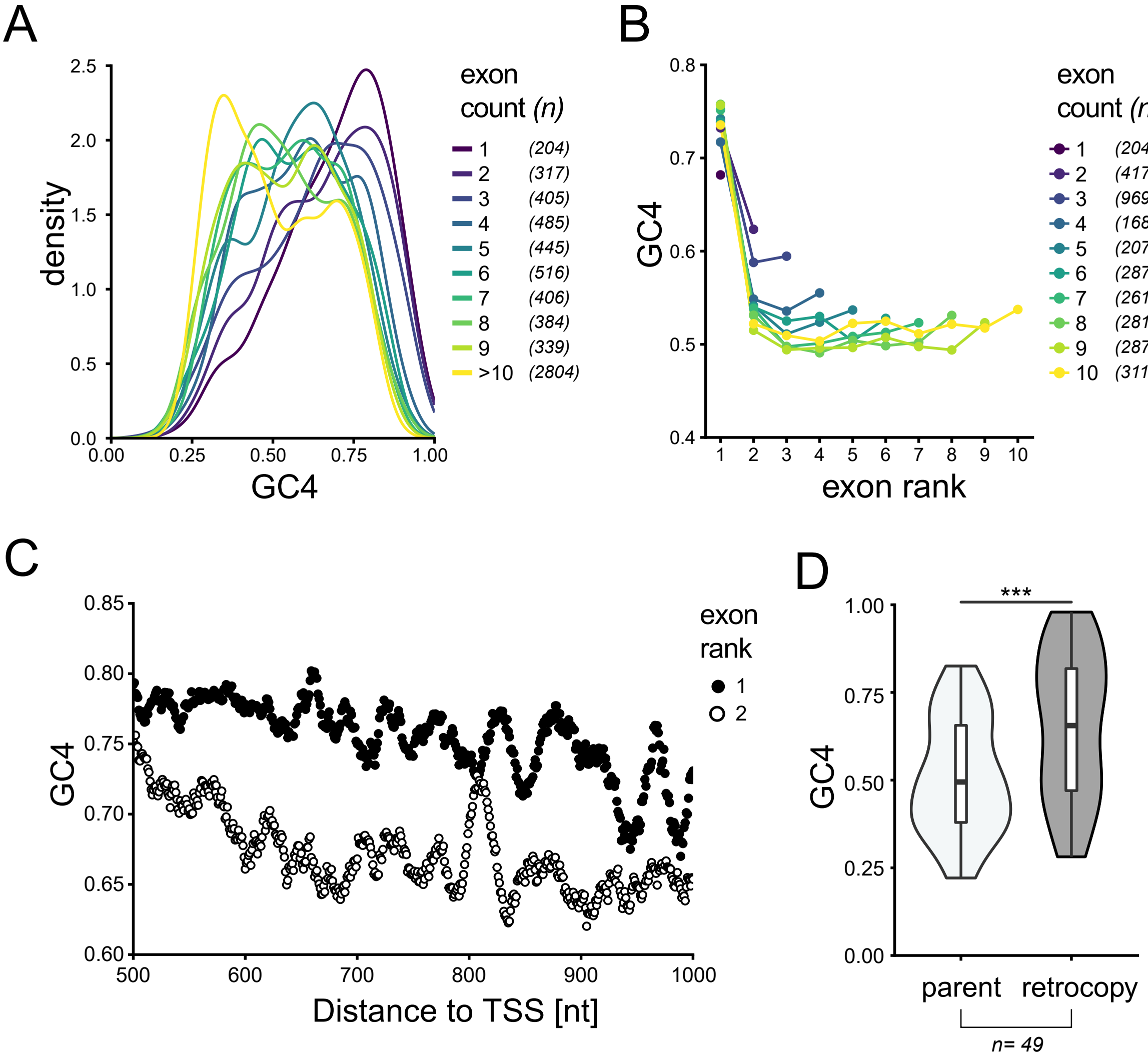
A
2.5
2.0
1.5
1.0
0.5
0.0
density
0.00
0.25
0.50
0.75
1.00
GC4
exon count (n)
1 (204)
2 (317)
3 (405)
4 (485)
5 (445)
6 (516)
7 (406)
8 (384)
9 (339)
>10 (2804)
B
0.8
0.7
0.6
0.5
0.4
GC4
1 2 3 4 5 6 7 8 9 10
exon rank
exon count (n)
1 (204)
2 (417)
3 (969)
4 (1680)
5 (2070)
6 (2879)
7 (2616)
8 (2818)
9 (2873)
10 (3111)
C
0.85
0.80
0.75
0.70
0.65
0.60
GC4
500
600
700
800
900
1000
Distance to TSS [nt]
exon rank
● 1
○ 2
D
***
1.00
0.75
0.50
0.25
0.00
GC4
parent
retrocopy
n= 49

