Splicing buffers suboptimal codon usage in human cells

Christine Mordstein[1,2], Rosina Savisaar[2,3], Robert S Young[2], Jeanne Bazile[1], Lana Talmane[1], Juliet Luft[1], Michael Liss[4], 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, Scotland, UK

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

[3]School of Communication and Information Technology, Nile University, Giza, Egypt

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

4,024 words (main text)

6 Figures

6 Supplementary Figures

2 Supplementary Tables

*Corresponding author: Grzegorz Kudla (gkudla@gmail.com)

1   **Abstract**

2

3   Although multiple studies have addressed the effects of codon usage on gene

4   expression, such studies were typically performed in unspliced model genes. In

5   the human genome, most genes undergo splicing and patterns of codon usage are

6   splicing-dependent: guanine and cytosine (GC) content is highest within single-

7   exon genes and within first exons of multi-exon genes. Intrigued by this

8   observation, we measured the effects of splicing on expression in a panel of

9   synonymous variants of GFP and mKate2 reporter genes that varied in

10   nucleotide composition. We found that splicing promotes the expression of

11   adenine and thymine (AT)-rich variants by increasing their steady-state protein

12   and mRNA levels, in part through promoting cytoplasmic localization of mRNA.

13   Splicing had little or no effect on the expression of GC-rich variants. In the

14   absence of splicing, high GC content at the 5' end, but not at the 3' end of the

15   coding sequence positively correlated with expression. Among endogenous

16   human protein-coding transcripts, GC content has a more positive effect on

17   various expression measures of unspliced, relative to spliced mRNAs. We

18   propose that splicing promotes the expression of AT-rich genes, leading to

19   selective pressure for the retention of introns in the human genome.

20

21   **Introduction**

22

23   Mammalian genomes are characterised by large regional variation in base

24   composition (Bernardi, 1993). Regions with a high density of G and C nucleotides

25   (GC-rich regions) are in an open, transcriptionally active state, are gene-dense,

26   and replicate early. In contrast, AT-rich regions are enriched with

27   heterochromatin, contain large gene deserts and replicate late (Arhondakis et al.,

28   2011; Lander et al., 2001; Vinogradov, 2003). The mechanisms that give rise to

29   this compositional heterogeneity have been under debate for years and many

30   researchers believe that the pattern originates from the process of GC-biased

31   gene conversion (Duret and Galtier, 2009), though other neutral and selective

32   mechanisms have been proposed as well (Eyre-Walker, 1991; Galtier et al., 2018;

33   Plotkin and Kudla, 2011; Sharp and Li, 1987b).

34

35   The sequence composition of mammalian genes correlates with the GC-content

36   of their genomic location. Thus, introns and exons of genes located in GC-rich

37   parts of the genome are themselves GC-rich. This can potentially influence gene

38   expression in multiple ways: nucleotide composition affects the physical

39   properties of DNA, the thermodynamic stability of RNA folding, the propensity of

40   RNA to interact with other RNAs and proteins, the codon adaptation of mRNA to

41   tRNA pools, and the propensity for RNA modifications, such as m6A (Dominissini

42   et al., 2012) and ac4C (Arango et al., 2018). Strikingly, studies of the effects of

43   nucleotide composition on gene expression in human cells have led to opposing

44   conclusions. On the one hand, heterologous expression experiments typically

45   report large positive effects of GC content on protein production in a wide

46   variety of transgenes, including fluorescent reporter genes, human cDNAs, and

47   viral genes (Bauer et al., 2010; Kosovac et al., 2011; Kotsopoulou et al., 2000;

48   Kudla et al., 2006; Zolotukhin et al., 1996). As a result, increasing the GC content

49   of transgenes has become a common strategy in coding sequence optimization

50   for heterologous expression in human cells (Fath et al., 2011). On the other hand,

51   genome-wide analyses of endogenous genes typically show little or no

52   correlation of GC content with expression (Duan et al., 2013; Lercher et al., 2003;

53   Rudolph et al., 2016; Semon et al., 2005).

54

55 We hypothesized that the conflicting results in heterologous and endogenous
56 gene expression studies can be partially explained by RNA splicing. Most
57 transgenes used in heterologous expression systems have no introns, whereas
58 97% of genes in the human genome contain one or more introns. Splicing is
59 known to influence gene expression at multiple stages, including nuclear RNP
60 assembly, RNA export, and translation. If splicing selectively increased the
61 expression of AT-rich genes, it could account for the lack of correlation of GC
62 content and gene expression in previous genome-wide studies. We therefore
63 compared spliced and unspliced genes with respect to their (1) genomic codon
64 usage, (2) expression levels of reporter genes in transient and stable transfection
65 experiments and (3) global expression patterns in human transcriptome studies.
66 We show that splicing increases the expression of AT-rich genes, but not GC-rich
67 genes, in part through effects on cytoplasmic RNA enrichment.

68

69 **Results**

70

71 **Codon usage of human protein-coding genes depends on RNA splicing**
72 We first analysed the relationship between the nucleotide composition of human
73 genes and splicing. GC4 content (GC content at 4-fold degenerate sites) correlates
74 negatively with the number of exons in humans (Figure 1A; Spearman's ρ =
75 -0.27; p < 2.2×10^{-16}; see also (Carels and Bernardi, 2000; Ressayre et al., 2015;
76 Savisaar and Hurst, 2016)). In addition, GC4 content is highest in 5'-proximal
77 exons (Figure 1B; Spearman's ρ = -0.18; p < 2.2×10^{-16}), and first exons have a
78 higher GC4 content than second exons (p < 2.2×10^{-16}, one-tailed Wilcoxon test).
79 Although these patterns could result from proximity to CpG-rich transcription
80 start sites (TSSs)(Zhang et al., 2004), we found that first exons have significantly
81 higher GC4 content than second exons even when controlling for the distance
82 from the TSS (Figure 1C). This suggests that splicing contributes to the observed
83 enrichment of G and C nucleotides in the 5'-proximal exons in human.

84

85 To understand the causal links between splicing and nucleotide composition, we
86 studied the compositional patterns of retrogenes. Retrotransposition provides a

87   natural evolutionary experiment of what happens when a previously spliced

88   gene suddenly loses its introns. We first analysed a set of 49 parent-retrogene

89   pairs for which both the parent and the retrocopy ORFs have been retained in

90   human and mouse. Strikingly, we found that the retrocopies had a significantly

91   higher GC4 content than their parents (median $GC4_{retrocopy}$ - $GC4_{parent}$ = 11.5%; p

92   = 2.1×10^{-4} from one-tailed Wilcoxon test; Figure 1D). It thus appears that after

93   retrotransposition, newly integrated intronless genes come under selective

94   pressure for increased GC content. In a comparison of 31 parent-retrogene pairs

95   retained between human and macaque, the median GC4 difference is not

96   significant (0.09%; p = 0.13, Wilcoxon test), but this may be explained by

97   duplication events in macaques being more recent (dS ~0.06) than in mouse (dS

98   ~ 0.5) and therefore less evolutionary time has passed to allow changes in GC

99   composition to have occurred. As a control, we analysed the GC4 content of

100 retrocopies classed as pseudogenes (Figure S1A) and found it to be significantly

101 lower compared to their parental genes (-2.963%; p < 2.2×10^{-16}, Wilcoxon test).

102 Furthermore, the genomic neighbourhood of functional retrocopies and

103 pseudogenes had significantly lower GC content than the neighbourhood of their

104 respective parental genes (Figure S1B). These observations suggest that

105 increased GC content is not intrinsically connected with retrotransposition, but

106 is required for maintaining long-term functionality of retrogenes. Taken

107 together, these results support a splicing-dependent mechanism shaping

108 conserved patterns of nucleotide composition across functional protein-coding

109 genes.

110

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

112 The above analyses show a connection between splicing and genomic GC content

113 of endogenous human genes. To test whether splicing differentially affects the

114 expression of genes depending on their GC content, we designed 22 synonymous

115 variants of GFP that span a broad range of GC3 content (GC content at the third

116 positions of codons) (Mittal et al., 2018) (Figure S2). The collection encompasses

117 most of the variation in GC3 content found among human genes. All variants

118 were independently designed by randomly drawing each codon from an

119 appropriate probability distribution, to ensure uniform GC content and statistical

120   independence between sequences. We cloned these variants into two
121   mammalian expression vectors: an intronless vector with a CMV promoter
122   (pCM3) and a version of the same vector with a synthetic intron located in the 5'
123   UTR (pCM4). The GC content profiles of the 5' UTRs were similar in both vectors
124   (Figure S2E,F). The vectors also encoded a far-red fluorescent protein, mKate2,
125   which we used to normalize GFP protein abundance (normalization reduced
126   measurement noise, but similar results were obtained with and without
127   normalization). Transient transfections of HeLa cells with three independent
128   preparations of each plasmid showed reproducible expression with a large
129   dynamic range: synonymous variants differed in GFP protein production 46-fold.
130   Consistent with previous studies, GFP fluorescence was strongly correlated with
131   GC3 content, both in spliced and unspliced genes (Figure 2A,B). Interestingly,
132   introduction of an intron into the 5' UTR increased the expression of most, but
133   not all variants. Typically, GC-poor variants experienced a large increase of
134   expression in the presence of an intron, whereas GC-rich variants were
135   unaffected or experienced a moderate increase (Figure 2C).

136

137   We obtained similar results in stably transfected HeLa and HEK293 cells (Figure
138   S3) and when expressing an independently designed collection of 25
139   synonymous variants of mKate2 in HeLa cells (Figure 2D-F, S2D). A Fisher's
140   exact test revealed that the percentage of variants whose expression was
141   increased by splicing significantly depended on GC3 content (p=0.02, N=47, GFP
142   and mKate variants combined). These experiments show that many AT-rich
143   genetic variants are expressed inefficiently in human cells, but low expression
144   can be partially rescued by splicing. Notably, the average GC content of the
145   human genome is 41% (Li, 2011). In our experiments, genes with GC content
146   below 41% are expressed extremely inefficiently, unless they contain an intron
147   (Figure 2). This may provide a strong selective pressure for the retention of
148   introns in many human genes.

149

150   To establish which stages of expression are responsible for this phenomenon, we
151   first measured mRNA abundance of GFP variants in transiently transfected HeLa
152   cells by quantitative RT-PCR (qRT-PCR). High GC content may introduce

6

153 unwanted bias in RT-PCR, so to allow fair comparison of all variants irrespective

154 of their GC content, qRT-PCR primers were placed in the untranslated regions,

155 whose sequence did not vary. Similar to protein levels, mRNA abundance varied

156 widely between synonymous variants of GFP. GC-poor variants experienced a

157 large increase of expression in the presence of an intron, whereas GC-rich

158 variants were less affected (Figure 2G-I). The range of variation in mRNA

159 abundance was much smaller in constructs with an intron than without intron

160 (Figure 2I), indicating that splicing buffers the effects of GC content on

161 expression.

162

163 We then asked if changes in mRNA abundance arose at transcriptional or post-

164 transcriptional levels. As a proxy for transcriptional efficiency, we measured the

165 abundance of intronic RNA for GFP variants expressed from the intron-

166 containing plasmid. GC content did not correlate with intronic RNA abundance

167 (Figure 2J), suggesting that the rate of transcription does not depend on GC

168 content of the coding sequence found downstream of the intron. Conversely, high

169 GC content was associated with stabilization in unspliced constructs (Figure 2K).

170 Taken together, these experiments show that splicing preferentially increases

171 the expression of GC-poor synonymous variants at a post-transcriptional level.

