Coding-Sequence Determinants of Gene Expression in Human Cells

Christine Mordstein

THE UNIVERSITY *of* EDINBURGH

Doctor of Philosophy
The University of Edinburgh
2016

Declaration

This dissertation is the result of my own work unless otherwise stated. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma or other qualification.

Signed:

Date:

Acknowledgments

The work presented here would not have been possible without the huge support of a number of people.

First of all, I would like to thank my supervisor, Greg Kudla, for the great trust that I was given throughout the years. His guidance and encouragement was integral to the success of this project. I particularly appreciate all the opportunities that I was given, from attending courses to conferences, as well as the indispensable benefit of his innovative out-of-the-box thinking, which all together helped me develop and greatly shaped me as a scientist.

The immense support that I received from my colleagues through all the ups and (especially) downs is priceless and I could not have wished for anything more. I would like to specifically thank all members of the Kudla and Cáceres labs, past and present, for all the help and scientific discussions. I feel privileged to have worked within such an incredibly stimulating environment, in which daily exchange of experience helped me grow not only as a researcher, but also as a person. I am grateful to have had such immensely dedicated, enthusiastic and inspiring colleagues across the institute; in particular Jim Brindle, Fiona Haward, Nele Hug, Yatendra Kumar, Dasa Longman, Ewelina Macech, Roza Masalmeh, Magda Maslon, Pragya Mittal, Olga Puchta and Hans-Joachim Sonntag without whom everything would have been only half the fun and twice the pain. Thanks also to everyone else who shared their time with me in one way or another (TGIF!), my summer students Jeanne Bazile and Miriam Pedron for their hard work, and all other PhD students for making the IGMM a very special place.

I greatly appreciate the help from all collaborators involved in this research; the groups of Martin Taylor, Andrzej Dziembowski and Laurence Hurst, as well as GeneArt. I have also hugely benefitted from the excellent in-house technical services, especially Lizzie Freyer (FACS) and Agnes Gallagher (DNA sequencing/bioanalyzer).

I am thankful to Nick Hastie and Madan Babu for kindly agreeing to review this thesis, and to my thesis committee, Javier Cáceres, Richard Meehan and Sascha Kagansky, for critically assessing my progress.

Thanks to my parents and brother for their support over the years. I would particularly like to thank Christin Bruchhagen for always believing in me (wer hätte gedacht, dass ein Bewegungs-Günther es mal soweit bringen würde!).

Lastly, none of this would have been possible without the endless support of Samir Watson - words cannot describe how thankful I am.

Abstract

The human genome is highly heterogeneous in its GC composition. How codon usage affects translation rates has been extensively studied and exploited to increase protein expression. Although effects on virtually all other steps in gene expression have been reported as well, so far no systematic approach has been taken to quantitatively measure the contribution of each to overall protein levels in human cells. Here, I utilise a library of several hundred synonymous variants of the Green fluorescent protein (GFP) to characterise the influence of codon usage on gene expression in human cells.

In an initial small-scale screen, I show that protein levels are largely correlated with codon-usage and particularly GC-content. Additionally, I demonstrate that these changes can already be seen on the RNA level, confirming more broadly previously published data from our lab (Kudla et al., 2006). In order to assess the consequences of randomised codon usage on a larger scale, I established and validated a high-throughput approach for the phenotypic profiling of reporter genes. Using a pool of cells stably expressing >200 GFP variants, I measured multiple parameters simultaneously, such as protein levels, translational state, RNA levels, stability and export. Data from these experiments confirm a strong relationship between GC-content, protein levels, as well as RNA export, reproducibly in two cell lines. Low expression of especially GC-poor variants could not be rescued by splicing, but increased nuclear-to-cytoplasmic RNA ratio, suggesting further mechanisms important for efficient gene expression. These effects are even more pronounced when the distribution of GC is spread evenly along the coding sequence. Interestingly, our data also suggests that high GC within the first 200nt is more predictive of efficient gene expression, contrasting studies performed on bacteria, in which strong secondary folding near the ribosomal binding site was shown to be non-permissive for translation (Kudla et al., 2009).

By relating experimentally derived parameters to sequence features known to inhibit expression, I demonstrate that cryptic splicing is a major factor leading to decreased levels of particularly GC-poor GFP variants. An attempt to quantitatively assess the relative contribution of several sequence features (e.g. tAI, GC3, CpG) using multiple regression analysis lead to inconclusive results, leaving the requirement for the exploration of alternative approaches in order to dissect the role of individual parameters, as well as to identify novel determinants of gene expression.

Lay summary

How genetic information is encoded in our DNA has been studied extensively over the last few decades, especially since the genetic code was solved in the 1960s. The genetic code is written with just 4 letters – A, T, G and C. The cell reads the code in triplets, which means that 3 letters together encode one amino acid, the building block of every protein. As there are 64 possible triplets but only 20 different amino acids, several triplets can code for the same amino acid, making it possible to write protein sequences in multiple ways. For long it was assumed that this would have no effect on protein levels, however, it has recently become clear that the choice of triplets can have profound consequences for a cell.

In order for a gene to be made into a protein, the DNA code is first copied into a messenger molecule called RNA. RNA is then transported to ribosomes, the protein factories of cells, where the message gets translated into proteins. Besides the genetic code, which tells the cell what kind of protein to make, there is also a second code, the regulatory code, which defines how much of each protein should be made. In this PhD project, I tried to decipher this code to better understand the importance of being able to write the same piece of information in many different ways.

Here, I show that the triplet choice for a particular protein can have a strong effect on how much RNA and protein is produced. By encoding the same model protein in hundreds of different ways, it is also possible to look at the various steps that lead to the expression of a gene, e.g. from reading the code, to transporting the message and the actual protein production. All this information allows us to measure how the triplet code matters. Interestingly, I found that the more G and C are used in a gene, especially at the beginning of a message, the more protein will be made.

Studying this code further will help us understand diseases that are caused by proteins which are encoded in an alternative way, leading to either too much or too little protein being made. The knowledge gained from this study can thus be used to modulate protein levels in a controlled manner, simply by changing the triplet code.

List of abbreviations

Acronym	Meaning
3' UTR	3' untranslated regio
5' UTR	5' untranslated region
A	Adenine
A2UCOE	Ubiquitous Chromatin Opening Element derived from the human *HNRPA2B1-CBX3* locus
C	Cytosine
CBC	Cap-binding complex
cDNA	complementary DNA
CDS	coding DNA sequence
CFTR	Cystic fibrosis transmembrane conductance regulator
CRISPR	Clustered regularly interspaced short palindromic repeats
DMEM	Dulbecco's modified Eagle's Medium
DNA	deoxyribonucleic acid
DRD2	Human dopamine receptor D2
EJC	exon junction complex
FACS	Fluorescence activated cell sorting
FCS/FBS	Fetal Calf serum/Fetal Bovine serum
G	Guanine
gDNA	genomic DNA
GFP	Green fluorescent protein
GPCR	G protein-coupled receptor
HF	high-fidelity
IRGM	Immunity-related GTPase family M protein
miRISC	microRNA-induced silencing complex
miRNA	microRNA
N	Any'; either Adenosine, Thymine, Guanosine or Cytosine
NGD	No-Go decay
NMD	Nonsense-Mediated Decay
PBS	Phosphate buffered saline
PCI	Phenol-Chloroform-Isoamyl alcohol
PCR	Polymerase chain reaction
qRT-PCR	quantitative reverse transcription followed by polymerase chain reaction
RBP	RNA-binding protein
RCC	Relative cytoplasmic concentration of RNA
RIN	RNA integrity number
RNA	Ribonucleic acid
RPFs	Ribosome protected fragments
RT	Reverse transcription
RTase	Reverse transcriptase
RT-PCR	Reverse transcription followed by polymerase chain reaction
RTX	Reverse Transcriptase Xenopolymerase

sSNP	synonymous single nucleotide polymorphism
T	Thymine
TALEN	Transcription activator-like effector nuclease
TF	Transcription Factor
TGIRT	Thermostable group II intron reverse transcriptase
TSS	Transcription Start Site
ZNF	Zing finger nuclease

List of figures and tables

Table of Contents

1. Introduction

1.1. The genetic code

Soon after the discovery of the structure of DNA followed the identification of the 64 nucleotide triplets encoding 20 amino acids, the building blocks of proteins, as well as 3 stop or non-sense codons which correspond to the end of a polypeptide chain (Figure 1, Crick et al., 1961; Watson and Crick, 1953). Since there are more codons than naturally occurring amino acids, 18 of the 20 amino acids are encoded by up to 6 different codons. This redundancy of the genetic code is termed degenerate. In most cases, changes at the third position of a codon do not affect the encoded amino acid and are therefore often referred to as 'silent', whereas changes at the first and second position may lead to a change in the peptide sequence. The genetic code is almost universal which means that a given gene is most likely translated into the same protein in any species. However, the frequency of use of particular codons might vary strongly between species. With increasing data from comparative analyses of codon usage in different organisms, it has become clear that synonymous codons are not randomly distributed along genomes and genes, and that certain organisms seem to prefer some over others (Grantham et al., 1980; Plotkin and Kudla, 2011; Sharp and Li, 1986).

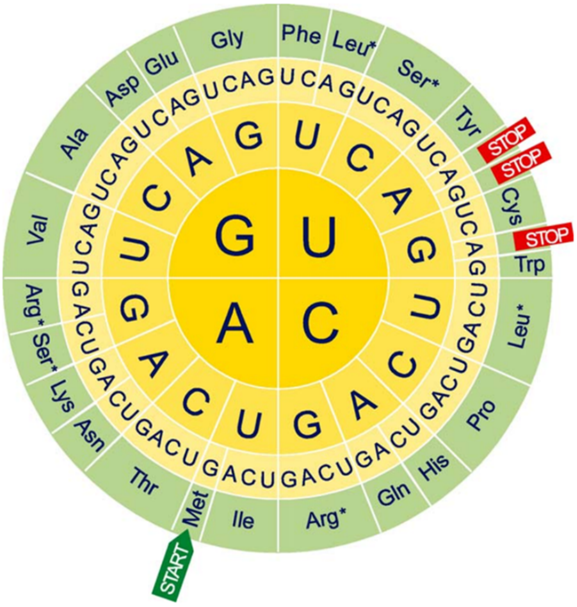


Figure 1. The genetic code in form of the codon sun (from Nierhaus, 2006).

This phenomenon is termed codon usage bias and can be very strong in different species, with some avoiding certain codons almost completely (Plotkin and Kudla, 2011; Sharp and Li, 1986, Figure 2). Even though in most cases, base changes at the third position of a codon do not alter the sequence of amino acids in a protein, it has in recent years been recognised that synonymous changes are not as silent as once assumed. Newly emerging evidence indicates their importance in several steps in gene expression regulation with various studies implicating synonymous mutations in disease (reviewed in Sauna and Kimchi-Sarfaty, 2011).

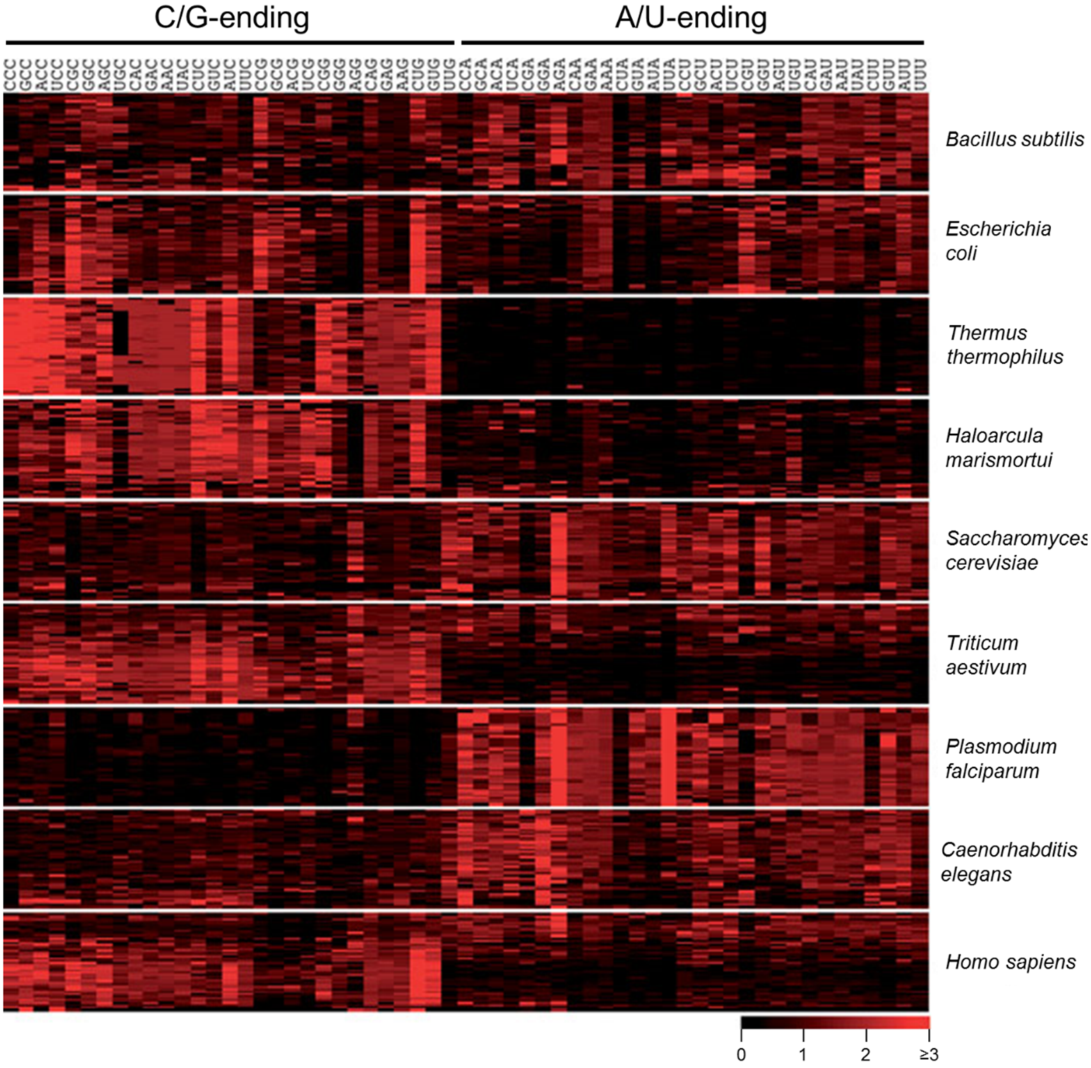


Figure 2. Codon usage bias within and between genomes.
The relative synonymous codon usage (RSCU) is plotted for 50 randomly selected genes from each of 5 species. RSCU ranges from 0 (codon absent), through 1 (no bias) to 6 (a single codon is used in a six-codon family). Genes are in rows and codons in columns. (Adapted from Plotkin and Kudla, 2011)

The underlying causes of codon usage bias have been attributed to two main mechanisms: mutational drift and natural selection (Plotkin and Kudla, 2011). Mutational variations are neutral as differences in codon usage arise from underlying mutational processes, such as biases in DNA repair (Galtier et al., 2001) or biases in nucleotides produced by point mutations (Kimura, 1980), without any effects on the organism's fitness. In contrast, processes involving natural selection postulate effects on fitness through synonymous codon changes, causing such to either be promoted or repressed within a population. In organisms with large population sizes and short generation times, e.g. bacteria and yeast, small deleterious mutations can therefore be acted on by natural selection, whereas in mammals, population sizes are much smaller and synonymous mutations are therefore assumed to be neutral (Kreitman, 1996). It is however likely that both mechanisms play a role in shaping codon usage bias between as well as within genomes (Chamary and Hurst, 2005). Across species, genomic GC content has been found to be a major determinant of codon usage (Chen et al., 2004). These are caused by mutational mechanisms affecting the whole genome. However in some species, such as mammals, mutation rates might vary depending on the sequence context, e.g. the hypermutability of CpG dinucleotides, in which case the often methylated cytosine is frequently mutated to a thymine (Bird, 1980).

Within genomes, the most studied factor affecting codon usage is the selection for efficient gene expression. There is a positive relationship between high codon bias and high expression levels in prokaryotes as well as eukaryotes which cannot be explained by mutational biases alone (Eyre-Walker, 1991; Sharp and Li, 1987). A more selectionist explanation for such a correlation is the adaptation of codon usage to tRNA abundance to deliver more efficient and/or accurate protein synthesis (Akashi, 1994). Although it is unclear how this is applicable to higher eukaryotes, since codon usage of genes expressed in different tissues or developmental stages may vary as well as the relative tRNA levels (Dittmar et al., 2006), which are technically challenging to measure accurately, and their expression might not correlate well with gene copy number. In mammals and in particular humans, there is still much controversy whether coupling of codon usage with tRNA abundance is an active mechanism shaping gene expression patterns of particular sets of genes, or whether codon usage of such is mainly driven by underlying sequence features such as GC content (Gingold et al., 2014; Plotkin et al., 2004; Rudolph et al., 2016).

Although codon bias has been primarily studied in the context of translation, it has in recent years become more evident that the effects of synonymous changes may already be observed at the RNA level. However, due to a lack of understanding of the underlying mechanisms, it is difficult to predict the effect of synonymous substitutions on mRNA processing, stability or translation. The aim of this thesis is therefore to systematically and quantitatively characterise the influence of coding-sequence changes on gene expression in order to investigate sequence properties associated with efficient gene expression in human cells.

1.2. Sequence composition of the human genome

Long before genome sequencing allowed the visual interpretation of the compositional heterogeneity, experimental approaches using e.g. CsCl ultracentrifugation already showed a relatively high resolution picture of the base compositions of mammalian DNAs (Corneo et al., 1968). These gradient approaches revealed that genomes of human cells contained large regions (>300kb) of high GC content uniformity, termed isochores, which are absent in genomes of lower organisms (Bernardi, 1993). With advances in whole genome sequencing and the eventual release of the finished sequence of the human genome by the International Human Genome Sequencing Consortium, GC heterogeneity could, for the first time, be directly visualised along all chromosomes (Lander et al., 2001). One of the most striking findings is that even though the genomic GC content averages at around 38%, genes cover a surprisingly broad range of GC content with a significantly higher average at around 46% (Lander et al., 2001). When looking at the relative gene density within the different families of isochores, it also became clear that gene distribution is very non-uniform in the genome and divided into two genome spaces: the genome core, composed of the two GC-richest isochore families H2 and H3, which comprise more than half of the genes although only representing about 15% of the total genome, and the genome desert, represented by the GC-poor isochore families L1, L2 and H1 (Bernardi, 2012). Both spaces were associated with different basic characteristics, such as the correlation between isochores families with recombination, replication timing, location and in particular chromatin structure in interphase nuclei; the chromatin of the genome core is “open”, but rather “closed” in the genome desert (Saccone et al., 2002). A study by Constantini and Bernardi investigated the frequencies of di- and tri-nucleotides and revealed large differences among all five isochore families which were found to be responsible for many of the basic properties of the human genome, such as differences in codon usage (Costantini and Bernardi, 2008). GC level was found to explain ~50% of variation in nucleosome occupancy *in vitro* which, considering wide-spread variation in nucleotide frequencies across different parts of genomes in different species, suggested the direct influence on chromatin structure (Tillo and Hughes, 2009). Later it was found that trinucleotide patterns were non-randomly distributed within the genome and strongly influenced nucleosome positioning (Arhondakis et al., 2011). Furthermore, a preference for certain regulatory sequences such as transcriptional start sites were found between isochore families L1 and H3. Whereas the regulatory sequences in L1 predominantly belong in a "TATA-box" model, in H3 they fit rather a "GC-rich" model. Consequently, the transcription factors bound by either GC-poor or GC-rich isochores are very different, which indicates that genes in different isochore families may be functionally different. Indeed, a few decades earlier, it was proposed that GC-rich isochores were richer in housekeeping genes and GC-poor isochores richer in tissue-specific genes (Mouchiroud et al., 1987, 1991). This was later confirmed by the finding that housekeeping genes are on average GC-richer (Vinogradov, 2003). The nucleotide composition of an isochore is therefore the best predictor for the nucleotide content at the synonymous site and hence, codon usage bias across genes.

1.3. Mechanisms by which codon usage affects gene expression

Eukaryotic messenger RNA does not exist by itself but rather in large complexes consisting of multiple protein factors and small or long non-coding RNAs, together forming large messenger ribonucleotide particles (mRNPs). The combination of molecules ultimately decide on the fate of each transcript by influencing virtually every step in gene expression. Some of these mechanisms, starting from mRNA transcription, are described in more detail below.

1.3.1. Transcription

The regulation of gene expression by modulation of the transcription process has long been recognised, however, how codon usage within coding regions may influence protein levels has only recently started to emerge. It was recently shown that about 14% of codons within 86% of all human coding genes contain transcription factor (TF) binding sites, as determined by DNase I footprinting experiments across 81 different cell types (Stergachis et al., 2013). This was thought to provide evidence for a TF 'regulatory' code overlapping the genetic code, suggesting that codon choice is not only constraint by protein structure and function, but also by TF binding. However, it should be noted that this study did not address the question of how TF binding within exons functionally shapes gene expression and how this could mechanistically be achieved (Weatheritt and Babu, 2013). It was further suggested that there may be a synonymous codon bias, with TF-bound regions being more enriched in G/C-ending codons, ultimately linking such sequence constraints to protein evolution and fitness (Stergachis et al., 2013). These results and the implied assumptions however, remain a topic of controversy (Agoglia and Fraser, 2016; Xing and He, 2015).

Intragenic CpG content has recently been implicated in regulating transcription rate as shown in run-on assays comparing CpG-enriched and CpG-depleted synonymous gene variants of viral and cytokine reporter genes and correlated with long-term expression in stable human cell lines (Bauer et al., 2010). Later it is was found that nucleosome positioning differed between these variants *in vitro* which may also be the cause for differences in chromatin accessibility measured *in vivo* likely being the cause for differences measured in RNA polymerase II elongation rates (Krinner et al., 2014). Furthermore, high CpG near the 5'end of reporter genes was shown to correlate with gene expression, and high CpG downstream of the transcription start site was suggested to be a general feature of highly expressed genes. The number of variants utilised in these studies were, however, very low (4 per reporter gene) and therefore too low to distinguish between effects caused by high CpG or high GC-content. In case of genome-wide data presented within the same study, again no clear distinction was made between GC, GC3 and CpG content, nor between transcription start site and start codon, which may introduce a systematic skew due to fundamental differences in different genomic backgrounds (e.g. UTR sequences tend to be shorter in GC-rich isochores). Therefore the generalisability of these findings are unclear and further investigation is required.

1.3.2. mRNA folding and stability

Although most synonymous changes at the third codon position have no effect on amino acid sequence, in respect to mRNA folding energy and hence mRNA secondary structure, they are not all equivalent. The (G+C)/(A+U) ratio together with the availability of Watson-Crick base pairing are the two major factors determining RNA folding energies. The GC-content strongly affects the total folding energy (FE) of a mRNA molecule, resulting from the fact that there are 3 H-bonds between G and C but only two between A and U. Computational analysis of mRNA sequence composition and total folding energy suggest an up to 4-fold change in folding energy by changing only the wobble bases (Biro, 2008). Since synonymous codons are not used in equal frequencies and do not occur randomly (see 1.1), this suggests a regulatory role of wobble bases in determining mRNA folding energy and thus mRNA secondary structure. Indeed, sequence elements rich in adenosine and uridine, called AU-rich elements (AREs), although unlikely to exhibit strong secondary structures, are known to affect mRNA stability in mammalian cells (Fan et al., 1997). Such sequence motifs were originally discovered in the 3'UTR region of mRNAs and were found to cause rapid degradation through deadenylation (Chen and Shyu, 1995) and are primarily found in genes which require very tight spatial and temporal regulation, e.g in cell proliferation or as a response to environmental stimuli (Barreau et al., 2005). AU-rich elements have been estimated to influence the expression pattern of up to 8% of human genes (Bakheet et al., 2001).

The folding structure of an mRNA molecule influences the interaction with regulatory molecules through *cis*-regulatory elements modulating the transcripts stability. Several studies on synonymous single nucleotide polymorphisms (sSNPs) within coding-sequences highlight the importance of epistatic interactions between nucleotides and the dramatic consequences of synonymous site changes on mRNA stability and its link to disease. Duan et al. found that several human G protein-coupled receptor (GPCR) genes diverted in their GC3 content significantly from their genomic non-coding GC background, urging further investigation into possible reasons leading to selection at these positions (Duan et al., 2003). A particular sSNP (957T) in human dopamine receptor D2 (DRD2), one of the proteins identified using this approach, causes a dramatic change in the mRNA folding structure, affecting transcript stability as well as translation rate. Interestingly, it could also be shown that the co-occurrence of another common disease-associated sSNP (1101A) leads to a secondary structure which closely resembles the wild-type structure and thus compensates for the otherwise detrimental effects of a single sSNP (positive epistasis). More recently, a genome-wide analysis on seven human lymphoblastoid cell lines measured the transcript-wide RNA stability by using 4sU-labelling of nascent transcripts. By calculating the ratio of nascent to total RNA (Duan et al., 2013), a positive correlation of RNA half-life with coding GC (r=0.141, p=9e^{-10}) as well as coding GC3 (r=0.224, p=2.8e^{-12}) was found, suggesting a more global role of codon usage in determining RNA stability. Kudla et al. directly measured mRNA stability between a very GC-poor and GC-rich coding-sequence variants of the Green Fluorescent Protein (GFP; GC3=0.35 and 0.96), both encoding the same final protein product and placed in the same genomic context (Kudla et al., 2006). No striking changes in mRNA half-lives were observed after a transcriptional block using Actinomycin D. However, Actinomycin D preferentially intercalates into GC-dinucleotide regions of genes, which causes this standard assay to be sub-optimal for the comparison of genes with extreme differences in GC composition, leaving the requirement for further investigation using an unbiased approach.

Structural changes caused by sSNP within the coding region have also been shown to interfere with the efficiency of microRNA (miRNA) binding. miRNAs are short regulatory RNAs which function through the formation of a miRNA-induced silencing complex (miRISC) with Argonaute proteins, which is targeted to complementary miRNA binding sites on transcripts, leading to their decay (Bartel, 2009). Target sites of miRNA are predominantly reported in

3'UTR sequences, although have also been reported in coding regions of mammalian mRNAs but with reduced inhibitory functions (Forman and Coller, 2010). Since target inhibition is likely due to interference with RNA structure and/or the translation machinery, it could be argued that miRNA binding sites in coding-sequences were evolutionarily selected for due to lower substitution rates (Hurst, 2006). The biological function of coding-sequences targeted by miRNAs is supported by multiple examples, one of which was identified in a GWAS analysis. The immunity-related GTPase family M protein (*IRGM)* gene is an interferon inducible GTPase that, if not regulated properly, causes susceptibility to Crohn's disease. A sSNP was found within the seed region of the miR-196 binding site which was shown to lead to insufficient miRNA binding and thus ineffective downregulation of *IRGM*, leading to a Crohn's disease associated inflammatory response within the intestinal epithelia (Brest et al., 2011).

In contrast to RNA degradation pathways mediated by the binding of specific RNA-binding proteins, the ribosome recognises some features on the mRNA which ultimately trigger transcript degradation. It was recently shown in yeast that the optimality of codons is a strong determinant of mRNA stability (Presnyak et al., 2015). It was suggested that slow-decoding codons lead to the recruitment of the CCR4-NOT deadenylase complex as well as the decapping enzyme Dcp2, independent of other ribosome-dependent degradation pathways, such as nonsense-mediated decay (NMD) or no-go decay (NGD). The same group recently established Dhh1p (DDX6) as a sensor for codon optimality (Radhakrishnan et al., 2016). Dhh1p was shown to accumulate on transcripts when ribosomes progress slowly on stretches of non-optimal codons, leading to the recruitment of CCR4-NOT and ultimately faster mRNA decay (Radhakrishnan et al., 2016). A further study showed the generality of such pathways across eukaryotes by demonstrating that uncommon codons as well as 3'UTR length determine the mRNA stability of maternal transcripts in zebrafish (Mishima and Tomari, 2016).

1.3.3. Pre-mRNA splicing

Mammalian genes consist of coding (exonic) and non-coding (intronic) regions. In order for mRNA to mature into a functional message, multiple mechanisms are in place which combined are required to assure the correct removal of introns. The nucleotide sequence of pre-mRNA determines the affinity as well as the recognition of spliceosomal factors and thus, any nucleotide changes might affect the final mRNA product.

sSNP can create cryptic splice-sites caused by dinucleotides falsely being recognised as intronic ends (Eskesen et al., 2004). Other splicing-control elements have more recently gained higher importance as well through their effect on the recruitment of spliceosomal proteins. Such sequences include Exonic splice enhancers (ESEs) as well as silencers (ESSs) and have been shown to be required for correct exon selection (Fairbrother et al., 2002; Wang et al., 2004). For example Pagani et al. showed that multiple synonymous mutations in exon 12 of the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene disrupt a composite exonic regulatory element of splicing (CERES) leading to exon skipping. The same group later demonstrated that about 25% of synonymous changes within exon 12 result in similar exon skipping phenotypes (Pagani et al., 2005). To better understand the sequence determinants of alternative splicing, Rosenberg et al. created large synthetic mini-gene libraries with degenerate, alternative 5' or 3' splice sites and quantitatively measured the isoform ratio (Rosenberg et al., 2015). The data was used to train a predictive model which was shown to be highly predictive of naturally occurring variants, such as in case of the *CFTR* gene (60% of the effects of sSNPs could be explained). A later study used a similar high-throughput approach but focussed on the effects of all single and double mutants of an entire exon and demonstrated that >90% of base changes altered the exon inclusion ratio (Julien et al., 2016) but also argued that splicing regulatory features are not organised into defined motifs but dispersed along the entire sequence. The importance of sequence composition over larger sequence stretches (as opposed to short *k*-mers in the case of more discreet regulatory motifs) might therefore be directly connected to the large genomic GC variation.

Several studies on the human genome in regards to its GC content heterogeneity revealed that gene structures differ significantly between regions of high GC and low GC content (Amit et al., 2012). It was found that regions of high GC contain relatively short introns, as is the case with e.g. human housekeeping genes. Mechanisms are therefore required to ensure correct splice site recognition while still remaining flexible to accommodate the varying sequence features between regions of differential GC content. Two different recognition models were proposed: the exon definition and intron definition (Figure 3). Both models primarily rely on the differences in exon/intron lengths in different genomic regions. As introns in GC-poor regions tend to be long (>250bp), it was proposed that exon-flanking introns act as "flags" to aid the splicing machinery in recognising the location of exons as the splicing machinery is unable to detect large introns directly (Robberson et al., 1990). On the other hand, in lower eukaryotes, introns tend to be shorter (<250bp) and are thus more likely to be recognised directly by the splicing machinery. A study by Amit et al. systematically studied the differences between exonic and intronic GC content in GC-poor and GC-rich regions (Amit et al., 2012). It was shown that long introns contained a markedly higher GC content than their flanking exons, whereas short introns do not differ compared to their surrounding exons. This can be seen across all higher eukaryotes with similar genome composition to humans, however not in lower organisms which also do not show differences in intron lengths. It was also demonstrated how splice-site mutations leading to disease-associated exon skipping or intron retention correlate strongly to GC-content. Since it was known that nucleosome positioning tends to be more precise in exons compared to flanking introns due to their preference for GC-rich sequences (Tillo and Hughes, 2009), it was additionally shown that nucleosome occupancy is much higher in exons of low GC regions compared to those in high GC regions, further confirming a link between chromatin structure and splicing (Schones et al., 2008; Schor et al., 2009).

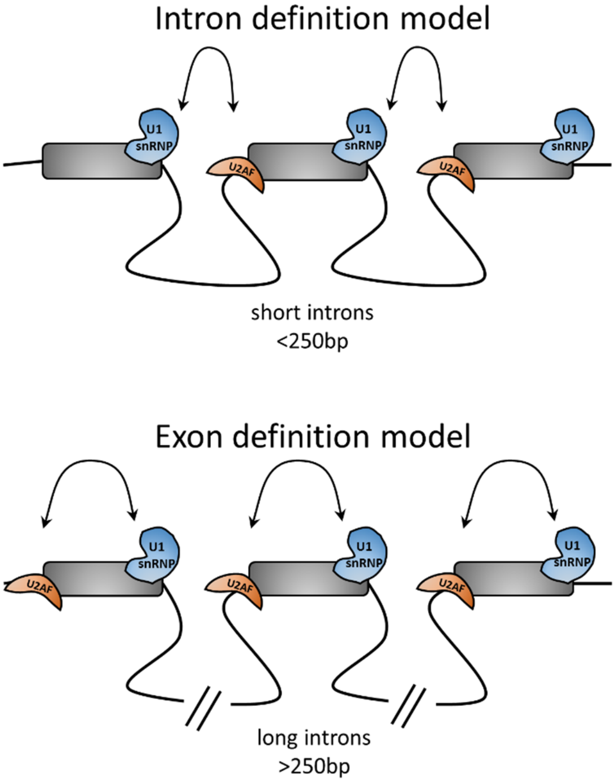


Figure 3. Intron and Exon definition models of pre-mRNA splicing.
The top panel illustrates the intron definition model when introns are short (<250bp) enough to be directly recognised by the splicing machinery. The bottom panel shows how splicing factors communicate across exons when introns are long (>250bp). (Adapted from De Conti et al., 2013).