Figure 1

Figure 2

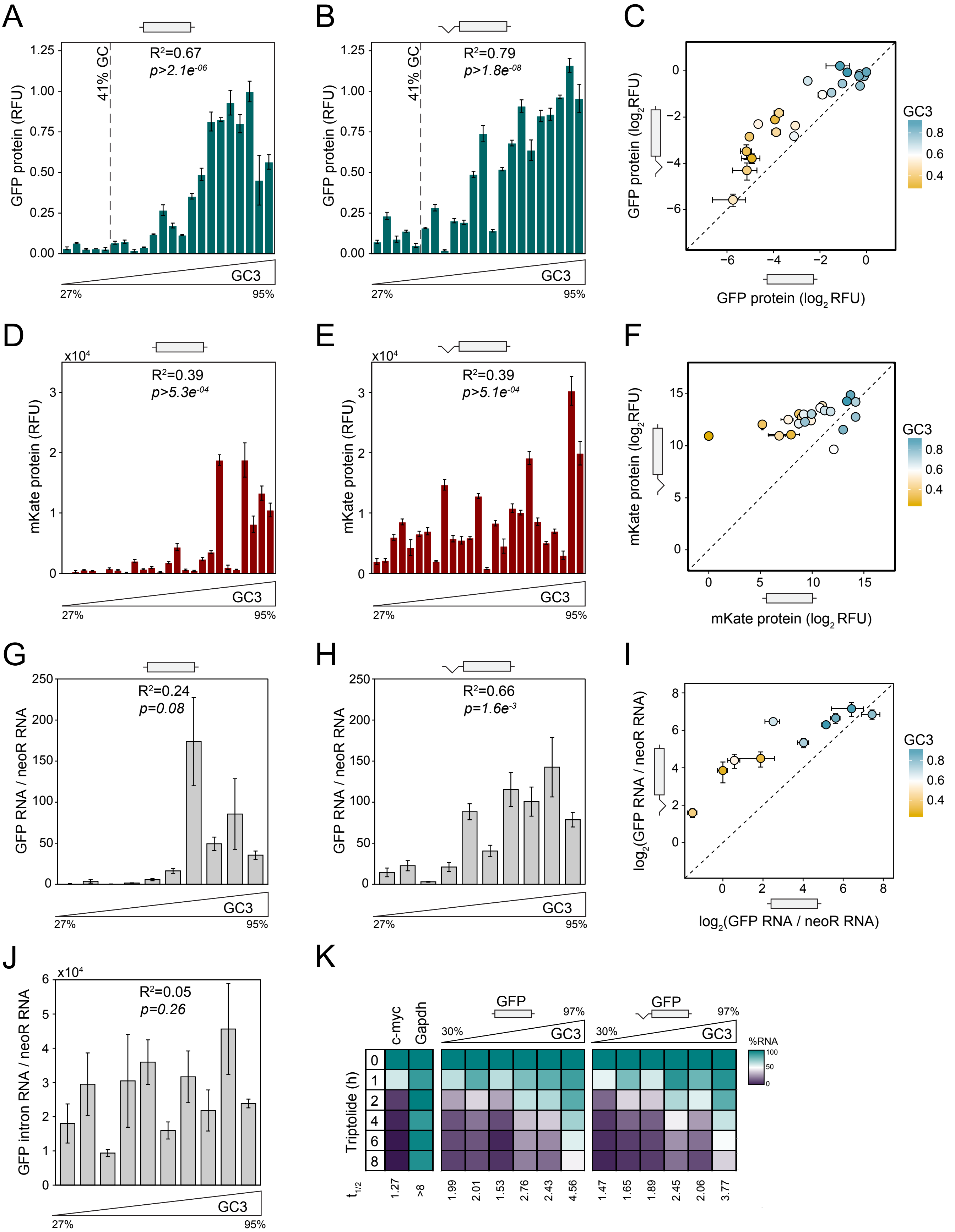

Figure 2

Figure 3

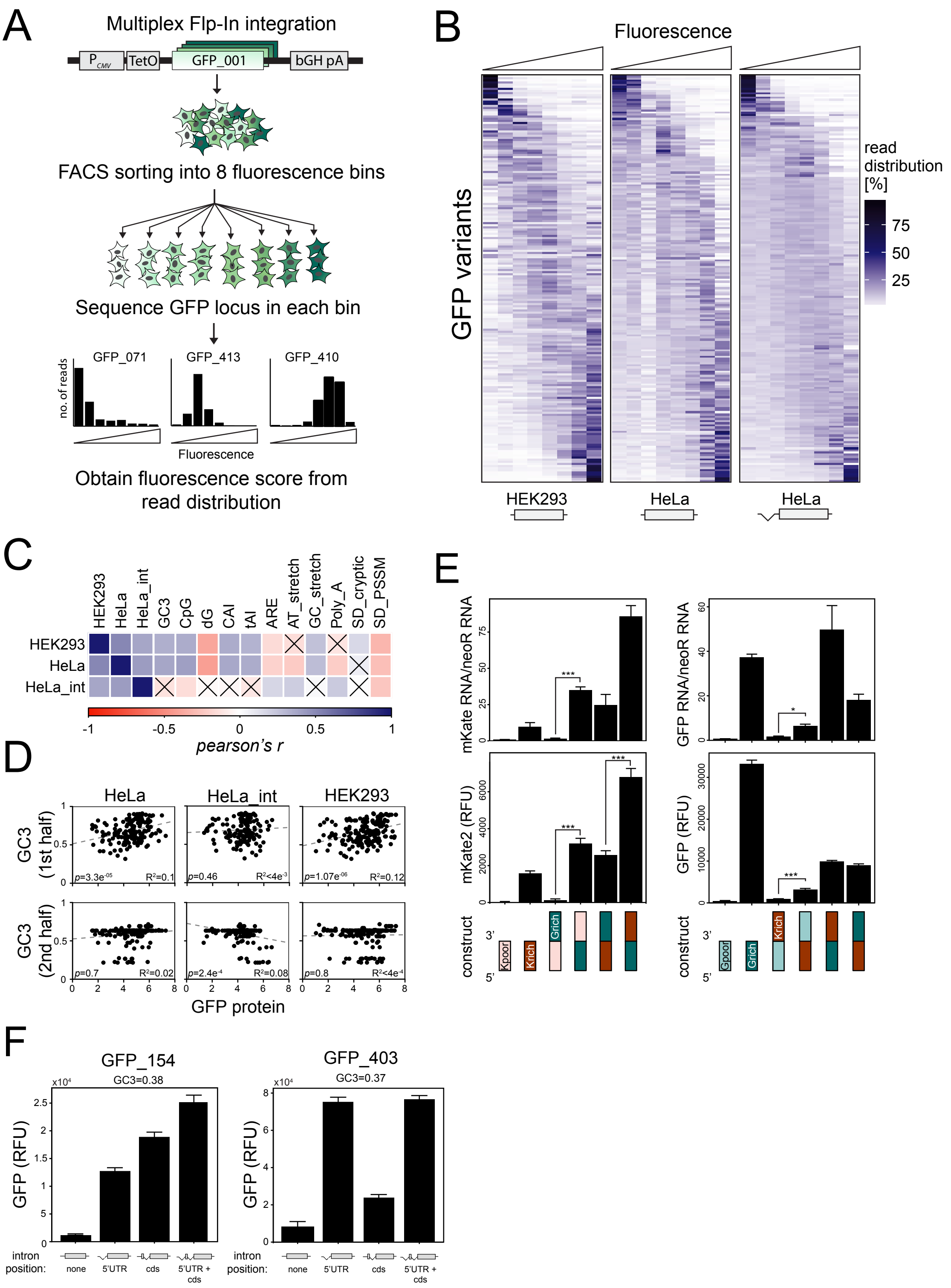

Figure 3

Figure 4

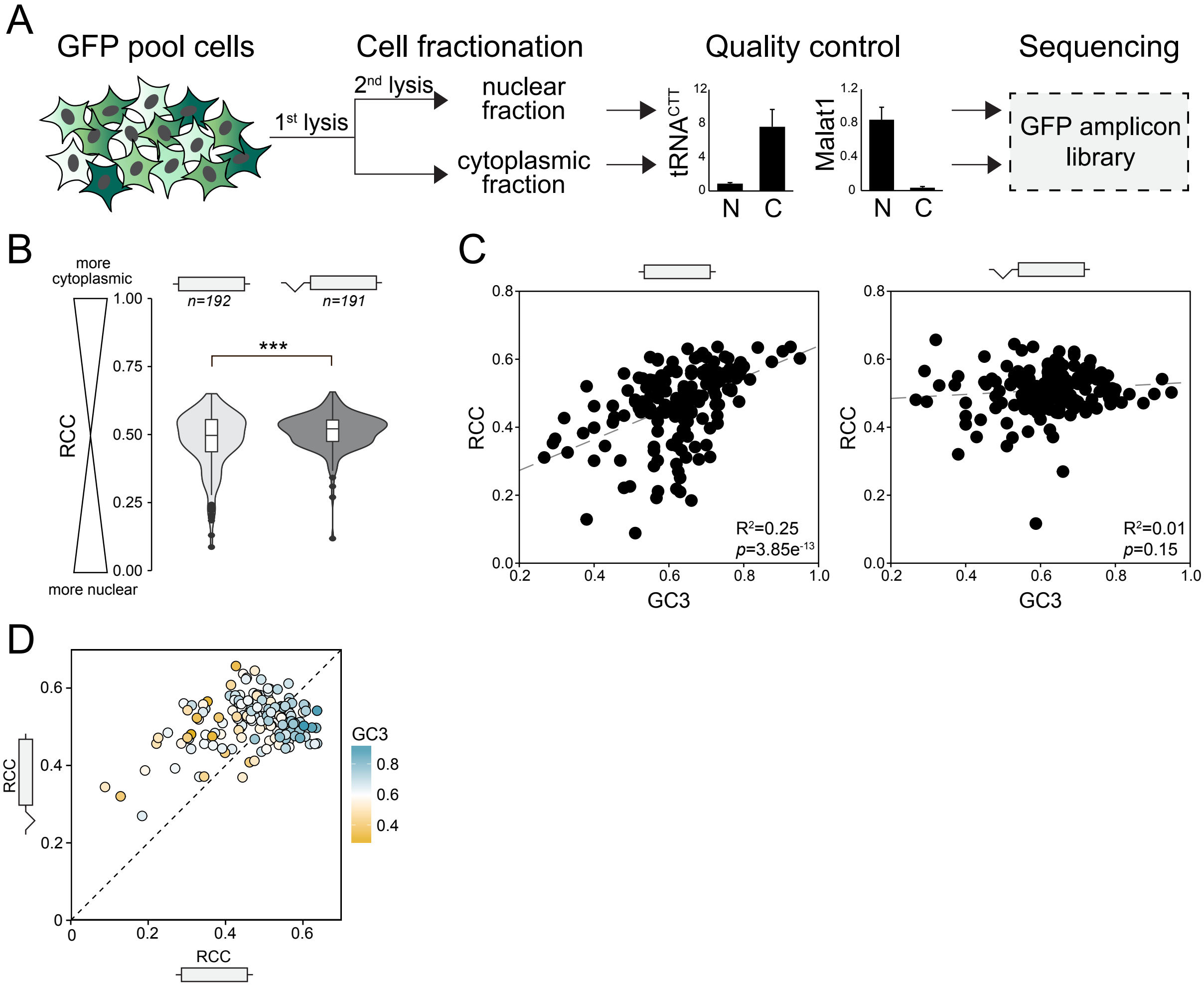

Figure 4

Figure 5

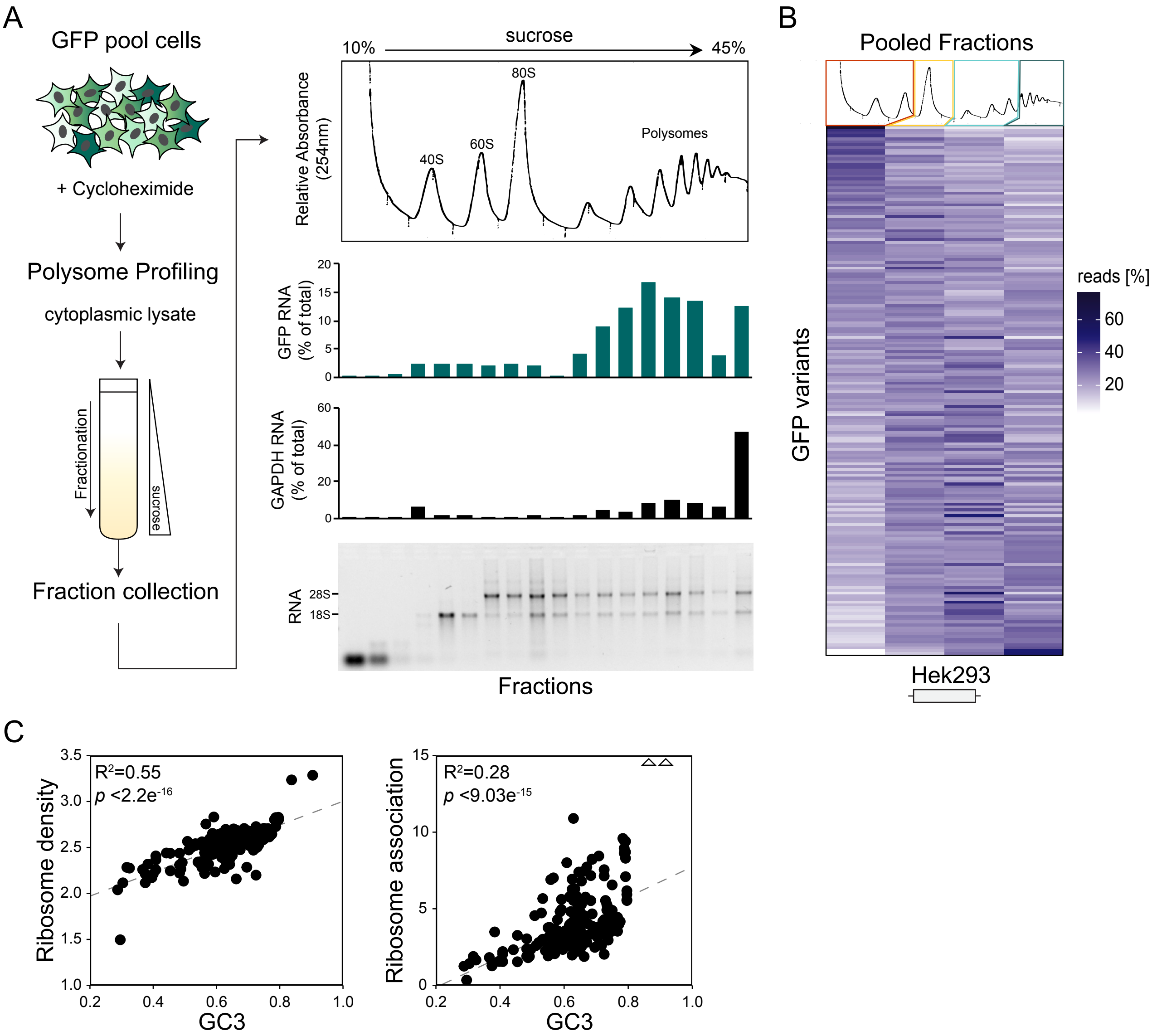

Figure 5

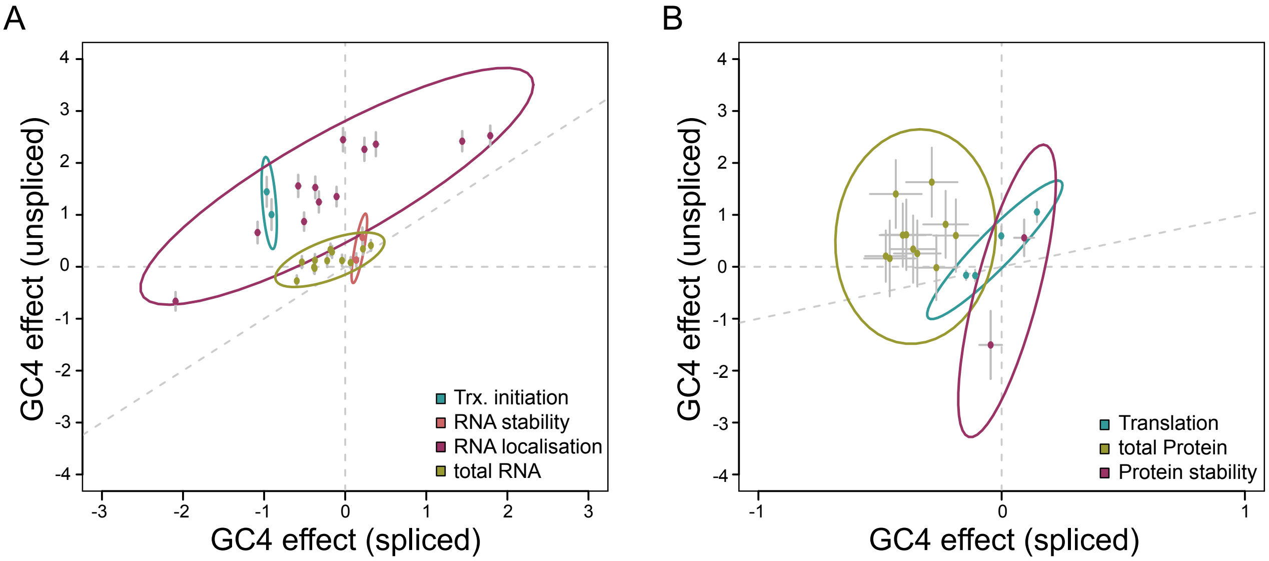
A
GC4 effect (unspliced)
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
GC4 effect (spliced)
🟩 Trx. initiation
🟥 RNA stability
🟪 RNA localisation
🟨 total RNA
B
GC4 effect (unspliced)
4
3
2
1
0
-1
-2
-3
-4
-1
0
1
GC4 effect (spliced)
🟩 Translation
🟨 total Protein
🟪 Protein stability

Figure 6

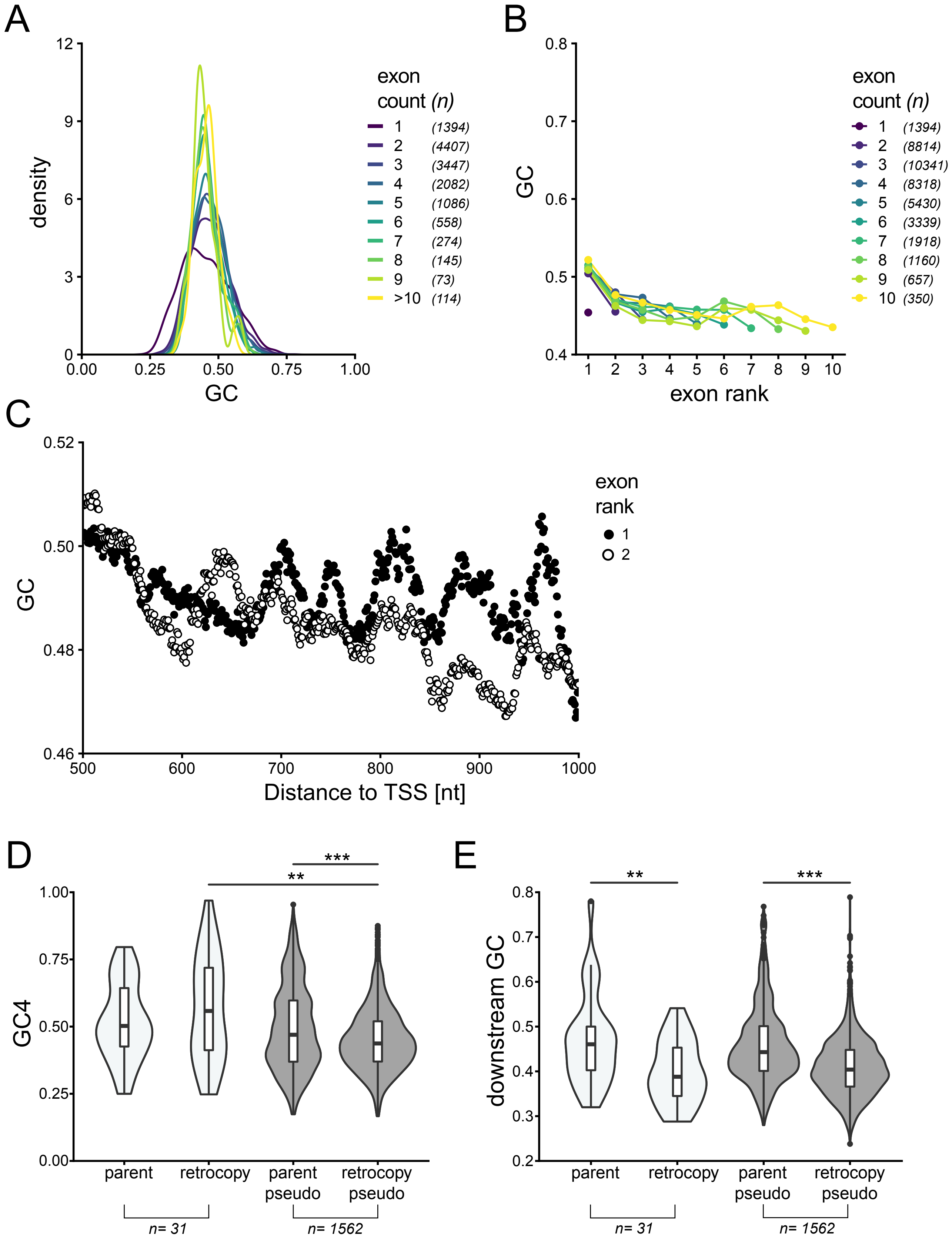