172

173 **High GC content at the 5' end correlates with efficient expression**

174 To further explore the sequence determinants of expression, we assembled a

175 pool of 217 synonymous variants of GFP that included the 22 variants studied

176 above, 137 variants from our earlier study (Kudla et al., 2009), and 58 additional

177 variants. We cloned the collection into plasmids with and without a 5' UTR

178 intron. We then established pools of HeLa Flp-In T-REx cells that stably express

179 these constructs from a single genomic locus under a doxycycline-inducible

180 promoter and measured the protein levels of all variants by Flow-Seq (Kosuri et

181 al., 2013). We also performed Flow-Seq in HEK293 cells using the intronless

182 constructs only. In Flow-Seq, a pool of cells is sorted by FACS into bins of

183 increasing fluorescence and the distribution of variants in each bin is probed by

184 amplicon sequencing to quantify protein abundance (Figure 3A). All variants

185 could be quantified with good technical and biological reproducibility and high

186   correlation was found between Flow-Seq and spectrofluorometric measurement

187   of individual constructs (Figure S4). Most variants showed the expected

188   unimodal distribution across fluorescence bins, but some variants showed

189   bimodal distributions, possibly indicative of gene silencing in a fraction of cells.

190

191   All Flow-Seq experiments showed substantial variation of expression between

192   synonymous variants of GFP (Figure 3B). GFP protein levels in HeLa cells (with

193   intron), HeLa cells (without intron), and HEK293 cells (without intron) were all

194   correlated with each other, but the moderate degree of correlation (r=0.51

195   HEK293 (without intron) vs HeLa (without intron); r=0.36 Hela (with intron) vs

196   HeLa (without intron)) suggests that the effects of codon usage on expression are

197   modulated by splicing and by cell line identity - in agreement with prior

198   observations of tissue-specific codon usage (Burow et al., 2018; Gingold et al.,

199   2014; Plotkin et al., 2004; Rudolph et al., 2016). Flow-Seq of unspliced variants in

200   HeLa and HEK cells confirms the positive correlation of synonymous site GC-

201   content with expression (Figure 3C). In contrast to the results reported by us and

202   others in bacteria and yeast (Goodman et al., 2013; Kudla et al., 2009; Shah et al.,

203   2013), but consistently with the positive correlation between GC content and

204   expression, strong mRNA folding near the beginning of the coding sequence

205   correlated with increased expression (Spearman's ρ = 0.27 in HeLa cells; ρ = 0.4

206   in HEK293 cells). Expression was positively correlated with CpG content and

207   codon adaptation index (CAI), and negatively correlated with the estimated

208   density of ARE elements or cryptic splice sites. Because of the strong correlation

209   between GC content, CpG content, CAI and mRNA folding energy, a multiple

210   regression analysis could not resolve which of these properties was causally

211   related to expression. Interestingly, for intron-containing variants, there was no

212   correlation, or a weak negative correlation, between expression and GC content,

213   CpG content, CAI, and mRNA folding energy (Figure 3C).

214

215   Some of the variants analysed by Flow-Seq featured large regional variation in

216   GC content (Figure S5A) and we asked whether the localization of low-GC and

217   high-GC regions within the coding sequence influences expression. We found

218   that the GC3 content in the first half of the coding sequence (nt 1-360), but not in

219   the second half (nt 361-720), was positively correlated with expression of

220   intronless GFP variants in the HeLa and HEK293 cells (Figure 3D). The GC3

221   content in the first half of the gene showed no correlation with expression in the

222   intron-containing constructs.

223

224   To further test whether the GC content at the 5' end of genes has a particularly

225   important effect on expression, we constructed in-frame fusions between GC-

226   rich and GC-poor variants of GFP and mKate2 genes and quantified their protein

227   and mRNA abundance in transient transfection experiments. Expression levels

228   showed a striking dependence on the GC content profile. mKate2_GCpoor

229   showed undetectable expression on its own or as a 5'-terminal fusion with GC-

230   rich GFP, but it was efficiently expressed as a 3'-terminal fusion with GC-rich GFP

231   (Figure 3E). By contrast, mKate2_GCrich was efficiently expressed both as 5'-

232   terminal and 3'-terminal fusion. Analogous experiments with GC-rich and GC-

233   poor variants of GFP fused to mKate2_GCrich led to similar conclusions (Figure

234   S5B). The differences in protein levels between the fusion constructs could be

235   explained by differences in mRNA abundance (Figure 3E).

236

237   **High GC content leads to cytoplasmic enrichment of mRNA and higher**

238   **ribosome association**

239   Using the pooled HeLa cells used in Flow-Seq, we then analysed the effects of GC

240   content on mRNA localization. We separated the pools into nuclear and

241   cytoplasmic fractions, isolated RNA and performed amplicon sequencing of each

242   fraction to analyse mRNA localization of each GFP variant. Analysis of fractions

243   showed the expected enrichment of the lncRNA MALAT1 in the nucleus, and of

244   tRNA in the cytoplasm, confirming the quality of fractionations (Figure 4A). For

245   each GFP variant, we calculated the relative cytoplasmic concentration of its

246   mRNA (RCC) as the ratio of cytoplasmic read counts to the sum of reads from

247   both fractions (RCC = c_cyto / (c_cyto+c_nuc); Figure 4B). A value of 0 therefore

248   indicates 100% nuclear retention, whereas a value of 1 indicates 100%

249   cytoplasmic localisation. In the absence of splicing, RCC scores ranged from 0.09

250   to 0.64 and RCC correlated significantly with GC content (r=0.51, p=3.85×10^{-13},

251   Figure 4C). In the presence of a 5' UTR intron, we observed a significant increase

252   in RCC score for GFP variants with low GC content, but no increase in RCC for GC-

253   rich variants (Figure 4D). GC3 content at the beginning of the coding sequence

254   was significantly correlated with RCC in the absence of splicing (r=0.5, p=2.0×10^{-11}), but not in the presence of splicing (R<0.01, p=0.48; Figure S5). Thus, high GC

256   content at the 5' end of genes increases gene expression in part through

257   facilitating the cytoplasmic localization of mRNA.

258

259   To assess whether GC content also affects translational dynamics, we performed

260   polysome   profiling   on   HEK293   GFP   pool   cells   using   sucrose   gradient

261   fractionation (Figure 5A). qRT-PCR analyses of RNA extracted from all collected

262   fractions showed a broad distribution of GFP across fractions, with enrichment

263   within   polysome-associated   fractions.   In   order   to   determine   distribution

264   patterns of individual GFP variants, RNA from several fractions was pooled (as

265   indicated   in   Figure   5B)   and   subjected   to   high-throughput   sequencing.   The

266   resulting   read   distribution   indicates   that   GC-rich   variants   are   associated   with

267   denser polysomal fractions (ribosome density, Figure 5C, left panel; R^2=0.55, p <

268   2.2×10^{-16}) and are more likely to be translated (ribosome association, Figure 5C,

269   right panel; R^2=0.28, p<9.03×10^{-15}), compared to GC-poor variants. This suggests

270   that   enhanced   translational   dynamics   also   contribute   to   more   efficient

271   expression of GC-rich genes.

272

273   The expression fate of endogenous RNA depends on splicing, nucleotide

274   composition, and cell type

275   To test whether splicing- and position-dependent effects of codon usage can also

276   be observed among human genes, we turned to genome-wide measurements of

277   expression at endogenous human loci and related these measurements to codon

278   usage and splicing. Although the correlations between GC content and expression

279   depended on the experimental measure and type of cells under study, we find

280   that   GC4   content   usually   has   a   more   positive   effect   on   gene   expression   in

281   unspliced genes relative to spliced ones (Figure 6, Table S1). In particular,

282   unspliced mRNAs show a more positive/less negative correlation of GC4 with

283   transcription initiation (GRO-cap data); cytoplasmic stability (exosome mutant);

284   RNA   (whole   cell   RNA-seq);   cytoplasmic   enrichment   (cell   fractionation),

10

285 translation rate (ribosome profiling vs whole cell RNA-seq); and protein amount

286 (mass-spec). These analyses suggest that GC4 content has an effect on the RNA

287 abundance of intronless mRNA molecules, which is carried through to the

288 protein expression. Taken together, these genome-wide analyses support our

289 observation of a splicing-dependent relationship between codon usage and

290 expression in human cells.

291

292 **Discussion**

293

294 We have shown that the effects of GC content on gene expression in human cells

295 are splicing-dependent (the effect is larger in unspliced genes compared to

296 spliced genes) and position-dependent (the effect is larger at the 5' end of genes

297 than at the 3' end). In addition, human genes show striking patterns of codon

298 usage, which differ between spliced and unspliced genes and between first and

299 subsequent exons. Our results have implications for the understanding of the

300 evolution of human genes and the functional consequences of synonymous

301 codon usage.

302

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

304 Specific patterns of codon usage have previously been found at the 5' ends of

305 genes in bacteria, yeast and other species (Gu et al., 2010; Kudla et al., 2009;

306 Tuller et al., 2010). In bacteria and yeast, strong mRNA folding near the start

307 codon prevents ribosome binding and reduces translation efficiency, resulting in

308 selection against strongly folded 5' mRNA regions (Kudla et al., 2009; Shah et al.,

309 2013). In addition a "ramp" of rare codons has been observed near the 5' end of

310 RNAs in multiple species, with a possible role in preventing a wasteful

311 accumulation of ribosomes on mRNAs (Tuller et al., 2010) or reducing the

312 strength of mRNA folding (Bentele et al., 2013). These phenomena cannot

313 explain our results in human, because both the folding energy and codon ramp

314 models predict low GC content near the start codon, whereas we observe high GC

315 content within first exons of human protein-coding genes (Figure 1B).

316 Furthermore, our experiments show that high GC content near the start codon

317   increases expression, whereas the folding energy and codon ramp models would

318   predict low expression.

319

320   We propose instead that splicing- and position-dependent effects of GC content

321   are explained by co-transcriptional or early post-transcriptional events in the

322   lifetime of an mRNA. Using matched reporter gene libraries, we show that most,

323   but not all, variants show an increase in expression when spliced. Splicing

324   typically increases the expression of AT-rich variants, but it does not further

325   increase the expression of GC-rich transcripts, which suggests that splicing and

326   high GC content influence expression through at least one common mechanism.

327   Splicing increases transcription (Kwek et al., 2002), prevents nuclear

328   degradation (Nott et al., 2003), facilitates nuclear-cytoplasmic mRNA export

329   through the Aly/REF-TREX pathway (Muller-McNicoll et al., 2016), and

330   stimulates translation (Nott et al., 2004). High GC content might increase RNA

331   polymerase processivity (Bauer et al., 2010; Zhou et al., 2016); GC-rich variants

332   are less likely to contain cryptic polyadenylation sites (consensus sequence:

333   AAUAAA) or destabilizing AU-Rich Elements (AREs) (Higgs et al., 1983); and high

334   GC content near the 5' end may also facilitate cytoplasmic localisation of mRNA.

335   GC-rich sequence elements of endogenous unspliced genes were previously

336   shown to route transcripts into the splicing-independent ALREX nuclear export

337   pathway, allowing efficient cytoplasmic accumulation (Palazzo et al., 2007). In

338   agreement with this, low expression caused by inhibitory sequence features

339   (such as low GC-content) can be rescued by extending the mRNA at the 5'end

340   with a GC-rich sequence (Figure 3D,E, Figure S5). This may act as a

341   compensatory mechanism when gene expression cannot rely on the positive

342   regulatory effects of splicing (Palazzo and Akef, 2012). In contrast, it was

343   recently shown that binding of HNRNPK to the GC-rich SIRLOIN motif leads to

344   nuclear enrichment of lncRNAs (and also some mRNAs) (Lubelsky and Ulitsky,

345   2018). Our genomic analyses of lncRNA sequences do not show the same

346   splicing-dependent compositional patterns as observed in mRNAs and it is

347   therefore likely that antagonistic pathways act simultaneously in shaping the

348   RNA expression landscape. Thus, we propose that the genomic patterns and their

349   consequences on gene expression reported here are general features of protein-
350   coding genes.