1.3.4. Nucleo-cytoplasmic mRNA export

A fundamental step in the expression of mature RNA molecules is the export from the nucleus into the cytoplasm. All RNA species produced in the nucleus travel through the nuclear pore complex via export factors. This is a highly regulated process which relies on the assembly of large ribonucleoprotein (RNP) particles. Transcription and export are linked through the co-transcriptional recruitment of the THO/TREX complex along with the adaptor protein Aly/REF. Spliced mRNA associates with the major export receptor NXF1 (TAP) and its heterodimeric partner NXT1 (p15) (Müller-McNicoll and Neugebauer, 2013), whereas the export of naturally intronless mRNAs has been shown to occur through the TREX/NXF1 pathway in a transcription-independent, but polyadenylation-dependent manner (Lei et al., 2011). While this pathway is utilised by most mRNAs, a subset of endogenous mRNAs that encode proteins important for cell proliferation and survival, use chromosome region maintenance 1 protein homologue (CRM1), which is the main export factor for proteins (Culjkovic-Kraljacic and Borden, 2013). CRM1 cannot directly interact with RNA, but with adaptor proteins or other RBPs, such as Hu-antigen R, which binds to AU-rich elements (Brennan et al., 2000). Similar mechanisms are also required in the case of intronless viral RNAs which additionally rely on the interaction with either specific cellular or viral proteins to aid the recruitment to the cellular export machinery.

The replication cycle of Human immunodeficiency virus type I (HIV-1) is divided into two phases, early and late. Late genes require the early viral adaptor protein Rev for efficient expression (Figure 4). Rev recognises an RNA-element, named Rev-responsive element (RRE), on all unspliced or singly spliced viral transcripts (Fischer et al., 1994; Malim et al., 1989). Rev contains both a nuclear localisation signal which can be recognised by importins (Pollard and Malim, 1998) as well as a nuclear export signal which can be recognised by the export receptor CRM1 (Fischer et al., 1995; Ossareh-Nazari et al., 1997), allowing Rev to shuttle through the nuclear pore complex between nucleus and cytoplasm. Singly spliced viral genes should therefore theoretically be able to utilise both NXF1 and CRM1-dependent RNA export pathways. A study by Taniguchi et al. however showed that Rev bound to viral RNA actively suppresses the NXF1-dependent pathway through both the accumulation of Rev along the transcript and the interaction with the Cap-binding complex (CBC) which inhibits the association of Aly/REF to the 5’ terminus of the transcript (Taniguchi et al., 2014). Viral genes, especially of the lentivirus family, are remarkably rich in adenine and relatively poor in cytosine (van der Kuyl, 2012). Inhibiting CRM1 function with the drug leptomycin B was shown to have a negative effect on cytoplasmic levels of rev-dependent transcripts (Graf et al., 2000), suggesting CRM1-dependent nuclear export. It was later demonstrated that increasing the GC-content of these viral genes circumvents the requirement for Rev (Kotsopoulou et al., 2000). Similar observations have been made with the HPV virus for which it was shown that it is possible to efficiently express AU-rich viral genes in the cytoplasm using vaccinia virus T7 RNA polymerase expression system (Sokolowski et al., 1998; Tan et al., 1995).

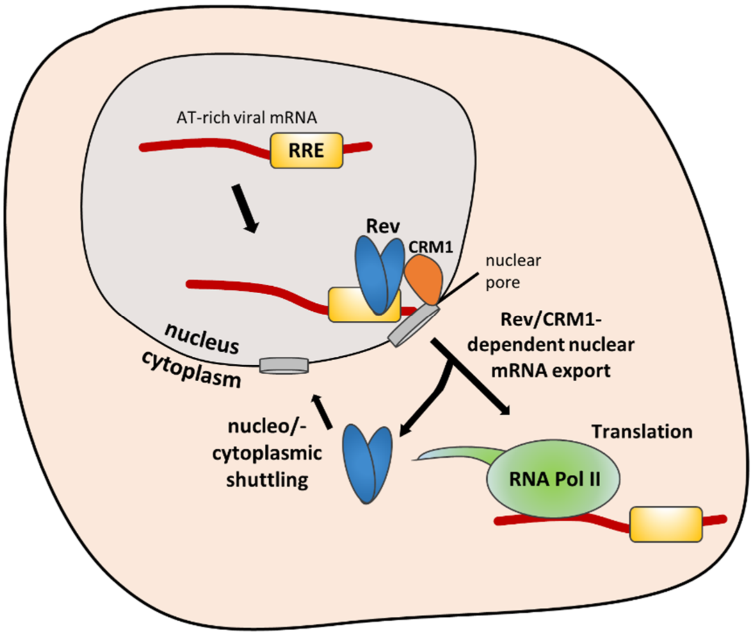


Figure 4. Rev-mediated nucleo-cytoplasmic RNA export of late viral genes.
The early viral protein Rev is required for mRNA export of late viral genes. Rev recognises the Rev-response element within the 3'UTR of late viral transcripts and promotes efficient nucleo-cytoplasmic export.

However, from this study it is not completely clear whether changing codon usage increases cytoplasmic RNA abundance due to increased nucleo-cytoplasmic export or whether the consequential removal of potentially destabilising elements, such as AU-rich elements (ARE) located on some viral genes (e.g. *gag*), leads to increased transcript stability. It was suggested that instability might not necessarily be conferred by the presence of AREs, but rather the overall AU content which may lead to increased nuclear degradation since changing the codon usage of certain viral genes which are not predicted to have any ARE-like motifs also increased transcript stability (Nguyen et al., 2004). Removing such inhibiting signals through sequence mutagenesis of *gag* removed the dependency on Rev for efficient expression (Graf et al., 2000; Schneider et al., 1997). Furthermore, changing the codon usage of the viral *vpu* and *vif* genes to closer resemble human genes increased their stability, as well as allowed the transcripts to be exported via a CRM1-independent pathway while still remaining biologically active (Nguyen et al., 2004). The same study also demonstrated that increased mRNA levels were not caused by increased transcriptional efficiency. It was furthermore shown that changing the codon usage of the Green Fluorescent Protein (GFP) reporter gene to mimic viral codon usage decreased GFP RNA levels substantially in the absence and presence of Rev without large effects on transcription or translation *in vitro*. This suggests that lowered expression is directly caused by the difference in codon usage introducing functionally undesirable features (Graf et al., 2006). A later study by Shin *et al.* demonstrated that differential codon usage of the viral glycoprotein genes gp160 (Rev-dependent) and gH (ORF57-dependent) are required for the correct temporal regulation of their expression (Shin et al., 2015). The authors also demonstrate how this dependency on adaptor proteins for efficient expression can be either flipped between both proteins by reversing their codon usage patterns, or can be imposed onto another, non-viral protein (luciferase). Overall, these finding suggest that codon usage can actively influence mRNA export by either affecting transcript stability, or the requirement for export adaptor proteins to facilitate interactions with the export machinery, however the driving mechanisms remain unclear and further studies are required to directly address such.

1.3.5. Codon usage in translation and protein folding

mRNA lies at the interphase between transcription and translation. Once fully processed and matured, mRNA is translated into protein by the ribosome. Thus, not only regulation on the post-transcriptional level, but also during translation plays a significant role in modulating protein yield. A vast amount of research around synonymous codon usage has focussed on the translation processes as well as the consequences on protein folding and functionality. Multiple different types of codon biases have been proposed to contribute to the variation of synonymous codon usage seen amongst genes within same genomes (Cannarozzi et al., 2010; Coleman et al., 2008; Tuller et al., 2010) and with further advances in methods able to explore translation dynamics on a genome-wide level, it has become clearer that differential codon usage may act as an additional layer for fine-tuning gene expression (reviewed in Quax et al., 2015).

Translation can be divided into three separate consecutive steps: Initiation, the process of assembly of the ribosomal subunits together with over 70 auxiliary factors; elongation, the actual synthesis of a peptide chain; and termination, which entails the dissociation of the full-length protein and removal of the ribosome from the mRNA transcript. The first step in protein synthesis relies on successful translation initiation. The Shine-Dalgarno sequence in prokaryotes and the Kozak sequence in eukaryotes are both involved in the interaction between mRNA and ribosomal complex. The strength of folding around the start codon can therefore influence the efficiency of translation initiation. In *E.coli* as well as *S.cerevisiae* it was shown using reporter gene libraries that a significant proportion of variation seen in protein expression could be explained by mRNA secondary structure formation around the start codon (Bentele et al., 2013; Goodman et al., 2013; Kudla et al., 2009) . In both cases, strong mRNA folding structures in the 5' end of mRNAs lead to reduced translation initiation. A computational study by Gu et al. further demonstrated that reduced mRNA stability near the translation-initiation site is a universal trend that can be seen across several species of prokaryotes and eukaryotes (Gu et al., 2010).

Although translation initiation is commonly assumed to be the rate limiting step in translation, several studies have also shown the involvement of synonymous codon usage in determining ribosomal speed in translation elongation. An essential step in protein synthesis is the successful base paring of the codon with the anticodon of its corresponding tRNA. However, no organism contains the full set of tRNAs with the complementary anticodons to all 61 naturally occurring amino acids due to the redundancy of the genetic code. Several tRNAs are isoaccepting, meaning they can recognise several codons using Watson-Crick base pairing at the first and second position of a codon, whereas "wobbling" is possible at the third position (Crick, 1966). It was assumed that more frequently used codons recognised by more abundant tRNAs lead to more efficient translation (Berg and Kurland, 1997). Some suggested that codons translated by rare tRNAs are decoded slower, thus resulting in lower elongation rates (Dana and Tuller, 2014). A study by Tuller et al. focussed on the distribution of rare and frequently occurring codons to study common patterns between genes and reported an evolutionary conserved codon ramp of between 30-50 rare codons which was proposed to initially slow down ribosomes to increase the overall efficiency of protein synthesis by preventing ribosome traffic jams (Tuller et al., 2010). This finding emphasised the importance of the 5' end of an mRNA transcript in determining translation efficiency. Several studies tried to address similar questions by investigating the mechanisms linking tRNA availability to particular patterns of codon arrangements within genes. Cannarozzi et al. suggested that fewer tRNA changes, i.e. increased use of multivalent tRNAs within the same gene, favours tRNA recycling. It was proposed that tRNAs remain in close proximity of the ribosome and their rapid recharging with their corresponding amino acid allows them to be readily re-used at the next occurrence of the same or an isoaccepting codon, leading to up to 30% faster translation rates (Cannarozzi et al., 2010). An earlier study by Gutman et al. studied the frequency of non-synonymous codon pairs and found that certain codon pairs appear more often than others, regardless of the frequency of the single codons (Gutman and Hatfield, 1989). The reasons for that are not clear, but it is thought that codon pair usage affects ribosome translocation due to the structural properties of different tRNAs as they are likely to interact while occupying the A and P sites of the ribosome (Buchan et al., 2006). Coleman et al. utilized these findings and synthesised a poliovirus mutant with a low codon pair bias, which could successfully be used to immunize mice without showing any symptoms (Coleman et al., 2008). Additionally, the concentration of tRNA-synthetases have been shown to oscillate between different phases of the human cell cycle which may explain differences in codon usage of gene sets expressed at different cellular stages (Frenkel-Morgenstern et al., 2012). It was later shown that tRNA concentrations also differ between proliferating and differentiating cell types (Gingold et al., 2014), suggesting expression regulation of certain subsets of genes by changing the codon to anticodon tRNA pools. This was recently contradicted by a study showing that any given tRNA pools are equally well able to translate any category of genes, in both healthy and cancer tissues and it was suggested that previously reported differences are primarily caused by underlying sequence features, such as genomic GC context (Rudolph et al., 2016).

However, with the development of high-throughput techniques allowing the monitoring of ribosome density on a genome-wide level (Ingolia et al., 2009), it is now possible to study translational dynamics and its link to codon usage more globally. Ingolia et al employed a pulse-chase strategy to measure translation elongation rates globally. By combining the drug harringtonine, which stalls ribosomes at the initiation codon, followed by a short run-off period and cycloheximide treatment, stalling actively translating ribosomes, they acquired several snapshots of the global translational landscape in mouse embryonic stem cells (Ingolia et al., 2011). In contrast to previous studies which concluded that the decoding speed for different codons varies strongly and affects elongation, no similar relationships could be found, as no translational pauses at rare codons were observed, leading to overall very little effects of codon usage on elongation rates. Although this does not exclude the possibility that for particular genes this may still be the case. Another study by Pop et al. further investigated the relationship of tRNA abundance and sequence features using existing ribosome profiling data under physiological conditions in yeast and could not find a significant correlation with microarray tRNA measurements (Pop et al., 2014) nor with the tRNA adaptation index, a measure of codon usage based on tRNA gene copy number (dos Reis et al., 2004). Since these experiments were mostly conducted under physiological conditions, it is possible that in previous studies which utilised overexpression-based systems, the abundance of particular tRNAs may become rate-limiting in elongation and thus, codon choice might gain more weight in its importance for overall expression patterns.

Ribosomal speed also influences co-translational protein folding (Zhang et al., 2009). Differential translational speed caused by rare codon usage was suggested to influence protein secondary structure in both computational and experimental analysis. In the Multi-drug Resistance 1 (MDR1) gene, a SNP at a synonymous site alters the conformation of its protein product, the P-glycoprotein (P-gp), without affecting mRNA or protein levels (Kimchi-Sarfaty et al., 2007). It was suggested that the change from the more frequently used codon GGC (relative synonymous codon usage (RSCU) = 22.4) to the less frequently used GGT (RSCU = 10.8), leads to a translational pause site which affects the timing of co-translational folding, resulting in an altered protein structure. Other studies showing similar effects of abnormal codon usage on protein function were able to demonstrated that non-optimal codon usage can be required for normal protein function (Xu et al., 2013; Zhou et al., 2013, 2015) which was further underlined by computational studies defining the key role of optimal and non-optimal codon usage in certain mRNA transcripts as essential for correct domain folding (Pechmann and Frydman, 2013) as it was found that β sheets are enriched in frequent codons, whereas both rare and frequent codons are required for proper α loop formation (Pechmann and Frydman, 2013). Furthermore, it was shown that clusters of rare codons found in genes coding for membrane and secretory proteins cause translational pausing to allow binding site recognition by signal recognition particles required for membrane translocation of the protein (Fluman et al., 2014; Pechmann et al., 2014).

1.4. Codon optimisation and biomedical applications

Several protein defects are associated with human complex diseases. Over the last few decades, recombinant proteins and protein therapeutics have become a common approach in treating such. The originally preferred method was the purification of proteins directly from plant, animal or human tissue, but the rise of recombinant DNA technology has largely replaced such practices. The huge variation in preferred codon usage between different species was one of the major hurdles that had to be overcome in order to improve protein yield and to make the purification process more time and cost efficient.

1.4.1. Recombinant protein expression

Many host organisms can be used to produce heterologous proteins, yet up to today, the most preferred choice is *E.coli* due to is its well understood genetics, fast growth rate and low-cost maintenance. A major area that utilises heterologous protein expression is biomedical research. Fluorescent protein bio-markers are now a commonly used tool for the visualisation of proteins and/or protein-interactions *in vivo*. One particularly well-known example of such a protein is the Green fluorescent protein (GFP) isolated from the jellyfish *Aequoerea victoria* which was first established as a novel reporter in prokaryotes and animals (Chalfie et al., 1994). Due to large differences in optimal codon usage, expression levels of heterologous proteins can often be very poor. The most commonly known approach to circumvent this issue is to alter rare codons in a target gene so that they more closely reflect the codon usage of the host, without modifying the amino acid sequence of the encoded protein. For example, Zolotukhin et al. modified the nucleotide sequence of GFP heavily by introducing 92 base substitutions in 88 codons to adjust the sequence to more preferentially used codons in the human genome which led to efficient expression in human cells (Zolotukhin et al., 1996).

Recombinant protein expression has also gained more and more importance in the production of human therapeutics. The first mammalian peptide to be successfully expressed in *E.coli* was Somatostatin, a 14 amino-acid residue (Itakura et al., 1977). This was achieved without the actual knowledge of the mRNA sequence but by reverse-translating the amino acid sequence into codons most frequently used in *E.coli.* However, protein expression in bacteria also harbours multiple disadvantages. Limitations for protein size, complexity and the lack of certain post-translational modifications restrict the production to smaller peptides or single protein domains. For this reason, human therapeutics are often produced in cultivated mammalian cell lines. To achieve efficient expression, several gene optimisation methods with the aim to adjust sequence parameters that were previously shown to be unfavourable for expression, have been proposed. Two of the most common measures of codon optimality, the Codon Adaptation Index, CAI (Sharp and Li, 1986), and the tRNA Adaptation Index, tAI (dos Reis et al., 2004), are often applied in an attempt to increase the translational yield. The CAI is a species-specific measure of codon frequency which is derived from a set of highly expressed reference genes. The assumption is that those genes are highly adapted to the tRNA pool and therefore allow more efficient translation and thus, higher protein yield. As alternative to the CAI, a further translation-related score, the tAI, was proposed (dos Reis et al., 2004). The tAI was suggested based on the finding that available tRNA pools within the cell are highly correlated with their respective gene copy numbers within a given genome, which would therefore allow to score each codon independently of a group of reference genes (Duret, 2000; Percudani et al., 1997). Although both measures were shown to correlate with expression data, this is not always the case (Kudla et al., 2009). Some studies evaluated the relationship between codon usage and expression levels in mammalian cells and concluded that there are functional differences in codon usage between protein-coding genes (Gingold et al., 2014; Ma et al., 2014; Plotkin et al., 2004). In stark contrast however, other recent work challenged this view by concluding that variation in codon usage is primarily driven by the underlying genomic sequence composition, arguing that mammals are better optimised for more complex, multi-layered regulatory mechanisms and therefore do not rely on translational efficiency as a major regulatory mechanism (Rudolph et al., 2016). These opposing views emphasize the need for more thorough investigation into codon usage and its link to gene expression control important for the enhancement of biomedical research purposes.

1.4.2. Gene therapy and DNA/RNA vaccines

More recently, codon optimisation has also been in the focus of research into novel vaccines based on DNA and RNA rather than protein antigens from disease-causing microorganisms. Immunisation by DNA/RNA vaccines relies on the ability of the produced protein to stimulate a humoral and cellular response. Since high immunogenicity depends on effective transcription and translation of the antigen, increasing protein production by optimising the codon usage is desirable. This has been successfully employed in several studies, leading to enhanced T-cell responses (Gao et al., 2003) and antibody production (Narum et al., 2001).

Another area of intense medical research which relies on similar principles is gene therapy. The general idea of gene therapy is to either replace a faulty copy of a gene, or to supplement with a functioning version with the objective to counteract the adverse effects of the defective gene. Many mutated and disease causing genes utilise a high number of less favourable codons and are expressed at low levels. Merely reintroducing a wild-type copy might therefore not be sufficient to compensate for the lack of functionality of the mutated gene product, in particular since the transfection efficiency is the limiting factor and often very poor. Mutations in the *CFTR* gene affect the function of the encoded ion channel which controls the flow of H_2O and chloride ions in and out of lung cells, causing cystic fibrosis. To tackle the issue of poor expression, Varathalingam et al. incorporated 1010 synonymous base changes into the *CTFR* sequence, creating a novel cDNA that shares only 77.4% sequence identity with the standard CFTR cDNA (Hyde et al., 2008; Varathalingam et al., 2005) but encodes the same protein and exhibits higher expression levels. In addition to modifying the codon usage, it has also been reported as beneficial to deplete transgene sequences of CpG dinucleotides as it was shown to minimise inflammation and maintain prolonged gene expression (Chevalier-Mariette et al., 2003; Dalle et al., 2005; Hyde et al., 2008; Mitsui et al., 2009). However, one study demonstrated that codon optimisation of the model gene murine erythropoeitin (mEPO) containing 20 CpGs increased expression levels as well as prolonged expression in mice compared to a CpG-depleted version of the wild type gene (Kosovac et al., 2010), leaving the requirement for further investigation into the underlying mechanisms.

1.5. Aims of this thesis

The GC content varies greatly across genomes and also across genes, leading to large differences in codon usage between and within genes. The knowledge of the various factors affecting gene expression level have for many decades been exploited to enhance protein abundance for research and/or therapeutic purposes and codon optimisation has become a common approach in enhancing protein levels of otherwise poorly expressed genes. In recent years it has become apparent that although most approaches to optimise expression levels usually aim to enhance a genes translational rate, it is even more crucial to ensure high transcript levels. Findings by Kudla *et al.* have shown that differences in expression levels between GC-poor and GC-rich coding-sequence variants can already be seen on the mRNA level (Kudla et al., 2006). To date, only few systematic and quantitative analyses of sequences features and their relative contributions in several steps in gene expression have been conducted on a single-gene level in a controlled human cell line system.

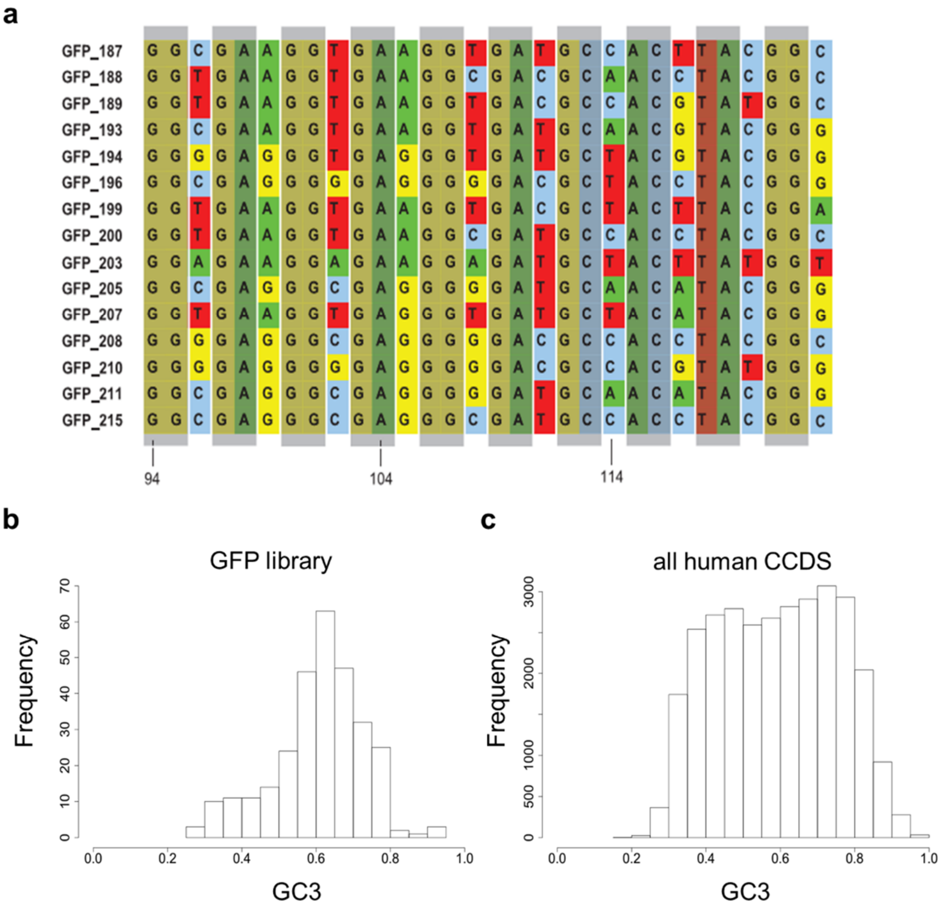
The PhD project presented here aims to address this with the following two objectives:

1) **Systematically and quantitatively measure the molecular phenotypes of several hundred coding-sequence variants of a fluorescent reporter gene *in vivo* in human cells**

2) **Investigate the coding sequence properties of genes that are associated with high mRNA stability, mRNA export, translation and protein yield**

1.5.1. Experimental system

In order to be able to directly compare the consequences of synonymous changes within the coding region of a gene, I utilised a cDNA library of the Green Fluorescent Protein (GFP) (Kudla et al., 2009). This library consists of several hundred synonymous coding sequence variants which differ only at the third position of each codon, encoding the exact same protein (Figure 5a). The synthetic GFP constructs differ between 1 to 180 silent base substitutions with an average of 114 substitutions between pairs of variants. The library was designed to span a wide range of GC3 content, varying between 25% - 97%, which compares well with the GC content variation of coding-sequences in human cells (Figure 5b+c).



```
GFP_187  G G C G A A G G T G A A G G T G A T G C C A C T T A C G G C
GFP_188  G G T G A A G G T G A A G G C G A C G C A A C C T A C G G C
GFP_189  G G T G A A G G T G A A G G T G A C G C C A C G T A T G G C
GFP_193  G G C G A A G G T G A A G G T G A T G C A A C G T A C G G G
GFP_194  G G G G A G G G T G A G G G T G A T G C T A C G T A C G G G
GFP_196  G G C G A G G G G G A G G G G G A C G C T A C C T A C G G G
GFP_199  G G T G A A G G T G A A G G T G A C G C T A C T T A C G G A
GFP_200  G G T G A A G G T G A A G G C G A T G C C A C C T A C G G C
GFP_203  G G A G A A G G A G A A G G A G A T G C T A C T T A T G G T
GFP_205  G G C G A G G G C G A G G G G G A T G C A A C A T A C G G G
GFP_207  G G T G A A G G T G A G G G T G A T G C T A C A T A C G G G
GFP_208  G G G G A G G G C G A G G G G G A C G C C A C C T A C G G C
GFP_210  G G G G A G G G G G A G G G G G A C G C C A C G T A T G G G
GFP_211  G G C G A G G G C G A G G G G G A T G C A A C A T A C G G G
GFP_215  G G C G A G G G C G A G G G C G A T G C C A C C T A C G G C
```

Figure 5. A synthetic library of gene variants of GFP with random codon usage.
a, Partial sequence alignment of several synonymous GFP variants which only differ at the third position of each codon (from Kudla et al., 2009) . **b**, Distribution of GC-content at the third codon position (GC3) for all GFP variants in the library or **c**, all human consensus coding sequences (CCDS).

2. Materials and Methods

2.1. Tissue culture

2.1.1. Cell lines used

HeLa (J. Caceres lab, MRC HGU); HeLa Flp-in (A. Jackson lab, MRC HGU); Hek293T (J. Caceres lab, MRC HGU); Hek293T Flp-in (Life Techologies); Hek293 Flp-in GFP000 & GFP001 (L. Lipinski, Warsaw). All cells were tested negative for Mycoplasma.

2.1.2. Maintenance of cell lines

All cell lines were cultured in Dulbecco's Modified Eagle's Medium (DMEM, Gibco) supplemented with 10% Fetal Calf Serum (FCS, Life technologies). HeLa Flp-in and HeK293 Flp-in cell lines were grown with Tetracycline-free FBS (Clontech). Cells were passaged regularly and were maintained at subconfluency (~70%).

For passaging, cells were washed with Phosphate buffered saline (Dulbecco's A, Oxoid) prior to treatment with 1x Trypsin/EDTA (Sigma) to detach cells before transferring into fresh culturing vessels. Generated Flp-in cell lines were maintained in media complemented with HygromycinB (HeLa Flp-in: 400mg/ml; Hek293 Flp-in: 100mg/ml; Life technologies) and 10ng/ml Blasticidin S (HeLa and Hek293 Flp-in; Thermo Fisher) to maintain selection for the Flp-in site as well as gene integration. Any cells used for fluorescence based assays were grown in phenol red-free media to lower background fluorescence (Biochrom DMEM F0475, supplemented with 2mM L-Glutamine, 10% Tetracycline free FBS (Clontech) and 1% Penicillin/Streptamycin). All cell lines were cultured at 37°C, 5% CO_2.

2.1.3. Plasmid Transfections

2.1.3.1. Reverse transfections

For GFP fluorescence screen:

Per well in a 96 well plate (Greiner), 70ng plasmid (GFP cloned into pCM3 or pCM4) were used in reverse transfections in triplicate. Enough transfection mix was made up for 4 wells. In brief, 280ng plasmid DNA was diluted in 40ul OptiMEM (Gibco). 1ul Lipofectamine2000 (life tech, 0.25ul per well) was diluted in in 40ul OptiMEM and incubated for 5min at RT. Both plasmid and Lipofectamine2000 dilutions were then mixed by pipetting up and down (total volume: 80ul) and incubated for 20-30min at RT. 20ul of the transfection complex was pipetted in 3 wells before adding 200ul of HeLa cells (45 000 cells/ml; 9 000 cells/well) grown in phenol red-free media (Biochrom, F0475, see also 2.1.4). Media was exchanged 3h post-transfection to reduce toxicity. Cells were incubated 48h at 37°C, 5% CO_2 before cell lysis.

For RNA isolation:

Cells were transfected as above in 24 well plates with 300ng plasmid DNA and 4ul Lipofectamine2000. Cells were incubated for 24h before harvesting RNA.

2.1.3.2. Stable transfection of Flp-in cell lines

Cells were grown in 6 well plates to 80% confluency. *pOG44* and *pCDNA5* were mixed in a 9:1 ratio to give 2ug DNA in total (e.g. 1.8ug *pOG44* + 0.2ug *pCDNA5*). 9ul/well Lipofectamine 2000 (LifeTechnologies) were mixed with 91ul Opti-Mem (Gibco) in an sterile 1.5ml tube and incubated at RT for 5’. Plasmid DNA was then added, mixed well and incubated at RT for 15’. After incubation, transfection mix was added dropwise to cells and incubated for 4h before changing media. Cells were incubated for 48h before chemical selection. 10ng/ml Blasticidin S (Thermo Fisher) and Hygromycin B (HeLa Flp-in: 400mg/ml; Hek293 Flp-in: 100mg/ml; Life technologies) were routinely added every 3rd passage for selection of cells with successful gene integrations. Clonal colonies were picked once they reached a reasonable size by picking them with a pipette and transferring them into 96 well plates. GFP pool cell lines were not clonally selected and stocks frozen once no significant cell death could be observed (complete selection).

2.1.4. Single GFP fluorescence screen

Plasmids expressing GFP variants were transfected into HeLa cells as described in 2.1.3.1 in 96 well plates (Greiner). Different standard media formulations were tested:

1) DMEM, high glucose; Gibco, cat no. 11965
2) DMEM, without phenol red; Biochrom, cat. no. F0475
3) DMEM, high glucose, HEPES, no phenol red; Gibco, cat. no. 21063
4) Fluorobrite DMEM; Gibco, cat. no. A18967-01

Due to low background fluorescence, **media 2)** was used for all subsequent experiments.

8h post-transfection, media was removed and cells lysed with 200ul cell lysis buffer. Recipes of cell lysis buffers tested are listed below. **Buffer 1** was used in the final protocol.

Buffer 1: 25mM Tris (pH 7.4), 150mM NaCl, 1% Triton X-100, 1mM EDTA (pH 8).

Buffer 2: 30mM Tris (pH 7.4), 150mM NaCl, 1mM EDTA (pH 8), 1% Triton X-100, 1mM DTT, 10mM NaF.

Buffer 3: 150mM NaCl, 20nM Tris (pH 7.5), 2mM EDTA (pH 8), 10% (v/v) glycerol, 1% (v/v) Triton X-100.

Cells in lysis buffer were incubated under gentle shaking for 15 min prior to fluorescence measurments. Fluorescence readings were obtained on a Tecan Infinite M200pro multimode plate reader. GFP: Ex 486nm/Em 515nm; mKate2 Ex 588nm/Em 633nm; Reading mode: bottom; Number of reads: 10 per well. Background subtraction: measurements from untransfected cells were subtracted from all other wells. This is followed by fluorescence normalisation: GFP signal was calculated relative to mKate2 signal. To compare plates from different days, GFP000, GFP001 and GFP034 were transfected on every plate to normalise data relative to these controls.

2.1.5. Transcription inhibition assay

Cells were grown to 80% confluency either in 6 well plates (GFP and EGFP cell lines) or 10cm plates (GFP pool cell lines). Media was replaced with fresh media containing 500mM Triptolide (Company name) or the equivalent volume of DMSO (Sigma, Cat). Cells were collected at multiple time points by removing media and adding Trizol reagent (Ambion; 1ml for 6 well plates or 3ml for 10cm plates). RNA was then extracted as outlined in (2.2.2).

2.1.6. Cell viability assay

Cell viability was assessed using alamarBlue (Thermo Fisher) containing resazurin, a blue, non-fluorescent compound. Upon uptake into cells, resazurin is converted to resofurin, a red, fluorescent compound. 9000 HeLa cells were reverse transfected in 96 well plates (2.1.3.1) and cell viability assessed 24 and 48hrs post-transfection by adding 20ul of 10x alamarBlue to each well. Cells were incubated for a further 4hrs before measuring fluorescence (Ex 560nm/Em 590nm). To calculate cell viability as percentage, the relative fluorescence value obtained from alamarblue added to media only was subtracted from all wells and viability calculated relative to the result of untreated cells (100%).