Figure S1. GC variation amongst lncRNAs and parent-retrogene pairs and their downstream sequence, related to Figure 1.

Figure S1 (continued) (A) GC distribution of human long non-coding RNA genes, grouped by number of exons per gene. The Y axis indicates the proportion of genes within a given range of GC, calculated using the ggplot2 geom_density() function.

(B) Mean GC content in non-coding exons, grouped by exon position (rank) and by number of exons per gene. (C) Mean GC within exons of rank 1 (black dots) or rank 2 (white dots) downstream of the transcription start site (TSS). (D) GC4 content distribution across parent and retrogene pairs conserved between human and macaque. White violins indicate pairs for which retrocopies are classed as functional (p=0.26, n=31, two-tailed Wilcoxon signed-rank test), whereas grey violins correspond to pairs in which the retrocopy is classed as non-functional pseudogene (p < 2.2×10-16., n=1562, two-tailed Wilcoxon signed-rank test). For the human-macaque set, the difference in GC4 between parents and functional copies is in the expected direction but not significant. (E) Violin plot showing GC content within a window between 2000 and 3000nt downstream from the stop codons of functional (white, p=9.27×10-4, n=31, two-tailed Wilcoxon signed-rank test) and non-functional (grey, p<2.2×10-16, n=1562, two-tailed Wilcoxon signed-rank test) parent-retrogene pairs conserved between human and macaque.