351

352   Recent studies also highlight patterns of codon usage as major determinants of
353   RNA stability in yeast (Presnyak et al., 2015), zebrafish (Mishima and Tomari,
354   2016) and other species (Bazzini et al., 2016). The usage of less common, 'non-
355   optimal' codons within transcripts was shown to control poly-A tail length and
356   RNA half-life in a translation-dependent manner through the coupled activity of
357   different CCR4-NOT nucleases (Radhakrishnan et al., 2016; Webster et al., 2018).
358   Consistent with these findings, we observed that CAI is positively correlated with
359   mRNA expression levels in human cells. However, it remains to be seen whether
360   the correlation of CAI with mRNA expression depends on translation. Because of
361   the strong correlation between GC content and CAI, it is difficult to disentangle
362   independent contributions of these variables. Additionally, we find that the
363   correlation between GC content (or CAI) and expression is position- and splicing-
364   dependent, whereas no evidence for such context-dependence has been reported
365   for the CCR4-NOT-mediated mechanism.

366

367   Other instances in which the effects of codon usage are context-dependent have
368   been described. Most notably, tRNA populations and transcriptome codon usage
369   patterns were shown to differ between mammalian tissues (Dittmar et al., 2006;
370   Gingold et al., 2014; Plotkin et al., 2004; Rudolph et al., 2016). Intriguingly, genes
371   preferentially expressed in proliferating cells and tissue-specific genes tend to be
372   AT-rich, whereas genes expressed in differentiated cell types and housekeeping
373   genes are more GC-rich (Gingold et al., 2014; Vinogradov, 2003). Although these
374   differences have been interpreted in terms of the match between codon usage
375   and cellular tRNA pools, it is plausible that translation-independent mechanisms
376   contribute to context-dependent effects of codon usage. Accordingly, in
377   Drosophila, codon optimality determines mRNA stability in whole cell embryos,
378   but not in the nervous system, independent of tRNA abundance (Burow et al.,
379   2018). Recently, it was shown that Zinc-finger Antiviral Protein (ZAP) selectively
380   recognises high CpG-containing viral transcripts as a mechanism to distinguish
381   self from non-self (Takata et al., 2017). We speculate that similar regulatory

13

382 proteins and mechanisms exist for cellular expressed genes. The cell lines used in
383 the present study, HeLa and HEK293, are both rapidly proliferating and
384 experimental results are correlated (r=0.36, Flow-Seq data), but divergent
385 expression of some GFP variants was also observed. Similarly, the effect size of
386 GC content on the expression of endogenously expressed genes varies with cell
387 type. It would be interesting to compare the expression of our variants in other
388 cell types to further address the question of tissue-specific codon usage and
389 adaptation to tRNA pools.
390

Implications for the evolution of protein-coding genes
392 The fact that long, multi-exon genes are often found in GC-poor regions of the
393 genome might result from regional mutation bias. However, an alternative
394 explanation is possible: GC-poor genes may be under selective pressure to retain
395 their introns, and intronless genes may experience selective pressure to increase
396 their GC content. These possibilities are supported by multiple observations:
397 Firstly, endogenous intronless genes are on average more GC-rich than intron-
398 containing genes. Secondly, the GC content of functional (but not non-functional)
399 retrogenes is higher compared to their respective intron-containing parental
400 genes, which cannot be explained by a systematic integration bias. Thirdly, in
401 genome-wide analysis, correlations between GC-content and expression are
402 generally more positive (or less negative) for unspliced compared to spliced
403 genes. Taken together, this suggests that for the long-term success of an
404 unspliced gene (i.e. stable conservation of expression and functionality) an
405 increase in GC content is essential. By contrast, splicing allows genes to remain
406 functional even when mutation bias or other mechanisms lead to a decrease of
407 their GC content.
408

Materials and Methods

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 same PCR-based method as 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 Supplementary Table 2 for list of all primers used), generating plasmid pCM1.

15

441   To be able to normalise spectrophotometric measurements from single GFP

442   transfection experiments, pCM1 and pCM2 were further modified to contain a

443   separate expression cassette driving the expression of a second fluorescent

444   reporter gene, mKate2. The mKate2 gene cassette from pmKate2-N (Evrogen)

445   was inserted via Gibson assembly cloning: First, the entire mKate2 expression

446   cassette was amplified using primers 'mKate2_gibs_F' and 'mKate2_gibs_R' which

447   add overhangs homologous to the pCM insertion site. Next, pCM1 and pCM2

448   were linearised by PCR using primers 'pCI_gib_F' and 'pCI_gib_R'. All PCR

449   products were purified using the Qiagen PCR purification kit and fragments with

450   homologous sites recombined using the Gibson assembly cloning kit (NEB)

451   according to manufacturer's instructions (NEB). Successful integration was

452   validated by Sanger sequencing. This generated plasmids pCM3 (-intron,

453   +mKate2) and pCM4 (+intron, +mKate2).

454

455   **Transient plasmid transfections for spectrofluorometric measurements**

456   Plasmids for transient expression of fluorescent genes were transfected into

457   HeLa cells grown in 96-well plates. Per plasmid construct, 3 technical replicates

458   were tested by reverse transfection. Enough transfection mix for 4 wells was

459   prepared by diluting 280ng plasmid DNA in 40ul OptiMem (Gibco). 1ul

460   Lipofectamine2000 (Invitrogen; 0.25ul per well) was diluted in 40ul OptiMem

461   and incubated for 5min at room temperature. Both plasmid and

462   Lipofectamine2000 dilutions were then mixed (80ul total volume) and further

463   incubated for 20-30min. 20ul of transfection complex was then pipetted into 3

464   wells before adding 200ul of HeLa cell suspension (45,000 cells/ml; 9,000

465   cells/well) in phenol red-free DMEM (Biochrom, F0475). Media was exchanged

466   3-4h post-transfection to reduce toxicity. Cells were then grown for a further 24h

467   or 48h at 37C, 5% CO2.

468   After incubation, cells were lysed by removing media and adding 200ul of cell

469   lysis buffer (25mM Tris, pH 7.4, 150mM NaCl, 1% Triton X-100, 1mM EDTA, pH

470   8). Fluorescence readings were obtained using a Tecan Infinite M200pro

471   multimode plate reader. The plate was first incubated under gentle shaking for

472   15min followed by fluorescence measurements using the following settings:

473   Ex486nm/Em 515nm for GFP and Ex588nm/Em633nm for mKate2; reading

474   mode: bottom; number of reads: 10 per well; gain: optimal.

475   For data analysis, measurements of untransfected cells were subtracted as

476   background from all other wells. For comparability of different plates within a

477   set of experiments, the same 3 genes were transfected on every plate to account

478   for technical variability. In the screen of individual GFP variants (see Figure 2),

479   GFP measurements were divided by mKate2 measurements from same wells to

480   reduce noise caused by well-to-well variation in transfection efficiency, but

481   similar results were obtained without normalisation.

482

483   **Transient transfections and RNA extraction for qRT-PCR analysis**

484   HeLa cells were reverse transfected in 12-well plates using 800ng plasmid DNA

485   and 2ul Lipofectamine 2000 (Invitrogen). DNA and Lipofectamine 2000 were

486   diluted in 100ul OptiMEM (Gibco) each, incubated for 5min, mixed and further

487   incubated for 20min. The transfection complex was then added to each well

488   before adding 10^5 HeLa cells. Cells were incubated for 24h at 37C, 5% CO2 before

489   harvesting. Cells were then harvested by adding 1ml Trizol reagent (Life

490   technologies). RNA was extracted according to manufacturer's instructions.

491   Resulting RNA was further treated with DNAse I using the Turbo DNase kit

492   (Ambion) to remove any residual plasmid and genomic DNA.

493

494   **qRT-PCR analysis**

495   cDNA for qRT-PCR analysis was prepared using SuperScript III Reverse

496   Transcriptase (Life technologies) according to the manufacturer's

497   recommendations with 500ng total RNA as template and 500ng random

498   hexamers (Promega). All qRT-PCRs were carried out on a Roche LightCycler 480

499   using Roche LightCycler480 SYBR Green I Master Mix and 0.3uM gene-specific

500   primers. Samples were analysed in triplicate as 20ul reactions, using 2ul of

501   diluted cDNA. Cycling settings: DNA was first denatured for 5min at 95°C before

502   entering a cycle (50-60x) of denaturing for 10sec at 95°C, annealing for 7sec at

503   55-60°C (depending on primers used), extension for 10sec at 72°C and data

504   acquisition. DNA was then gradually heated up by 2.20 °C/s from 65 to 95°C for

505   5sec each and data continuously collected (Melting curve analysis). Data was

506 evaluated using the comparative Ct method (Livak and Schmittgen, 2001). RNA

507 measurements from transient transfection experiments were normalised to the

508 abundance of neomycin RNA, which is expressed from the same plasmid, to

509 control for differences in transfection efficiency (primers 'Neo_F' and 'Neo_R').

510

511 **Subcellular fractionation**

512 This protocol is based on the cellular fractionation protocol published by

513 (Gagnon et al., 2014) but includes a further clean-up step using a sucrose cushion

514 as described by (Zaghlool et al., 2013) and a second lysis step as described by

515 (Wang et al., 2006). Cell lysis and nuclear integrity was monitored throughout by

516 light microscopy following Trypan blue staining (Sigma). Cells were grown in

517 10cm plates for 24h to about 90% confluency. Cells were then washed with PBS

518 and trypsinised briefly using 1ml of 1xTrypsin/EDTA. After stopping the reaction

519 with 5ml DMEM, cells were transferred into 15ml falcon tubes and collected by

520 spinning at 100g for 5min. Resulting cell pellets were resuspended in 500ul ice-

521 cold PBS, transferred into 1.5ml reaction tubes and spun at 500g for 5min, 4°C.

522 The supernatant was discarded and cells resuspended in 250ul HLB (10mM Tris

523 (pH 7.5), 10mM NaCl, 3mM MgCl2, 0.5% (v/v) NP40, 10% (v/v) Glycerol, 0.32M

524 sucrose) containing 10% RNase inhibitors (RNasin Plus, Life Tech) by gently

525 vortexing. Samples were then incubated on ice for 10min. After incubation,

526 samples were vortexed gently, spun at 1000g for 3min, 4°C, and supernatants

527 and pellets were processed separately as indicated in a) and b) below.

528 a) Cytoplasmic extract:

529 The supernatant was carefully layered over 250ul of a 1.6M sucrose cushion and

530 spun at 21,000g for 5min. The supernatant was then transferred into a fresh

531 1.5ml tube and 1ml Trizol was added and mixed by vortexing.