2.1.7. Polysome Profiling

Hek293 Flp-in GFP pool cell lines were grown to 90-95% confluency on 15cm dishes. Cells were treated for 20min with 100ug/ul Cycloheximide. After incubation, media was removed and plates washed 2x with ice-cold PBS before scraping cells into 1.5ml tubes. Cells were pelleted at 7000rpm, 4°C for 1min and cells carefully resuspended in 250ul RSB (10x RSB: 200mM Tris pH 7.5, 1M KCl, 100mM $MgCl_2$) containing 1/40 RNasin (40U/ul, Promega) until no clumps are visible. 250ul of Polysome extraction buffer are then added (1ml 10x RSB + 50ul NP-40 + 9ml H_2O + 1 complete mini EDTA-free protease inhibitor pill (Roche)) and lysate passed 5x through a 25G needle avoiding bubble formation. Lysate is then incubated on ice for 10min before spinning 10min at 10000g at 4°C. The supernatant is then transferred into a fresh 1.5ml tube and the OD measured at 260nm as RNA. Sucrose gradients (10–45%) containing 20 mM Tris, pH 7.5, 10 mM $MgCl_2$, and 100 mM KCl were made using the BioComp gradient master. 100ug of Lysate were loaded on sucrose gradiants and spun at 41 000rpm for 2.5h in a Sorvall centrifuge with a SW41Ti rotor. Following centrifugation, gradients were fractionated using a BioComp gradient station model 153 (BioComp

Instruments, New Brunswick, Canada) measuring cytosolic RNA at 254 nm and 18 fractions collected. The resulting fractions were pooled into 4 samples: Fractions 1- (A) Free Ribonucleoprotein (RNP) complexes, (B) monosomes, (C) light polysomes (2-4) and (D) heavy polysomes (5+) (Figure 36). RNA from all 4 samples was prepared for high-throughput sequencing and resulting reads processed and filtered as before (described in X.X). RNA was precipitated using 1 volume 100%EtOH and 1ul Glycoblue, before extracting RNA using the Trizol method described below (2.2.2).

2.2. Molecular Biology/Biochemical techniques

2.2.1. Agarose gel-electrophoresis

20xTBE: Tris Base (216g/l), Boric Acid (110g/l), 0.5M EDTA (pH 8) 80ml/l
A standard 1% agarose gel is prepared using 1g/100ml Agarose (life technologies) in 1xTBE (20xTBE stock: Tris Base (216g/l), Boric Acid (110g/l), 0.5M EDTA (pH 8) 80ml/l) and 3ul Ethidiumbromide (VWR) per 100ml gel. The percentage of gels varies depending on the size of fragments to be analysed.

2.2.2. Isolation of total RNA

Total RNA was isolated using Trizol reagent (Life tech) according to manufacturer's manual. Cells are lysed directly in the cell culture vessels after removal of cell culture media. 1ml Trizol is used per well in a 6 well plate or 500ul for wells of 24 well plates. Trizol samples are transferred into 1.5ml microtubes. 200ul Chloroform (Sigma) are added and tubes shaken vigorously for 15s before incubation on ice for 20min. For complete phase separation, tubes are spun at max speed for 20min at 4°C in a microcentrifuge. The upper (clear) phase is carefully removed and transferred into a fresh 1.5ml tube containing 1volume 100% Isopropanol and 15ug Glycoblue (Thermo Fisher, AM9515). Tubes are vortexed briefly. RNA is then precipitated for 20min at -20°C before pelleting by spinning 20min at max speed at 4°C. RNA pellets are washed twice with 1ml 75% Ethanol to avoid Phenol carryover. RNA pellets are then air-dried to remove all Ethanol before resuspension in sterile RNAse-free dH2O. RNA is then incubated for 10min at 55°C followed by quick-chilling on ice to remove secondary structures before storage at -80°C.

2.2.3. Subcellular fractionation of cells

This protocol is based on the cellular fractionation protocol published by Gagnon et al. (2014) but includes a further clean-up step using a sucrose cushion as described by Zaghlool et al. (2013) as well as a second lysis step as described by Wang et al. (2006). Cell lysis and nuclear integrity was monitored by Light Microscopy following Trypan blue staining. When starting using this protocol and it is recommended to monitor successful subcellular fractionation by fluorescence microscopy as described in Gagnon et al. (2014). Cells were grown in 10cm plates to about 90% confluency. Cells were then washed with PBS and trypsinised briefly using 1ml of 1xTrypsin/EDTA. The reaction was then stopped with 5ml of DMEM and cell suspension transferred into 15ml falcons. Cells were collected by spinning at 100g for 5min. Cell pellets were resuspended in 500ul ice-cold PBS and transferred into 1.5ml reaction tubes and spun at 500g for 5min at 4°C. The supernatant was then discarded and cells resuspended in 250ul HLB (10mM (pH 7.5), 10mM NaCl, 3mM MgCl2, 0.5% (v/v) NP40, 10% (v/v) Glycerol, 0.32M sucrose) containing 10% RNase inhibitors (RNasin Plus, Life Tech) by gently vortexing. Samples were then incubated on ice for 10min. After incubation, samples are vortexed gently and spun at 1000g for 3min at 4°C. Proceed to step a) and b) with the supernatant and pellet from this step.

a) Cytoplasmic extract:

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

b) Nuclear extract:

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

RNA from tubes with cytoplasmic and nuclear extracts was then extracted as described in 2.2.2.

2.2.4. Oligonucleotide sequences

qPCR primers	**5' → 3'**
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
U6_F	ATCTGATACGTCCTCTATCCGA
U6_R	GCAATACCAGGTCGATGCGT
MiSeq library + sequencing	
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
attB1_EG_F	GGGGACAAGTTTGTACAAAAAAGCAGGCTTCGAGCTCAAGCTTCGAATTCTG
attB2_E_R	GGGGACCACTTTGTACAAGAAAGCTGGGTGGCCGCTTTACTTGTACAGCTC
attB2_G_R	GGGGACCACTTTGTACAAGAAAGCTGGGTGGCCGCTTTACTTGTATAGTTC
mkate2_gibs_F	GATCCGCGTATGGTGGCCTTAAGATACATTGATGAG
mkate2_gibs_R	TGTAAGCGGATGCCGCACATGTTCTTTCCTGCG
pCI_gib_F	CGGCATCCGCTTACAGACAA
pCI_gib_R	CACCATACGCGGATCCTTATC
other primers	
pGK8_T7_F	TGCGTCCGGCGTAGAGGATC
pGK8_T7_R	GCCTCTTGCGGGATATCC
pCM_seq_F	CTGGGCTTGTCGAGACAGAG
pCM_seq_R	TGCAGCTTATAATGGTTACA

2.2.5. *In vitro* transcription assay

GFP variants 400, 403, 407, 412, 417 and 422 were *in vitro* transcribed from *pGK8* which contains a T7 promotor upstream of the coding sequence, using the MEGAscript T7 transcription kit (Ambion). First, GFPs including the T7 promoter were PCR amplified using primers "pGK8_T7_F" and "pGK8_T7_R" and 0.5ng *pGK8-GFPXXX* as template. PCR products were column-purified using the Qiagen PCR purification kit. For *in vitro* transcription reaction, 0.1-0.2ug DNA was mixed with 8ul nucleotide mix (equimolar ratios; 75mM each), 2ul 10x buffer, 2ul enzyme mix and made up to 20ul total volume with H2O. Reaction buffer was added last, before mixing gently and incubating samples at 37°C for 4hrs. 1ul of Turbo DNase was added, mixed and incubated for 15' at 37°C before adding 115ul H2O and 15ul Ammonium acetate stop solution. Samples were mixed thoroughly before extracting RNA using an equal volume of acid phenol/chloroform, followed by a second extraction using chloroform only. The aqueous phase was recovered and transferred into a fresh tube before precipitating RNA by adding 1 volume isopropanol. Samples were chilled for at least 15' at -20°C before spinning at maximum speed for 15'. Resulting RNA pellets were dried and resuspended in RNAse-free water. RNA quality and product length was assessed by agarose gel electrophoresis.

2.2.6. cDNA synthesis

cDNA prepared for qRT-PCR and semi-quantitative PCR analysis was synthesised by reverse transcribing 500ng RNA using SuperScript III (Life Technologies) and 500ng random hexamers (Promega). cDNA prepared from RNA after polysome profiling was prepared with oligo-dTs (Promega) to avoid reverse transcription of primarily ribosomal RNA. Generally, cDNA prepared for sequencing library preparation was synthesised using SuperScript III (Life tech) and 2 nmol gene specific primers ('S_indexX_right_PEPCR'). In addition, a new commercially available enzyme, TGIRT III, was additionally tested for high throughput experiments as it has been shown to be able more efficient in processing highly structured and highly modified RNAs (Mohr et al., 2013; Qin et al., 2015). 500ng RNA were mixed with 1uM gene-specific primer and heated at 70C for 10min. The reactions were quick chilled, and 4ul of 5x buffer,

2.2.7. Real-time Polymerase Chain Reaction (PCR)

All qRT-PCRs were carried out on a Roche LightCycler 480 using LightCycler480 SYBR Green I Master Mix according to manufacturer's protocol using 0.3uM gene-specific primers. Samples were analysed in 20ul reactions in triplicates. DNA was first denatured for 5min at 95°C before entering a cycle (50-65x) of denaturing for 10sec at 95°C, annealing for 7sec at

55-60°C (depending on primers used), extension for 10sec at 72°C and data acquisition. DNA was then gradually heated up by 2.20 °C/s from 65 to 95°C for 5sec each and data continuously collected (Melting curve analysis). The data from transient transfection experiments was analysed using the Pfaffl method (Pfaffl, 2001). All other data was evaluated using the comparative C_t method (Livak and Schmittgen, 2001).

2.2.8. Polymerase Chain Reaction (PCR)

PCR amplification from plasmid DNA was generally conducted using 0.5ng template and either Q5 Polymerase (NEB) or Accuprime Pfx (NEB) according to manufacturer's specifications. For PCR amplification during all next-generation sequencing library preparations, Accuprime Pfx (NEB) was used.

2.2.9. Sanger sequencing

All DNA cycle sequencing reactions were performed using the BigDye Terminator Kit (LifeTech) and are set up in 0.5ml thick-walled tubes using. 200ng Plasmid DNA, 0.8uM primer, big dye terminator mix and reaction buffer are made up to 20ul and placed into a thermocycler. Cycling program: Initial denaturation at 96°C for 3min, followed by 25 cycles of denaturing for 30sec at 96°C, primer annealing at 50°C for 15sec and extension at 60°C for 2min. Products are then precipitated by mixing reactions with 2ul of 3M Sodium acetate (pH 4.6) and 50ul of 100% Ethanol at incubation for 15min out of bright light before pelleting at maximum speed in microcentrifuge for 20min. Resulting pellets are washed with 150ul 75% Ethanol and spun again for 2min. Pellets are then air-dried out of light. The prepared sequencing reactions are processed by the in-house technical service on applied Biosystems Genetic analyser.

2.2.10. Determination of nucleic acid concentrations and quality

The concentrations of DNA and RNA were generally determined spectrophotometrically by measuring the optical density at 260nm using a NanoDrop (Thermo Scientific). DNA concentration of samples for high-throughput sequencing were determined using an Agilent Bioanalyzer 2100 and/or using Qubit Fluorometric Quantitation (Thermo Fisher). RNA quality and concentration of samples for high-throughput sequencing library preparation were assed using the Agilent Bioanalyzer system.

2.3. Plasmids and GFP library

2.3.1. GFP library

GFP000 ('GFP') and GFP001 ('EGFP') are taken from Kudla et al, 2006. The GFP library was published in Kudla et al., 2009.

2.3.2. Preparation of competent DH5α cells

A single colony of DH5α is picked from an LB agar plate (for 500ml: 5g NaCl, 5g Tryptone, 2.5g yeast extract, 7.5g Agar, dH_2O to 500ml) and used to inoculate 2.5ml of LB (no agar). The culture is incubated overnight at 37°C with shaking at 220rpm. The subculture is then diluted 1:100 by inoculating 2.5ml into 250ml of LB supplemented with 20mM MgSO4. Cells are then grown until the OD600 reaches between 0.4-0.6. The culture is then split into two 250ml centrifuge bottles and pelleted by centrifugation at 4500g for 5min at 4°C. The cell pellet is gently resuspended in 0.4 original volume of ice-cold TFB1 (50ml/bottle). The resuspended cells are then combined in one bottle and cells kept on ice for all following steps. Pipettes, tubes and flasks are pre-chilled on ice. Resuspended cells are incubated on ice for 5min at 4°C before centrifugation at 4500g for 5min at 4°C. The resulting cell pellet is resuspended in 1/25 of original volume of ice-cold TFB2 (10ml for a 250ml subculture). Cells are then incubated on ice 15-60min before snap-freezing in a dry-ice bath for long-term storage at -80°C.

2.3.3. Bacterial transformation for general plasmid propagation

50ul of competent DH5α are transformed with either 2ul of 10ul ligation reactions, 1ul of 5ul Gateway ligations or <1ng plasmid DNA. Competent bacteria and DNA are carefully mixed in a tube and incubated on ice for 15min, then heat-shocked for 45' at 42°C followed by quick-chilling on ice for 2min. Cells are diluted by adding 950ul SOC rich media and shaken at 230rpm at 37°C for 1h. For In case of plasmid transformations, 100ul are spread on LB-agar plates containing appropriate antibiotics. For transformations of ligation reactions, cell suspension are spun for 3min at 3000rpm, then resuspended in 100ul SOC and all spread on agar plates.

2.3.4. Expression vectors for single GFP transfections (*pCM1-4*)

p*CM1-4* are plasmids based on pCI-neo (Promega). pCI-neo contains a CMV immediate-early enhancer/promoter region allowing strong and constitutive expression in mammalian cells. Upstream if the multiple cloning site is a chimeric intron composed of the 5'-donor site from the first intron of the human beta-globin gene and the branch and 3'-acceptor site from the intron of immunoglobulin gene heavy chain variable region (pCI-neo vector technical bulletin, Promega). The Gateway-destination cassette (RfA) containing the attR1 and 2 recombination sites as well as the ccdB gene and chloramphenicol resistance cassette was cut enzymatically from 1ug *pBluescript-RfA* using 1ul *EcoRV (NEB)* and *SmaI (NEB). pCI-neo* was cut with the same enzymes which have their restriction site in correct orientation within the multiple cloning sites. The enzymatic digests were stopped by heat inactivation for 20min at 65°C. Digest products of *pBluescript-RfA* were dephosporylated using 1U Antarctic phosphatase (NEB) for 30min at 37°C followed by heat inactivation for 5min at 65°C. Ligation reactions were set up in a 3:1 insert to vector ratio using T4 DNA ligase (NEB) in 10ul reactions by incubation for at least 1h at RT before transformation of DH5α and selection on Ampicillin (50ug/ml) and Chloramphenicol (1ul/ml) LB-agar plates. The chimeric intron contained within the 5'UTR of the *pCI-neo* expression cassette was deleted by Phusion site-directed deletion mutagenesis (ThermoFisher) using Phusion-*Taq* and primers 'pCI_del_int_F' and 'pCI_del_int_R'. The resulting plasmid is referred to as *pCM2.*

The mKate2 gene expression cassette from *pmKate2-N* (Evrogen) was cloned into the backbone of *pCM1* and *pCM2* using Gibson assembly cloning (NEB) according to manufacturer's instructions: *pCM1/2* were linearised using primers 'pCI_gib_F' and 'pCI_gib_R'. Homologous ends were added to mKate2 using primers 'mkate2_gibs_F' and 'mKate2_gibs_R' in an PCR step followed by PCR purification (Qiagen PCR purification kit). The mKate2 PCR product was recombined into linearised *pCM1/2* using the Gibson assembly cloning kit (NEB) according to the manufacturer's instructions. Resulting plasmids were amplified in DH5α and colonies screened for the correct insert. This resulted in both *pCM3* (no intron, with mKate2) and *pCM4* (with intron, with mKate2). Since all four versions of the *pCM* vector contain the Gateway RfA destination cassette, any sequence contained within a Gateway entry vector can be sub-cloned into these vectors using the Gateway LR clonase reaction described in 2.3.5 (ThermoScientific).

GFP and EGFP were cloned from pGFP-N2 and pEGFP-N2 respectively into *pCM1* by cutting the gene sequences from each vector using restriction enzymes XhoI and NotI (both NEB, 1ul each per 1ng plasmid DNA) for 1h at 37°C. *pCM1* was digested in the same manner. Enzymes were heat-inactivated at 65°C, 20min. Resulting DNA digests were resolved on a 1% Agarose/TBE and bands corresponding to GFP and EGFP cut out and gel-purified (Qiagen gel purification kit) before desphosporylation using 1U Antarctic Phosphatase (NEB) for 30min at 37°C followed by heat inactivation at 65°C, 5min. Purified GFP and EGFP were ligated into *pCM1* using 200 U DNA ligase (NEB) in a 3:1 ratio according to manufacturer's instructions for 1h at RT. 2ul of the ligation mixed were transformed into DH5α as described in 2.3.3. Resulting colonies were screened for the presence of the correct insert and sequence-verified.

2.3.5. Gateway cloning

50ng gateway entry vector and 50ng gateway destination vector are mixed with 0.5 ul LR Clonase mix (Thermo fisher). The volume is adjusted to 4.5ul using TE and the reaction incubated at room temperature for 1hr. 0.5ul of Proteinase K are then added and incubated for 30min at room temperature. 1ul of the reaction is used to transform chemically competent bacteria.

2.3.6. Gateway Entry vectors containing GFP or EGFP

GFP (here referred to as 'GFP000') and EGFP ('GFP001') were enzymatically cut from *pGFP-N2* and *pEGFP-N2* (both Promega) using BamHI and EcoRI and ligated into *pGK3*, a gateway entry vector cut with the same enzymes. *pGK3-GFP* and EGFP respectively can be sub-cloned into any gateway destination vector, such as any *pCM* plasmid.

2.3.7. pcDNA5 gateway destination vectors

pcDNA5/FRT/TO/Dest was obtained from Ewelina Macech (Cancer Centre, Warsaw, Poland) and contains the Gateway-compatible attB destination cassette for subcloning the gene of interest from a gateway-entry vector. I modified this vector to additionally have a version containing the same 5'UTR intron vector sequence as in *pCI-neo* to allow direct comparison between expression experiments.

1ug pcDNA5/FRT/TO/Dest was digested with 1ul AflII (NEB) for 1h at 37°C followed by reaction clean-up using the Qiagen Purification kit according to the manufacturer's manual. The intronic sequences was amplified from *pCI-neo* by PCR using the primers 'Gib_intr_F' and 'Gib_intr_R' using Q5 High-Fidelity Polymerase (NEB) using the following reaction conditions: 1ng pCI-neo were mixed with 0.5ul 25mM dNTPs, 2.5ul 10uM Primer-mix, 35ul H2O and 1ul Q5 Polymerase. A PCR was performed using an annealing temperature of 59°C. The primers contain a 15nt overhang that is homologous to the ends of pcDNA5/FRT/TO/Dest when linearised with AflII.

2.3.8. Plasmid DNA preparation

Single bacterial colonies were picked from LB-agar plates and resuspended in 2ml of LB-broth containing appropriate antibiotics in 12ml snap-cap tubes and incubated overnight (~16h) at 37°C in a shaking incubator at 230rpm. Plasmid DNA is extracted from the bacterial overnight culture using the Qiagen Minispin Plasmid kit according to the manufacturer's instructions. Cells are collected by spinning 3min at 6000g in a 1.5ml microtube. The cell pellet is then resuspended in 250ul buffer PI before addition of 250ul buffer P2 and mixing by vigorously inverting the tube 4-6 times until the solution becomes clear. 350ul of buffer N3 are added and mixed immediately by inversion of the tube 4-6 times. Lysates are spun for

10min at max speed in a microcentrifuge. The supernatant is then applied to a QIAprep spin column and spun at max speed for 1min. The flow-through is discarded and the column washed by addition of 750ul of buffer PE and spinning for 1min at max speed. The flow-through is again discarded and the column spun an additional time to assure it is completely dry. The column is then transferred into a fresh 1.5ml tube and DNA eluted by adding 30-50ul of EB or H2O directly to the membrane, incubation for 1min at RT and spinning at max speed for 1min. Concentrations are assessed using a Nanodrop (2.2.10).

2.3.9. Multiplex Gateway LR reaction

Since all GFP variants are stored in gateway-compatible entry vectors, I modified the standard gateway LR reaction protocol to multiplex the recombination of 217 different GFP-containing entry vectors with 1 of 2 destination vectors (pCDNA5/FRT/TO/Dest and pCDNA5/FRT/TO/Dest+intron). In a standard LR reaction using (as described in 2.3.5) about 100-200 colonies would be expected in total when transforming all 5ul into 5x 50ul chemically competent DH5α. I scaled this reaction up by 10x to ensure all 217 GFP variants will enter the destination vectors. A pool of all 217 entry vectors was prepared with a concentration of 0.06ng of each GFP variant (total DNA concentration: 13.02ng/ul). For each destination vector, a separate multiplex LR reaction was set-up using 500ng destination vector, 5ul LR Clonase, 38ul entry-vector pool and TE (pH 8) up to 45ul in total. The reactions were incubated at 25°C overnight. 5ul of Proteinase K were then added to each and incubated for 10min at 37°C. The total 50ul reaction mix was then used to transform 2.5ml of DH5α in a 15ml Falcon by heat shocking cells for 2min 30s at 42°C before adding 10ml SOC medium and incubating while shaking for 1h at 37°C. After incubation, cells were spun down at 3000g for 3min. Resulting bacterial pellet was resuspended in 1ml SOC medium and 100ul plated onto 10x L-Ampicillin agar plates. Plates were incubated at 37°C overnight. After incubation, each plate contained >800 colonies each. Bacterial colonies were scraped off the plates and collected in a falcon tube. 2x 2ml LB-Amp cultures were inoculated with 200ul of cell pool for later preparation of glycerol stocks (11.5% glycerol in LB). Total plasmid DNA was extracted using a Qiagen Midiprep kit according to the manufacturer's instructions.

2.3.10. Restriction digest of single pcDNA5 clones

Successful integration of the GFP sequence into pCDNA5/FRT/TO/Dest was confirmed by restriction digest. Not*I* restriction sites are present just downstream and upstream of the gateway attB sites. 500ng Plasmid DNA was mixed with 0.5ul NotI (NEB) and 1ul buffer 1 (NEB) in a total reaction volume of 10ul. Samples were incubated for 1h at 37°C. 2ul of 6x orange gel loading dye (NEB) were added and fragments resolved by gel electrophoresis on a

1% agarose/TBE gel. In addition, 96 clones were picked and plasmid DNA used in Sanger sequencing to confirm the presence of several different variants.

2.3.11. Amplicon-library preparation for high-throughput sequencing

Libraries from genomic DNA were generated by PCR using primers specific for the GFP UTRs. These primers also contain the required adaptor sequences for MiSeq-sequencing, as well as a 6nt index for multiplexing multiple samples in one run of sequencing. Between 6-10ug of total genomic DNA were used in multiple PCR reactions (200ng per 50ul reaction). After PCR, all reactions of the same template were pooled together, and 1/3 of the reaction purified using the Qiagen PCR purification kit according to the manufacturer's instructions. DNA was eluted in 50ul. Size selection of the PCR products was performed using the Invitrogen E-gel system (Clonewell gels, 0.8% agarose). After gel purification, samples were purified using the Qiagen MinElute PCR purification kit according to the manufacturer's instructions. Sizes of the selected fragments were confirmed and quantified using the Agilent Bioanalyzer 2100. Up to 8 different libraries were multiplexed for 1 run of sequencing by mixing individual libraries in equimolar ratios.

2.4. Fluorescence activated cell sorting (FACS)

GFP expression of cells was induced for 24h prior to cell harvesting. Data was acquired on a BD FACS AriaII cell sorter or BD LSRFortessa cell analyser and a minimum of 50.000 events were recorded.

2.4.1. Flow-Seq

80x15cm cell culture plates of HeLa Flp-in GFP pool cells and 40x15cm cell culture plates of Hek293 Flp-in GFP pool cells were induced with 1ug/ml Doxycyline (Sigma, D9891) for 24h or 48h in phenol red-free DMEM (Biochrom, F0475) supplemented with 10% FCS (Sigma, F-7524) and 2mM L-Glutamine. After 24h or 48h cells were harvested by trypsinisation and cells sorted into 8 fluorescence bins. Polypropylene collection tubes were coated with 1%BSA/PBS and cushioned with 200ul 20%FBS/PBS. 10^7 cells were collected in each tube. Cell suspensions were decanted into 15ml Falcon tubes and spun for 5' at 500*g*. The supernatant was transferred into a fresh 15ml falcon and precipitated using 2volumes of EtOH/0.1 volume Sodium Acetate pH 5.3 and 10ul Glycoblue (Ambion). Tubes were shaken vigorously for 10s and then incubated at -20C for 15min before spinning at 3000g for 20min. Resulting pellets were air-dried and resuspended in 1ml digest buffer before combining with cell pellets which were also resuspended in 1ml digest buffer. 10ul/ml of RNase A (Qiagen) were added, lids sealed with parafilm and tubes rotated at 37C. After 1h, 10u/ml Proteinase K (20mg/ml, Roche) were added to sample before rotating a further 2h at 55C. DNA was then extracted using three times using 1 volume Phenol:Chloroform:Isoamyl alcohol (PCI, 25:24:1, Sigma). Each time samples were shaken vigorously for 10s after the addition of PCI and spun at 3000 g for (first) 20min or (following) 5min. The bottom layer including interphase were discarded before each PCI addition. After the last PCI extraction, the upper layer was transferred into a fresh 15ml tube and one extraction using 1 volume chloroform:Isoamyl alcohol (CI, 24:1, Sigma) was performed to ensure removal of all Phenol. After a 5min spin at 3000g, the upper layer was transferred into a fresh 15ml tube and DNA precipitated using 2volumes EtOH (100%) and 0.1 vol Sodium Acetate pH 5.3. After a 5min incubation on ice, tubes were spun for 30min at 3000g. The resulting DNA pellets were washed 2 times with 75% EtOH before being air-dried and resuspended in a suitable amount of Tris-EDTA (10mM) (~200ul). The quality of genomic DNA was assessed on a 0.8% Agarose/TBE gel.

2.5. High-throughput Sequencing and bioinformatics

High-throughput sequencing was conducted by Edinburgh Genomics (University of Edinburgh) and Imperial BRC Genomics facility (Imperial College London) using the Illumina MiSeq platform (2x300nt paired-end reads). Raw sequencing files (fastq files) were demultiplexed by 6nt indices by the respective genomics facility. To remove the plasmid sequence from the second read, reads were trimmed using flexbar (-as ATGTGCAGGGCCGCGAATTCTTA -ao 4 -m 15 -u 30). Reads were then mapped to the GFP library using bowtie2 (-X 750) and filtered using samtools (-f 99).

For Flow-seq data, only variants with a minimum of 1000 reads across all 8 sequencing bins were used for further analysis. For cell fractionation experiments, data with a minimum of 1000 reads across both fractions were used.

Open-source packages available for R were used for generating correlation matrices (corrplot), heatmaps (ggplot2), boxplots (graphics/ggplot2), Venn diagrams (VennDiagram) and multiple regression analyses (relaimpo). The GC3 of all human CCDS (assembly: GRCg38_hg38; only CDS exons) was calculated using R package ‘seqinr’.

The minimum free energy of predicted mRNA secondary structure for GFP variants or their portions was calculated using the hybrid-ss-min program version 3.8 (default settings: NA = RNA, t = 37, [Na+] = 1, [Mg++] = 0, maxloop = 30, prefilter = 2/2).

Assignment of Geneart sequence parameters to each GFP variant was performed by G. Kudla.

3. Dissecting the effects of coding-sequence GC content on protein expression

Codon usage has previously been reported to have extensive effects on multiple steps in gene expression. Here, I first describe the optimisation and validation of a fluorescence-based assay that allows reliable and reproducible measurements of GFP expression in human cells by spectrophotometry and present the results obtained from this screen. Additionally, I investigate the effects of codon usage on the RNA level in transiently and stably transfected HeLa and Hek293 cells and present data linking codon usage to RNA export as well as stability.

3.1. Plasmid design

To study the effects of codon usage on protein expression, I utilise a library of several hundred synonymous sequence variants of the Green Fluorescent Protein GFP published in Kudla et al. (Kudla et al., 2009) which has previously been used for expression studies in E.*coli* and yeast (Shah et al., 2013). The objective for the following experiment was to establish a reliable fluorescence-based assay to quantitatively measure GFP expression in human cells. Throughout this report I utilised two reference GFPs which were previously described (Kudla et al., 2006) and tested in similar experiments: GFP_000 (, same as "GFP" in Kudla et al., 2006; GC3=35%; poorly expressed) and GFP_001 (, same as "EGFP" in Kudla et al., 2006; GC3=96%; highly expressed).

For expression experiments in human cells, selected GFP variants were sub-cloned into *pCI-neo*¸ a commercially available mammalian expression vector for high expression (Promega). GFP expression is driven by a CMV promoter, allowing high and constitutive expression in human cells. This vector also contains a chimeric intron in the 5'UTR upstream of the GFP insertion site. The intron is composed of the 5'-donor site from the first intron of the human beta-globin gene and the branch and 3'-acceptor site from the intron of immunoglobulin gene heavy chain variable region (*pCI-neo* product manual, Promega). It is thought that the presence of introns enhances expression levels by increasing transcript stability and facilitating efficient RNA processing (Choi et al., 1991; Nott et al., 2003). Since not all human genes contain introns, about 12% are estimated to be intronless (Louhichi et al., 2011), I modified *pCI-neo* by removing the intronic sequence by site-directed mutagenesis to allow the separate assessment of codon usage on expression in both spliced and unspliced genes. Immediately downstream of the GFP sequence is a SV40 late polyadenylation signal. A neomycin resistance cassette is also present on the vector backbone, and is used as internal control in some data presented later on in this chapter. As the GFP variant library is stored in Gateway-compatible Entry vectors, I modified both versions of *pCI-neo* for convenience to Gateway Destination vectors which allows easy sub-cloning of GFP variants by homologous recombination (see methods 2.3.5 for details). For simplicity, the expression vectors will be referred to as *pCM3* (no intron) and *pCM4* (with intron) throughout this report. An outline of the basic protocol for this screen is depicted in figure 6. The development of the final assay protocol is described in detail in the following section.

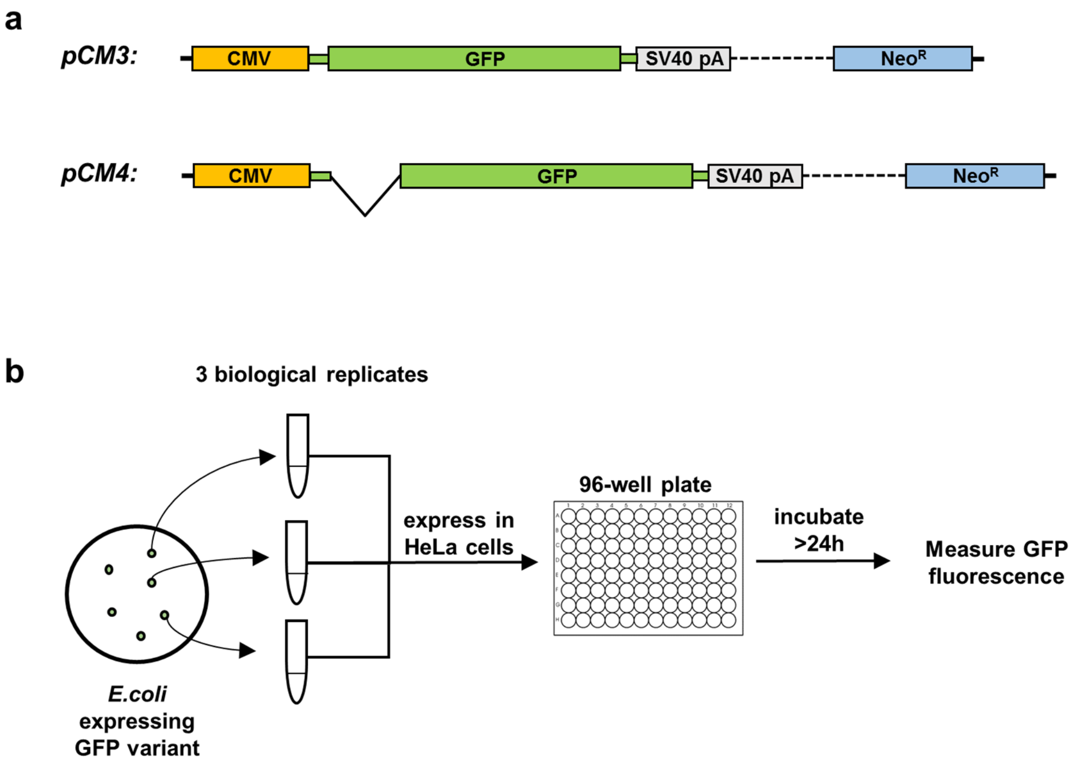


Figure 6. Outline of the protocol to measure expression of GFP variants with and without intron.
a, GFP variants are cloned into CMV-driven expression vectors either containing a chimeric intron in the 5'UTR (*pCM4*) or without intron (*pCM3*). **b**, Plasmid DNA from 3 GFP clones is purified and transiently transfected into HeLa cells in 96-well plates. GFP fluorescence is measured >24hrs post-transfection using a microplate spectrophotometer.

3.2. Assay optimisation

In order to establish a reliable and highly reproducible expression assay, I optimised multiple key parameters and conditions. Various transfection protocols were tested to achieve highest gene expression without trading in cell viability. The most commonly used transfection protocol requires the cells to be seeded one day prior to DNA transfection in order to have an actively dividing population of cells at the time of transfection. This method is commonly referred to as 'forward transfection' (Figure 7a). Although this approach works well for most adherent cell lines, for suspension cells and for high-throughput applications, a different protocol, termed 'reverse transfection', is preferred. In this protocol, cells are transfected at the time of seeding which reduces hands-on cell culture time by one day (Figure 7c). This also eliminates the common occurrence of uneven expression levels of the gene of interest within the culture well. Forward transfection efficiency tends to be highest around the area where the DNA-transfection complex first touches the cells which can lead to strong intra-well variation.