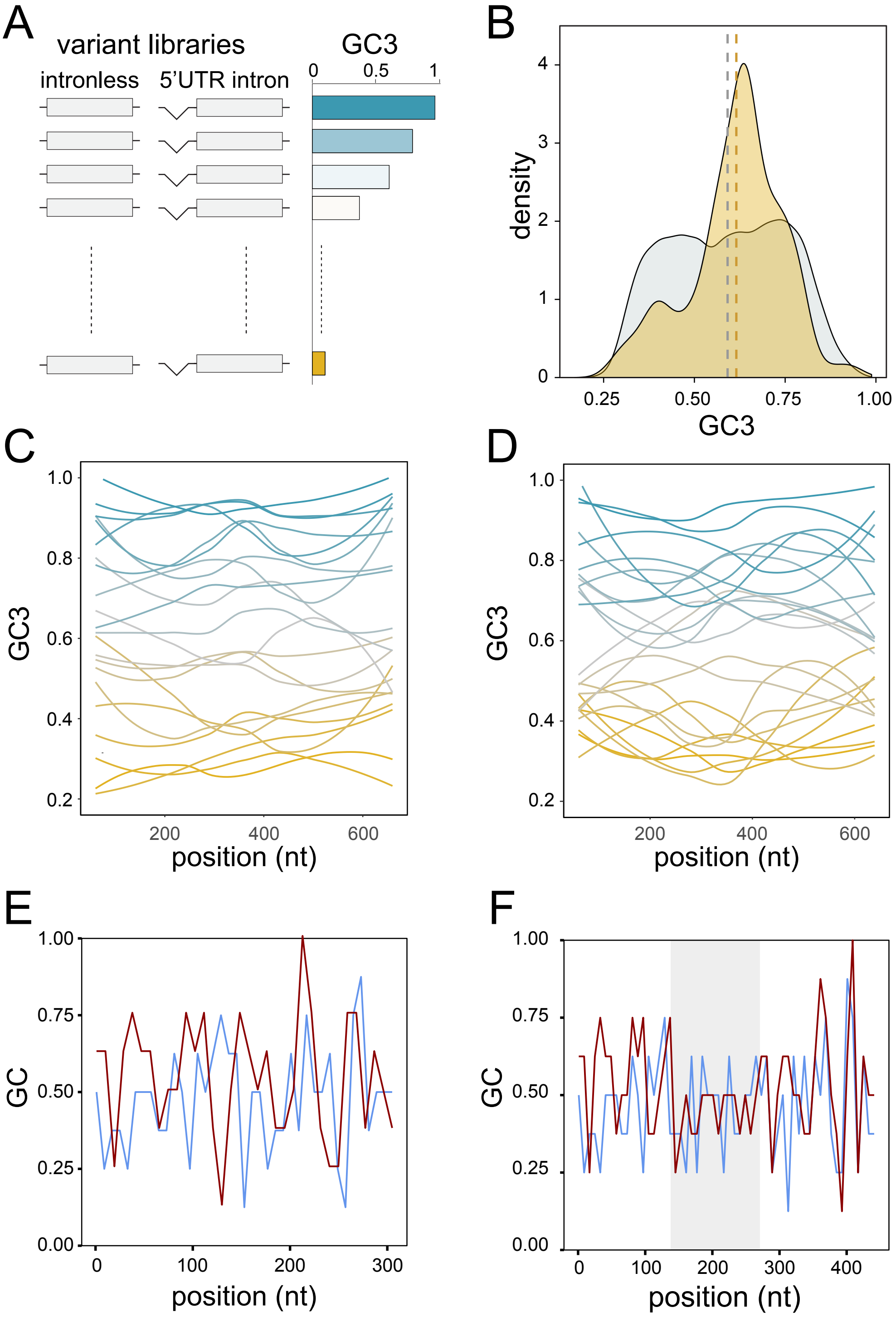

Figure S2. GC content variation amongst endogenous genes and reporter libraries, related to Figure 2. (A) Libraries of reporter genes with random synonymous codon usage were designed to cover a broad range of GC3 content variation. Variants were expressed with and without a synthetic 5' UTR intron. (B) GC3 content distribution amongst human consensus coding sequences (CDS; grey) in comparison to the GFP variant library used in this study (GFP lib; orange). Dashed lines indicate the mean GC3 for each data set. (C-D) Loess-smoothed GC3 profiles along the 22 GFP variants (C) and 23 mKate variants (D) that were analysed by spectrofluorometry (Figure 2). (E) Sliding window analysis of GC content in 5' UTRs of intronless expression cassettes utilised in this study. Blue: pCM3 (transient transfection, no intron); red: pcDNA5/FRT/TO/DEST (stable transfection, no intron). (F) As above, intron-containing expression cassettes. Blue: pCM4 (transient transfection, with intron); red: pcDNA5/FRT/TO/DEST/INT (stable transfection, with intron). Grey shading indicates the position of the synthetic intron.