532 b) Nuclear extract:

533 The pellets were washed 3 times with HLB containing RNase inhibitors by gently

534 pipetting up and down 10 times followed by a spin at 300g for 2min. After the

535 3rd wash, nuclei were resuspended in 250ul HLB and 25ul (10%) of detergent

536 mix (3.3% (wt/wt) sodium deoxycholate/6.6% (vol/vol) Tween 40) dropwise

537 added while vortexing slowly (600rpm). Nuclei were then incubated for 5min on

538 ice before spinning at 500g for 2min. The supernatant was discarded and pellets

18

539   resuspended in 1ml Trizol (Ambion) by vortexing. 10ul 0.5M EDTA are added to

540   each nuclear sample in Trizol and tubes heated to 65°C for 10min to disrupt very

541   strong Protein-RNA and DNA-RNA interactions. Tubes were then left to reach

542   room temperature and RNA was extracted following the manufacturer's

543   instructions.

544

545   **Transcription inhibition assay**

546   HeLa T-Rex Flp-in cell lines were grown to 80-90% confluency in 6 well for 24h

547   before treatment with 500nM Triptolide (Sigma). Cells were harvested at

548   indicated time points and RNA extracted using Trizol reagent (Ambion). Control

549   cells were treated with the equal volume of DMSO (drug carrier). To assess

550   transcript levels, qRT-PCR was performed as described above. GFP levels were

551   normalised to levels of 7SK, a RNA polymerase III-transcribed non-coding RNA,

552   whose expression levels are not affected by Triptolide treatment. Relative

553   transcript levels of c-Myc are shown as an example of a relatively unstable

554   transcript.

555

556   **Generation of stable Flp-in cell lines**

557   We adopted a multiplex-Gateway integration method to create a pool of 217 GFP

558   plasmids which are compatible with the T-Rex Flp-in system (Invitrogen) for

559   creating stable, doxycycline-inducible cell lines, in which each variant is

560   expressed from the same genomic locus, allowing direct comparison of

561   expression levels.

562   pcDNA5/FRT/TO/DEST (Aleksandra Helwak, University of Edinburgh) contains

563   the Gateway-compatible attB destination cassette to allow the subcloning of

564   genes from any Gateway-entry vectors. This plasmid was further modified to

565   contain the same 5'UTR intron sequence as in pCM4 used in transient expression

566   experiments using Gibson Assembly (NEB): the intronic sequence was amplified

567   from pCM4 by PCR using primers 'Gib_intr_F' and 'Gib_intr_R' using Q5 High-

568   Fidelity Polymerase (NEB). The primers added 15nt overhangs which are

569   homologous to the ends of pcDNA5/FRT/TO/DEST when linearised with AflII.

570   The Gibson assembly reaction was performed as per manufacturer's instructions

571   (NEB), generating pcDNA5/FRT/TO/DEST/INT.

19

572 217 individual GFP variants stored in Gateway-entry vector pGK3 were mixed

573 with a concentration of 0.06ng of each GFP variant. For each pcDNA5 destination

574 vector, a separate Gateway LR reaction was set-up in a total volume of 45ul using

575 500ng destination vector, 5ul LR Clonase enzyme mix, 38ul of the mixed 217

576 pGK3-GFP plasmids and TE (pH 8). The reactions were incubated at 25C

577 overnight followed by Proteinase K digest (5ul, LR Clonase kit) for 10min at 37C.

578 The total 50ul reaction mix was transformed into 2.5ml highly competent

579 DH5alpha in a 15ml Falcon tube by heat-shocking cells for 2min 30s at 42C,

580 followed by cooling on ice for 3min, before adding 10ml SOC medium and

581 incubating while shaking for 1h at 37C. After incubation, cells were spun down at

582 3000g for 3min and resulting bacterial pellets resuspended in 1ml fresh SOC.

583 10x100ul were plated onto L-Ampicillin agar plates and incubated overnight at

584 37C resulting in >800 colonies per plate. Bacterial colonies were scraped off the

585 plates and collected in a falcon tube. Plasmid DNA was extracted using a Qiagen

586 Midiprep kit according to the manufacturer's instructions, resulting in two

587 plasmid pools: pCDNA5/GFPpool and pcDNA5/INT/GFPpool. Both pools were

588 subjected to high-throughput sequencing to confirm the presence of different

589 GFP variants.

590 HeLa T-Rex Flp-in cells (gifted by the Andrew Jackson lab, The University of

591 Edinburgh) and Hek293 T-Rex Flp-in (Thermo Scientific) were grown to 80%

592 confluency in 6 well plates. For GFP plasmid pool transfections,

593 pCDNA5/GFPpool or pCDNA5/INT/GFPpool were mixed in a 9:1 ratio with the

594 Flp-recombinase expression plasmid pOG44 (Invitrogen) to give 2ug in total

595 (1.8ug pOG44 + 0.2ug pCDNA5) and diluted in OptiMEM (Gibco) to 100ul.

596 Transfections were performed with 9ul Lipofectamine2000 (Invitrogen) and

597 91ul OptiMEM per well by incubating 5min at room temperature before mixing

598 with plasmid DNA and a further 15min incubation. The transfection mix was

599 then added dropwise to the cells. Media were replaced with conditioned media

600 4h post-transfection. Cells were incubated for further 48h before chemical

601 selection to select for successful gene integration using 10ng/ul Blasticidin S

602 (ThermoFisher) and 400mg/ml (HeLa T-Rex Flp-in) or 100mg/ml (Hek293 T-

603 Rex Flp-in) Hygromycin B (Life Technologies). Successful selection was

604 determined by monitoring cell death in untransfected cells. Chemically resistant

20

605   cells represent pools of cell lines expressing different GFP variants from the
606   same genomic locus. High-throughput sequencing of the GFP integration site
607   within each generated cell line pool confirmed the successful integration of all
608   variants.
609   HeLa T-Rex Flp-in and Hek293 T-Rex Flp-in cell lines expressing two individual
610   GFP variants (GC3=96% and GC3=36%; see Supplementary Figure 3) as spliced
611   and unspliced transcripts were generated using the same protocol.
612
613   **Flow-Seq: FACS sorting and genomic DNA extraction**
614   80x15cm cell culture plates of HeLa T-Rex Flp-in GFP pool cells and 40x15cm cell
615   culture plates of Hek293 T-Rex Flp-in GFP pool cells were induced with 1ug/ml
616   Doxycyline (Sigma, D9891) in phenol red-free DMEM (Biochrom, F0475)
617   supplemented with 10% FCS (Sigma, F-7524) and 2mM L-Glutamine. After 24h
618   or 48h, cells were harvested by gentle trypsinisation and cells were sorted into 8
619   fluorescence bins using a BD FACS Aria II cell sorter. To define the range of GFP
620   positive signal, cells without stable GFP expression were used as negative
621   control. 80% of HeLa and 90% Hek293 GFP pool cells fell into the GFP-positive
622   range. Each fluorescence bin was chosen to comprise roughly 10% of the GFP-
623   positive population. The bin spacing was kept the same for the sorting of HeLa
624   cell pools expressing unspliced and spliced GFP variants to allow direct
625   comparisons of the fluorescence profiles of individual variants.
626   About 10^7 cells per bin were collected in Polypropylene collection tubes (Falcon)
627   coated with 1% BSA/PBS, cushioned with 200ul 20%FBS/PBS. Cell suspensions
628   were decanted into 15ml tubes and cells collected by spinning 5min at 500g. The
629   supernatant was transferred into fresh 15ml tubes and precipitated using 2
630   volumes of 100% EtOH/0.1 volume Sodium Acetate (pH 5.3) and 10ul Glycoblue
631   (Ambion). Tubes were shaken vigorously for 10s before incubating at -20C for
632   15min, followed by spinning at 3000g for 20min. Resulting pellets were air-
633   dried, resuspended in 1ml digest buffer (100mM Tris pH 8.5, 5mM EDTA, 0.2%
634   SDS, 200mM NaCl) and then combined with the respective cell pellet. 10ul RNAse
635   A (Qiagen, 70U) was added and samples gently rotated at 37C. After 1h, 1ul/ml
636   Proteinase K (20mg/ml, Roche) was added to the samples before rotating a
637   further 2h at 55C. Genomic DNA was purified 3 times by using 1 volume

638 Phenol:Chloroform:Isoamyl alcohol (PCI, 25:24:1, Sigma). After each addition of
639 PCI, samples were shaken vigorously for 10s before spinning at 3000g for 20min
640 (first extraction) or 5min (all following). The resulting bottom layers including
641 the interphase were removed before each PCI addition. After the last PCI
642 extraction, the upper layer was transferred into a fresh 15ml tube and 1
643 extraction performed using 1 volume chloroform:isoamyl alcohol (CI,24:1,
644 Sigma). After a 5min spin at 3000g, the upper layer was transferred into a fresh
645 15ml tube and DNA precipitated using EtOH/Sodium Acetate as before. After a
646 5min incubation on ice, DNA was collected by spinning for 30min at 3000g. The
647 resulting DNA pellets were washed 2 times with 75% EtOH before air-drying and
648 resuspending in 200ul Tris-EDTA (10mM). The quality of the extracted genomic
649 DNA was assessed on a 0.8% Agarose/TBE gel.
650

Polysome profiling

652 Hek293 Flp-in GFP pool cell lines were grown to 90% confluency on 15cm
653 dishes. Cells were treated for 20min with 100ug/ul Cycloheximide before
654 harvesting cells by removing media, washing with 2x ice-cold PBS followed by
655 scraping cells into 1ml PBS and transferring into 1.5ml tubes. Cells were pelleted
656 at 7000rpm, 4°C for 1min and resulting cell pellet carefully resuspended by
657 pipetting up and down in 250ul RSB (10x RSB: 200mM Tris (pH 7.5), 1M KCl,
658 100mM MgCl2) containing 1/40 RNasin (40U/ul, Promega), until no clumps
659 were visible. 250ul of polysome extraction buffer was then added (1ml 10x RSB
660 + 50ul NP-40 (Sigma) + 9ml H2O + 1 complete mini EDTA-free protease inhibitor
661 pill (Roche)) and lysate passed 5x through a 25G needle avoiding bubble
662 formation. The lysate was then incubated on ice for 10min before spinning
663 10min at 10,000g, 4°C. The supernatant was then transferred into a fresh 1.5ml
664 tube and the RNA concentration estimated by measuring the OD at 260nm.
665 Sucrose gradients (10–45%) containing 20 mM Tris, pH 7.5, 10 mM MgCl2, and
666 100 mM KCl were made using the BioComp gradient master. 100ug of Lysate
667 were loaded on sucrose gradients and spun at 41,000rpm for 2.5h in a Sorvall
668 centrifuge with a SW41Ti rotor. Following centrifugation, gradients were
669 fractionated using a BioComp gradient station model 153 (BioComp 23

670   Instruments, New Brunswick, Canada) by measuring cytosolic RNA at 254 nm

671   and collecting 18 fractions.

672   RNA from all fractions was precipitated using 1 volume of 100% EtOH and 1ul

673   Glycoblue (Ambion), before extracting RNA using the Trizol method (Life

674   Technologies). Equal volumes of RNA of each fraction was run on a 1.3%

675   Agarose/TBE gel to assess the quality of fractionation and RNA integrity.

676   Additionally, equal volumes of RNA of each fraction were used in cDNA synthesis

677   using SuperScript III (ThermoFisher) and 2uM gene-specific primers for GFP

678   ('pcDNA5-UTR_R') and GAPDH ('GAPDH_R') followed by qRT-PCR analysis. For

679   high-throughput sequencing, total RNA from collected fractions was combined in

680   equal volumes into 4 pools (as indicated in Figure 5B; free ribonucleoprotein

681   (RNP) complexes, monosomes, light polysomes (2-4) and heavy polysomes (5+))

682   before amplicon library preparation (as described below).