This can be an important factor to consider when assays are to be conducted in multi-well plates such as 96- or 384-well plates, as the highest expression efficiency would then usually correspond to the well centre due to the limited area size (schematically depicted in Figure 7b and d). This may cause strong well-to-well variation if only one or few fluorescence readings are taken from each well. For these reasons, I chose to optimise and implement a reverse transfection approach.

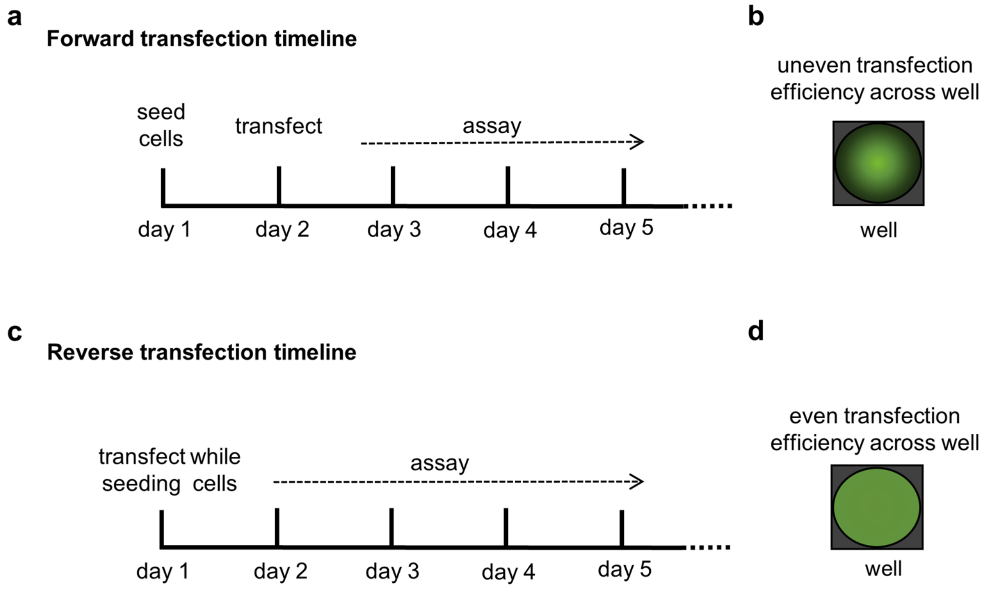
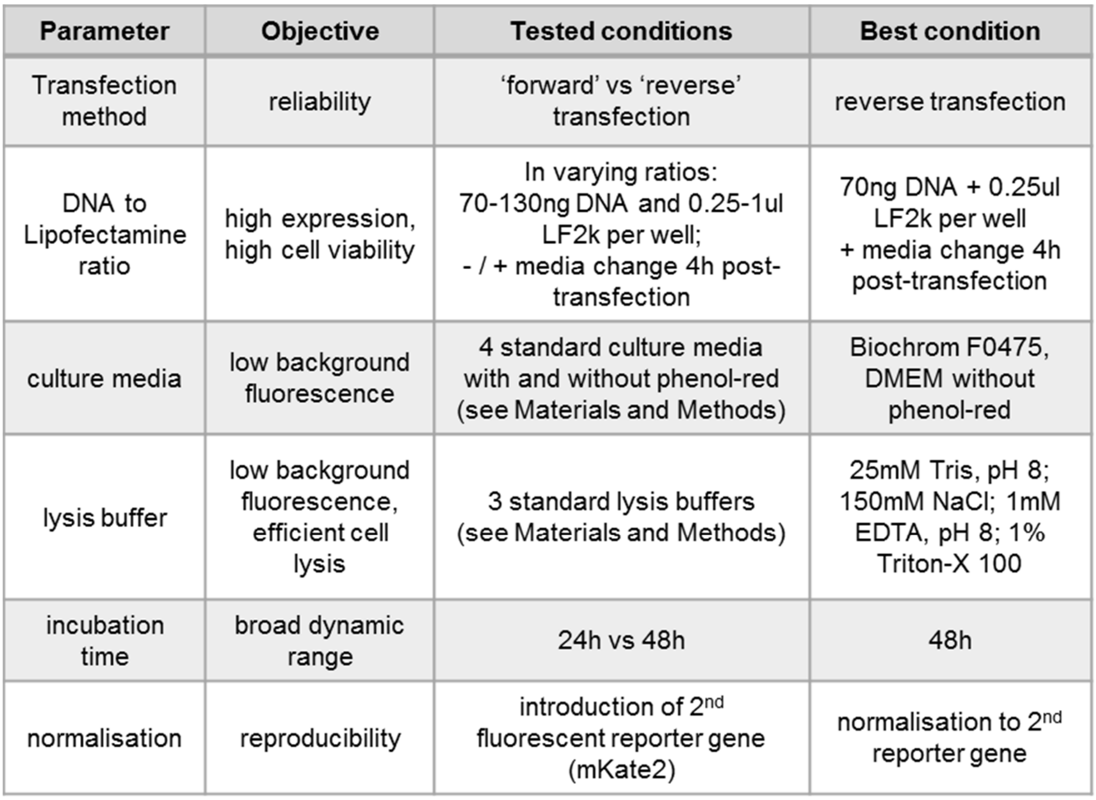

a
Forward transfection timeline
seed cells
transfect
assay
day 1
day 2
day 3
day 4
day 5
b
uneven transfection efficiency across well
well
c
Reverse transfection timeline
transfect while seeding cells
assay
day 1
day 2
day 3
day 4
day 5
d
even transfection efficiency across well
well

Figure 7. Comparison of forward and reverse transfection.
a, In forward transfections, cells are seeded on day 1 and transfections performed on day 2, whereas in **b,** reverse transfections, cells are transfected at the time of seeding on day 1, cutting down the assay time by one day.

To achieve high reproducibility, I optimised further parameters, such as the ratio of DNA to transfection reagent, to ensure high expression without strong effects on cell viability, various culture media and cell lysis buffers to reduce background noise caused by auto fluorescence of buffer components, as well as the incubation time to increase the dynamic range of this assay. An overview of optimised parameters and tested conditions is shown in table 1.



Parameter	Objective	Tested conditions	Best condition
Transfection method	reliability	'forward' vs 'reverse' transfection	reverse transfection
DNA to Lipofectamine ratio	high expression, high cell viability	In varying ratios: 70-130ng DNA and 0.25-1ul LF2k per well; - / + media change 4h post-transfection	70ng DNA + 0.25ul LF2k per well + media change 4h post-transfection
culture media	low background fluorescence	4 standard culture media with and without phenol-red (see Materials and Methods)	Biochrom F0475, DMEM without phenol-red
lysis buffer	low background fluorescence, efficient cell lysis	3 standard lysis buffers (see Materials and Methods)	25mM Tris, pH 8; 150mM NaCl; 1mM EDTA, pH 8; 1% Triton-X 100
incubation time	broad dynamic range	24h vs 48h	48h
normalisation	reproducibility	introduction of 2nd fluorescent reporter gene (mKate2)	normalisation to 2nd reporter gene

Table 1. List of optimised parameters.

A commonly faced issue in fluorescence-based assays is the signal-to-noise ratio caused by buffer components exhibiting auto fluorescence. I first tested whether such noise could be reduced by testing various commercially available standard DMEM (Dulbecco's Modified Eagle Medium) formulations. Media often contains phenol red to allow monitoring of the pH of a growing culture, however, it also strongly interferes with fluorescence measurements due to auto fluorescence. Therefore, I tested several other DMEM formulations that do not contain phenol red to avoid this issue (Figure 8a). Expression levels of GFP_000 were measured in cells grown in four different standard media: *a*, DMEM containing phenol-red ("Red"), *b*, DMEM buffered with HEPES, no phenol-red ("HEPES"), *c*, similar formulation as *a,* no phenol-red ("Biochrom"), and *d*, DMEM marketed as particularly good for imaging applications, no phenol-red ("Fluorobrite"). The signal-to-noise ratio for mKate2 fluorescence is highest when measurements are taken in Biochrom media (Figure 8a, black bars). For GFP, the signal-to-noise ratio is higher when measured in Biochrom and Fluorobrite compared to the other tested media (Figure 8, grey bars). I therefore decided to use Biochrom media for all following experiments as well as due to its reduced cost (about 50% below Fluorobrite).

Fluorescence measurements are also strongly depended on the distribution of cells across the entire well; e.g. since cells are not grown to 100% confluency to avoid contact inhibition, if no cells happen to be attached in the particular area that is excited by the light beam, the obtained result would not necessarily be representative of the entire well. The multiwell plate reader used in this study allows to make multiple independent measurements across the well which I utilised to obtain an average result for each well. Another possible method to avoid the effects of uneven growth is to lyse cells before measuring fluorescence which will release most protein from the cells and, with the inclusion of a shaking step, aids the even distribution of GFP molecules in solution across the well. I tested three standard cell lysis buffers to see whether the sensitivity of measurements would increase. I transfected cells grown in Biochrom media with the poorly expressed GFP_000 and highly expressed GFP_001 variants. Background fluorescence was subtracted from transfected wells and the resulting fluorescence values used to measure the dynamic range of this assay using the ratio between GFP_001 and GFP_000 before and after lysis (Figure 8b). A low dynamic range could be indicative of e.g. high experimental background. The dynamic range was consistently increased after lysis, confirming that the addition of a cell lysis step can increase the sensitivity of this assay further. No prior wash-step is included before the addition of lysis buffer to reduce cell-to-cell variation caused by cells detaching in the process. For this reason, using media with low auto fluorescence is still a crucial feature of the overall protocol.

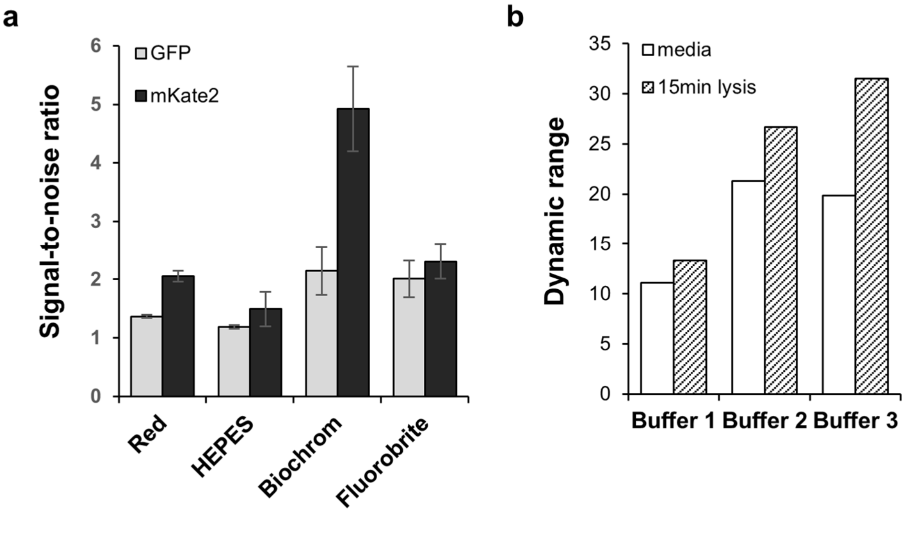


Figure 8. Optimisation of fluorescence measurements.
a, HeLa cells were transfected with GFP_000 and fluorescence measured 24h post-transfection in different culture media (see materials and methods 2.1.4). Bars represent the ratios between GFP or mKate2 signal and their respective background (media only) measurements. Error bars denote the standard deviation. n=3. **b**, The dynamic range was calculated as the ratio of GFP_000 and GFP_001 signal after background normalisation as measured in HeLa cells 24h post-transfection, either measured in media or after 15min lysis using 3 different lysis buffers (see materials and methods 2.1.4). Shown are the means of n=3.

Despite all the optimisation that was undertaken, some variability between replicate wells persisted, in particular between different plasmid preparations (Figure 9a). This is most likely caused by differences in DNA amounts, impurities and/or general transfection efficiency caused by e.g. uneven cell growth. To be able to account for such variation, I cloned a further fluorescent reporter gene into the backbone of the expression vectors as an internal control for transfection efficiency. I chose mKate2 (Evrogen), a far-red fluorescent protein unrelated to GFP, which allows very good spectral separation from GFP and does not cause any cross-excitation during photometric measurements (Figure 9b+c). Normalising GFP expression to mKate2 levels significantly decreases the noise between replicate transfections and thus further increases the dynamic range of this assay (Figure 9d).

A further important parameter for expression-based screens is the actual transfection procedure. The optimisation of transfection protocols is a constant compromise between high expression levels and cell viability as due to the nature of the procedure a certain degree of toxicity is always expected. I therefore conducted a cell viability assay on cells 24h and 48h post-transfection to assess the effects of the transfection procedure on cell health. Figure 10 shows the results obtained for one of such optimisation experiments in which I varied plasmid DNA amount to assess the trade-off of viability for higher GFP expression.

To decrease variation of measurements, I tested three DNA preparations of each GFP variant in triplicate transfections in a 96-well plate (9 wells measured in total per variant). The fluorescence was measured 48h post-transfection and data from all wells was averaged to obtain a relative fluorescence score for each GFP variant. To ensure comparability between different assay plates, three GFP variants were selected and transfected on every plate. By normalising the relative fluorescence values of all measured variants on one plate to these controls, it is possible to directly compare all assayed plates to one another, decreasing error due to technical differences between experiments.

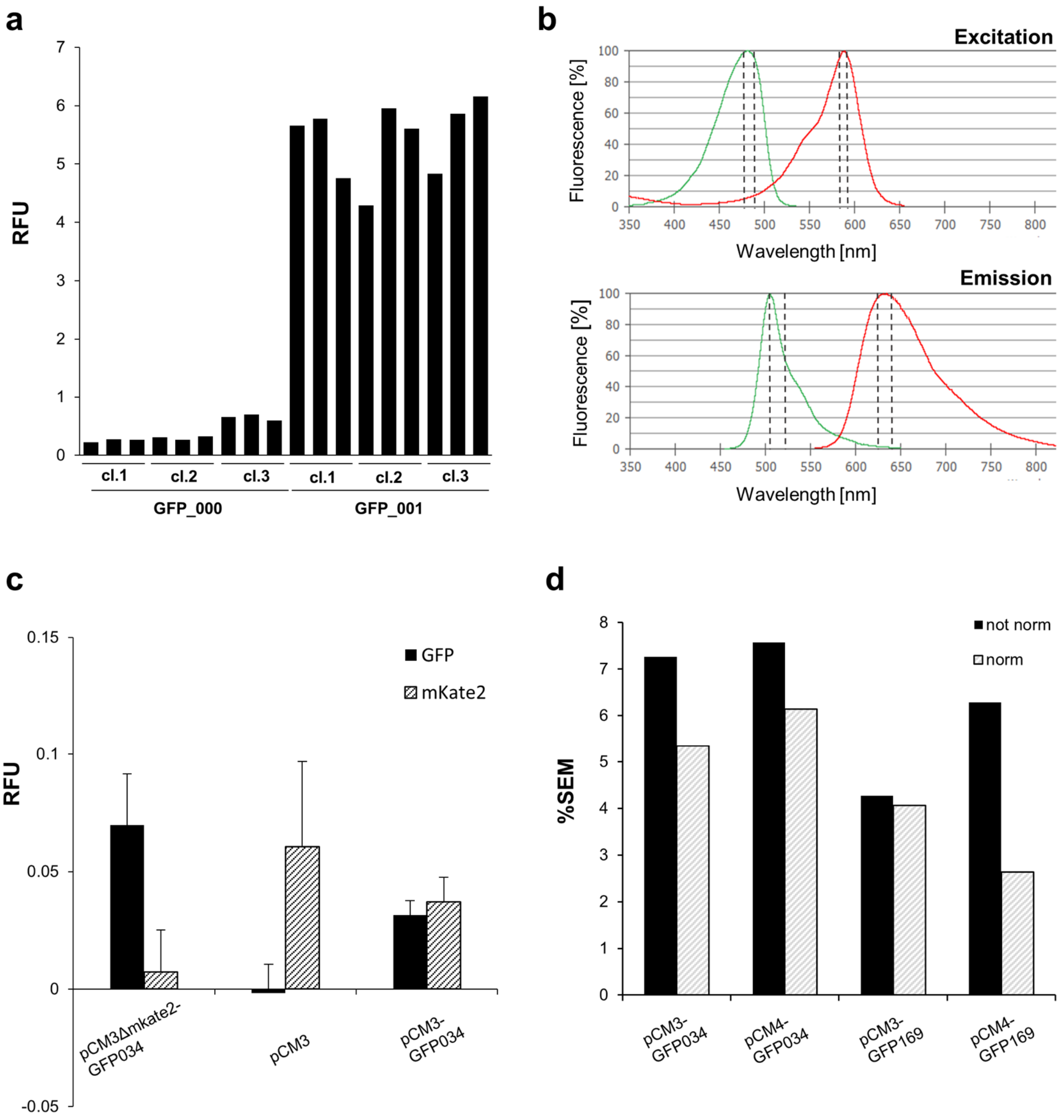


Figure 9. Decreasing well-to-well variability using an internal control.

a, Variability in fluorescence levels between replicates of transfected Hek293 cells. Three separate plasmid preparations of GFP and EGFP were transfected into 3 wells each. Each bar represents the fluorescence measured in one well. **b**, Excitation and Emission spectra of GFP and mKate2. Dotted vertical lines indicate bandwidths used for measurements in this study (GFP Ex/Em: 480-490/505-525; mKate2 Ex/Em: 583-593/623-643). **c**, mKate2 fluorescence of cells transfected with plasmids carrying either GFP only (*pCM3ΔmKate2-GFP034*), mKate2 only (*pCM3*) or both (*pCM3-GFP034*). Error bars=SEM. **d**, Standard error of fluorescence averaged of data from 4 different plasmid preparations of GFP034 and GFP169 in *pCM3* or *pCM4*, each transfected in triplicate. The proportion of the standard error of the mean was calculated as percentage of the average fluorescence value (%SEM) obtained from all measured wells with (white bars) and without normalisation to mKate2 (black bars).

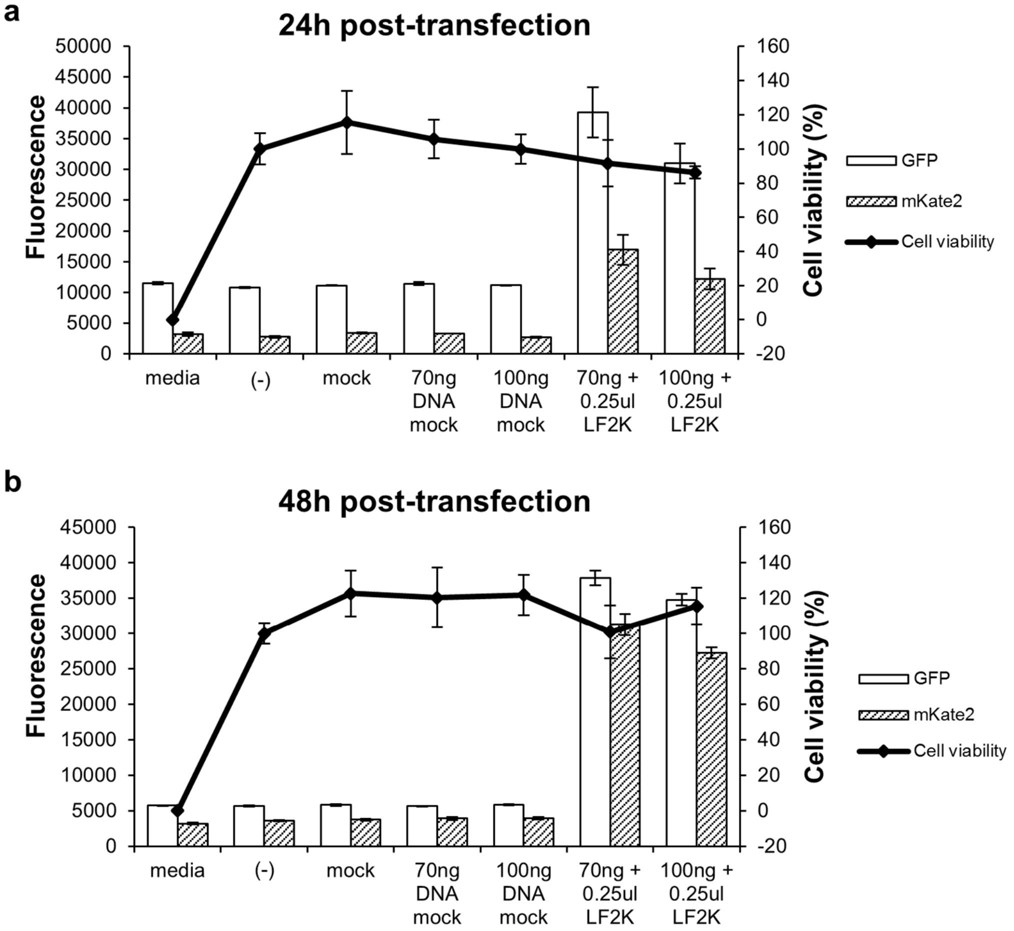


Figure 10. Fluorescence and cell viability optimisation.
9 000 HeLa cells were reverse transfected with indicated amounts of *pCM3* and 0.25ul Lipofectamine 2000 in 96 well plates (3 wells each condition). GFP (white) and mKate2 (pattern) fluorescence was assessed **a**, 24h or **b,** 48h post-transfection in media (Biochrom) before assessing cell viability (black line) using alamarBlue (Invitrogen). Error bars denote the standard deviation.

3.3. High GC content increases protein levels of GFP with and without introns

For an initial screen, I selected 38 synonymous GFP variants which cover a very broad range of GC3 content (0.27-0.97%). Individual GFP variants were cloned into *pCM3* and *pCM4.* To assess expression of each, plasmid DNA from three individual bacterial clones was purified and used in transient transfections in 96-well plates. For each plasmid preparation, three wells of HeLa cells were transfected. Cells were incubated for 48hrs post-transfection before lysing cells and measuring GFP fluorescence using a microplate spectrophotometer to obtain an average expression score for each variant (Figure 11). The tested variants cover a very broad range of expression levels, varying up to 140-fold. Notably, the highest expression was seen for GFP_001 (Figure 11, denoted with an asterisk), which is a highly codon- and expression-optimised version of GFP ("EGFP" in Kudla et al., 2006). None of the other tested variants reached similar levels. A clear trend between GC content and protein levels can be seen.

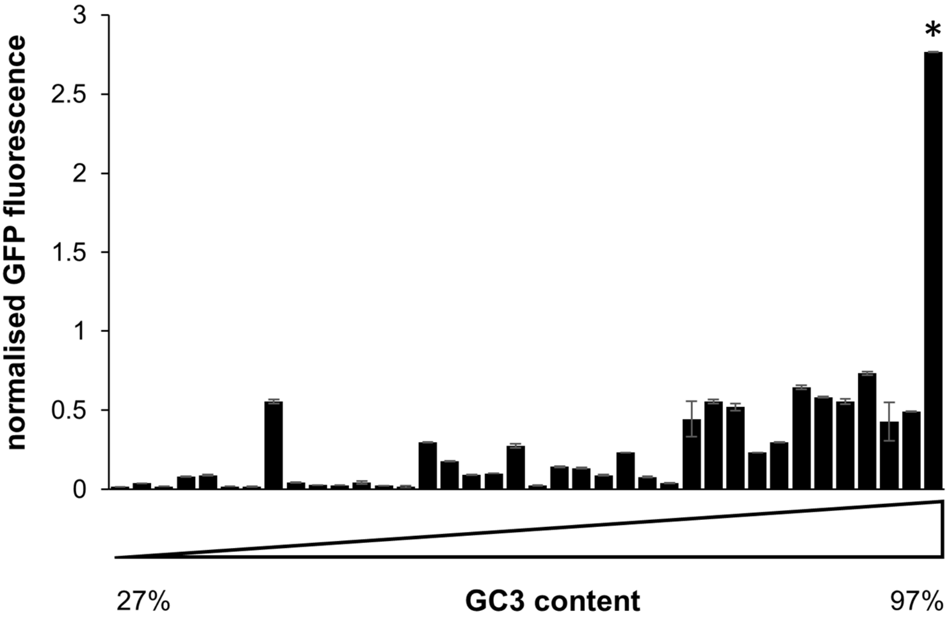


Figure 11. Fluorescence levels of 38 GFP coding-sequence variants.
Variants are arranged by their GC3 content from low to high. The asterisk indicates GFP001. 3 plasmid preparations for each variant were prepared and each measured in 3 wells. Error bars = SEM.

GFP is a naturally intronless gene, however, the vast majority of transcripts encoded in the human genome contain at least one intron. The process of splicing has been well established as an important step in efficient gene expression, primarily due to its significant role in gene regulatory steps such as transcript stabilisation and nucleo-cytoplasmic mRNA export (Furger et al., 2002; Gupta et al., 2013; Valencia et al., 2008). For this reason, I also expressed the same GFP variants from a modified version of the original vector which additionally contains a chimeric intron in the 5'UTR (*pCM4*, see Figure 6a in 3.1). The objective of this experiment is to test whether the large differences in the expression of unspliced GFP variants can be rescued by the inclusion of an intron as it would be expected to increase expression of particularly poorly expressed variants. As can be seen in figure 12, inclusion of an intron exhibits varying effects on fluorescence levels, ranging from no change to an up to 10-fold increase in expression. Overall, the strong relationship between GC content and protein levels remains (R^2=0.5395, *p*=3e^{-8}).

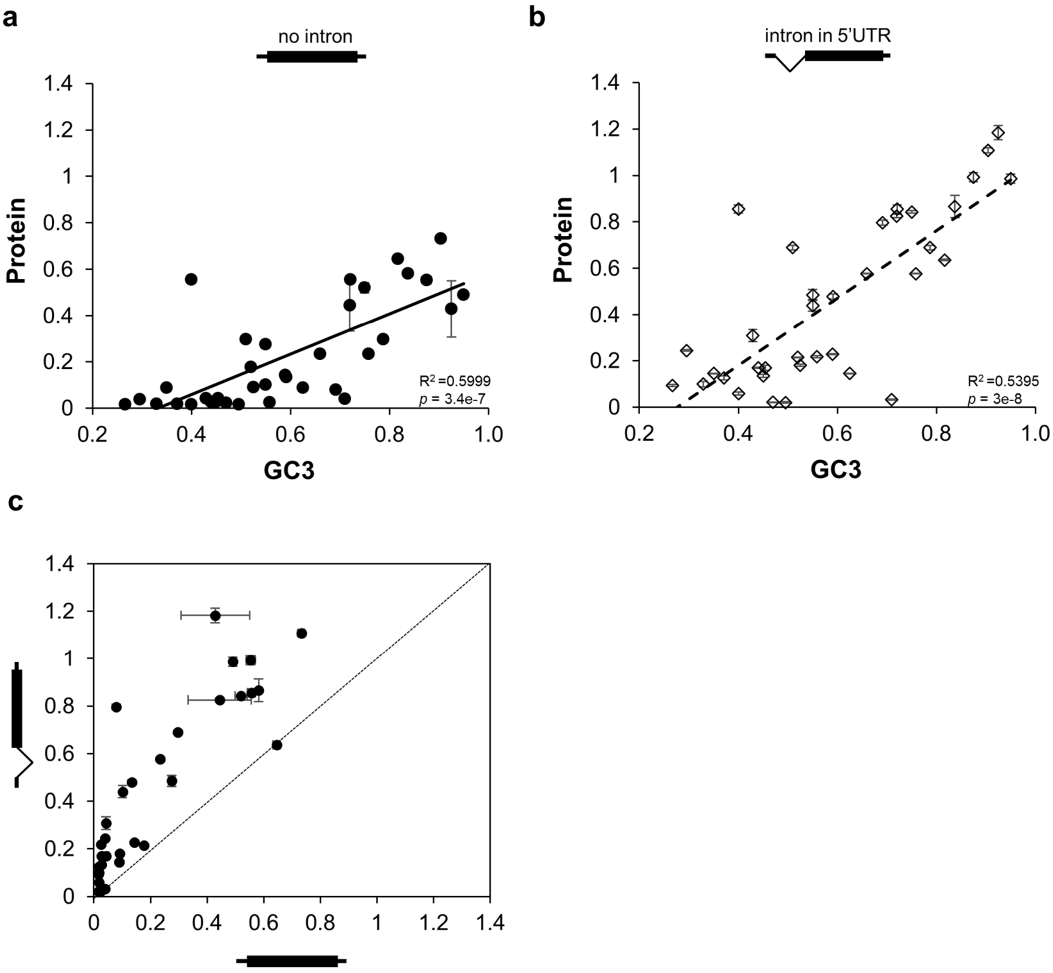


Figure 12. Protein levels of 36 GFP variants.
HeLa cells were transfected with plasmids **a**, *pCM3* (no intron, black circles) or **b**, *pCM4* (with intron in 5'UTR, white diamonds) encoding 36 GFP variants using the protocol outlined above and fluorescence used as proxy for protein levels plotted against GC3 content. **c**, The effect of an intron in the 5'UTR on protein levels of GFP. The diagonal line illustrates x=y. Error bars = SEM. n=9.

As the above experiments are based on strong overexpression of GFP from plasmids, it was important to confirm that these effects can also be seen in a stable cell line system under more physiological conditions, and to confirm that previously published data is reproducible (Kudla et al., 2006). To do so, I established stable HeLa and Hek293 cell lines expressing two GFP variants – GC-poor GFP_000 (GC3=0.33) and GC-rich GFP_001 (GC3=0.97), both with or without an intron from the same genomic locus, to avoid context-dependent effects. As shown in figure 13, both variants exhibit similar differences in expression levels as would be expected from transient transfection experiments, both in HeLa and Hek293 cells. The presence of an intron increases expression of GFP_000 in both cell lines, whereas in case of GFP_001, protein levels are either unaffected (HeLa) or slightly decreased (Hek293).

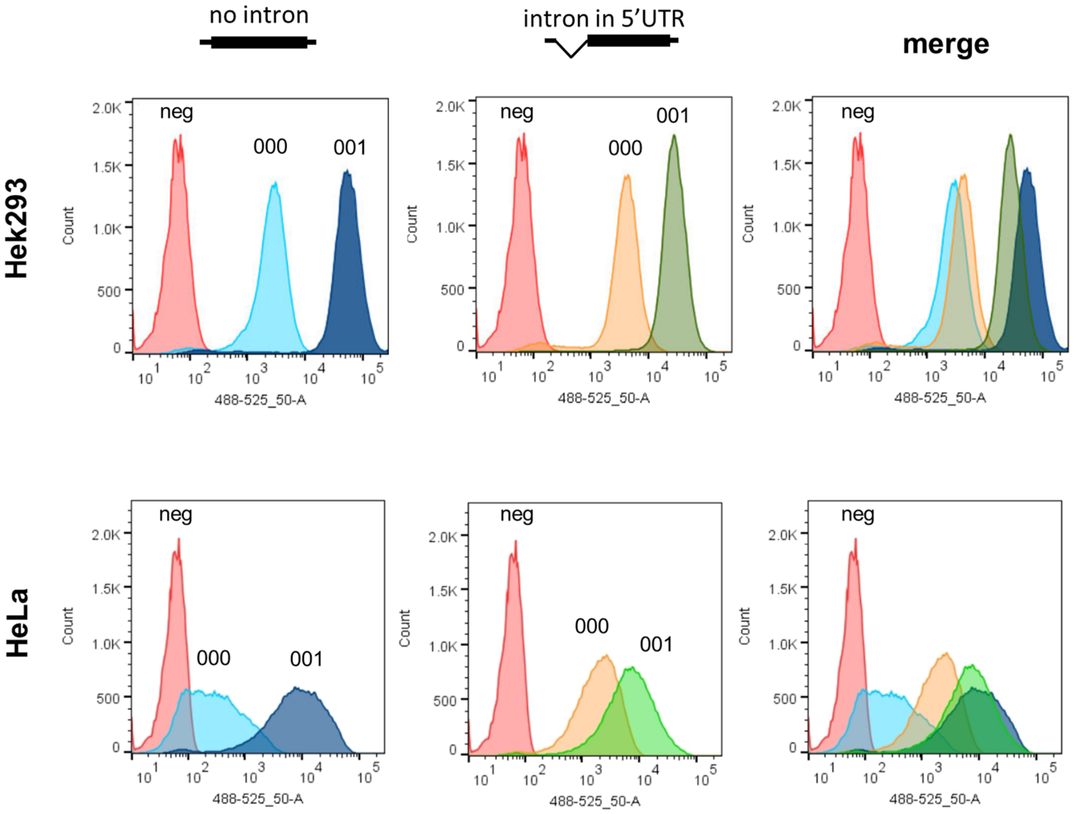


Figure 13. Expression of GFP_000 and GFP_001 in stable cell lines.
Stable Hek293 (top row) and stable HeLa Flp-in (bottom row) expression GC-poor GFP_000 and GC-rich GFP_001 without (left column) or with (centre column) intron in their 5'UTR. Expression measured 24hrs post-induction by Flow-cytometry.

3.4. High GC content increases mRNA levels of GFP

Previous published data established that the GC content of genes not only increases protein yield but also correlates with increased mRNA levels (Kudla *et al*., 2006). To test whether these changes can be seen across the GFP library, I selected 24 GFP variants which cover a broad range of GC-content and for which protein levels were previously quantified (see 3.3). These were transiently transfected into HeLa cells and mRNA levels assessed by quantitative Real-time PCR using primers schematically depicted in figure 14a.