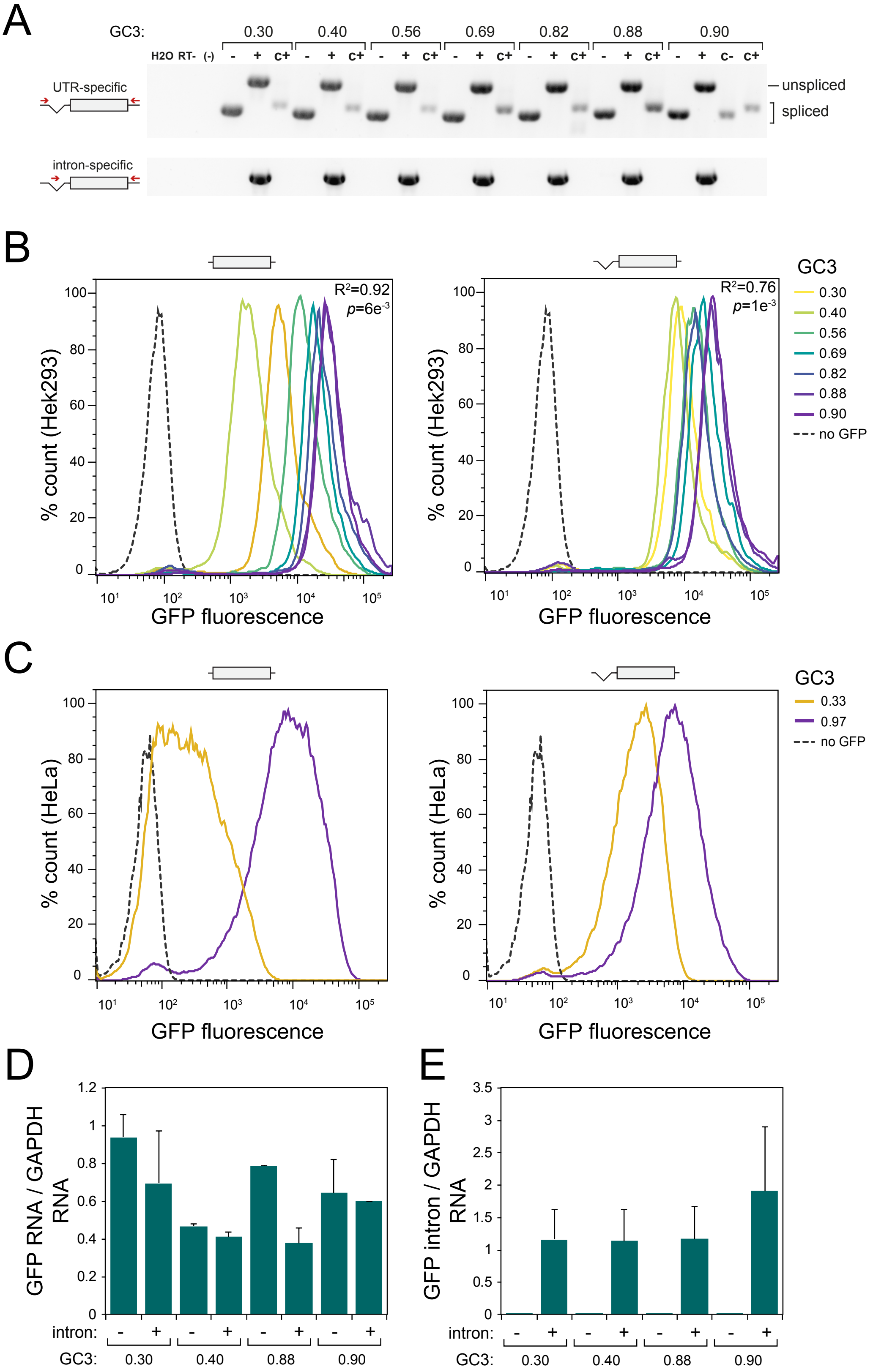

Figure S3. Effect of GC content on expression of fluorescent reporter genes in stably transfected cell lines, related to Figure 2.

Figure S3 (continued). (A) RT-PCR using total RNA from HEK293 Flp-In cell lines stably expressing several variants of GFP with a broad GC3 range (GC3 range: 0.3 – 0.9) and containing the same 5' UTR intron as used throughout this study. PCR was performed using either UTR-specific primers that detect spliced as well as unspliced GFP transcripts (upper gel, labelled 'UTR-specific)), or primers that exclusively detect unspliced transcripts (lower gel, labelled 'intron-specific'). Plasmids containing the respective GFP expression cassettes, both with or without UTR intron, are shown as controls. (B-C) Flow cytometry measurements of GFP variants covering a broad range of GC3 variation in stably transfected HEK293 Flp-in (B) and HeLa Flp-in (C). (D-E) qRT-PCR measurements of nascent RNA isolated using 4sU labelling from 2 GC-poor (GC3=0.3 and 0.4) and 2 GC-rich (GC3=0.88 and 0.9) GFP variants, expressed as unspliced or spliced constructs. GFP RNA levels were measured using 3' UTR specific primers (D, full length transcripts) and intronic RNA levels (E, pre-mRNA). Data points represent the mean of 2 independent experiments, -/+ SD.

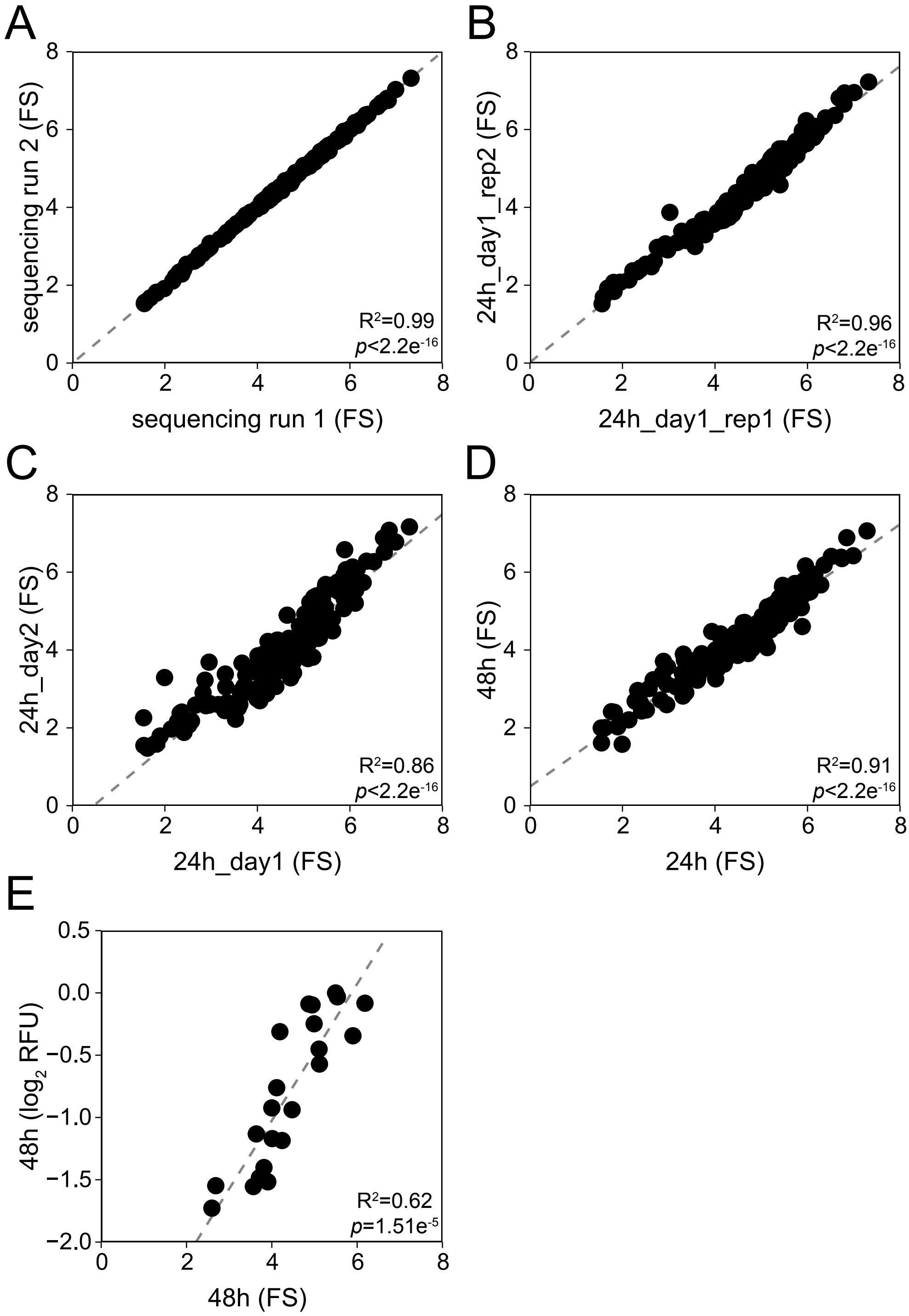

Figure S4. Reproducibility of Flow-seq experiments in HeLa cells (unspliced GFP variants), related to Figure 3.