683

684   **High-throughput library preparation and sequencing**

685   Sequencing libraries were generated by PCR using primers specific for GFP

686   amplification (Supplementary Table 2) which carry the required adaptor

687   sequences for paired-end MiSeq sequencing, as well as 6nt indices for library

688   multiplexing. Between 6-10ug of total genomic DNA were used in multiple PCR

689   reactions (200ng per 50ul reaction). All PCRs were performed using Accuprime

690   Pfx (NEB) according to manufacturer's recommendations using 0.4ul Accuprime

691   Pfx   Polymerase   and   0.3uM   of   each   primer   ('PE_PCR_left'   and

692   'S_indexX_right_PEPCR'). The cycling conditions were as follows: Initial

693   denaturation at 95C for 2min, followed by 30 cycles of denaturation at 95C for

694   15sec, annealing at 51C for 30sec, extension at 68C for 1min. The final extension

695   was performed at 68C for 2min. After PCR, all reactions of the same template

696   were pooled and 1/3 of the reaction purified using the Qiagen PCR purification

697   kit according to the manufacturer's instructions. DNA was eluted in 50ul H2O.

698   Library size selection was performed using the Invitrogen E-gel system

699   (Clonewell gels, 0.8% agarose) followed by Qiagen MinElute PCR purification.

700   Correct fragment sizes were confirmed and quantified using the Agilent

701   Bioanalyzer 2100 system.

702 For library preparation of RNA samples, 500ng RNA was first converted into
703 cDNA using 2nmol GFP-specific primers ('S_indexX_right_PEPCR') using
704 SuperScript III (Life technologies) according to manufacturer's protocol, using
705 50C as extension temperature. Resulting cDNA was then treated with 1ul
706 RNaseH (NEB) for 20min at 37C, followed by heat inactivation at 65C for 5min.
707 Samples were diluted 1:2.5 before using 2ul as template in PCR for library
708 preparation. A minimum of 8x50ul PCR reactions were set up and pooled for
709 each sample before PCR purification, followed by E-gel purification as described
710 above.

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

714

715 **Analysis of GFP pool experiments**

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

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

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

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

729 For polysome profiling, only variants with a minimum of 1000 reads across all 4
730 sequencing bins were used for further analysis. To estimate ribosome density,
731 for each GFP variant, the number of reads in each bin (n(i)) was multiplied by the
732 respective polysomal fraction index (i) before taking the sum and dividing by the
733 total sum of reads across all fractions:

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

735   Ribosome association for each variant was calculated as the sum of reads (n) in

736   light polysomes, heavy polysomes and monosomal fractions, divided by the sum

737   of reads found in the free RNP fraction:

738   Ribosome association(variant) = (n(monosomes) + n(light polysomes) +

739   n(heavy polysomes)) / n(free RNPs)

740

741   **Definition of calculated sequence features**

742   GC3: GC content in the third position of codons

743   CpG: number of CpG dinucleotides

744   dG: The minimum free energy of predicted mRNA secondary structure around

745   the start codon was calculated using the hybrid-ss-min program version 3.8

746   (default settings: NA = RNA, t = 37, [Na+] = 1, [Mg++] = 0, maxloop = 30, prefilter

747   = 2/2) in the 42-nt window (-4 to 38) as in (Kudla et al., 2009).

748   CAI: Codon Adaptation Index (*H. sapiens*) (Sharp and Li, 1987a) was calculated

749   using a reference list of highly expressed human genes collected from the EMBL-

750   EBI expression atlas https://www.ebi.ac.uk/gxa.

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

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

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

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

755   Poly_A: number of times the position-specific scoring matrix

756   ((47,3,0,50)(18,6,9,67)(53,12,12,23)(59,6,0,35)(70,6,6,18)) was identified in

757   each sequence.

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

759   SD_PSSM: number of times the position-specific scoring matrix

760   ((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

761   6,17,21,46)) was identified in each sequence.

762

763   FIMO (http://meme-suite.org) was calculated to identify and count sequence

764   motifs. Open-source packages available for R were used for generating

765   correlation matrices (corrplot), heatmaps (ggplot2), boxplots

766   (graphics/ggplot2), The GC3 of all human coding sequences (assembly:

767   GRCg38_hg38; only CDS exons) was calculated using R package 'seqinr'.

768

769   **Computation methods for analysis of endogenous gene expression**

770   **Data Collection – see also Supplementary Table 1**

771   1. GC4 content was calculated for each protein-coding transcript annotated

772   in GENCODE version 19 as the GC content of the third codon position

773   across all fourfold-degenerate codons (CT*, GT*, TC*, CC*, AC*, GC*, GA*,

774   CC*, GC*). The core promoter of each transcript is further defined as -300

775   bp/+100 bp around the annotated TSS.

776   2. The level of transcription initiation was quantified in K562 and Gm12878

777   cells as the number of GRO-cap reads from the same strand which overlap

778   the core promoter.

779   3. Nuclear stability was assessed using CAGE data obtained in triplicate from

780   Egfp, Mtr4 and Rrp40 knockdowns (GSE62047; (Andersson et al., 2014)).

781   Similarly to the approach used for the GRO-cap data, we calculated the

782   RPKM across core promoters for each library separately. The baseMean

783   expression for each treatment was quantified using DESeq2, where

784   promoters with no reads across any replicate were first removed from

785   each comparison. Nuclear stability was then assessed as the fold-change

786   between the Egfp and Mtr4 knockdown and cytoplasmic stability by the

787   estimated fold-change between the Mtr4 and Rrp40 knockdowns.

788   4. The level of the mature mRNA was quantified using RNA-seq libraries

789   from whole cell samples (prepared as described elsewhere for HEK293

790   cells and downloaded from

791   http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncod

792   eCshlLongRnaSeq for Gm12878, HepG2, HeLa, Huvec and K562 cells).

793   Reads were pseudoaligned against GENCODE transcript models using

794   Kallisto, set with 100 bootstraps. All other parameters were left at their

795   default. Transcript expressions were extracted as the estimated TPM

796   (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, $p < 2.2 \times 10^{-16}$), 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 Supplementary Figure 6. 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

27

828   bootstrapped by sampling with replacement and recalculating the regression

829   coefficients for spliced and unspliced transcripts. The 95% confidence interval of

830   these coefficients (discounting the standard error in these estimations) obtained

831   by 1,000 samplings of this type was used to draw the ellipses shown in Figure 6.

832

833   **Analysis of GC content variation in the human genome**

834   The GRCh38 sequence of the human genome, as well as the corresponding gene

835   annotations (Ensembl release 85), was retrieved from the Ensembl FTP site

836   (Zerbino et al., 2018). The full coding sequences (CDSs) of protein-coding genes

837   were extracted, filtered for quality and clustered into putative paralogous

838   families (see (Savisaar and Hurst, 2016) for full details). For all analyses, a

839   random member was picked from each putative paralogous cluster. In addition,

840   only one transcript isoform (the longest) was considered from each gene. Note

841   that exon rank was always counted from the first exon of the gene, even if it was

842   not coding. For Figure 1C, GC4 was averaged across all sites that were at the

843   same nucleotide distance to the TSS and within an exon of the same rank. For the

844   functional retrocopies analysis, the parent-retrocopy genes derived in (Parmley

845   et al., 2007) were used. Pseudogenic retrocopies were retrieved from

846   RetrogeneDB (Rosikiewicz et al., 2017). Retrocopy annotations were filtered to

847   only leave human genes with a one-to-one ortholog in *Macaca mulatta*. Next,

848   only ortholog pairs where both the human and the macaque copy were

849   annotated as not having an intact reading frame and where the human copy was

850   annotated as *KNOWN_PSEUDOGENE* were retained. For the analyses reported in

851   Supplementary Figure 1, the functional retrocopies were also retrieved from

852   RetrogeneDB, as we could not access genomic locations for the (Parmley et al.,

853   2007) set. The functional retrogenes were retrieved similarly to pseudogenes,

854   except that both the human and the macaque copy were required to have an

855   intact open reading frame and the human copy could not be annotated as

856   *KNOWN_PSEUDOGENE*.

857   Python 3.4.2. was used for data processing and R 3.1.2 was used for statistics and

858   plotting (R Development Core Team, 2005).

859

860   **Acknowledgments**

874

875

876

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

879 (A) GC4 distribution of human protein-coding genes, grouped by number of
880 exons per gene.

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

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

885 (D) GC4 distribution of functional retrogenes (dark grey) and their
886 corresponding parental genes (light grey) conserved between mouse and human
887 (p=2.1×10⁻⁴, from one-tailed Wilcoxon signed rank test, n=49).

888
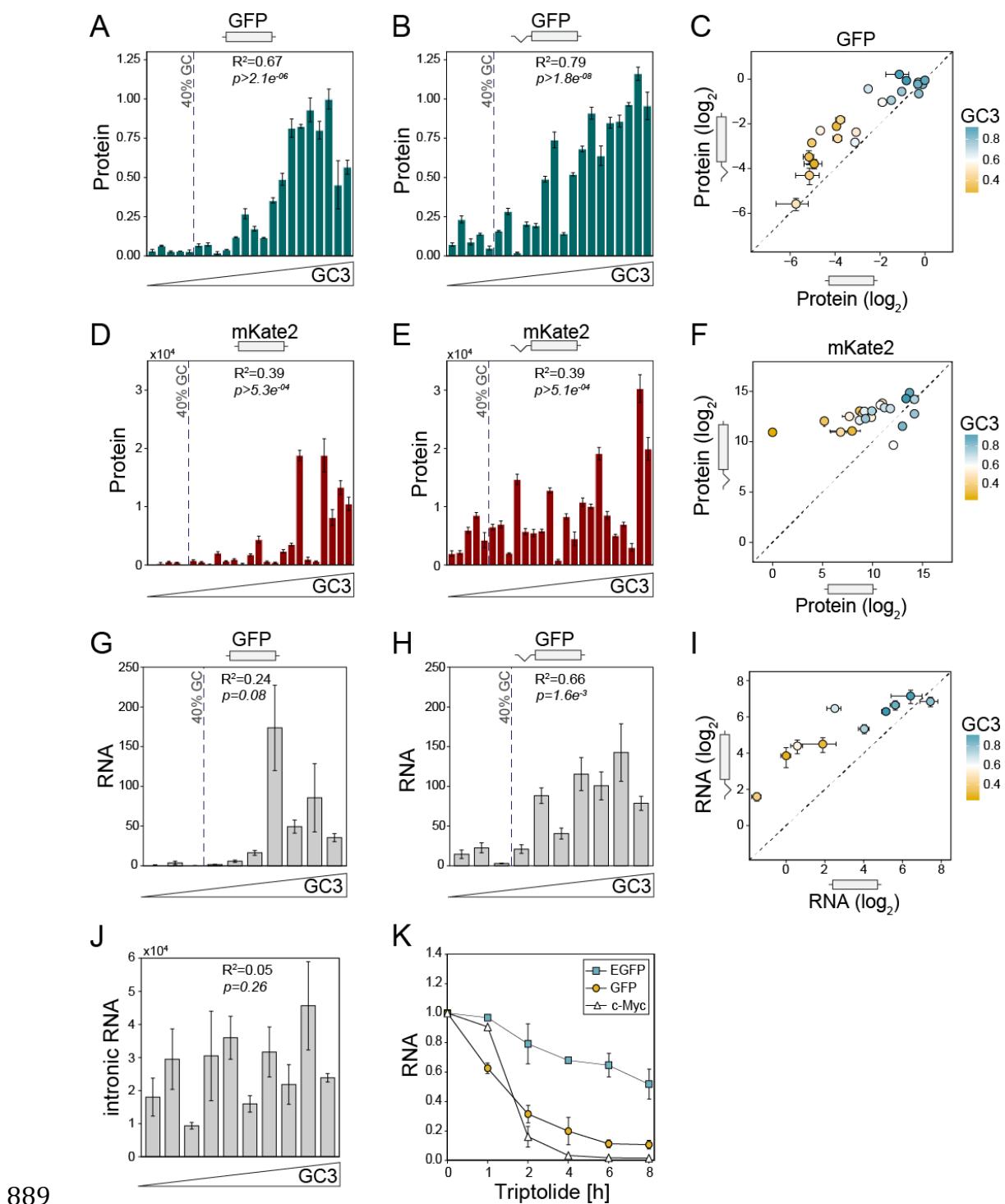
Figure 2. The effect of GC content on gene expression depends on splicing.