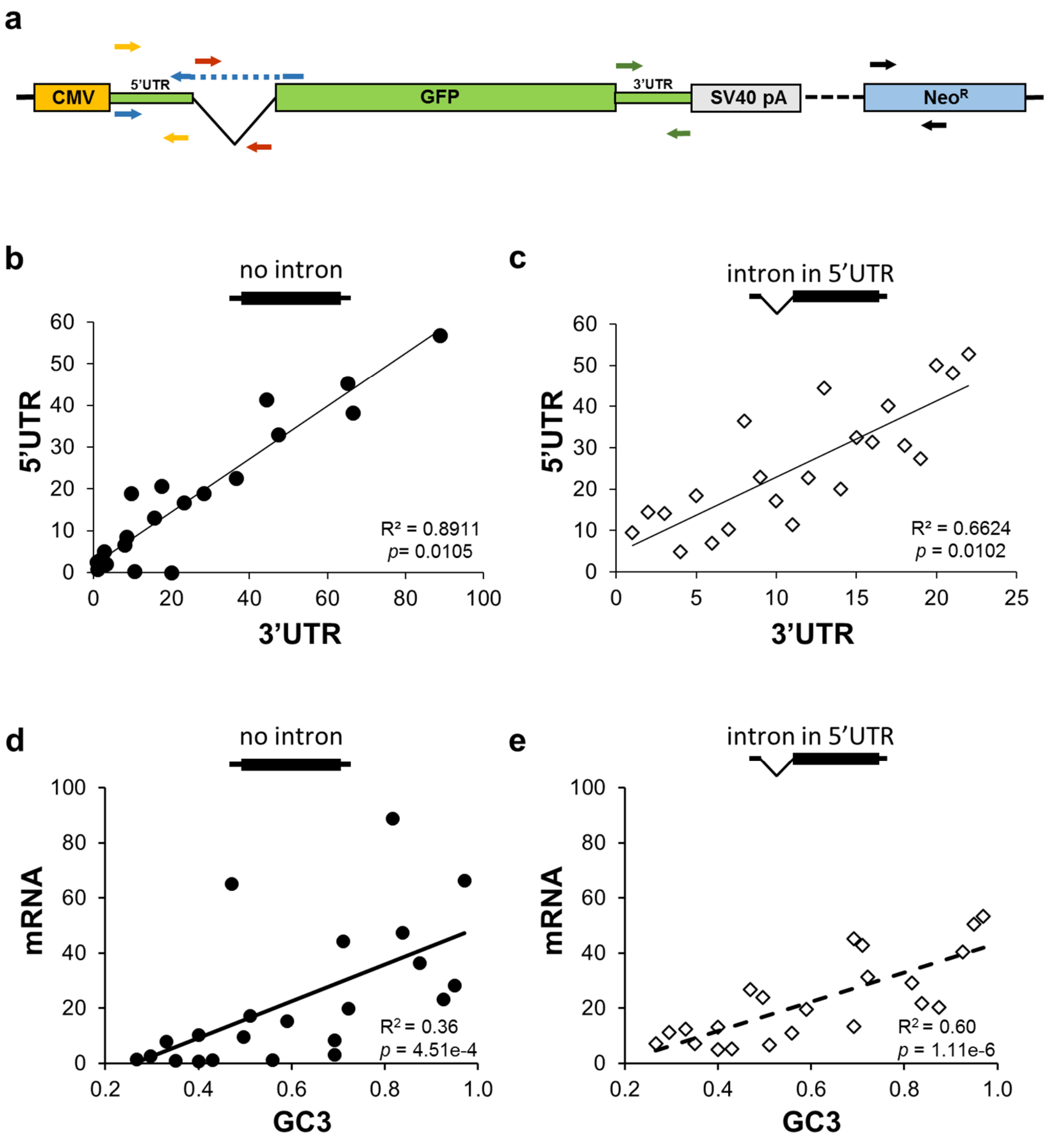


Figure 14. RNA levels of GFP variants. a, Schematic representation of annealing sites of primers used to measure the abundance of 5'UTR (orange), 3'UTR (green), spliced RNA (blue) and unspliced RNA (red). **b** + **c**, Correlations of results obtained with either 5'UTR or 3'UTR primers of 24 GFP variants. **d**, Relative RNA levels of 24 GFP variants expressed from *pCM3* (black circles) or *pCM4* (white diamonds) using primers annealing in the 5'UTR (indicated by orange arrows in **a**).

Similarly as with protein levels, high GC3-content correlates with high mRNA levels (Figure 14d, R^2=0.36, p=4.51e^{-4}). The same variants were expressed and quantified in an intron-containing version of the expression vector as before (see Figure 12), to test whether the inclusion of an intron could rescue the low mRNA levels of particularly GC-poor GFPs. However, the correlation with GC content remains (R^2=0.60, p=1.11e^{-6}).

In order to visualise the correlation between RNA and protein levels, I plotted data for those variants, for which I obtained measurements for both parameters, in figure 15. For both variants expressed with and without intron, the data correlates, however only marginally significantly for those without intron (R^2=0.2093, *p*=0.04259) and insignificantly for those with intron (R^2=0.184, *p*=0.05916). To further illustrate the effect of an intron on expression level, the fold change in translational yield (fluorescence/mRNA) that occurs when an intron is introduced, is depicted in figure 15c. The effects vary across all tested variants and no clear trend is visible; however, most variants with a positive fold change are in the upper GC3 range.

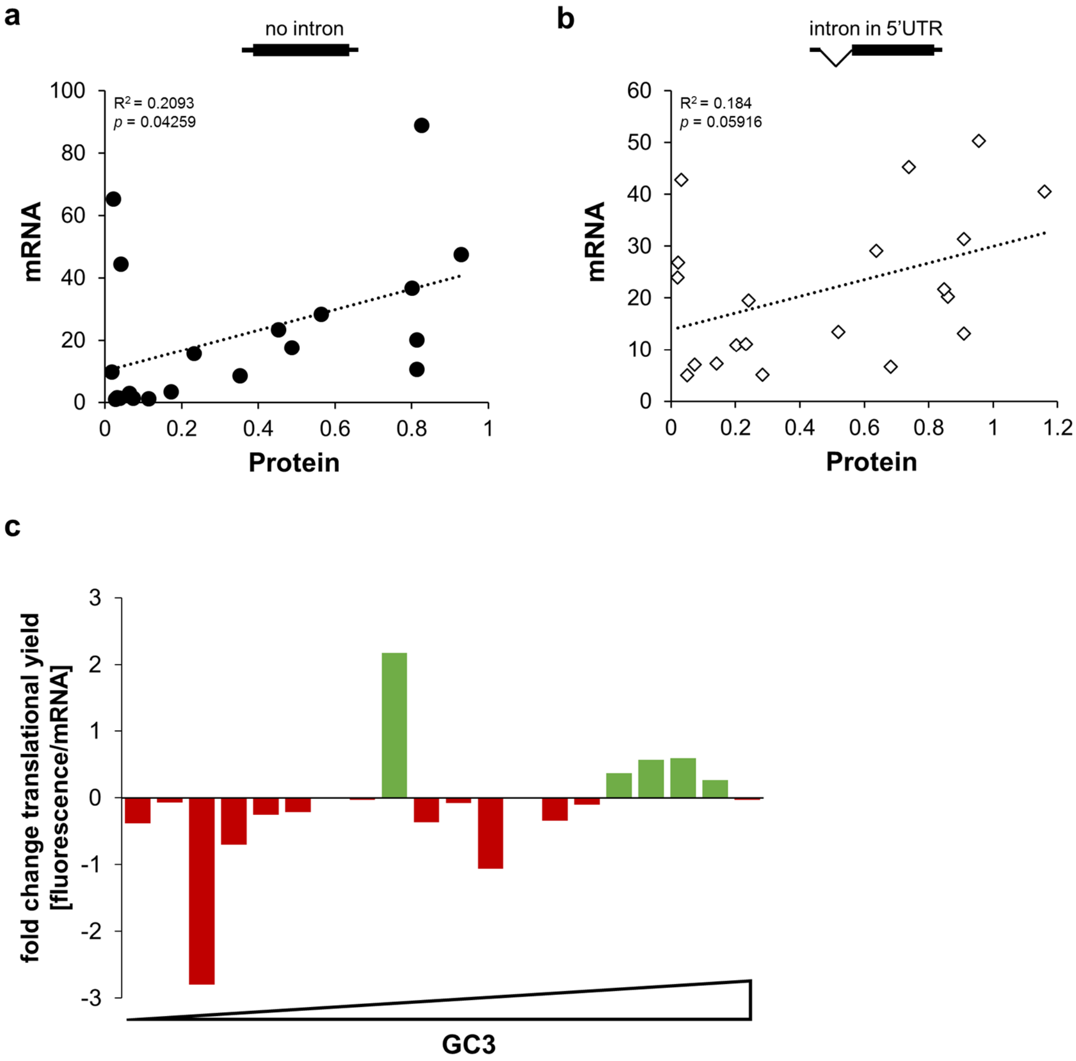


Figure 15. Relationship between RNA and protein levels of 20 GFP variants.
a, b, RNA levels are plotted against fluorescence measurements shown in figure 11 for 20 variants expressed without intron (**a**) or with intron (**b**). **c**, Shown is the fold change in translational yield (fluorescence/mRNA) that occurs when an intron is present. Variants are arranged by their GC3 content as indicated by the triangle.

To confirm previously published results on GFP_000 and GFP_001 showing a comparable decrease in mRNA levels as on protein levels (Kudla et al., 2006), I quantified mRNA expression of those variants in HeLa and Hek293 cells. Results in figure 16 confirm that in both cell types GFP_001 shows several-fold higher mRNA levels than GFP_000. In addition, I also measured expression in the presence of an intron which shows an increase of overall mRNA levels of GFP_000 in both HeLa and Hek293 by 2-fold, however only increases GFP_001 in HeLa, but not Hek293. Semi-quantitative PCR of plasmid preparations using primers specific for spliced (Figure 14, blue arrows) or unspliced RNA (Figure 14 , red arrows) showed similar PCR amplification efficiencies of GFP_000 (GC3=0.33) and GFP_001 (GC3=0.97), confirming that the previous results are not influenced by a potential bias in PCR amplification efficiency (Figure 17).

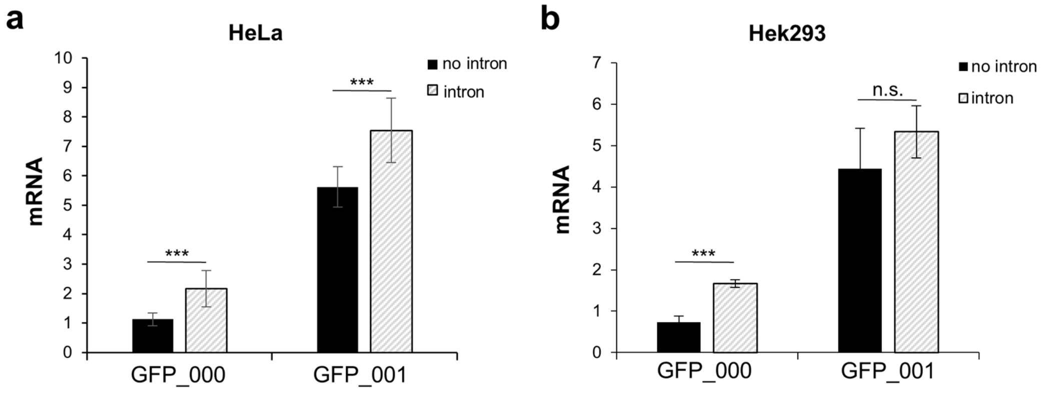


Figure 16. RNA levels of GFP_000 and GFP_001 in HeLa and Hek293 cells.
RNA expression levels of GFP_000 and 001 with (white bars) or without (black bars) intron in 5'UTR, measured in **a,** HeLa and **b**, Hek293 cells 24hrs post-transfection. Splicing increases expression in HeLa cells (GFP_000: *p*=3e^{-4}, GFP_001: *p*=3e^{-3}), but only for GFP_001 in Hek293 cells (GFP_000: *p*=3.96e^{-5}). Error bars = StDev; n.s.= not significant. *n*=3.

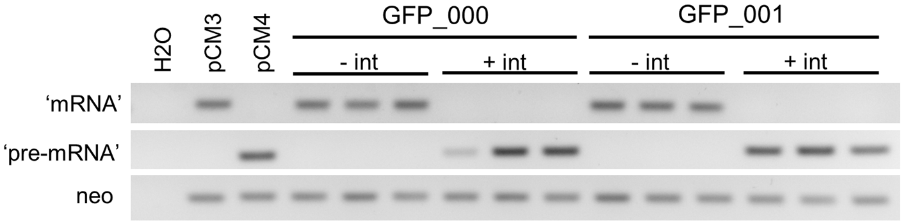


Figure 17. PCR amplification from vectors used in transfection experiments. GC-poor GFP_000 and 001 were cloned into *pCM3* (-int) or *pCM4* (+int) and 3 separate plasmid preparations of each used as template in PCR amplification using primers used for mRNA and pre-mRNA quantification in figure 14. The neomycin resistance gene (neo) is used as internal control. PCR products were resolved on a 2% agarose/TBE gel. The result is representative of three experiments.

3.5. GC content does not affect pre-mRNA levels of spliced GFP variants

The differences in RNA levels between GFP variants could indicate changes in transcription initiation and/or elongation rates. To further investigate the possibility of differential transcription rates, pre-mRNA levels of several additional variants containing an intron were measured using intron-specific primers (Figure 18). No large differences between GFP variants and no correlation between GC-content and pre-mRNA levels can be seen. These results suggest that transcription is not the main factor contributing to overall mRNA levels. However, these results could be interpreted as showing the abundance of intron-containing transcripts only and not full-length GFP transcripts. Therefore, these results do not take into account the possibility of 5' and 3' truncated transcript populations. To test this, I previously compared the correlation of transcript measurements using primers specific for either 3'UTR or 5'UTR (Figure 14b+c) which correlated well for unspliced variants (r^2=0.88, p=0.0105) as well as spliced variants (R^2=0.6624, p=0.0102), suggesting that there is no over-abundance of 5' or 3' truncated RNA species. Similar measurements were conducted with GFP_000 and 001 in HeLa and Hek293 cells (Figure 18b+c). In both cell lines, pre-mRNA levels of GFP_001 are significantly increased by 1.5-fold in HeLa (p=0.0262) and 2.8-fold in Hek293 (p=2.576e^{-12}), similarly to some of the transiently expressed variants shown in figure 18a. Overall the results from these experiments suggest that changes in GC-content leave pre-mRNA levels largely unaffected.

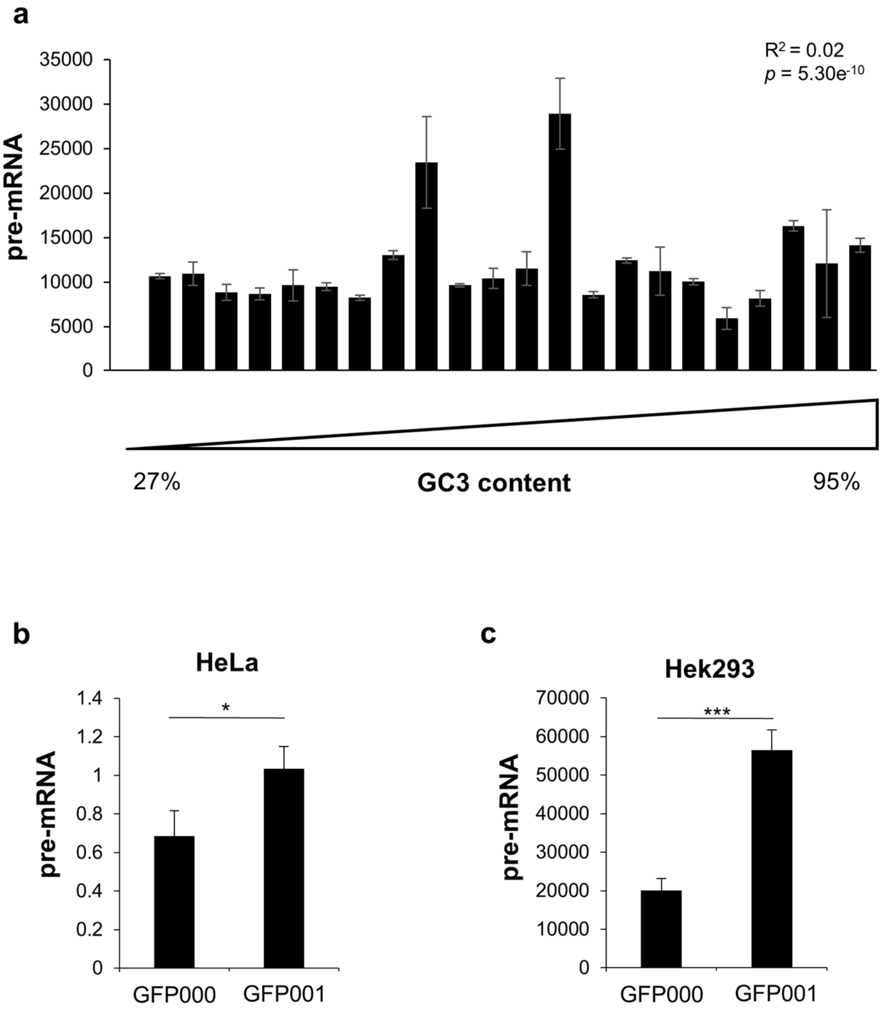


Figure 18. Pre-mRNA levels of GFP variants.
a, HeLa cells were transfected with *pCM4* encoding 24 GFP variants. 24h post-transfection, total RNA was extracted and pre-mRNA levels analysed by qRT-PCR using primers as depicted in figure 14 (red arrows). Variants are arranged by their GC3 content as indicated by the triangle from low to high. **b,** RNA from HeLa and Hek293 cells transfected with either GFP_000 or GFP_001 with introns was extracted 24h post-transfection and analysed by qRT-PCR as in **a**. Error bars = SEM.

3.6. GC content affects RNA localisation of GFP_000 and GFP_001 mRNA

Several studies have previously shown that the expression of very GC-poor late viral genes requires the activity of an early viral adapter protein, Rev, which facilitates nuclear RNA export (Malim et al., 1989). This requirement can be circumvented by increasing the GC-content of those genes (Kotsopoulou et al., 2000; Tan et al., 1995). We therefore hypothesised that this is a more universal effect and that mRNA export could be a rate-limiting step in the expression of particularly GC-poor GFP variants.

One assumption that can be made when considering mRNA export as a bottleneck in gene expression is that the nuclear to cytoplasmic ratio for a particular RNA will be higher compared to RNA which is not limited by export. Therefore, I performed cellular fractionation of Hek293 cells stably expressing either GC-poor GFP_000 (GC3=0.33) or GC-rich GFP_001 (GC3=0.97) prior to RNA extraction and quantification by qRT-PCR (Figure 19). Two endogenous genes were used as controls to account for the quality of the cell fractionation: U6 snRNA, which is mainly localised in the nucleus, and tRNA-Lys(CTT), which is predominantly localised to the cytoplasm. GFP mRNA levels were normalised to the respective controls in each fraction. The fold difference between GFP_000 and GFP_001 is greater in the cytoplasmic fraction (5-fold, p=1.378e-^{5}) than in the nucleus (n.s., p=0.0717).

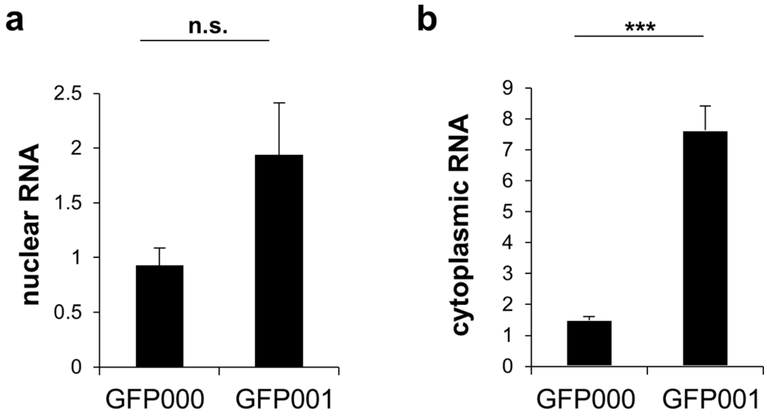


Figure 19. Subcellular localisation of GFP000 and GFP001 RNA.
Stable Hek293 cells expressing either variant were induced for 24h before cellular fraction and qRT-PCR analysis. **a,** No difference between GFP_000 and GFP_001 can be seen in the nuclear fraction (*p*=0.0717). **b,** GFP_001 is 5-fold more abundant than GFP_000 in the cytoplasmic fraction (*p*=1.378e^{-5}). Error bars = SEM. n.s. = not significant. *n*=3.

3.7. GC content affects RNA stability of GFP_000 and GFP_001

RNA stability assays on GFP_000 and GFP_001 were previously published and reported no significant difference in half-lives ("GFP" and "EGFP" in Kudla *et al.*, 2006). However, these assays were conducted by blocking transcription using Actinomycin D, which is known to intercalate with GC-rich sequences and is therefore not suitable for comparing GC-poor genes with very GC-rich genes (Bailey et al., 1993). Because of the known sequence bias, as well as to verify previous results, I repeated these experiments using Triptolide, a sequence-independent reagent (Titov et al., 2011). Unlike previous published data, RNA stability of GC-rich GFP_001 ($t_{1/2}$=8.6h) is about 3.5-fold higher than of GC-poor GFP_000 ($t_{1/2}$=2.4h) (Figure 20).

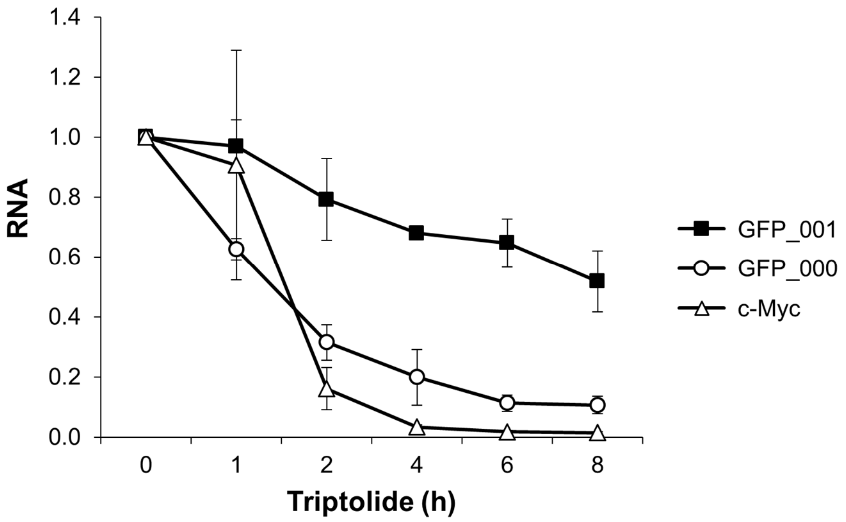


Figure 20. RNA stability GFP_000 and 001 in stable Hek293 cell lines.
Cells were induced for 24h before treatment with 500mM Triptolide. At the indicated times, RNA was isolated and quantified by qRT-PCR. RNA levels were normalised to 7sk, a RNA polymerase III transcribed gene which remains unaffected by Triptolide. RNA levels of c-Myc (triangles) are shown as an unstable RNA control ($t_{1/2}$=1.2h). The half-life of GC-rich GFP_001 ($t_{1/2}$=8.6h, black squares) is about 3.5-fold longer than of GC-poor GFP_000 ($t_{1/2}$=2.4h, circles). Error bars = SEM. *n*=2.

3.8. Changes in codon usage can lead to aberrant splice-site recognition

As synonymous codon changes are often associated with splicing defects, it is possible that similar effects could be seen across our GFP library. To study this, I performed a semi-quantitative RT-PCR using primers which are specific for the 3’ and 5’UTR of GFP to amplify all transcripts which contain both UTR sequences. This should lead to the amplification of full length GFP only, however, if varying codon usage causes cryptic splicing of GFP, more than one product will be amplified. When analysing PCR products by agarose electrophoresis, for most GFP variants tested, additional smaller products are visible, indicating the presence of splice isoforms (Figure 21a). To verify that these products are indeed GFP and not unspecific by-products, some of these products were Sanger-sequenced (Figure 21b). A sequence alignment confirms that these products are derived from GFP but noticeably often seem to lack the same part, starting just after the start codon and reaching up to roughly 217nt into the sequence. These sites happen to coincide with regions that are conserved in most GFP variants at these positions. When looking at sequence features around these sites more closely, a similarity to the consensus exon-intron and intron-exon boundary sequence can be seen. This finding suggests that variation in codon usage around those conserved sites strongly affects their false recognition as exon-intron boundaries.

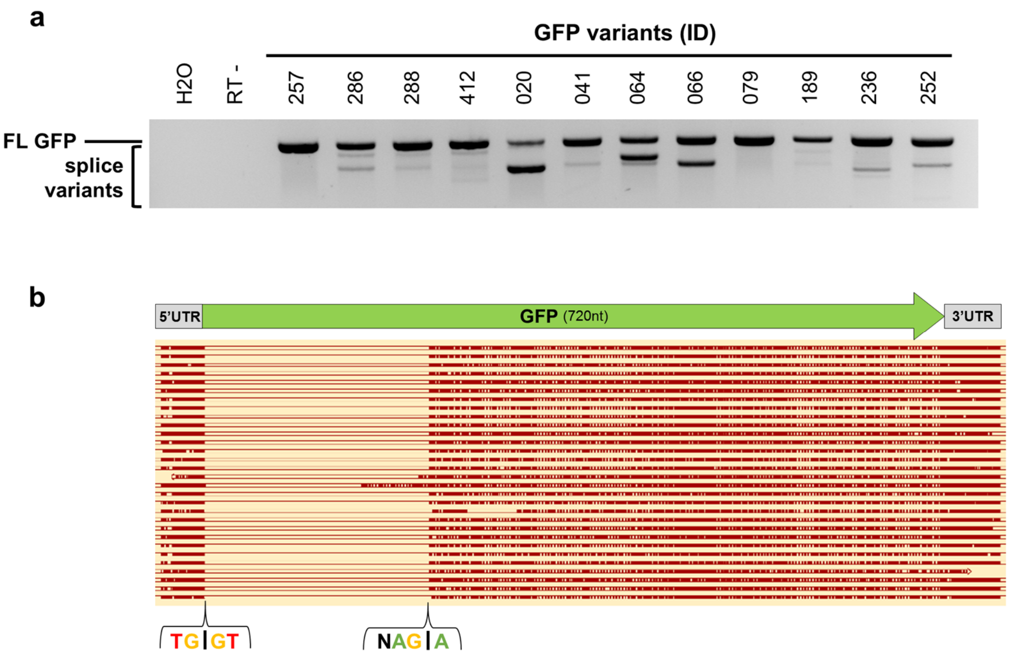


Figure 21. Cryptic splicing of GFP variants.
a, Hek293 Flp-in cells were transfected with *pCM3* encoding several GFP variants. RNA was extracted 24h post-transfection and analysed by RT-PCR using GFP UTR specific primers. Products were resolved on a 1% TBE/agarose gel. **b,** Sequence alignment of transcript isoforms of several GFP variants. Indicated are the conserved nucleotides directly at the cryptic splice site boundaries. N = any.

3.9. Discussion

In this chapter I describe experiments that show the effects of codon usage on several stages of expression in a set of synonymous variants of the naturally intronless GFP gene in human cells. Data presented here confirm and expand results from a previously published study (Kudla et al., 2006) by demonstrating that high GC content leads to increased RNA levels, likely due to increased transcript stability, and ultimately higher protein expression across many coding-sequence variants of GFP. I additionally show that by introducing an intron into the 5'UTR, the poor expression of particularly GC-poor GFP variants can only partially be rescued. I also provide evidence that codon variation leads to cryptic splicing at defined sites.

3.9.1. The effects of splicing on GFP expression

Fluorescence measurements of 36 GFP synonymous coding variants show a high correlation between GC3 content and protein levels (Figure 11). One possible explanation for the low expression of particularly GC-poor GFP variants could be the lack of an intron as it has often been suggested that mRNA splicing is required for efficient gene expression. The presence of introns has been demonstrated to significantly contribute to gene expression through transcript stabilisation (Choi et al., 1991; Nott et al., 2003). It would therefore be expected that the presence of an intron should at least rescue the low RNA levels of poorly expressed variants. Although splicing does increase mRNA levels of some variants, the majority of effects are very small (Figure 14). GC-content remains highly correlated with RNA levels across all tested variants, despite the presence of an intron. It should be noted that the qRT-PCR quantifications of mRNA levels shown here do not take into account the possibility of truncated transcript populations as only the presence of UTR sequences is measured. The measurements of 5'UTR and 3'UTR fragments for unspliced variants were highly correlated (R^2=0.8911, *p*=0.0105, Figure 14b), however there is considerably more variation for spliced GFPs (R^2=0.6624, *p*=0.0102, Figure 14c). The observed differences in the abundance of 5'UTR and 3'UTR sequences might be indicative of transcript degradation. It is likely that the amount of degradation products will also vary as individual sequence features would be expected to lead to changes in mRNP composition and thus differences in stability. It is unclear however, whether the trends seen here are due to a general increase in transcript stability with increasing GC-content, or whether this is directly coupled to an increase in ribosome occupancy (Nott et al., 2004).

It was previously shown that the deposition of the exon-junction complex (EJC) on an mRNA promotes the assembly of mRNPs which either inhibit the association with translational repressors, or alternatively promote the formation of translationally more active mRNP which e.g. facilitate translation initiation (Abaza and Gebauer, 2008; Nott et al., 2004). By introducing an intron into the 5'UTR of the GFP gene, the low protein levels of some variants could partially be rescued, but still remain relatively low compared to those that already exhibited medium to high expression (Figure 12c, R^2=0.7716). No clear trend can be seen when comparing translational yields between variants expressed with and without intron as effects vary from moderate positive or negative effects to up to 3- fold differences. However, variants with high GC3 seem more likely positively affected by the intron compared to those with low GC3 (Figure 15c), although many exceptions are visible, such as GFP_236, which has the highest increase in translational yield (2.17-fold) but is only moderately GC-rich (GC3=0.51), as well as GFP_422, which is the variant with highest GC3 of the tested variants (GC3=0.95) but has a negligible decrease in translational yield (-0.03-fold). Taken together, it is not clear whether these results are caused by a general effect of GC on the translational yield. Northern blotting with probes targeted to 3'UTR and 5'UTR could provide a better overview of all RNA species present for particular variants. Combining these results with polysome profiling to observe changes in the translational state of unspliced compared to spliced variants, may provide a more comprehensive insight of the effects of splicing on GFP expression. Overall, these results suggest that splicing may increase the translational output of already highly expressed GC-rich variants but not of overall poorly expressed GC-poor variants. It is possible that poor expression of these variants in particular is caused by undesirable sequence features already affecting processes further upstream (addressed in chapter 5).

Since splicing occurs co-transcriptionally, the possibility of decreased transcription rates leading to lower mRNA levels cannot be excluded. For this reason I assessed pre-mRNA levels for variants expressed with an intron which showed no large changes across the 24 variants tested (Figure 18a). In this experiment, the amount of intronic sequence was quantified, not the full length pre-mRNA transcript. Nonetheless, this data suggests that transcription of all variants is actively initiating, regardless of sequence-composition. *In vitro* testing of the ability of DNA polymerases to amplify a GC-poor and a GC-rich variant showed no obvious differences, indicating that DNA topology of the plasmid constructs is not significantly obstructing transcription (Figure 17). In contrast, transcription of several GFP variants covering a broad range of GC-content *in vitro* lead to significant differences in the RNA yield obtained for each (data not shown) suggesting that for at least some variants, there might be effects on either transcription initiation or elongation that could possibly also affect RNA levels *in vivo*. It has been shown that transcriptional silencing of transgenes is a common occurrence in stable cell line systems which is of particular interest for gene replacement studies. It was demonstrated that expression downregulation is coupled to nucleosome re-positioning towards the 5'end of transgenes, leading to a less accessible chromatin structure prohibiting efficient transcription (Bauer et al., 2010). In the same study, it was also suggested that this process is linked to CpG content, with a CpG depleted GFP variant (CpG=0; GC3=0.8) exhibiting 1.6 to 2-fold lower protein levels compared to a CpG enriched variant (CpG=60; GC3=0.96) in Hek293 cells. In contrast, I present data of two GFP variants with same CpG contents as in the study by Bauer et al., though differing significantly stronger in GC3 (GC3=0.33 vs GC3=0.96) which results in an up to 20-fold difference in protein expression (Figure 13). qRT-PCR data for pre-mRNA levels suggest only a 2.5-fold difference (Figure 18b) whereas Bauer et al. observed an 7-fold increase in newly synthesised mRNA in nuclear run-on assays. Whether transcription is indeed a major contributing factor of differential expression of GFP variants used here, a nuclear run-on assay should be performed in order to assess the transcription dynamics further and to test whether any differences are facilitated by either GC3 or CpG content, or a combination of both. Taken together, these results suggest that differences in transcription may be involved in differential expression of GFP variants, however, since no large variation could be seen on the pre-mRNA level in transiently transfected cells, changes on GFP expression are unlikely to be mediated by the same mechanisms, i.e. nucleosome re-positioning, as suggested by Bauer et al. To further confirm my findings however, Northern blotting should be performed to assess full-length pre-mRNA levels more quantitatively ideally in stable cell lines.