(A-E) GFP Flow-Seq fluorescence scores (FS), calculated as described in the Methods section. (A) Re-sequencing of the same amplicon-library. (B-C) Replicate Flow-seq experiments performed on the same day (B) or different days (C). (D) Flow-Seq experiments performed on the same pool of cells, 24h and 48h after the induction of GFP expression. (E) Correlation between fluorescence measurements of 22 GFP variants obtained in the HeLa GFP pool cell line by Flow-Seq (X axis) and in transiently transfected HeLa cells by spectrofluorometry (Y axis, data from Figure 2).

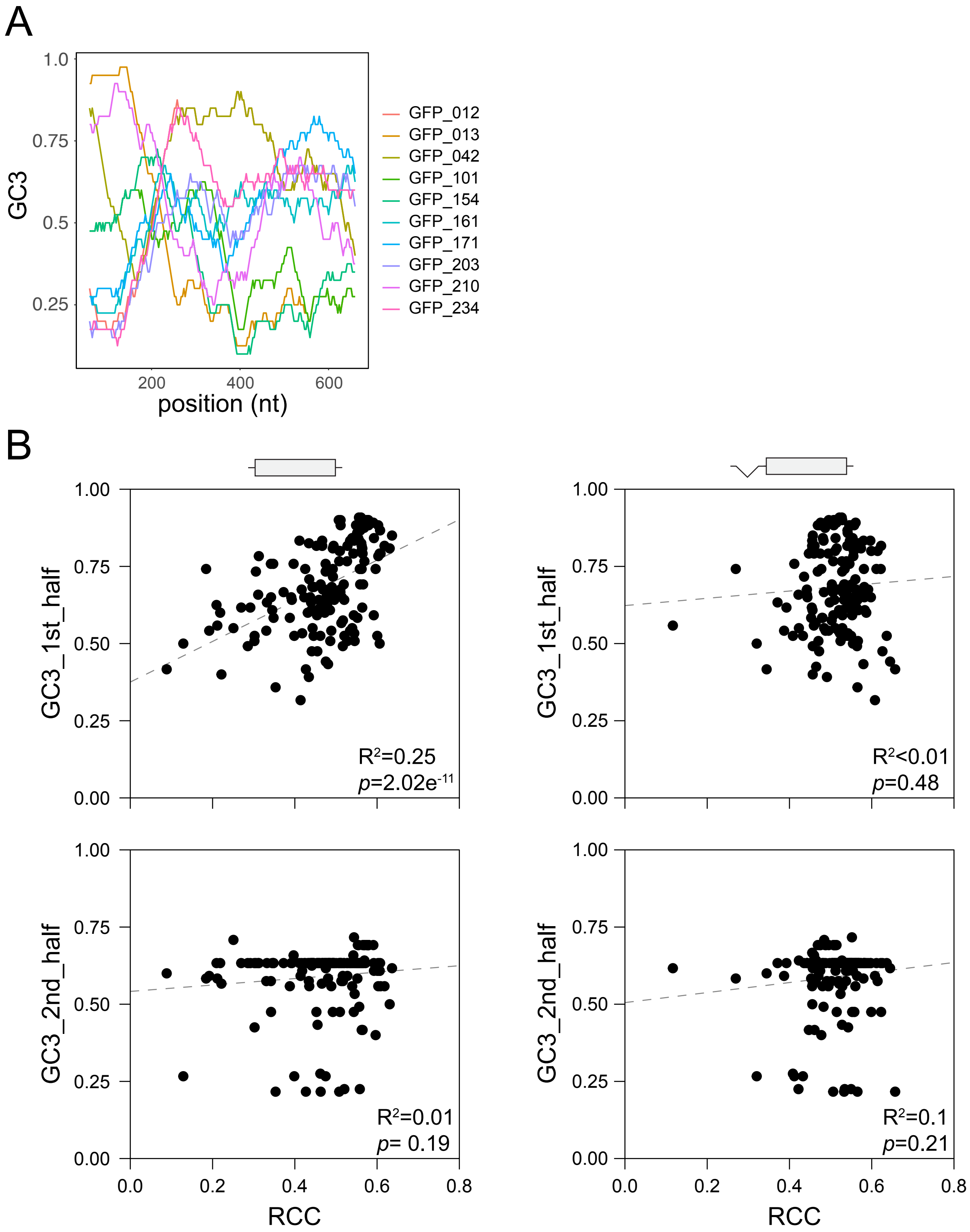

Figure S5. Position-specific effects of GC content on expression, related to Figures 3 and 4. (A) Sliding window analysis of GC3 content in selected GFP variants used in the pooled amplicon sequencing experiments. (B) Correlations between the GC3 content in the 1st (nt 1-360) and 2nd (nt 361-720) halves of GFP variants and their relative cytoplasmic mRNA concentrations (RCC).