(A-B) Protein levels of 22 GFP variants when transiently expressed as unspliced (A) or spliced (B) constructs were expressed in HeLa cells and quantified by spectrofluorometry. Each data point represents the mean of 9 replicates, +/- SEM.

896   (C) Correlation of protein levels between unspliced and spliced variants of GFP

897   (n=22, R^2=0.69, p=9.0×10^{-7}). The dashed line indicates x=y.

898   (D-E) Protein levels of 23 mKate2 variants in the absence (A) or presence (B) of

899   splicing. Each data point represents the mean of 9 replicates.

900   (F) Correlation of protein levels between unspliced and spliced variants of

901   mKate2 (n=23, R^2=0.29, p=2.8×10^{-4}).

902   (G-H) mRNA levels of 10 GFP variants when transiently expressed as unspliced

903   (G) or spliced (H) constructs in HeLa cells and quantified by qRT-PCR. Data

904   points represent the mean of 3 replicates, +/- SEM.

905   (I) Comparison of mRNA expression from spliced and unspliced GFP variants.

906   (J) Intronic RNA levels of GFP variants measured by qRT-PCR.

907   (K) RNA stability time course of GC-rich (97% GC3; $t_{1/2}$=8.6h) and GC-poor (33%

908   GC3, $t_{1/2}$=2.4h) GFP variants in stably transfected HeLa Flp-In cells after blocking

909   transcription with 500nM Triptolide. Results represent the averages of 2

910   independent experiments, +/- SD.

911
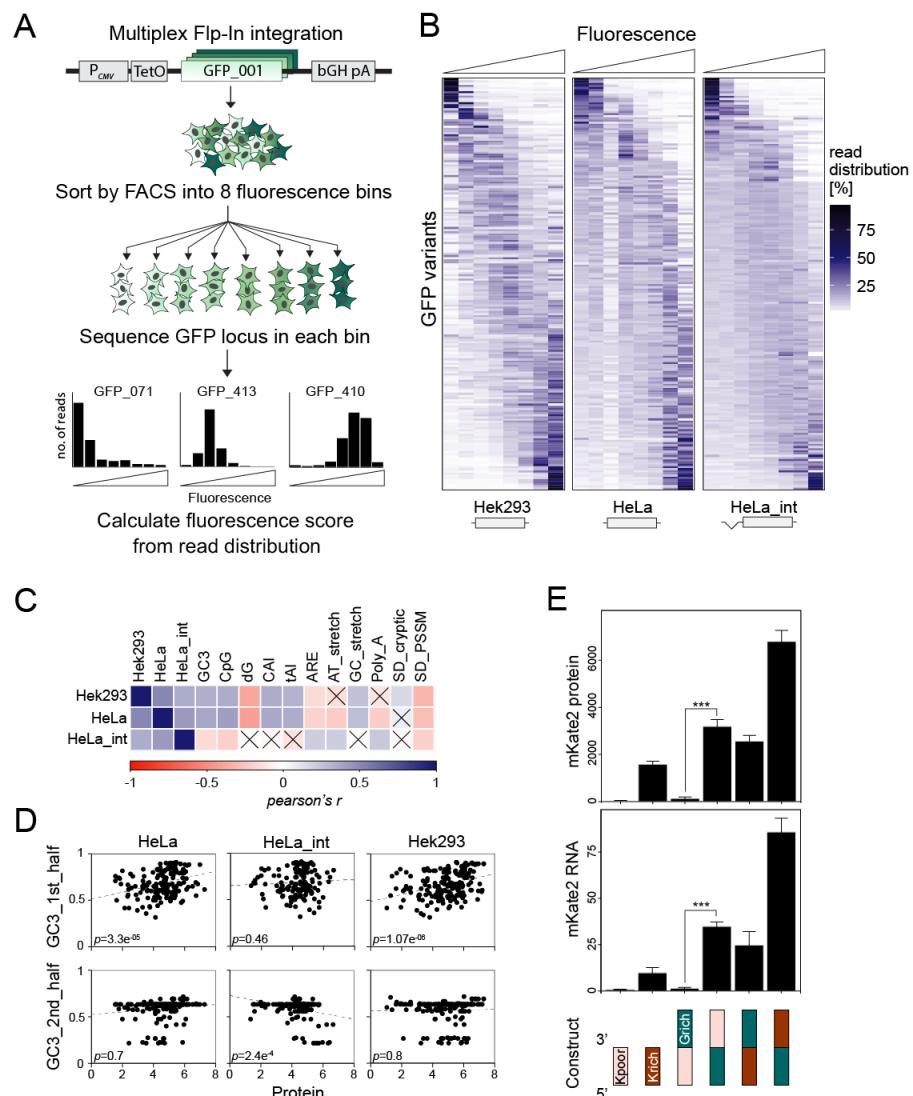
Figure 3. Splicing- and position-dependent effects of codon usage on protein production.

(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. 24h post-induction, cells are sorted by FACS into 8 fluorescence bins, genomic DNA extracted followed by high-throughput sequencing of the GFP locus. Individual fluorescence scores for each variant are calculated from normalised read distributions. (See Methods and 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 bin

33

926 equals about 100-fold. Data shown represents the average of 3 Flow-Seq

927 measurements for HeLa cells, the average of 3 Flow-Seq experiments for HeLa

928 with intron and 1 experiment for HEK293 cells.

929 (C) Pearson's correlation matrix of experimental measurements obtained by

930 Flow-Seq and sequence covariates. The colour of squares indicates the

931 correlation coefficient; crosses indicate non-significant correlations.

932 (D) Correlations between Flow-Seq measurements and GC3 content of 1st (nt 1-

933 360) and 2nd (nt 361 - 720) halves of GFP sequences.

934 (E) Protein and mRNA measurements of translational fusion constructs between

935 GC-poor (30% GC3, Kpoor) and GC-rich (85% GC3, Krich) variants of mKate2

936 with a GC-rich variant of GFP (97% GC3, Grich). Data represents the mean of 3

937 replicates + SEM (see also Figure S5).

938
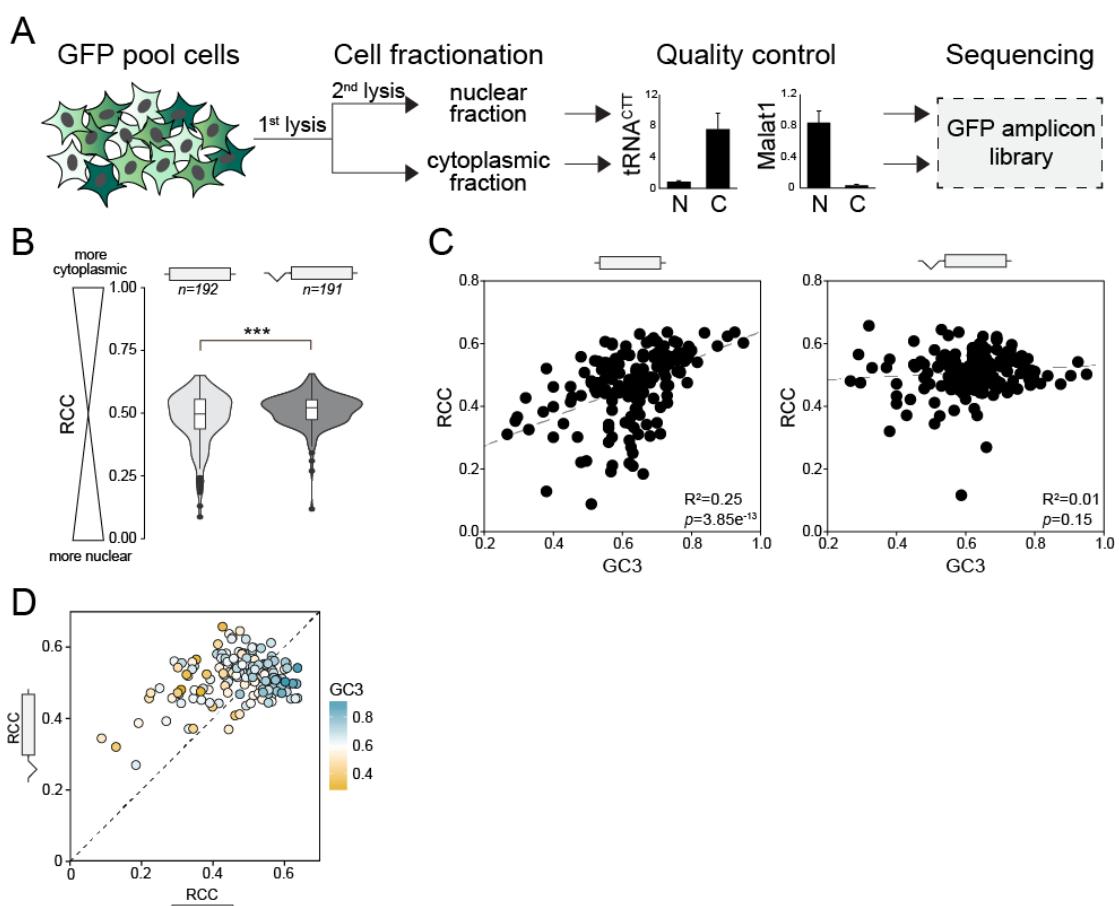
Figure 4. High GC content increases cytoplasmic localisation of mRNA.

(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 (see also Methods).

(B) Relative cytoplasmic concentration (RCC) of unspliced and spliced GFP variants. Data represents the mean of 2 replicates. ***$p=2\times10^{-6}$.

(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 ($R^2=0.1$, $p=2.6\times10^{-5}$).

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.

Figure 6. Splicing-dependent codon usage shapes global gene expression.
Plots showing the effect of GC4 content on the expression of unspliced (x-axis) and spliced (y-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.
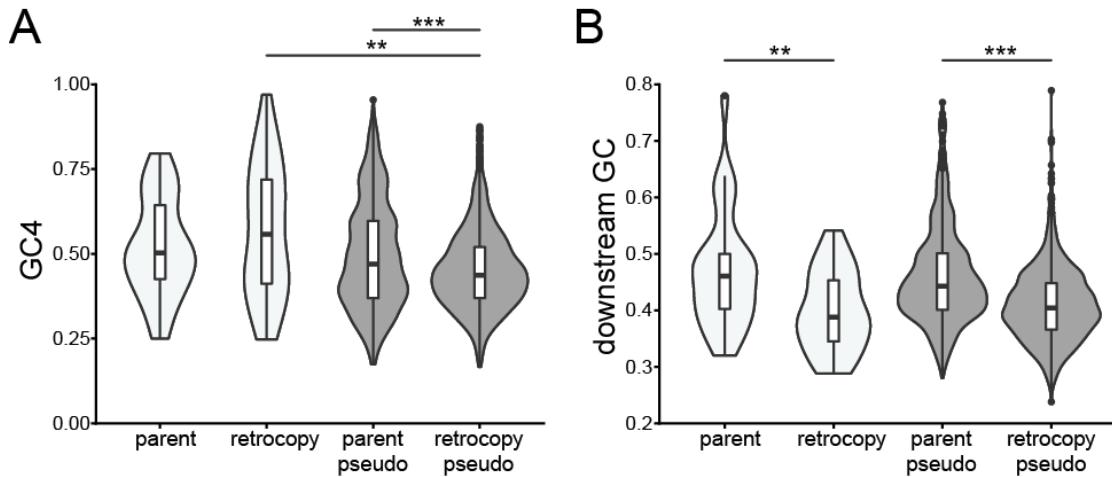
Supplementary Figure 1. GC4 variation amongst parent-retrogene pairs and their downstream sequence.