3.9.2. Codon usage and cryptic splicing

The effects of single synonymous nucleotide polymorphisms (sSNPs) on splicing have been extensively studied as well as the effects of the local GC content on alternative splicing (Amit et al., 2012). In case of stable Hek293 cell lines expressing GFP_001, the introduction of a 5'UTR intron leads to a decrease in fluorescence. This is surprising as so far no study has shown any negative effects of splicing on gene expression. In this particular case, the splicing process could either be interfering with other mechanisms, e.g. structural changes and the binding of splicing factors could lead to spatial competition with other RNBPs important for high expression, or, the presence of cryptic splice acceptor sites within the coding region could in combination with the strong splice donor site of the 5'UTR intron be acting as alternative splice site, leading to non-functional transcript isoforms. Evidence for the latter scenario is given by my finding that for various GFP variants more than one transcript isoform can be detected by RT-PCR (Figure 21), suggesting that cryptic splicing is indeed occurring. I observed that for many of those the same sequence fragment is often missing (Figure 21b). This fragment is located between two sites that are preserved across most of the GFP sequences: the 5' splice site is located right after the ATG and the 3' splice site overlaps with an *Xba*I restriction site which was utilised for the initial assembly of the variant library (Kudla et al., 2009). These sites happen to weakly resemble consensus splice donor and splice acceptor sequences (GU/AG) and are therefore likely the cause of the aberrant removal of this particular sequence fragment. Since this particular transcript isoform is not observed in all variants and also varies in the extent of occurrence (e.g. for GFP_020 it represents the majority of transcripts, see Figure 21b), the codon choice surrounding these sites are possibly mediating the strength of the cryptic splice site recognition.

A study by Amit et al. (2012) focussed on the importance of GC-content between exons and introns for splice site selection and more specifically how exon skipping and intron retention is controlled by defined GC-boundaries (Amit et al., 2012). I hypothesise that similar mechanisms are acting in my system and in order to further investigate the causes for aberrant GFP splicing, it would therefore be interesting to select GFP variants that primarily differ in GC-content within the fragment that is most often removed. If GC content is the main driver of the cryptic splice phenotype, variants with low and high GC should exhibit clear differences. Due to the high overall expression levels of GC-rich variants, I would expect those variants to be less likely spliced cryptically than others. This does not exclude the possibility of particular sequence motifs playing an additional role in splice site selection. Amit et al. further describe a link between chromatin architecture and the recruitment of the splicing machinery by demonstrating how nucleosome positioning correlates with GC content and ultimately with splice site recognition (Amit et al., 2012). It is however unclear whether the here presented observation could be mediated by similar mechanisms as despite literature suggesting that transfected plasmid DNA could be assembled into nucleosome-like structures, it is not known to what extent these structures reflect a stable cellular genomic DNA context (Jeong and Stein, 1994; Mladenova et al., 2009; Reeves et al., 1985).

3.9.3. Codon usage and mRNA export

Several studies have previously investigated the effects of codon usage on mRNA export, primarily in the context of viral gene expression. Cellular spliced mRNA is exported via the

NXF1 export receptor aided by the Aly/REF adaptor proteins. Viral late-expressing genes however, rely on early-expressed viral export adaptor proteins, such as Rev, which recognises a particular RNA-element (RRE) on the viral transcript and mediates mRNA export via the CRM1-dependent pathway normally used for nuclear export of unspliced RNA (Fischer et al., 1994; Malim et al., 1989). It could be shown that increasing the GC-content of usually very AT-rich viral genes circumvents the requirement for Rev (Kotsopoulou et al., 2000). We hypothesised that this could be a more general effect and potentially a bottleneck for the expression of particularly GC-poor genes. Subcellular fractionation and quantification of GC-poor GFP_000 and GC-rich GFP_001 showed that there are indeed differences in RNA localisation. Quantification of nuclear RNA levels showed no difference in GFP_000 and GFP_001 levels (Figure 19), further indicating that overall poor expression is not necessarily caused only by lower rates of transcription (Bauer et al., 2010).

When comparing RNA levels in the cytoplasm, GFP_001 is 5-fold more abundant than GFP_000 (p=1.378e^{-5}). This is in contrast to data obtained on CpG-variants by Bauer et al. who could not find any significant differences in RNA localisation (Bauer et al., 2010). It is not clear whether the differences observed here are due to lower nucleo-cytoplasmic export rates of GFP or caused by differences in RNA turnover rates. I therefore measured RNA half-lives of both variants (Figure 20) to re-assess previously published data from our lab which concluded that there is no difference in transcript stability between GFP_000 and GFP_001 (Kudla et al., 2006). Similar conclusions were drawn by Bauer et al (2010). However, in both studies, the experiments were conducted using Actinomycin D which exhibits its function by intercalating in DNA with a strong sequence-preference for GC-rich regions. It is therefore not possible to reliably compare the stability of genes with very different GC content. I repeated these experiments with an alternative, sequence-independent reagent, Triptolide, which acts at the level of transcription initiation rather than elongation (Leuenroth and Crews, 2008; Titov et al., 2011; Wang et al., 2011). Results show that there is indeed a significant difference of about 3.5-fold between GFP_001 ($t_{1/2}$=8.6h) and GFP_000 ($t_{1/2}$=2.4h), suggesting that increased GC-content also leads to increased transcript stability (Figure 20). This difference however seems unlikely to be sufficient to explain the much larger difference between these two variants on the protein level (34-fold). The experiment presented here was conducted on total RNA, but considering the previous differences in RNA localisation between variants, it would be interesting to repeat such stability measurements followed by subcellular fractionation to monitor RNA decay in separate cellular compartments, to further explore the possibility of differential recognition of sequence features by different components of the RNA degradation machineries. Furthermore, information from this would be useful in determining whether the GFP RNA export rate is decreased or whether the observed changes are mainly due to differential RNA stability. These experiments could also be performed on cells expressing GFP variants with an intron as this would give further clues about the role of splicing and how it might be exhibiting positive effects on expression levels (if any at all), i.e. through transcript stabilisation or by facilitating export, and why a complete rescue of expression levels cannot be achieved.

4. A high-throughput approach for the phenotypic profiling of reporter genes

The previous chapter described the systematic investigation of the effects of GC content by measuring several molecular phenotypes of coding-sequence variants of GFP. To be able to identify more general and possibly more subtle effects and to relate such to various sequence properties, the number of measured variants needs to be high enough for meaningful statistical analyses. Therefore, we designed a high-throughput sequencing-based approach which allows the measurement of multiple parameters of many GFP variants simultaneously in stable human cells lines. The following chapter describes the experimental outlines, assay design and data validation.

4.1. Phenotypic profiling of fluorescent reporter genes

Several previous studies that were interested in how codon choice affects gene expression did so using high-throughput approaches. Most of these studies were performed in bacteria (Goodman et al., 2013; Kosuri et al., 2013; Kudla et al., 2009) or yeast (Dean and Grayhack, 2012; Gamble et al., 2016; Presnyak et al., 2015; Shah et al., 2013) that readily allow the screening of thousands of variants. Short life cycles, large population sizes and easy genetic manipulations make such organisms perfect for high-throughput studies. Fewer studies tried to approach similar questions in mammalian systems (Gingold et al., 2014; Rudolph et al., 2016) and primarily did so by assessing global gene expression changes, focussing predominantly on effects on the translational level. However, few used a controlled experimental set-up in mammalian cells to measure the consequences of codon usage directly on a single-gene level across the whole coding sequence for several hundred synonymous variants simultaneously at multiple stages in gene expression. Here, I outline the experimental design and validation of a human cell line system which allows the measurement of multiple molecular phenotypes for over 200 GFP sequence-variants quantified using next-generation sequencing (Figure 22). Due to the large variation in sequence features of the genes to be tested, e.g. ranging from the lower to the upper end of GC-content, as well as the various phenotypes to be measured (from RNA to protein), the experimental design required the careful consideration of a few key aspects which are outlined and discussed in the following sections.

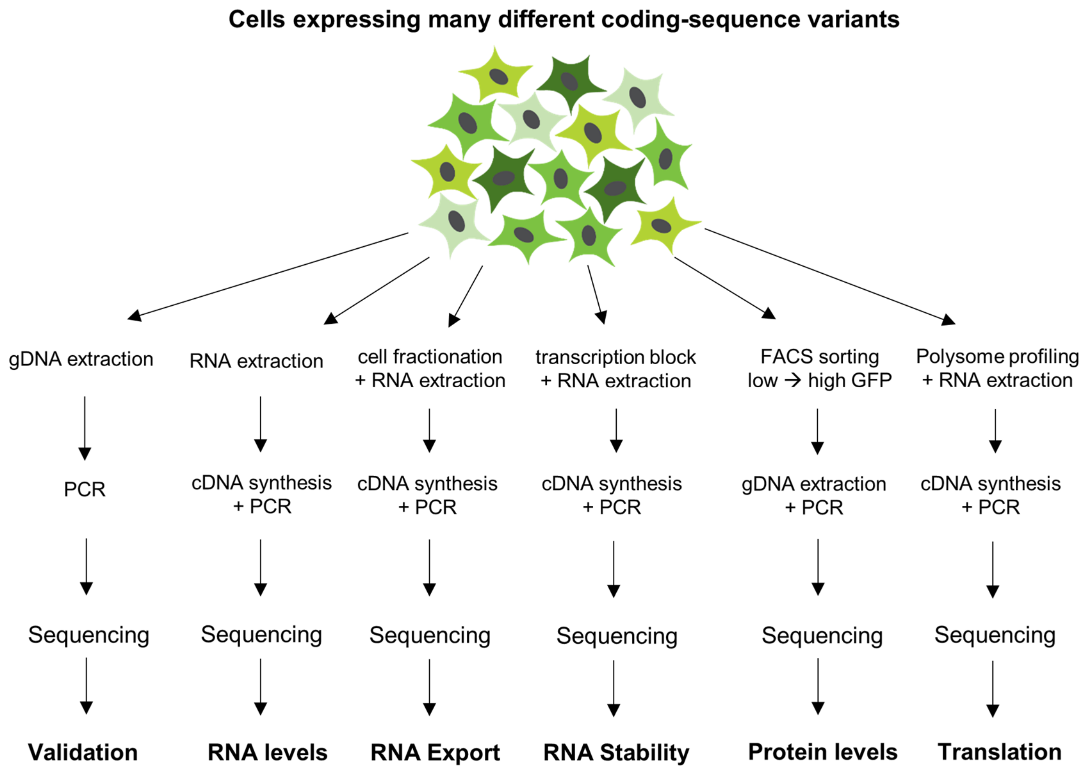


Figure 22. Overview of experiments measuring various phenotypes of many GFP sequence variants.
Cells expressing many GFP variants can be used to measure multiple molecular phenotypes simultaneously, starting from RNA or gDNA.

4.1.1. Choosing and establishing a human cell line system

One of the main considerations for the phenotypic screening of codon usage variants is which type of cellular expression system to use. This strongly depends on the purpose and types of measurements that are to be conducted. For example, if protein levels are to be measured, it must be ensured that each cell is only expressing one particular variant at a time to allow protein quantification for each variant independently rather than the cumulative effect of many within one cell. This rules out transient transfection systems in which multiple copies of plasmids and hence, multiple variants are introduced into each cell. In contrast, if the main focus is on RNA phenotypes, this would not necessarily be a limitation as RNA levels could be normalised to the total DNA content (genomic or plasmid) as e.g. in a study by Puchta et al., 2016, in which unique 20nt barcodes were placed in an untranscribed region just downstream of the gene variants, or, in case of transient transfections, total RNA expression can be normalised to total DNA transfected. Here however, the aim is to build an overview of the effects of codon usage on many stages in gene expression, both on the RNA as well as protein level. This led to the conclusion that a stable cell line system would be the most appropriate approach for the purpose of this study.

Classical approaches for stable gene integration include infection with viruses produced by viral packaging cell lines, which are used to modify the virus genome to carry the gene of interest and produce the viral particles, or by utilising the “copy and paste”-mechanism of retrotransposon sequences, such as the *piggyback* (Ding et al., 2005) or *sleeping beauty* systems (Ivics et al., 1997, 2009). Although both approaches are highly efficient, the integration into the genome is not targeted to a specific locus and will occur, in respect to transcriptionally active regions in the genome, close to random (Yant et al., 2005). Additionally, such integration events are not limited in their frequency of occurrence per cell, leaving the possibility of multi-copy insertions. As a consequence, a genetic screen would have to follow to *a*, assess successful integration, usually using antibiotic markers, *b*, verify that only one gene copy is present in the genome and *c*, ensure the integration site is indeed transcriptionally active.

Over the last decades, several more targeted approaches have emerged, revolutionising genome engineering. Zinc finger nucleases (ZFNs, Urnov et al., 2010) as well as transcription activator-like effector nucleases (TALENs, Joung and Sander, 2013) are able to cut double-stranded DNA *in vivo* at specific sites. Both are made up of multiple modules, each recognising either three to four bases (in case of ZFNs) or single nucleotides (TALENs), making it possible to target any particular sequence by mix-and-matching. The gene insertion, replacement or deletion is mediated by the cell’s own DNA repair mechanisms by homologous recombination. More recently, another genome editing system has become a widespread method of choice for genome engineering: the CRISPR/Cas9 system (Cong et al., 2013; Jinek et al., 2013). CRISPR, which is the acronym for “Clustered regularly interspaced short palindromic repeats”, does not rely on a protein-DNA recognition interface, as in case of TALENs and ZNFs, but instead uses a short RNA to guide the Cas9 nuclease to a specific genomic sequence. Cas9 cleaves the DNA and the repair is again mediated via homologous recombination if a repair template is provided. However, nucleases such as Cas9 and TALEN can exhibit off-target effects by cutting in unwanted places (Fu et al., 2013; Hsu et al., 2013) and can also suffer from low repair efficiency (Mao et al., 2008), requiring rigorous single-clone screening.

To overcome such time-consuming limitations, several commercial cell systems have become available which allow efficient stable integration into a single defined genomic locus, without the need for genotyping, as off-target integrations are rare and successful integration can be ensured by chemical selection. Such cell lines were engineered to carry flags in a defined, highly transcriptionally active genomic site, suitable for homologous recombination-based insertions. By placing the gene into a plasmid with sites homologous to the genomic target region and co-transfection with a vector carrying a site-compatible DNA recombinase, the gene will be integrated by homologous recombination into this particular locus only, providing that both vectors are successfully delivered into the same cell. One example of such a system is the “Flp-in” system by Invitrogen (www.invitrogen.com) which I utilised in the following experiments and is described in more detail in the following section.

4.1.1.1. Generating a pool of Flp-in cell lines

To be able to dissect the sequence effects of a large cDNA library on gene expression, it is crucial that positional effects caused by random genomic integration sites can be eliminated as a source of error (as discussed above). I utilise the ease of the Flp-in system (Invitrogen) for establishing stable cell lines which expresses a gene of interest from one specific locus from a tetracycline inducible promoter (Figure 23). To establish such stable Flp-in cell lines, a standard plasmid transfection with a mix of 2 plasmids is performed: The plasmid carrying the gene of interest cloned between recombination sites homologous to the “Flp-in” sites in the genome of the parental Flp-in cell line (in this case: *pcDNA5*), and another carrying the expression cassette of the required Flp-in recombinase which mediates the homologous gene cassette exchange (*pOG44*). Whereas the usual procedure would involve the integration of just one gene per stable transfection, for the purpose of this study, the integration of several different gene variants, i.e. different variants integrating into different cells, was required. As the main limitation for the successful genomic integration is the successful delivery of the plasmid carrying the gene of interest as well as the vector carrying the recombinase and its efficient expression, theoretically, vectors carrying any number of different genes can be co-transfected.

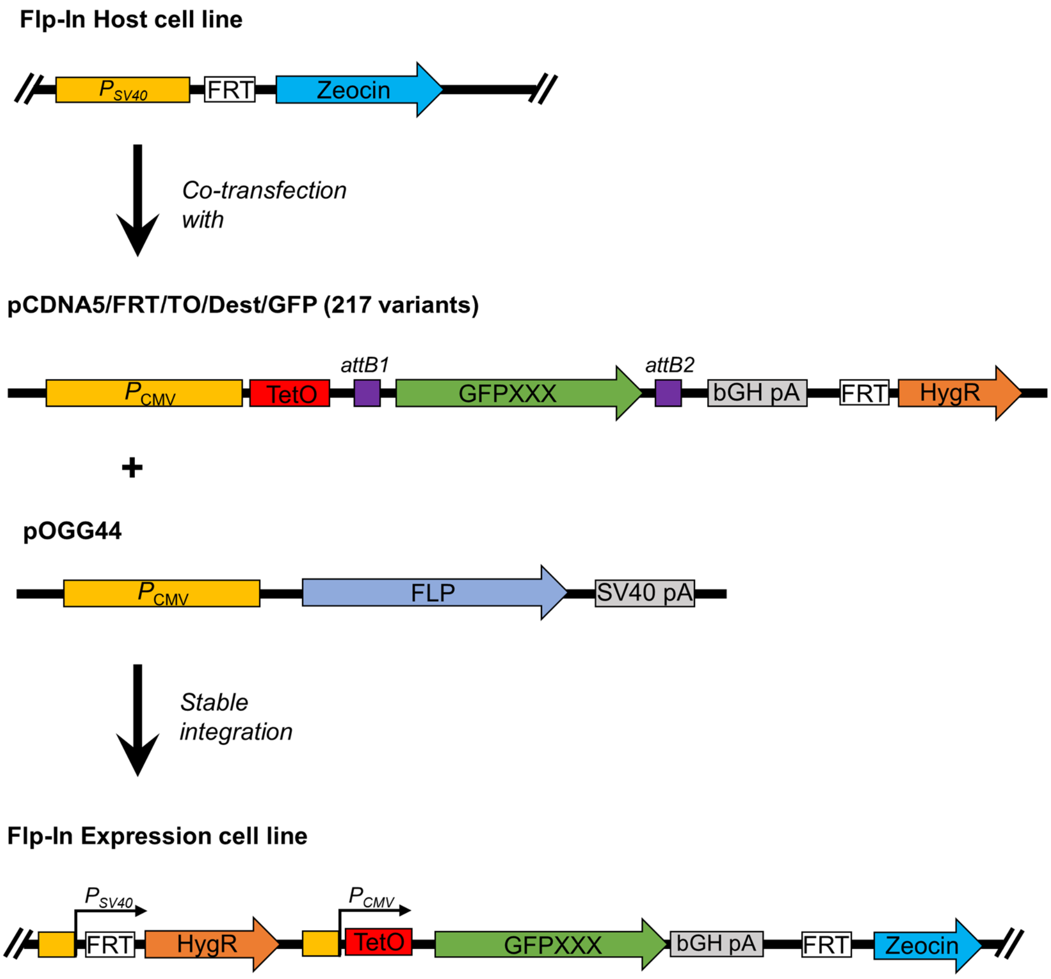


Figure 23. Schematic of the Flp-in expression system.
Flp-in host cell lines contain an FRT site, which serves as binding and cleavage site for the Flp recombinase, and a Zeocin resistance cassette. Cells are co-transfected with the Flp recombinase expression vector *pOGG44* and *pcDNA5/FRT/TO/Dest/GFP* for the expression of a pool 217 GFP variants. *pcDNA5* also carries a Hygromycin resistance cassette. Stable integration into Flp-in host cell lines results in Flp-in expression cell lines with constitutive expression of the Hygromycin and Zeocin resistance cassettes, as well as inducible GFP expression under the control of a tetracycline-regulated CMV promoter.

To be able to achieve several hundred different gene integrations, all variants had to be cloned into *pcDNA5*. To achieve this, I made use of the Gateway cloning system, which is a different homologous recombination-based system, commonly used to facilitate the sub-cloning of genes between Gateway-compatible vectors. Since the GFP cDNA library is conveniently stored within Gateway-Entry vectors (*pGK3*), I modified the Flp-in vector *pcDNA5* to be a compatible Destination vector to easily multiplex the Gateway cloning procedure and sub-clone many different sequences into *pcDNA5* in one reaction (outlined in Figure 24a).

Using this approach, I established a *pcDNA5* vector pool with 217 different GFP variants. I also generated a version of *pcDNA5* which contains the same chimeric intron in the 5'UTR of the expression cassette as previously used in single GFP experiments discussed in chapter 3 (Figure 12) to further study the involvement of splicing on expression. Using this approach, it is theoretically possible to create a plasmid pool with any number of sequence variants, providing an efficient recombination reaction in an appropriate scale. This vector pool was then used to establish stable Flp-in cell lines. To which extent the established cell lines represent the original vector pool depends not only on the copy number of plasmids in the pool, which I kept at equimolar ratio in the Gateway-reaction, but also on high transfection efficiency, delivering both GFP vector as well as the Flp-recombinase encoding vector *pOGG44* into cells, and lastly, on successful homologous recombination events. Using the *pcDNA5* GFP vector pool, I established several batches of Hek293 and HeLa Flp-in cell lines.

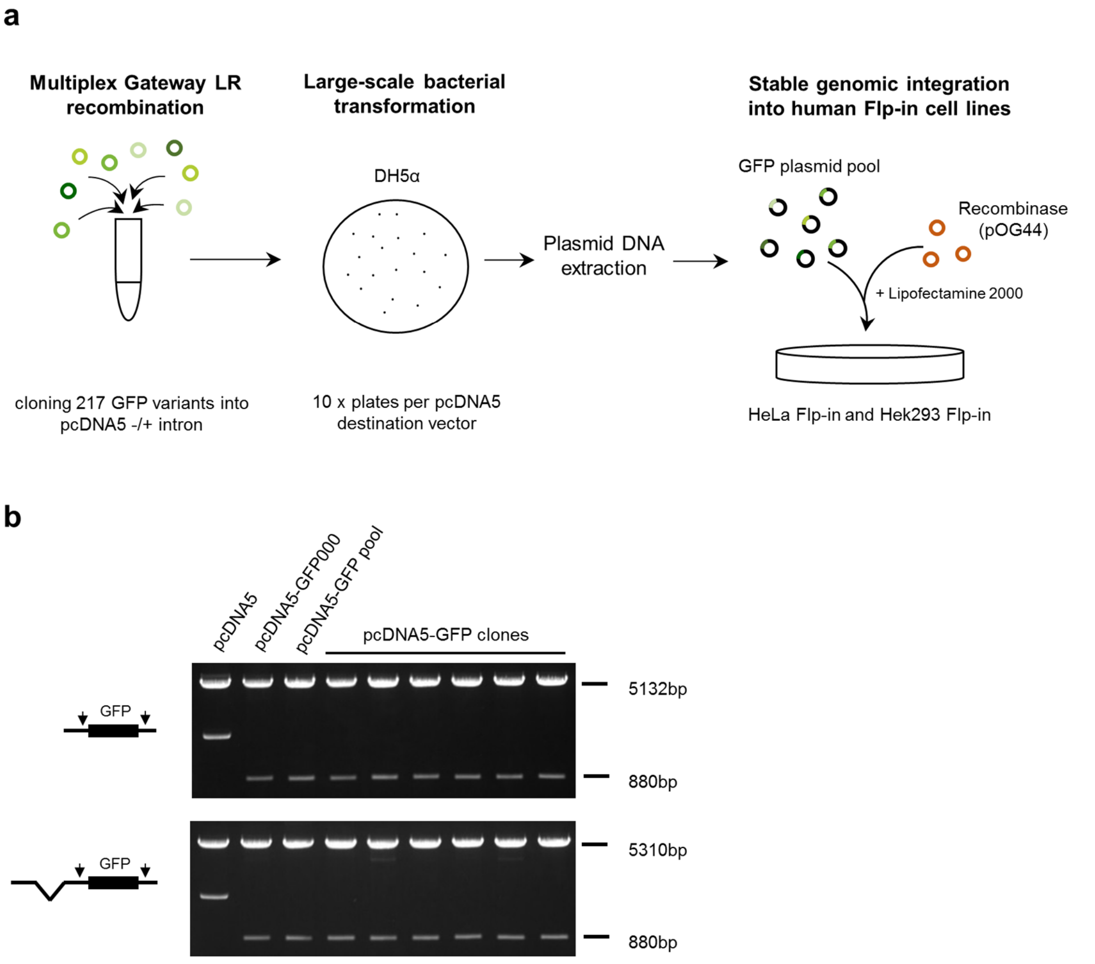

a
Multiplex Gateway LR recombination
Large-scale bacterial transformation
Stable genomic integration into human Flp-in cell lines
DH5α
Plasmid DNA extraction
GFP plasmid pool
Recombinase (pOG44)
+ Lipofectamine 2000
cloning 217 GFP variants into pcDNA5 -/+ intron
10 x plates per pcDNA5 destination vector
HeLa Flp-in and Hek293 Flp-in
b
pcDNA5
pcDNA5-GFP000
pcDNA5-GFP pool
pcDNA5-GFP clones
GFP
5132bp
880bp
GFP
5310bp
880bp

Figure 24. Multiplex Gateway cloning.
a, 217 variants of GFP stored in Gateway-Entry vectors are mixed in equimolar ratio and simultaneously recombined into Gateway-compatible Destination vectors *pcDNA5* and *pcDNA5+intron*. The LR-recombination mixture is used to transform DH5α, followed by overnight incubation on LB-agar plates containing 50mg/ml ampicillin at 37°C. Plasmid DNA from resulting colonies was extracted and used in co-transfections with the Flp-in recombinase carrying vector pOG44 to establish stable HeLa and Hek293 GFP pool cell lines. **b**, Plasmid DNA from 8 colonies from large-scale LR transformations described in **a,** were screened for the presence of GFP by restriction digest with *NotI* to confirm successful integration. Products were resolved on a 1% TBE/agarose gel.

4.1.2. Sequencing library design and optimisation

For experiments in which the main focus lies in changes on a single gene level rather than global gene expression changes, targeted amplicon libraries for high-throughput sequencing can be prepared. However, multiple technical considerations should be made for the library design, taking current technical limitations and possibilities in sample preparation and data generation into account. To avoid the necessity to obtain the full gene sequence, some studies utilise short barcode sequences that are added to the sequence of interest. However, whether this is appropriate depends on the types of experiments to be conducted. For experiments looking at RNA or protein phenotypes, barcodes included in transcribed regions of the gene might affect the folding structure of the molecule, thus likely influencing stability, regulation, as well as functionality. If however the experimental design allows conclusions to be made by deep sequencing of DNA as starting material, barcodes can be included in untranscribed portions of the locus as done in e.g. Puchta et al., 2016, in which the influence of random sequence mutations in an essential RNA on cell fitness was measured by the frequency of gene copies in the cell pool. Whether or not barcodes should be used will also strongly depend on the length of the studied gene as amplicon length is a factor that needs to be considered when choosing the sequencing platform. Current technologies allow either single-end or paired-end reads. Paired-end reads will give higher confidence in low frequency variation and can also be used when the amplicon length exceeds the maximum read length. The two resulting reads from either end of the amplicon can then be paired and used to map the sequence back to a reference. If large numbers of samples are to be sequenced, it might be possible to multiplex samples for sequencing runs using different short indices. Whether this can be done, depends on how many variants are expected and how many reads for each are needed (sequencing depth) to be confident.

Besides the library design, experimental factors for the actual library preparation need to be considered, as well as technical biases that are likely to be introduced at several stages. Regardless whether the starting material is gDNA or cDNA, the target sequence is amplified in a PCR reaction. For this, a high-fidelity (HF) DNA polymerase should be used to avoid amplification-induced sequence variation. However, many studies have shown that HF-polymerases do not all perform equally well in library preparations, probably due to varying degrees of processivity, allowing some to proceed through strong secondary structures more readily than others (Aird et al., 2011; Cline et al., 1996; Dabney and Meyer, 2012; Miura et al., 2013; Oyola et al., 2012). This tends to affect very AT- or GC-rich sequences more strongly, leading to lower read coverage of such.

If the starting material for library preparation is RNA, its quality should first be measured using a Bioanalyzer (Agilent) to obtain the RNA integrity number (RIN; Schroeder et al., 2006) to assess the degree of RNA degradation before proceeding with the cDNA synthesis step. Since this step relies on the enzymatic activity of a reverse transcriptase (RTase), it can further introduce sequence-specific biases due to the lack of proof-reading ability and/or low processivity (Mohr et al., 2013). However, in the last few years, RTases have been genetically engineered to exhibit higher processivity and to be able to perform at higher temperatures, decreasing the likelihood of secondary structure formation (Mohr et al., 2013; Nottingham et al., 2016; Qin et al., 2016). The library design chosen in this study and the optimisation for library preparation steps that I have performed are described in the following sections.

4.1.2.1. Amplicon library design and data analysis

As some of the measurements that are to be conducted with the pool cell lines are focussed on the RNA-phenotype without the possibility to quantify genomic DNA content as a proxy (e.g. in sub-cellular fractionation experiments), the GFP sequences were not barcoded to avoid interference of such additional sequences with RNA expression. To uniquely identify each GFP variant, we use the Illumina MiSeq sequencing platform which can generate up to 20million 300nt reads from the 5' and 3'end (paired-end). We chose this platform because of the combination of read lengths, accuracy, number of read counts and cost. As GFP is 720nt long, it is consequently not possible to cover the full length sequence using this approach. Among the >400 GFP variants available in the lab, 217 can be uniquely identified by sequencing 300nt from each end, and I therefore only selected these variants for establishing the cell line pools as described above. I introduced appropriate sequencing adaptors required for the attachment of sequences to the flow cell, sequencing primer binding sites, as well as indices for multiplexing libraries as overhangs on primers used in first strand synthesis and subsequent PCR amplification (schematic representation in Figure 25). Following high-throughput sequencing, raw reads are processed using an analysis pipeline outlined in figure 26.

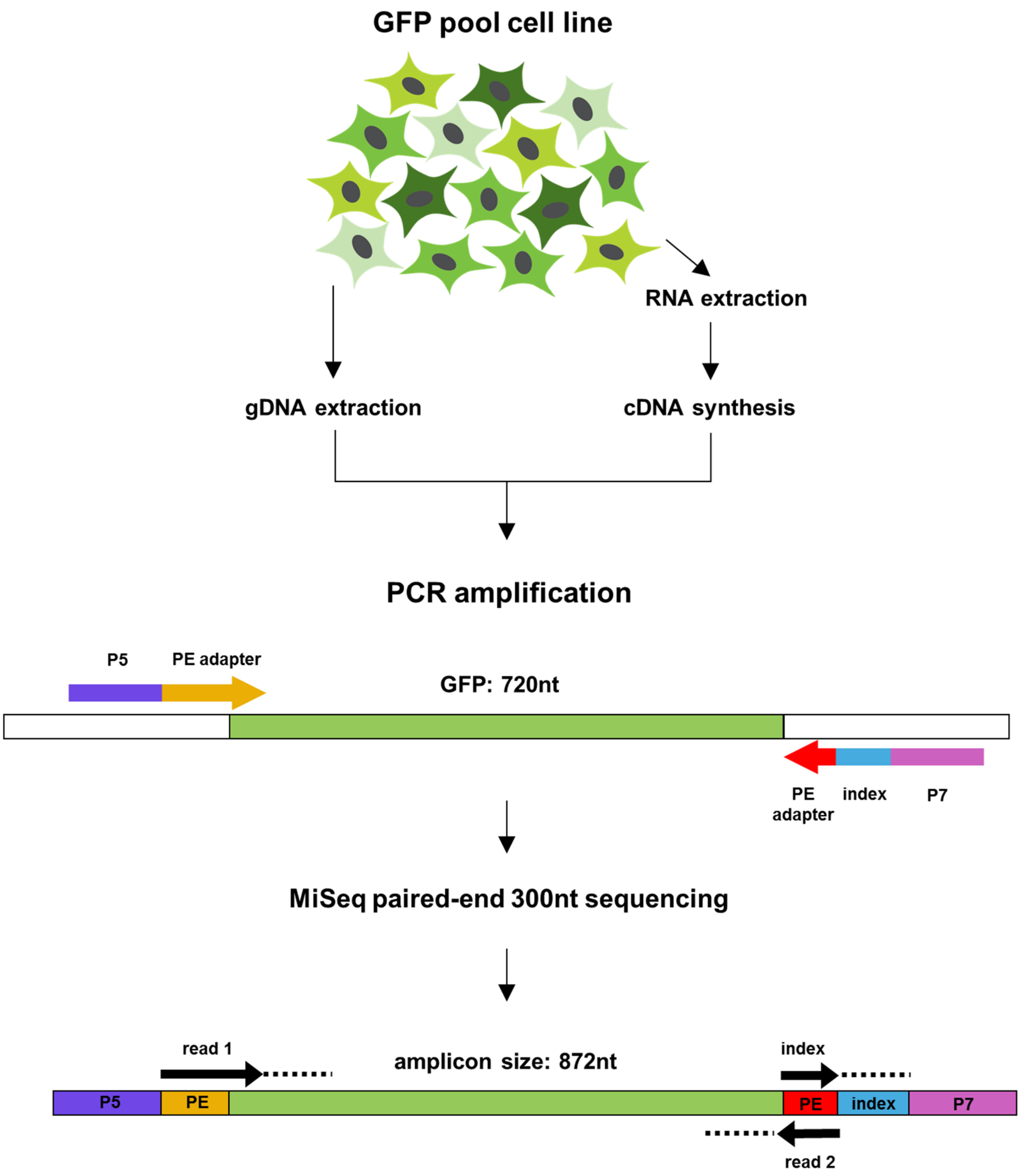


Figure 25. Schematic of amplicon library and sequencing primers.
GFP sequences are amplified from gDNA or cDNA using PCR primers with overhangs attaching the P5 and P7 adaptor sequences required for annealing to complementary oligos on the flow cell surface, as well as primer annealing sites (PE) for both paired-end read primers and index read primer (reverse complement to read 2 primer).