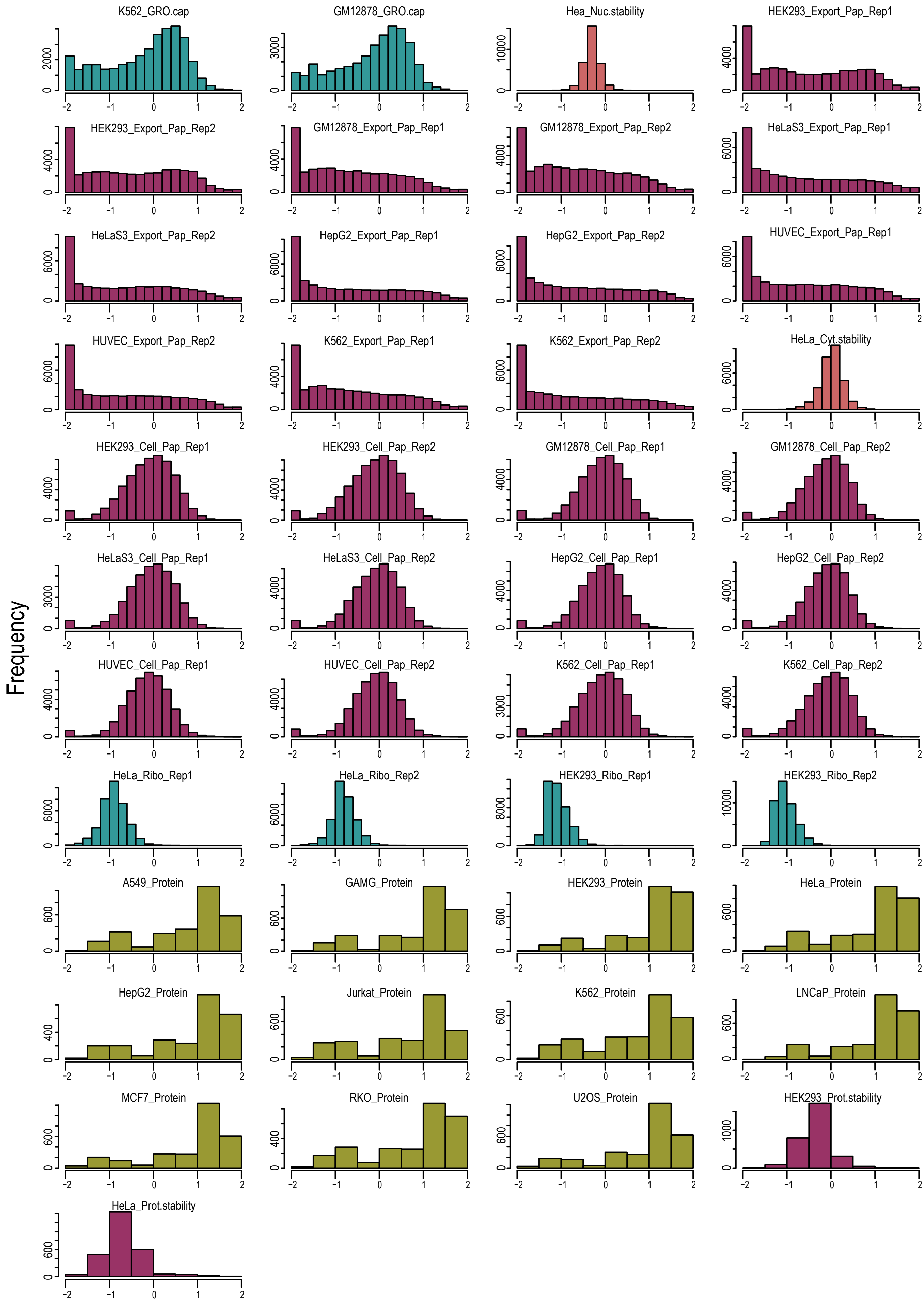

Figure S6. Distribution of RNA and protein expression data used in regression modelling, related to Figure 6.

Figure S6 (continued) Human RNA and protein expression data were extracted from various databases, filtered and normalized as described in Table S1 and STAR Methods. The histograms show the distributions of preprocessed expression measurements.

Table S1. Sources of human gene expression data, related to Figure 6. The cellular process to be quantified is indicated above the table, and the experimental techniques and data sources are indicated below. Each dot indicates an experimental replicate measurement.

	Transcription	nuclear stability	cytoplasmic stability	RNA levels	RNA export	Translation	Protein levels	Protein stability
K562	●			●●	●●		●	
Gm12878	●			●●	●●			
HeLa		●	●	●●	●●	●●	●	●
Hek293				●●	●●	●●	●	●
Huvec				●●	●●			
HepG2				●●	●●		●	
A549							●	
GAMG							●	
Jurkat							●	
LnCap							●	
MCF7							●	
RKO							●	
U2OS							●	
data type	GRO-cap	CAGE-seq: Mtr4 KD/ EGFP KD	CAGE-seq: Rrp40 KD/ Mtr4 KD	RNA-seq	RNA-seq	Ribo-seq	Mass-spec	Mass-spec/Ribo-seq
data source	ENCODE	Andersson et al., 2014	Andersson et al., 2014	Hek293: this study; all others: ENCODE	Hek293: this study; all others: ENCODE	ENCODE	Geiger et al., 2012	Geiger et al., 2012; ENCODE

Supplementary Table 1

Table S2. List of primer sequences, related to STAR methods.

MiSeq library + sequencing	**5' → 3'**
PE_PCR_left	AATGATACGGCGACCACCGAGATCTACACGCTGGCACGCGTAAGAAGGAGATATAACCATG
S_index1_right_PEPCR	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGTGCAGGGCCGCGAATTC
S_index2_right_PEPCR	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGTGCAGGGCCGCGAATTC
S_index3_right_PEPCR	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGTGCAGGGCCGCGAATTC
S_index4_right_PEPCR	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGTGCAGGGCCGCGAATTC
S_index5_right_PEPCR	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGTGCAGGGCCGCGAATTC
S_index6_right_PEPCR	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGTGCAGGGCCGCGAATTC
S_index7_right_PEPCR	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGTGCAGGGCCGCGAATTC
S_index8_right_PEPCR	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGTGCAGGGCCGCGAATTC
Read1_seq_primer_GFP	GCTGGCACGCGTAAGAAGGAGATATAACCATG
cloning primers	
pCI_del_int_F (phospho)	GTGTCCACTCCCAGTTCAAT
pCI_del_int_R (phospho)	CTGCCCAGTGCCTCACGACC
mkate2_gibs_F	GATCCGCGTATGGTGGCCTTAAGATACATTGATGAG
mkate2_gibs_R	TGTAAGCGGATGCCGCACATGTTCTTTCCTGCG
pCI_gib_F	CGGCATCCGCTTACAGACAA
pCI_gib_R	CACCATACGCGGATCCTTATC
qPCR primers	
pcDNA5-UTR_F	GTTGCCAGCCATCTGTTGTT
pcDNA5-UTR_R	CTCAGACAATGCGATGCAATTTCC
pc5_5UTR_F	CCGGGACCGATCCAGCCTCC
pc5_3UTR_R1	GCAAACAACAGATGGCTGGC
pc5_3UTR_F	TAAGAATTCGCGGCCCTGC

pc5_INT_F	GAAGTTGGTCGTGAGGCACTG
pCI-UTR_F	CTTCCCTTTAGTGAGGGTTAATG
pCI-UTR_R	GTTTATTGCAGCTTATAATGGTTAC
pCI-mRNA_F	GCTAACGCAGTCAGTGCTTC
pCI-mRNA_R	ACACCCAGTGCCTCACGAC
pCI-premRNA_F	GAGGCACTGGGCAGGTAAGTATC
pCI-premRNA_R	GTGGATGTCAGTAAGACCAATAGGTG
Gapdh_F	GGAGTCAACGGATTTGG
Gapdh_R	GTAGTTGAGGTCAATGAAGGG
Neo_F	CCCGTGATATTGCTGAAGAG
Neo_R	CGTCAAGAAGGCGATAGAAG
LysCTT_F	TCAGTCGGTAGAGCATGAGAC
LysCTT_R	CAACGTGGGGCTCGAACC
Malat1_F	CAGACCCTTCACCCCTCAC
Malat1_R	TTATGGATCATGCCCACAAG
cMyc_F	CTCCTACGTTGCGGTCACAC
cMyc_R	CCGGGTCGCAGATGAAACTC