(A) 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). Note that a different retrogene dataset was used than in the main text (see Methods for details). For the human-macaque set, the difference in GC4 between parents and functional copies is in the expected direction but not significant.

(B) 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.

999

1000
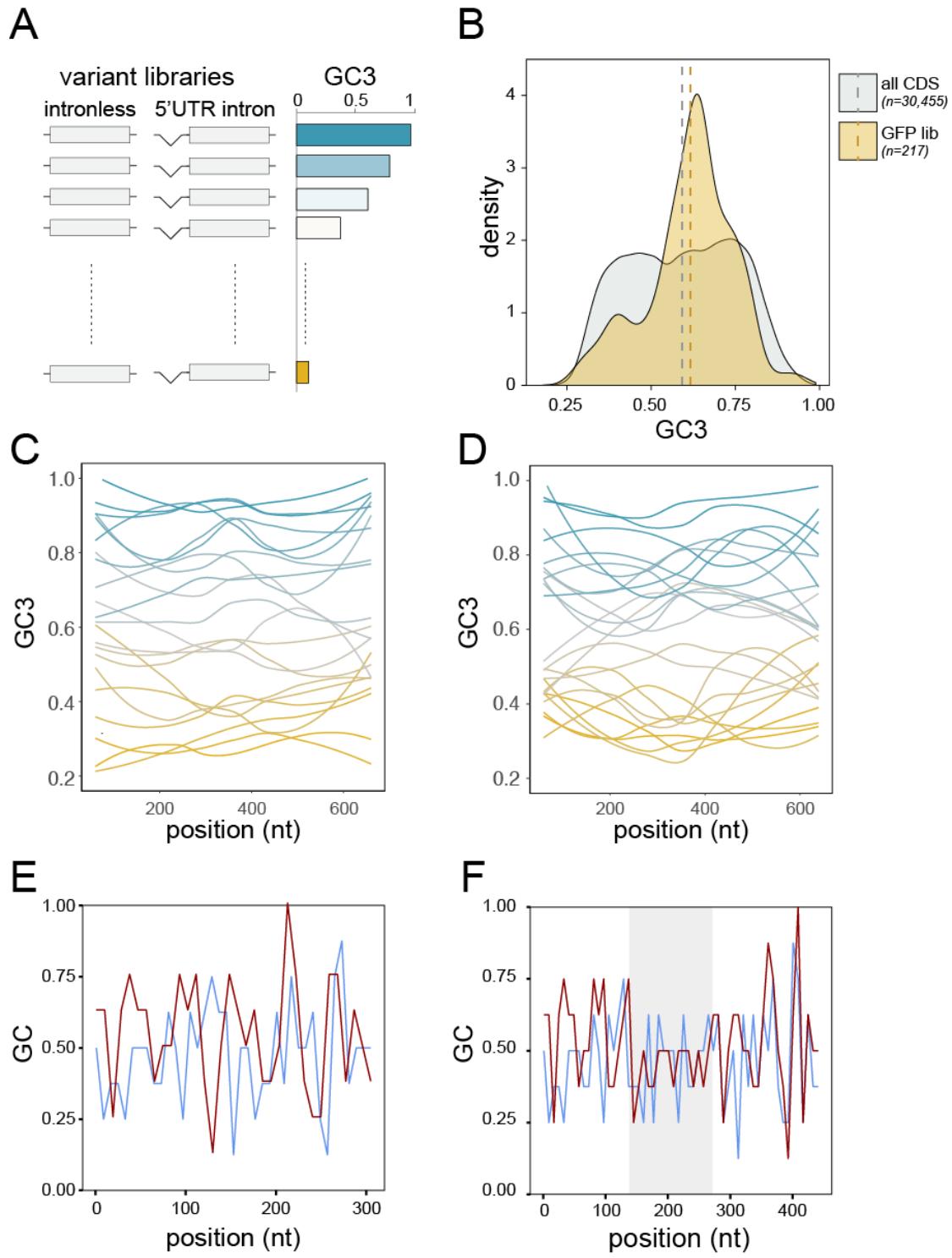
Supplementary Figure 2. GC content variation amongst endogenous genes and reporter libraries.

(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.

1006 (B) GC3 content distribution amongst human consensus coding sequences (CDS;
1007 grey) in comparison to the GFP variant library used in this study (GFP lib;
1008 orange). Dashed lines indicate the mean GC3 for each data set.

1009 (C-D) Loess-smoothed GC3 profiles along the 22 GFP variants (C) and 23 mKate
1010 variants (D) that were analysed by spectrofluorometry (Figure 2).

1011 (E) Sliding window analysis of GC content in 5'UTRs of intronless expression
1012 cassettes utilised in this study. Blue: pCM3 (transient transfection, no intron);
1013 red: pcDNA5/FRT/TO/DEST (stable transfection, no intron).

1014 (F) As above, intron-containing expression cassettes. Blue: pCM4 (transient
1015 transfection, with intron); red: pcDNA5/FRT/TO/DEST/INT (stable transfection,
1016 with intron). Grey shading indicates the position of the synthetic intron.

1017

Supplementary Figure 3. Effect of GC content on expression of fluorescent reporter genes in stably transfected cell lines.

Flow cytometry measurements of two GFP variants in stably transfected HEK293 Flp-in (A) and HeLa Flp-in (B). GC poor = 33% GC3; GC rich = 97% GC3; (-)ctrl = untransfected cells. Data shows representative results of at least 3 experiments.
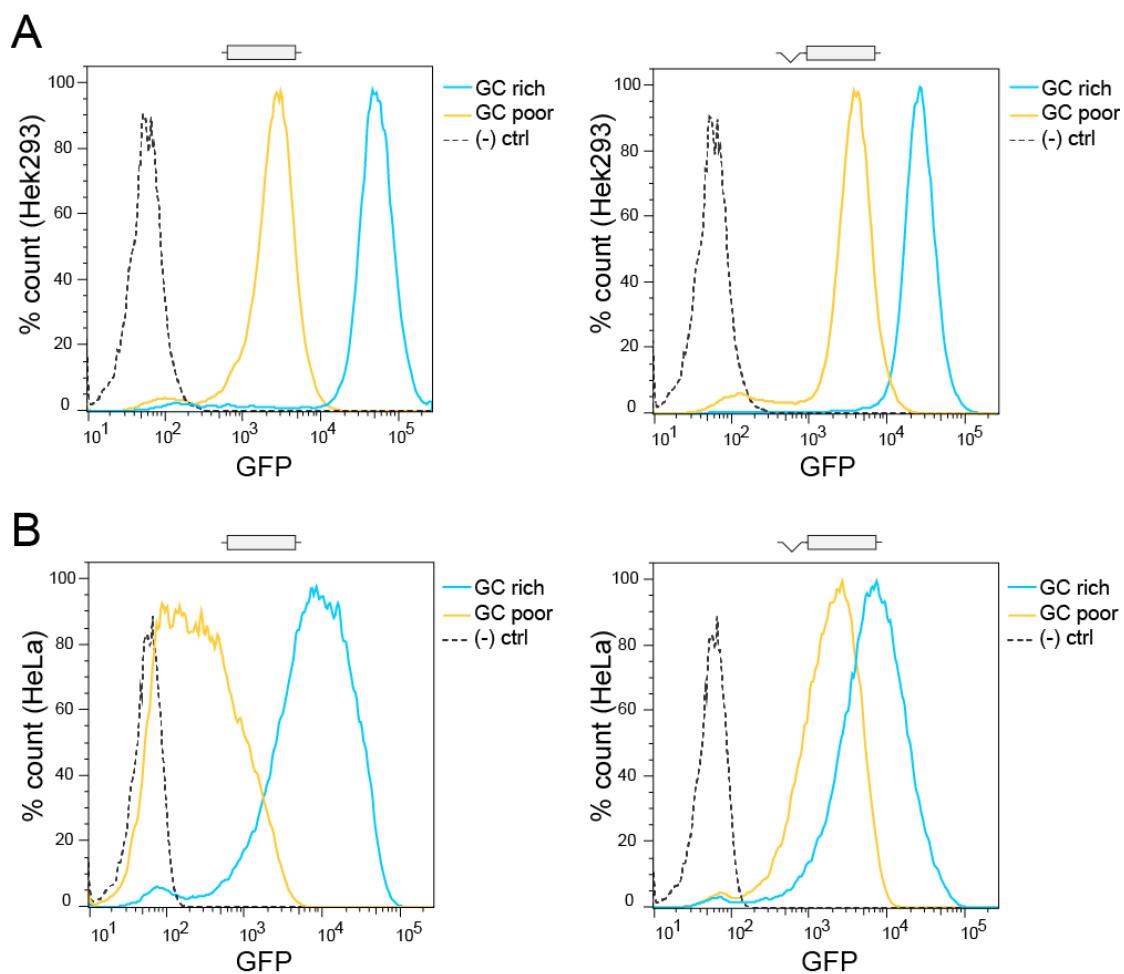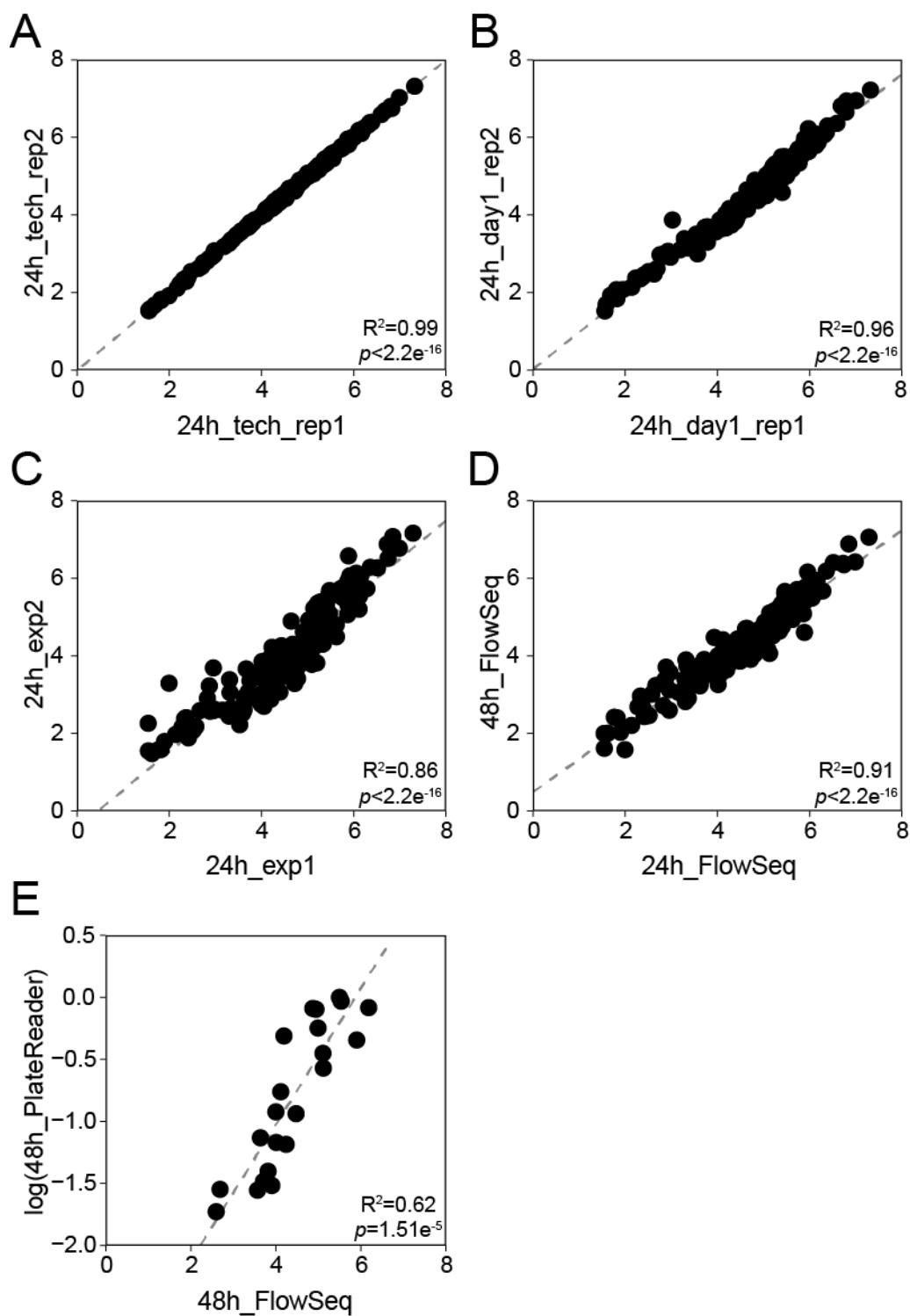
Supplementary Figure 4. Reproducibility of Flow-seq experiments in HeLa cells (unspliced GFP variants).