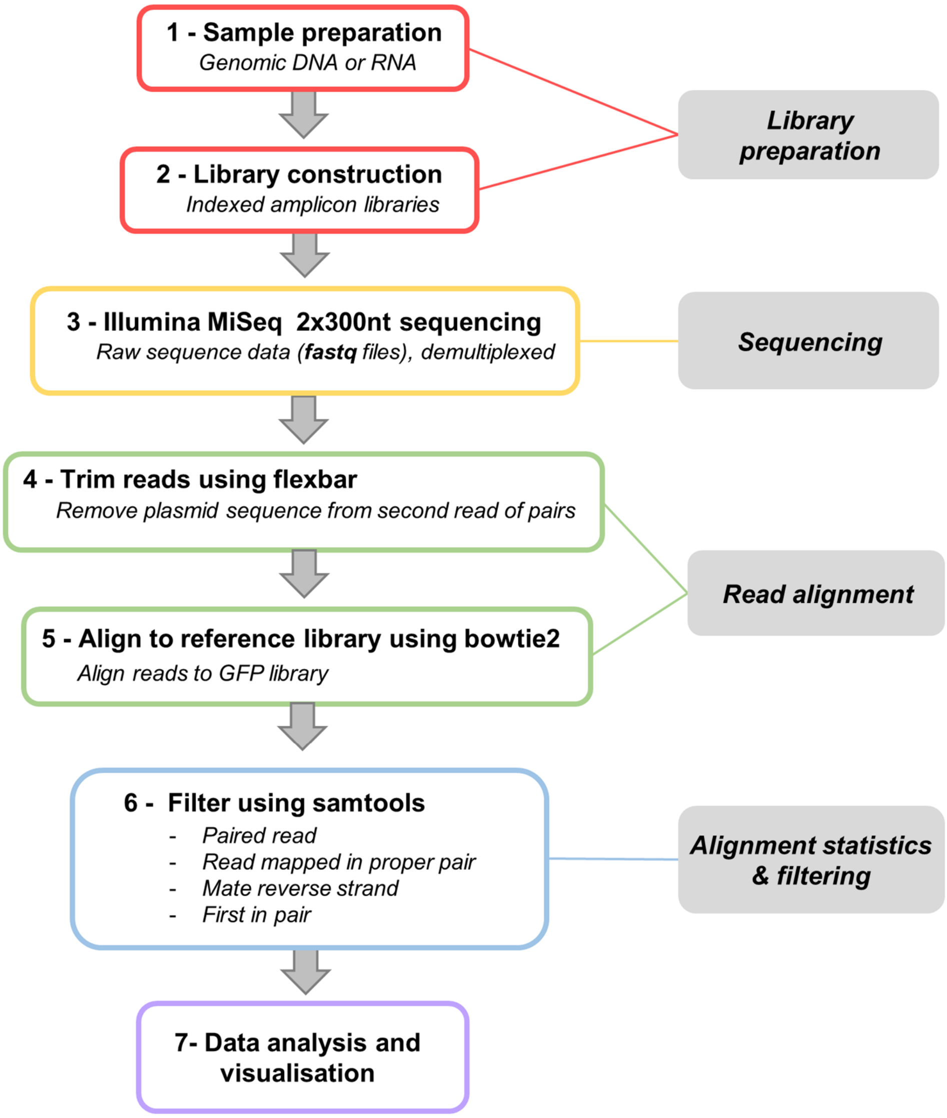

1 - Sample preparation
Genomic DNA or RNA
2 - Library construction
Indexed amplicon libraries
Library preparation
3 - Illumina MiSeq 2x300nt sequencing
Raw sequence data (fastq files), demultiplexed
Sequencing
4 - Trim reads using flexbar
Remove plasmid sequence from second read of pairs
5 - Align to reference library using bowtie2
Align reads to GFP library
Read alignment
6 - Filter using samtools
- Paired read
- Read mapped in proper pair
- Mate reverse strand
- First in pair
Alignment statistics & filtering
7- Data analysis and visualisation

Figure 26. GFP amplicon library sequencing and data analysis pipeline.

4.1.2.2. Reducing DNA amplification bias

When preparing sequencing libraries, one of the main concerns are biases that can arise due to sequence composition. Several published studies investigated the effects of various DNA polymerases on the sequence complexity of sequencing libraries and showed a dramatic drop in the abundance of particularly GC-rich reads after PCR amplification (Aird et al., 2011, Dabney and Meyer, 2012). As it is crucial for the purpose of this study to reduce any biases that can be introduced in course of the sequencing library preparation, since such could skew or even mask relevant relationships, conditions minimising sequence-related biases had to be determined.

To assess whether sequence biases are introduced during the PCR step, I selected 3 GFP variants with varying GC content (GFP_400=36%, GFP_407=43%, GFP_422=59%) as templates in semi-quantitative PCR using different DNA polymerases. I chose the enzymes based on their performance in published studies (e.g. Herculase II (Agilent) and AccuPrime Pfx (ThermoFisher) were the best performers in Dabney and Meyer, 2012), or because they are marketed as highly efficient in the amplification of difficult and GC-rich templates (Q5, NEB). I tested these against the polymerase most commonly used for high-throughput library preparations, Phusion HF (NEB), as this is the polymerase recommended by Illumina (Illumina paired-end sample preparation guide, www.support.illumina.com).

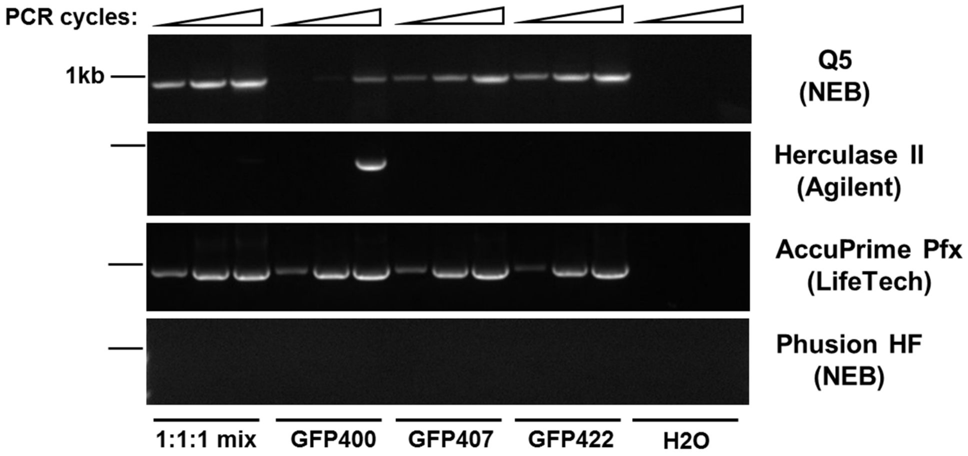


Figure 27. Comparison of PCR efficiency of various DNA polymerases on GFP variants. GFP variants 400 (GC=36%), 407 (GC=43%) and 422 (GC=59%) were amplified from *pCM3* using equal amounts of template in all reactions. Various numbers of PCR cycles were chosen to better assess the efficiency. All reactions were set-up and performed in the conditions specified by the manufacturers. Resulting PCR products were resolved on a 1% agarose/TBE gel. *Note*: for Phusion HF primer dimers are visible on higher exposure which indicates that the reactions did not fail, but rather that the amplification of the template was not efficient.

As can be seen in figure 27, different DNA polymerases exhibit various amplification efficiencies. It should be noted that no particular positive control was included in this experiment as all templates could be considered as positive controls since all were sequence-verified and vary only in the GFP sequence, not at the primer-annealing sites. The recommended polymerase for sequencing library preparation by Illumina is Phusion HF polymerase. Therefore it is surprising that the amplification failed and no products are visible, not even after an excessive 40 cycles (Note: primer dimers are visible on a higher exposure, indicating that the PCR reaction per se did not fail; data not shown). In other published studies, the use of various available variations of Phusion polymerase led to a decrease in average library length as well as a marked increase in average %GC compared to polymerases from other manufacturers (Dabney and Meyer, 2012). This indicates that Phusion does suffer from a severe sequence-preferences which can, in worst case, introduce large biases when amplifying a mix of templates in the same reaction, such as in this experiment. Q5 polymerase shows a distinct bias towards templates with higher GC content. This is not unexpected as Q5 is marketed as performing well with GC-rich templates (Q5 High-fidelity DNA polymerase manual, www.neb.com). Herculase II and Accuprime Pfx were both polymerases which performed very well in the study by Dabney and Meyer (Dabney and Meyer, 2012). However, Herculase II also failed to amplify all 3 templates efficiently. A product for GFP_400 can only be detected after 40 cycles (a faint band is also visible in the pool after 40x at higher exposure), but none for any other GFP. The only polymerase in this test that amplified all GFPs equally well with no strongly visible biases and high product yields is Accuprime Pfx (ThermoFisher). Hence, I chose to use this polymerase for all library preparations.

4.1.2.3. Increasing reverse transcription efficiency

The starting material for the sequencing library preparation for some of the here described experiments is RNA. In such cases, a first strand synthesis reaction needs to be performed first to yield cDNA which is then used as PCR template. However, since this reaction also relies on the activity of an enzyme, this step could potentially be biased by difficult sequence features, similar as in PCR, which may lead to pre-mature termination of the cDNA synthesis reaction. I therefore compared 2 commercially available enzymes, Superscript II and Superscript III (both ThermoFisher), in regards to their processivity. To do so, I selected 6 different GFP variants spanning a broad range of GC content (36-59%) and *in vitro* transcribed those using T7 polymerase. Equal amounts of RNA were used in first strand synthesis reactions with GFP-specific primers using the different enzymes, followed by PCR amplification with Accuprime Pfx. The rationale is that any biases that will be seen after the

PCR reaction should reflect the combined ability of both reverse transcriptase and DNA polymerase to efficiently yield the relevant product. As can be seen in figure 28a, Superscript II exhibits less efficient reverse transcription ability than Superscript III, as the product yield decreases gradually with increasing GC content. This can be explained by the lower incubation temperature, which is likely not high enough to break strong secondary structures between or within RNA molecules, however, is the optimal for this particular enzyme. Superscript III was tested at two different temperatures, 50°C and 55°C, as recommended by the manufacturer (ThermoFisher). At 50°C, the obtained yields are more comparable between templates.

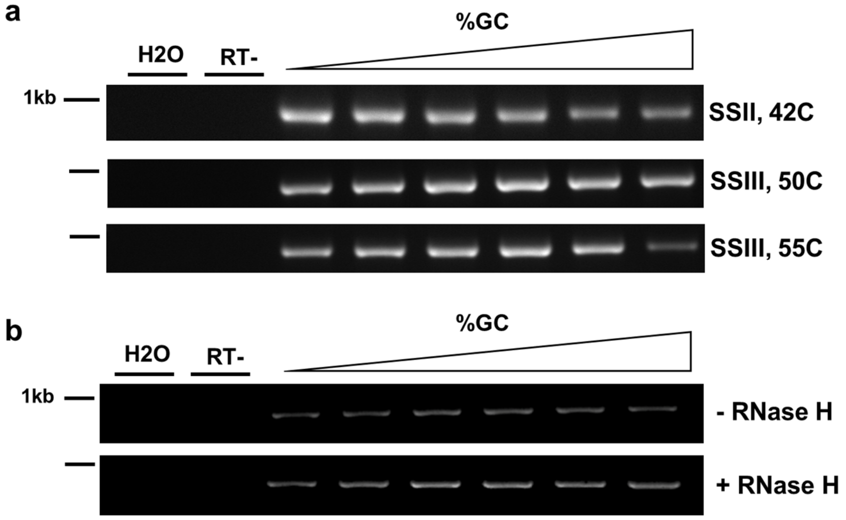


Figure 28. Optimisation of PCR amplification of several GFP variants.
a, Superscript II and Superscript III were compared for processivity and optimal conditions and resulting cDNA used in PCR amplification with Accuprime Pfx (Agilent). Optimal recommended incubation temperature for SSII is 42°C, whereas SSIII can be used at higher temperatures. Only 50°C (lowest) and 55°C (highest) were tested. Equal amounts of RNA were used to make cDNA and equal volumes were then used in subsequent PCR amplification. **b**, RNA of GFP variants was reverse transcribed using Superscript III at 50°C and either treated with RNase H for 20min at 37°C (bottom) or left untreated (top) before PCR amplification with Accuprime Pfx. PCR products were resolved on a 1% TBE/agarose gel.

Following on from this experiment, I also tested whether the inclusion of an RNase H digest before PCR could have a positive effect on PCR yield (Figure 28b). RNase H specifically degrades the RNA strand in RNA:DNA hybrids that might form after cDNA synthesis. It is thought that strong binding of complementary RNA to the DNA template can negatively affect PCR efficiency (Jeanty et al., 2010; Kitabayashi and Esaka, 2003) and I would expect that such effects are greater in GC-rich sequences due to the stronger binding kinetics. I therefore used the same RNA templates as above in first strand synthesis with Superscript III at 50°C and compared yields with or without treatment with RNase H. The additional RNase H digest resulted in overall higher yield (Figure 28b).

4.2. Assay Validation

4.2.1. A fluorescent reporter cell line pool

After establishing stable GFP cell lines, it needs to be validated whether the integrated sequences indeed include all expected variants and also whether each is present in roughly equal frequency across the cell pool as this is important to assure that sequencing results obtained with these pools will be statistically useable and that results for some variants will not have to be dismissed due to stochastic limitations.

In first instance, I tested whether the cells are indeed inducible. To do so, I performed a Doxycycline induction time course to quantify the inducibility of the cells and their mean fluorescence using Fluorescence-activated cell sorting (FACS). Representative data for a HeLa GFP pool cell line is plotted in Figure 29a. I also analysed the fluorescence range of all established cell line pools (HeLa, Hek293, each with or without intron) in comparison to two clonal GFP cell lines, GFP_000 (GC3=0.27) and GFP_001 (GC3=0.95) (Figure 29b). These two variants define the likely lower and upper fluorescence limits that would be expected to be covered if the established pool cell lines do contain a range of different GFP variants. It should be noted that these two variants are not themselves included in the pool. Indeed, as can be seen in figure 29b, all pool cell lines cover a broad fluorescence range, covering both the lower as well as the upper fluorescence boundaries as defined by the clonal cell lines. Since this is not a confirmation but rather a positive indication that these cells actually express all 217 variants included in the original transfection mix, I proceeded to confirm this via high-throughput sequencing. I amplified the genomic locus of the GFP integration site of about 10^7 cells and prepared sequencing libraries as described previously (Figure 25). The sequencing results were analysed computationally using the pipeline described in figure 26. Several major observations could be made from this analysis: Most importantly, all 217 GFP variants could be detected by sequencing, both in vector pools, as well as cell lines, suggesting the successful integration into cells. The proportion of each variant present in the cell pools however varies strongly across cell lines (Figure 30). Since the frequency of variants within the HeLa pool without intron is very broad, this pool was in first instance not considered for any further experiments or analysis. For this reason, I prepared multiple batches of Hek293 as well as HeLa cell line pools. A summary of all GFP pool cell lines that I established in due course, as well as which experiments I conducted with each, is shown in table 2.

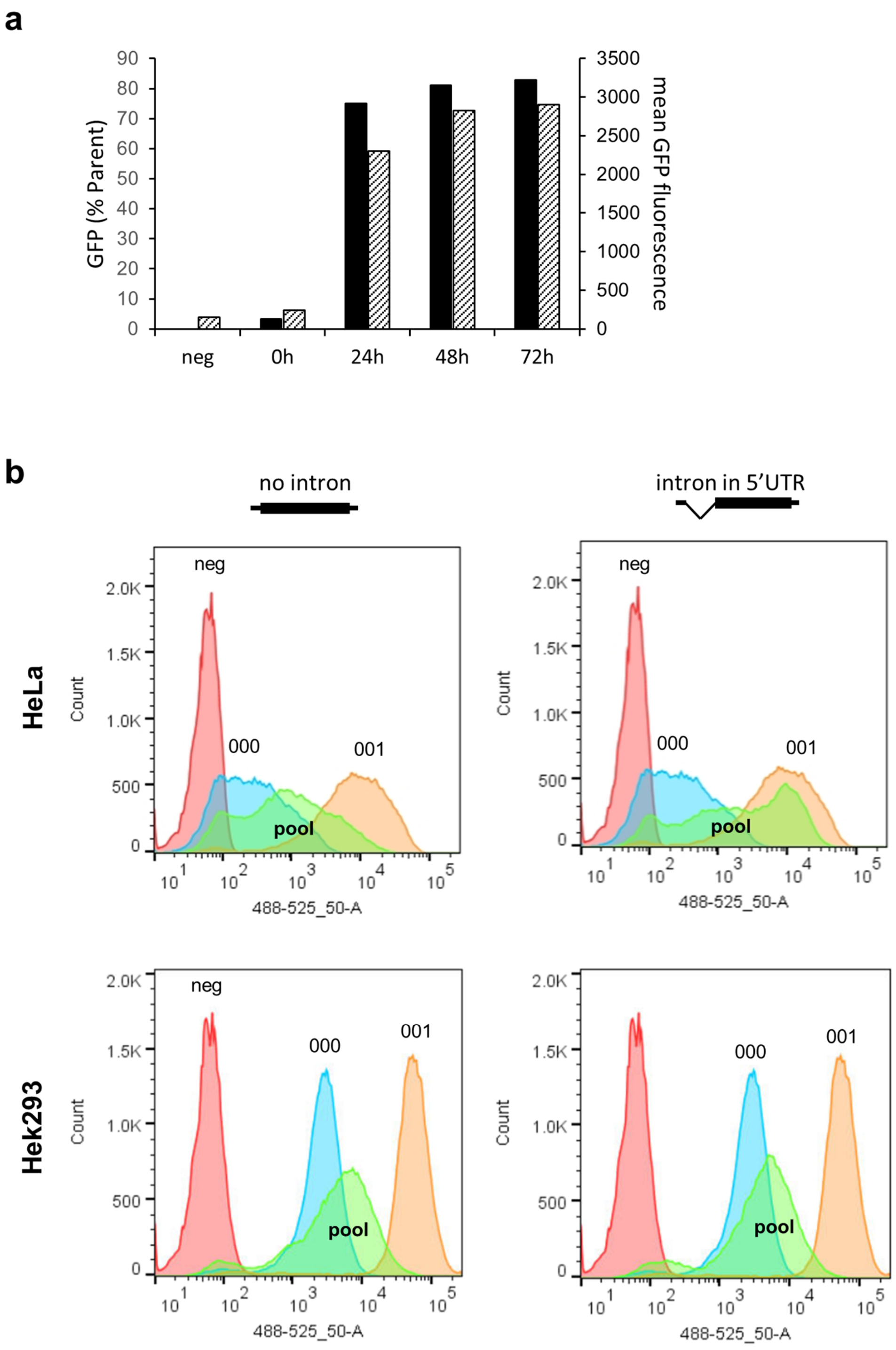


Figure 29. GFP expression of stable GFP pool cell lines.
a, GFP induction time-course of HeLa Flp-in GFP pool cell line. GFP expression was induced using 1ug/ml Doxycycline and assessed post-incubation by FACS. Plotted are the fractions of GFP-positive cells at each time point and the respective mean GFP fluorescence. **b**, Stable HeLa (top row) and Hek293 (bottom row) Flp-in cells expressing either GC-poor GFP000 (blue), or GC-rich GFP001 (orange), or a pool of 217 GFP coding variants (green). GFPs are expressed without (left column) or with (right column) an intron in the 5'UTR. GFP_000 and GFP_001 without introns are shown for comparison. Cells were induced for 24h prior to FACS-analysis.

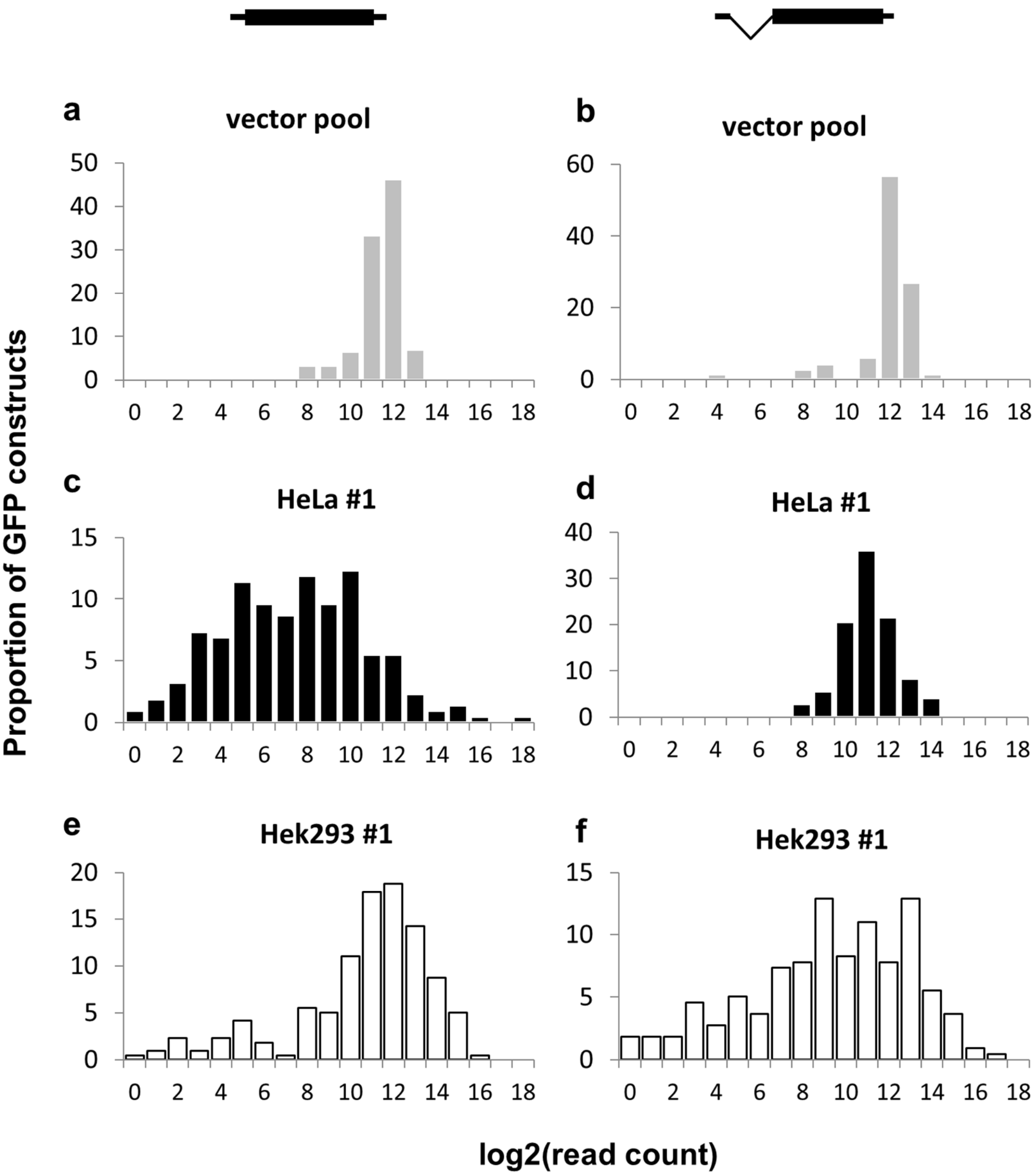
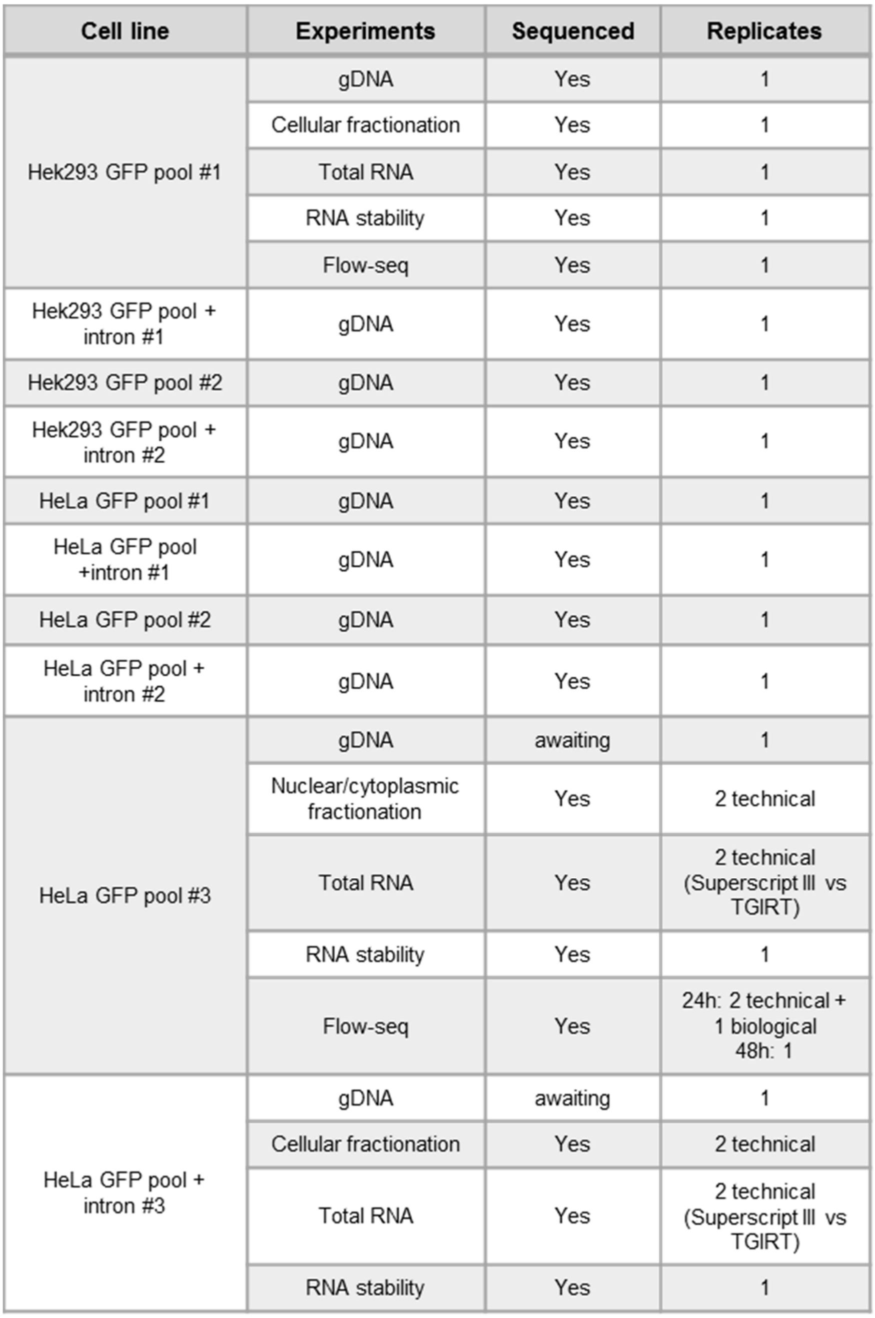


Figure 30. Proportion of GFP variants within plasmid and cell line pools.
Both *pcDNA5* vector pools (-/+intron), as well as all established Flp-in GFP pool cell lines were sequenced to calculate the proportion of each variant within the pools. The distribution varies strongly between cell lines.



Cell line	Experiments	Sequenced	Replicates
Hek293 GFP pool #1	gDNA	Yes	1
	Cellular fractionation	Yes	1
	Total RNA	Yes	1
	RNA stability	Yes	1
	Flow-seq	Yes	1
Hek293 GFP pool + intron #1	gDNA	Yes	1
Hek293 GFP pool #2	gDNA	Yes	1
Hek293 GFP pool + intron #2	gDNA	Yes	1
HeLa GFP pool #1	gDNA	Yes	1
HeLa GFP pool +intron #1	gDNA	Yes	1
HeLa GFP pool #2	gDNA	Yes	1
HeLa GFP pool + intron #2	gDNA	Yes	1
HeLa GFP pool #3	gDNA	awaiting	1
	Nuclear/cytoplasmic fractionation	Yes	2 technical
	Total RNA	Yes	2 technical (Superscript III vs TGIRT)
	RNA stability	Yes	1
	Flow-seq	Yes	24h: 2 technical + 1 biological 48h: 1
HeLa GFP pool + intron #3	gDNA	awaiting	1
	Cellular fractionation	Yes	2 technical
	Total RNA	Yes	2 technical (Superscript III vs TGIRT)
	RNA stability	Yes	1

Table 2. GFP pool cell lines and experiment documentation.

Figure 30 shows modest differences in GFP variant copy numbers in the vector pool, and large differences of GFP variant copy numbers in the cell line pools. The variation could represent stochastic differences in cloning and/or genomic integration of GFP variants, or systematic biases, possibly resulting from different sequence composition. To analyse the origins of this variation, I correlated total read counts of all GFP variants found in the vector pool to their respective GC content (Figure 31a). All GFP variants were added in equal ratios to the recombination mix, so it would be expected that sequence biases would lead to a skew in the total read counts obtained for all variants. However, no significant correlation can be seen, suggesting that no particular sequence bias was introduced during library preparation and sequencing run. To test whether all GFPs are integrated equally well into vectors, as well as integrated to similar extends into cell lines, I correlated total read counts for both vector pools to each other (Figure 31b), as well as the read counts of established cell lines to their respective vector pools (Figure 31c+d). Both vector pools, with and without intron, correlate well to each other (Figure 31b, $R^2=0.5283$, $p<10e^{-5}$) which is expected since the same Gateway entry vector pool was used in both reactions. The GFP read counts in each of the Hek293 pool cell lines correlate weakly with the vector pool (Figure 31c+d; no intron: $R^2=0.0539$, $p=5.33e\text{-}4$; with intron: $R^2=0.0244$, $p=2.07e\text{-}2$). This is also expected since the variation in the cell pool is mostly dependent on random integration rather than on the variation in the vector pool, though variants with low coverage in the vector pool also have low coverage in pool cell lines.

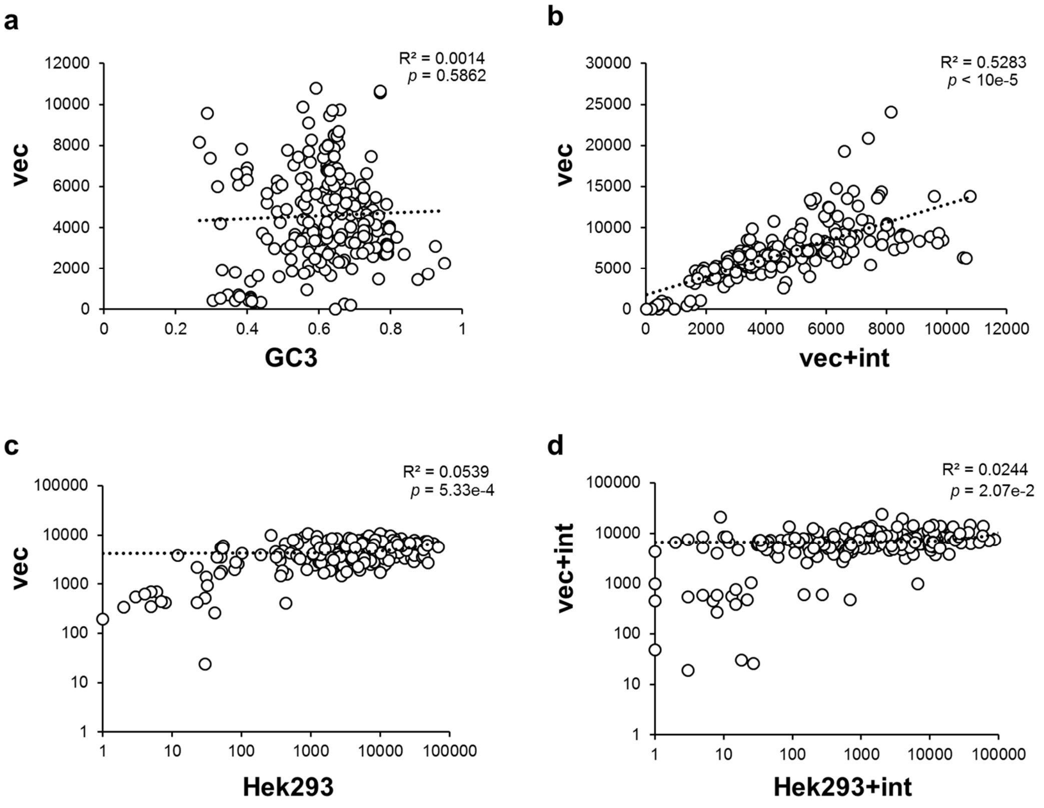


Figure 31. Comparison between sequenced vector pools, GC3 and pool cell lines.
a, Read counts of the sequenced *pcDNA5* vector pool are plotted against the GC3 of GFP variants. **b**, Both *pcDNA5* vector pools, with and without intron, are correlated (R^2=0.5283, p<10e^{-5}). **c,** and **d,** GFP read counts from sequenced cell line pools do not correlated well with their respective vector pools(**c,** vec**=**without intron: R^2=0.0539, p=5.33e-4; **d, vec+int=**with intron: R^2=0.0244, p=2.07e-2).