(A) Re-sequencing of the same amplicon-library.

(B-C) Replicate Flow-seq experiments performed on the same day (B) or different days (C).

1033   (D) Flow-Seq experiments performed on the same pool of cells, 24h and 48h

1034   after the induction of GFP expression.

1035   (E) Correlation between fluorescence measurements of 22 GFP variants obtained

1036   spectrofluorometry of transiently transfected HeLa cells and by Flow-Seq of

1037   HeLa GFP pool cell line.

1038
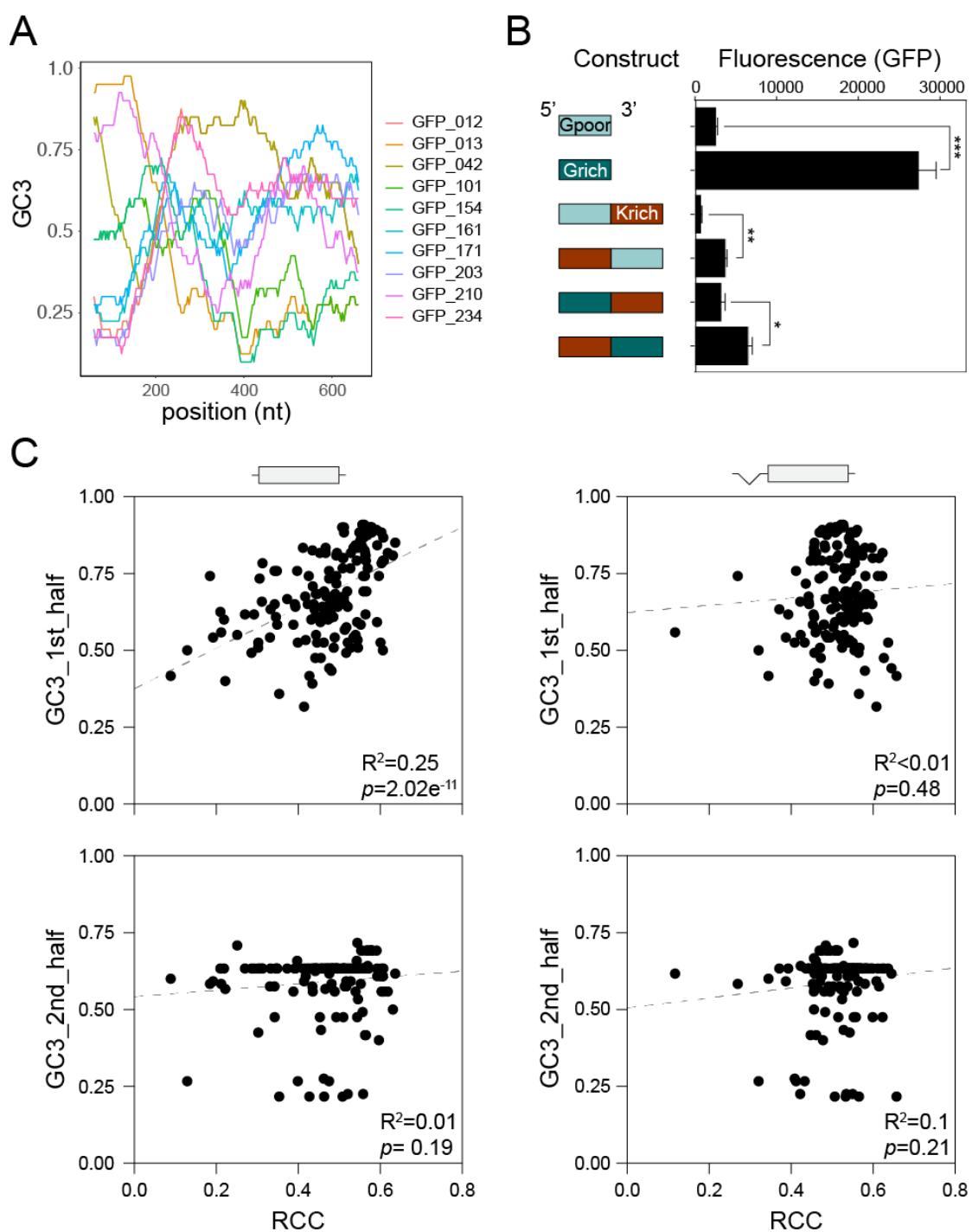
Supplementary Figure 5. Position-specific effects of GC content on expression.

(A) Sliding window analysis of GC3 content in selected GFP variants used in the pooled amplicon sequencing experiments.

(B) Protein measurements of translational fusion constructs between GC-poor (33% GC3, Gpoor) and GC-rich (97% GC3, Grich) variants of GFP with a GC-rich

44

1047   variant of mKate2 (85% GC3, Krich), upon transient transfection into HeLa cells.

1048   Data represent the mean of 3 replicates + SEM.

1049   (C) Correlations between the GC3 content in the 1st (nt 1-360) and 2nd (nt 361-

1050   720) halves of GFP variants and their relative cytoplasmic mRNA concentrations.

1051

1052

1053

Supplementary Figure 6. Distribution of RNA and protein expression data used in regression modelling.

Human RNA and protein expression data were extracted from various databases, filtered and normalized as described in Supplementary Table 1 and in the

46
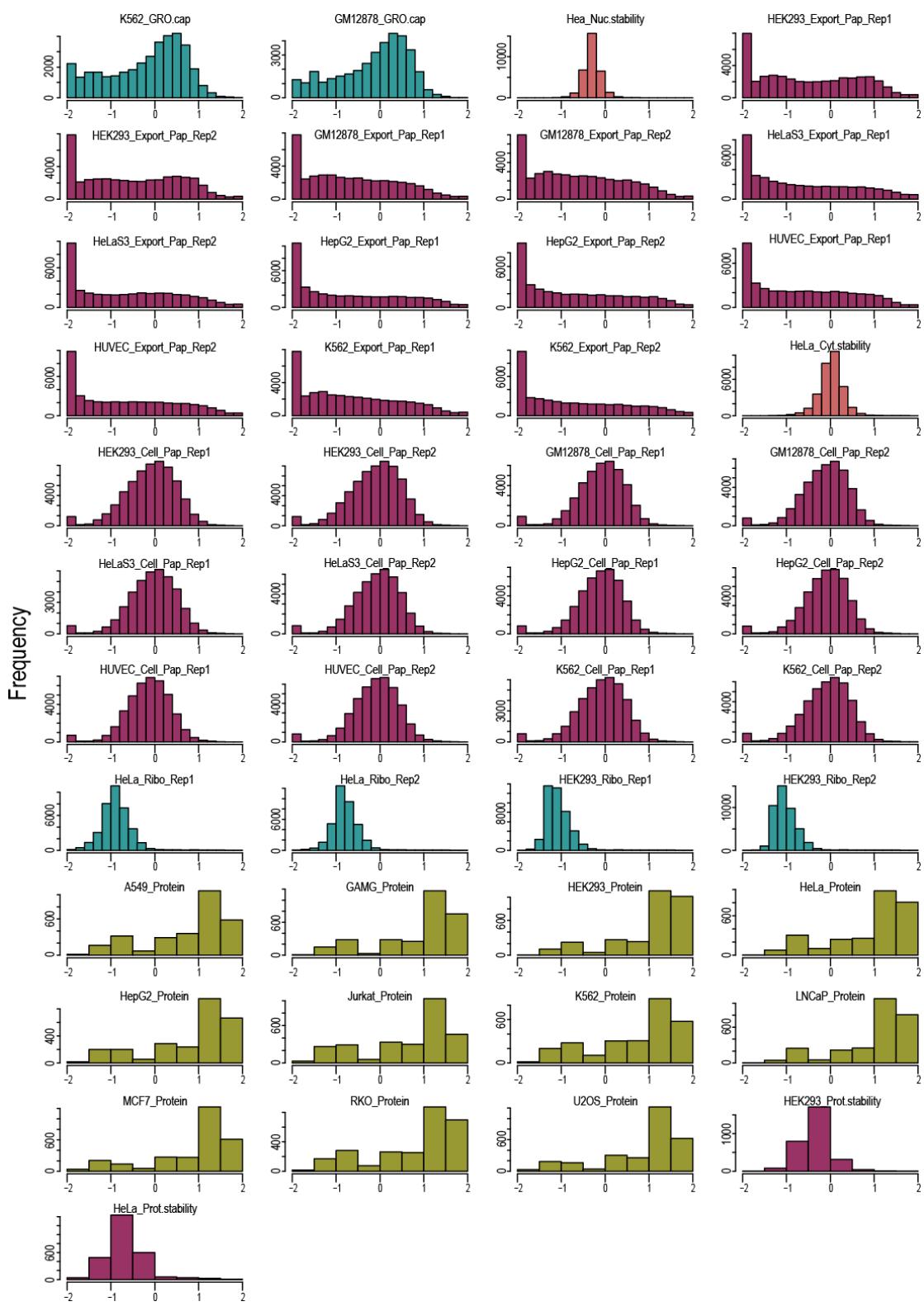
1058   Methods section. The histograms show the distributions of preprocessed

1059   expression measurements.

1060

1061 Supplementary Table 1. Sources of human gene expression data.

1062 The cellular process to be quantified is indicated above the table, and the

1063 experimental techniques and data sources are indicated below. Each dot

1064 indicates an experimental replicate measurement.

1065

	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

1066
1067

Supplementary Table 2. List of primers used.

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
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

1070

1071

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.

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

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.

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.

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.

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.