4.2.2. Estimating protein levels by Flow-seq

To obtain estimates of protein levels for each GFP variant, we used an approach based on Flow-Seq, a method published by Kosuri et al. (2013) used to screen for novel regulatory sequences in a large library of ribosomal binding site (RBS) and promoter variants in *E.coli*. Protein levels are first estimated by FACS and cells sorted into several tubes according to their fluorescence properties. Genomic DNA from the collected cells is then extracted and high-throughput sequencing libraries prepared to detect the frequency of each variant in every collection tube. The schematic outline of the adapted method is shown in figure 32.

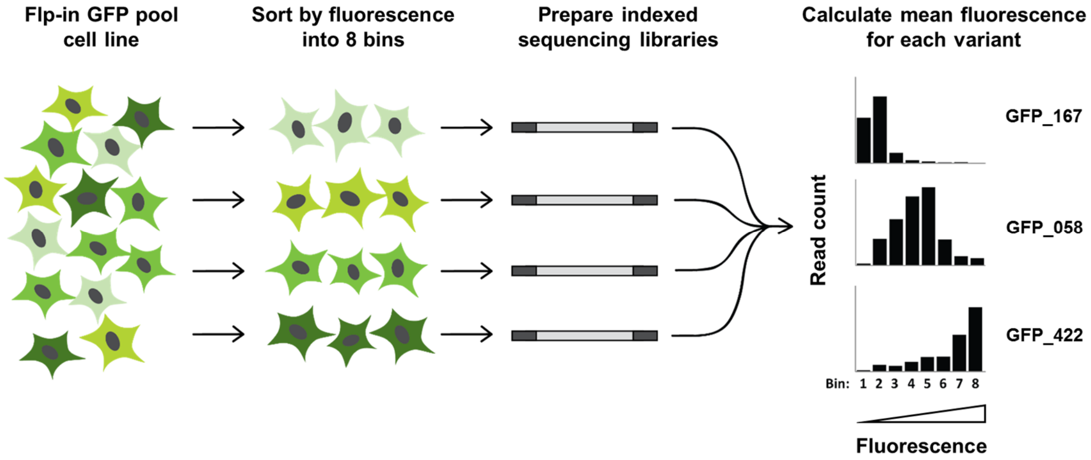

Flp-in GFP pool cell line
Sort by fluorescence into 8 bins
Prepare indexed sequencing libraries
Calculate mean fluorescence for each variant
Read count
GFP_167
GFP_058
GFP_422
Bin: 1 2 3 4 5 6 7 8
Fluorescence

Figure 32. Overview of adapted Flow-seq method.
Stable cell lines expressing many GFP variants are analysed by FACS and sorted into 8 bins according to their fluorescence levels from low to high. Each bin corresponds to 8-11% of the total log-fluorescence range. Genomic DNA is extracted from collected cells and used as template to create indexed sequencing libraries. The frequency of each GFP is calculated for all bins and the mean localisation calculated as a score for protein expression.

For initial experiments, GFP expression was induced in both Hek293 and HeLa pool cell lines without introns for 24h hours prior to FACS analysis. The fluorescence range covered in both cell lines is similarly broad, differing by up to 100 fold (Figure 33). GFP-positive cells were sorted into 8 expression bins, from low to high fluorescence, each gate comprising roughly equal numbers of cells. For each bin, I collected 1x10^6 cells and extracted genomic DNA which was then used as template to create indexed high-throughput sequencing libraries.

Computational analysis assigning sequencing reads to respective GFP sequences confirmed that all 217 variants were present in varying frequencies across all 8 collected bins (Figure 34). As in the GFP histograms above (Figure 33), most variants exhibit intermediate fluorescence levels, in both Hek293 and HeLa data (Figure 34). Most variants show well-defined peaks in neighbouring bins. In some cases the distribution is relatively broad and in fewer exceptions even U-shaped (Figure 34c). For later data analysis, these variants might therefore have to be filtered out as calculating an average fluorescence score for such might potentially introduce more noise into the data. To obtain a fluorescence value for each variant, I multiply the number of reads (n) within each bin with their respective bin number (i) before taking the sum and dividing by the total number of reads across all bins (equation shown below).

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

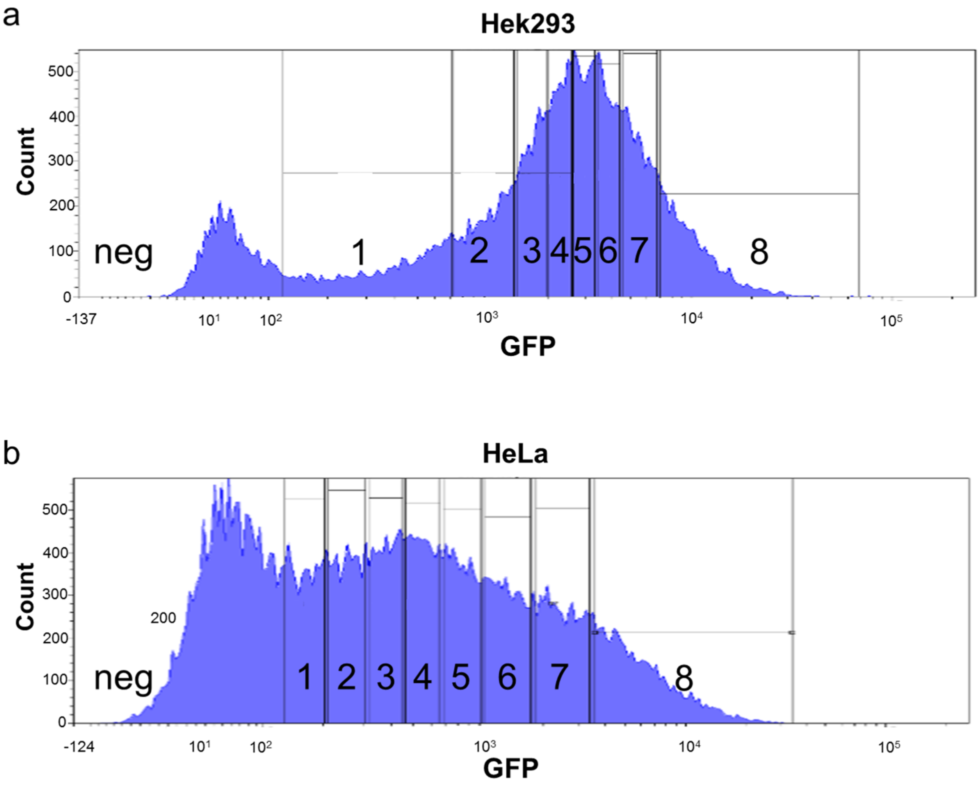


Figure 33. Fluorescence binning using Fluorescence-activated cell sorting (FACS). Hek293 (**a**) and HeLa (**b**) GFP pool cells were induced with 1ug/ml Doxycycline for 24h before FACS analysis. Cells were sorted into 8 fluorescence bins and roughly equal numbers of cells collected (~1x10^6). GFP negative cells ('neg') were excluded.

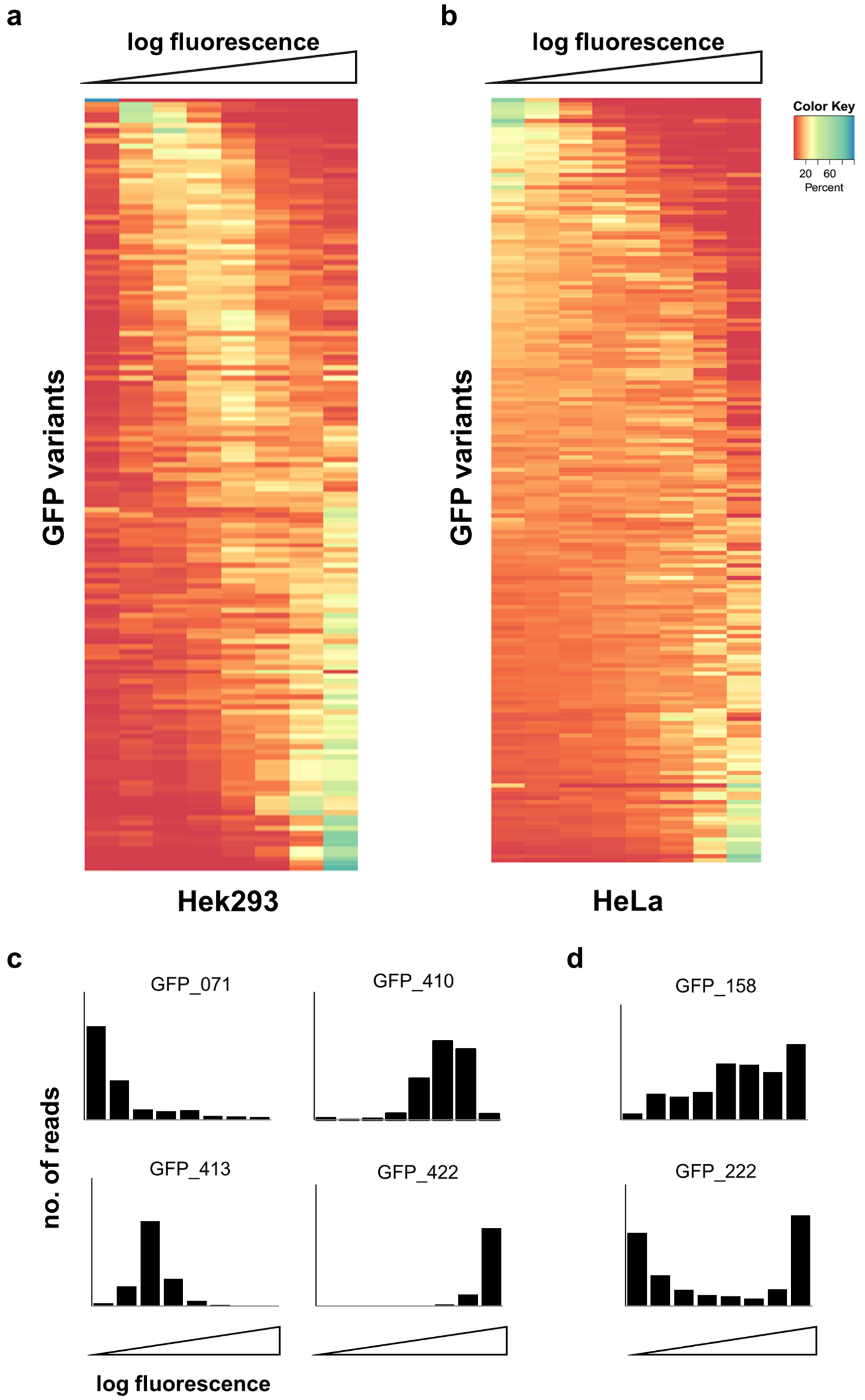


Figure 34. Fluorescence levels of GFP variants in Hek293 cells.
a + b, Distribution of GFP variants across all 8 fluorescence bins. For every bin, the frequency of each variant is shown as the proportion (%) of reads relative to the total number of reads for this variant across all bins. Shown are only variants with more than 1000 reads in total across all bins. The fold change in fluorescence between the lowest and highest bin is about 100-fold. **c**, GFP variants generally fall into well-defined bins. In a few cases, variants have unusual distribution patterns (**d**).

For HeLa cells, I acquired multiple Flow-seq data sets after 24h GFP induction denoted as 24_A, 24_B and 24_C. 24_A and 24_B are technical replicates, i.e. the same batch of cells FACSed twice on the same day followed by independent library preparations. 24_C is a biological replicate, i.e. a different batch of cells FACSed on a different day followed by independent library preparation. The prepared library of sample 24_A was sequenced twice; Figure 35a shows the high reproducibility of sample re-sequencing for this library (R^2=0.9989). The reproducibility between the FACS sorting experiments and library preparations are very high, as can be seen when comparing data from both technical replicates 24_A and 24_B (Figure 35b, R^2=0.9843). Furthermore, the reproducibility of data between different batches of cells FACSed on different days is high as well (Figure 35c; R^2=0.8583).

To further validate the Flow-Seq results, I compared the Flow-seq data with plate-reader fluorescence measurements of single GFP variants (see chapter 3.3). Since the relationship between FACS results and bin number is exponential (Figure 35d, R^2=0.9426), I compared Flow-seq results with the plate reader data by fitting an exponential curve which shows that both data sets correlate (Figure 35d; R^2=0.7965, $y=41.679e^{0.627x}$). I therefore proceeded to transform data for all GFP variants measured by Flow-seq by using the coefficient given by this exponential fit ($y=0.0011e^{1.193x}$) in order to calibrate the data of all variants and to convert them into a linear scale.

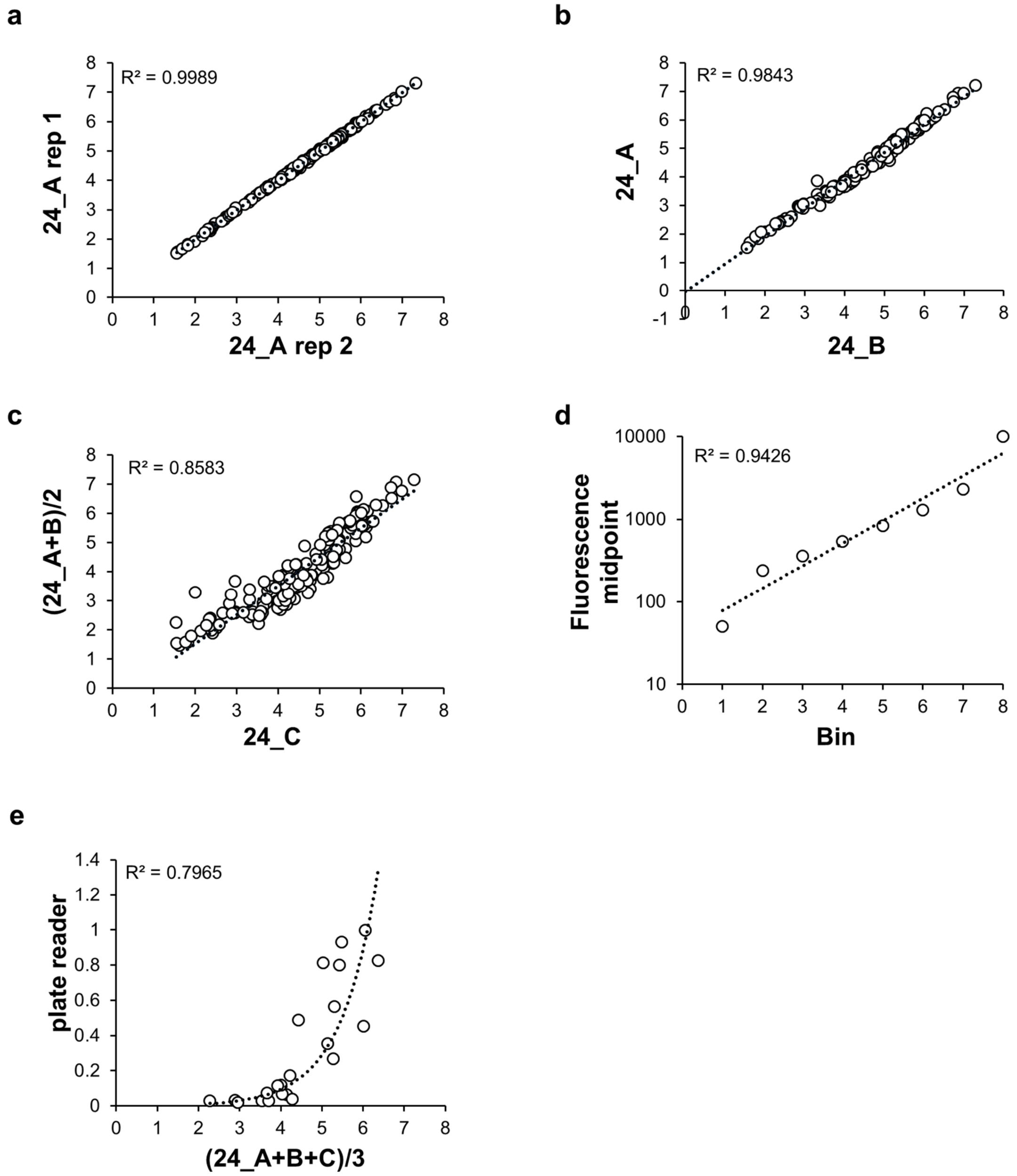


Figure 35. Flow-seq reproducibility within and between experiments.
Correlations between **a**, sequencing runs (R^2=0.9989), **b**, technical replicates (R2=0.9316) as well as between **c**, biological replicates (R^2=0.8583) are highly correlated. **d**, Fluorescence midpoints of the 8 FACS bins show an exponential fit (R^2=0.9426, $y=41.679e^{0.627x}$). **e**, Correlation between plate reader measurements and Flow-seq data (R^2=0.7965, $y=0.0011e^{1.193x}$).

4.2.3. Investigating translation dynamics using polysome profiling

To study translational dynamics, I performed polysome profiling followed by high-throughput sequencing. Polysome profiling utilises sucrose gradient centrifugation to allow the separation of light and heavy molecular weight molecules (Figure 36). As ribosomal RNA is the most abundant RNA species in cells, fractionation and the simultaneous monitoring of the UV absorption profile allows the visualisation of ribosomal subunits, as well as monosomes and polysomes along the gradient. Using this distribution profile, the translation dynamics of mRNA molecules can be studied by visualisation of their association with ribosomal proteins (same sedimentation rate), by e.g. Northern blotting, RT-PCR or, in case of transcriptome-wide studies, using microarrays or high-throughput sequencing (Arava et al., 2003; Ingolia et al., 2011). Quantifying the association of ribosomes to mRNAs may give hints to the translational state of a particular transcript.

Hek293 GFP pool cells were treated with cycloheximide to immobilise elongating ribosomes on transcripts (Schneider-Poetsch et al., 2010). Cytoplasmic cell lysates were then subjected to polysome profiling as outlined in figure 36. The UV absorption profile shows good peak separation between ribosomal subunits and polysomes. RNA extracted from all collected fractions further recapitulates the separation of RNA species by molecular weight – small RNA species (e.g. tRNA, 5S rRNA) are found almost exclusively in low sucrose fractions, whereas 18S and 28S rRNA are only present in medium to high sucrose fractions. To further assess the quality of fractionation, the distribution of GAPDH mRNA was measured using quantitative RT-PCR. As a highly abundant housekeeping gene, GAPDH would be expected to be primarily associated with polysomes. The results show that the vast majority of GAPDH RNA is indeed found in heavier polysomal fractions. The same qRT-PCR analysis was performed using primers specific for the 5’ UTR of GFP to monitor the distribution of GFP mRNA across all fractions. The distribution of GFP is broader compared to GAPDH, with some RNA not associated with assembled ribosomes, and the bulk associated along all polysomal fractions, which most likely reflects the variation in translational states between different variants within the pool.

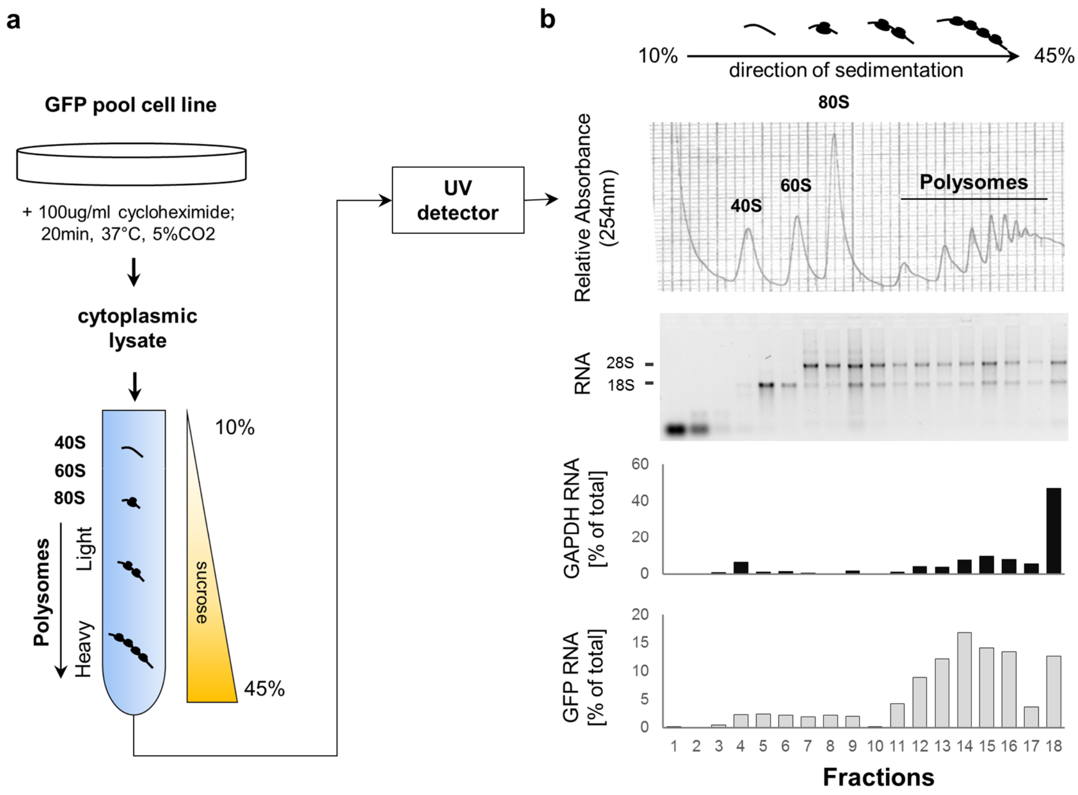


Figure 36. Measuring translation dynamics of GFP variants using polysome profiling. **a**, Schematic overview of sucrose gradient centrifugation. Cells are treated with 100ug/ml cycloheximide for 20min at 37°C, 5% CO2 to freeze actively elongating ribosomes. Ribo-/Polysomes are separated through sucrose gradient centrifugation (10%-45%) to allow separation of molecules by their specific sedimentation rate from low (top) to high (bottom) molecular weight. **b**, Output of polysome profiling. The UV absorption profile is measured for gradients described in **a,** starting from top (low molecular weight) to bottom (heavy molecular weight). Peaks corresponding to ribosomal subunits, as well as mono- and polysomes are indicated. The gradient is collected as 18 fractions, followed by RNA isolation. Relative RNA level of GAPDH and GFP are measured for all fraction using qRT-PCR.

4.2.4. Measuring RNA localisation

Many viral transcripts have been shown to rely on viral adaptor proteins, such as Rev, to mediate efficient nucleo-cytoplasmic export and expression (Malim et al., 1989). The requirement for such adaptors can be circumvented when the usually very AT-rich viral genes are codon optimised to closer reflect the preferred codon usage of the host cell, which for human cells consequently also implies an increase in average GC-content (Kotsopoulou et al., 2000; Shin et al., 2015; Tan et al., 1995). We therefore hypothesised that RNA export might be a bottleneck in the expression of particularly GC-poor GFP variants.

To estimate RNA export, I performed subcellular fractionations to measure cytoplasmic and nuclear RNA levels separately. Cellular fractionation protocols that rely on one or more cell lysis steps and which can be easily performed in small-scale are notoriously prone to cross-contamination between both fractions, either due to insufficient lysis leading to nuclear fractions containing large amounts of cytosolic components (e.g. ER), or damage to the nuclear membranes causing leakage of nuclear content into the cytoplasmic fraction, often leading to large variation in fractionation quality amongst replicates. More sophisticated protocols that allow the separation into different organelles are technically very challenging, require very large amounts of cells, and do not easily allow the processing of multiple samples at a time. I therefore optimised existing protocols to ensure reliable separation, yet allow multiple samples to be processed simultaneously in a technically reasonable manner. The method is described in detail in Materials and Methods 2.2.3 and is based on the cellular fractionation protocol published by Gagnon et al. (Gagnon et al., 2014), but modified to be performed in small-scale (1.5ml tubes) and includes a further clean-up step using a sucrose cushion as described by Zaghlool et al. (Zaghlool et al., 2013) to avoid nuclei carryover into the cytoplasmic fraction. I also added a second, stronger lysis step as described by Wang et al. (Wang et al., 2006) to ensure the more efficient removal of ER components surrounding nuclei. I assessed the quality of fractionation by monitoring the presence (or absence) of ribosomal precursor RNA (nuclear) and tRNAs (cytoplasmic) using the Agilent Bioanalyzer as well as by qRT-PCR by measuring the relative abundance of particular marker RNAs (as shown in Figure 37).

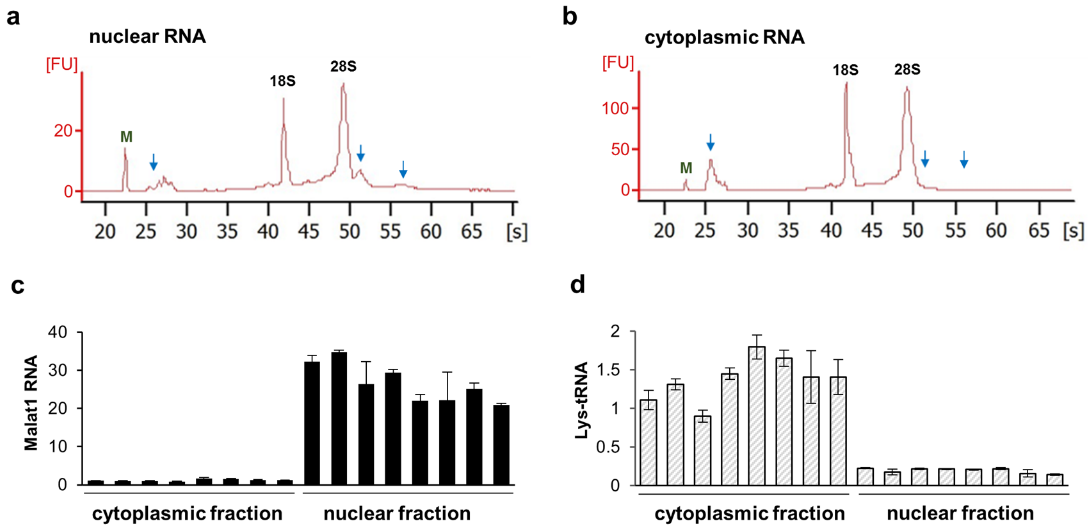


Figure 37. Validation of subcellular fractionation of Hek293 cells.
Subcellular fractionation was performed on Hek293 cells according to the protocol outlined in 2.2.3. RNA from both nuclear and cytoplasmic fractions was isolated and analysed using an Agilent bioanalyzer. The presence of ribosomal pre-cursor RNAs (52s and 56s) in the nuclear fraction (**a**), as well as an enrichment of tRNAs (26s) in the cytoplasmic fraction (**b**) are indicative of successful fractionation. M denotes the 25nt marker. qRT-PCR analysis is additionally used to measure the relative abundance of specific RNAs, e.g. chromatin associated non-conding RNAs such as Malat1 (**c**) or particular tRNAs (**d**, lysine tRNA-CTT).

GFP expression of Hek293 GFP pool cells was induced for 24hrs before performing cellular fractionation. Both nuclear and cytoplasmic fractions were prepared for high-throughput sequencing followed by data processing. For each GFP variant, the relative cytoplasmic concentration of its mRNA (RCC) was calculated as the ratio of cytoplasmic read counts to the sum of reads from both fractions ($RCC = \frac{c(cyto)}{c(cyto)+c(nuc)}$; Figure 38). A value of 0 therefore indicates 100% nuclear retention, whereas a value of 1 indicates 100% cytoplasmic localisation. Variants were filtered by read counts (>1000) and an RCC score calculated for each. In Hek293, scores ranged from 0.17 to 0.77, with the majority being more cytoplasmic than nuclear (RCC > 0.5, 62.8%). No variants localise exclusively to one or the other fraction only. Fractionations were also performed on HeLa pools with and without intron in the 5'UTR (2 replicates each). For fractionations of cells expressing GFP without an intron, the two replicates correlate (*r*=0.69) but show variation in the distribution pattern of RCC scores (Figure 38b, c and d). The replicate fractionations of the intron-containing pools also correlate (*r*=0.71) and cluster together (Figure 38b).

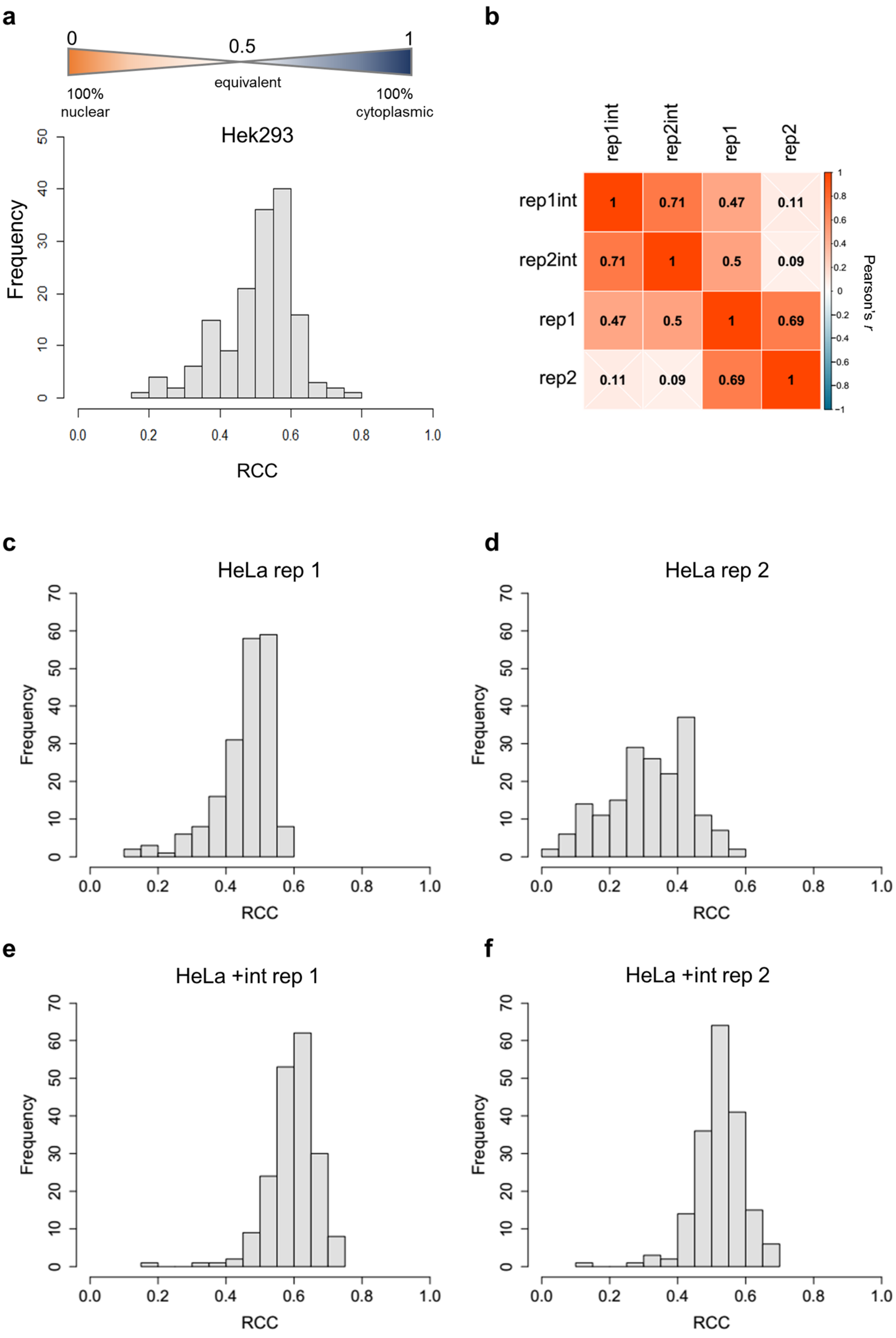


Figure 38. Relative cytoplasmic concentration of RNA of GFP variants expressed in Hek293 cells.

The RCC for each variant is the ratio of cytoplasmic read count to the sum of reads from both cytoplasmic and nuclear fraction. Histograms show the frequency of RCC scores for **a**, Hek293, **c+d**, Hela and **e+f**, Hela + intron cell line pools. **b,** Spearman correlation matrix of all HeLa replicates. Rep1/2 = no intron, Rep1int/2int = with intron.

4.2.5. Measuring RNA levels and stability

Codon usage has recently been shown to be a factor determining RNA stability and steady-state levels in yeast (Presnyak et al., 2015; Radhakrishnan et al., 2016) as well as zebrafish (Mishima and Tomari, 2016). Furthermore, I previously showed for two sequence variants of GFP with varying GC content that RNA half-life was increased with high GC (chapter 3.7, Figure 20). I therefore wanted to first test how GFP mRNA abundance varies across all GFP variants. To do so, GFP expression in the Hek293 GFP pool was induced for 24h followed by genomic DNA and RNA extraction and sequencing library preparation. Since the copy numbers of GFP variants vary within the pool, the read counts resulting from the RNA-seq were normalised to the read count from the gDNA sequencing. By doing so, the resulting normalised RNA levels between GFP variants vary by up to 1300-fold.

To analyse possible bias in RNA-seq library preparation (as discussed above), I transfected 12 GFP variants individually into Hek293 Flp-in host cell lines in order to validate the results obtained from high-throughput sequencing by qRT-PCR. As shown in figure 40a, RNA measurements between RNA-seq and qRT-PCR do not correlate, suggesting that either of the two quantification methods is biased. Results from qPCR analysis may not necessarily be representative of full-length mRNA measurements since the primers used anneal within the 3'UTR, whereas for RNA-seq library preparation, primers used anneal to both 5' and 3'UTR. On the other hand, bias could also arise in RNA sequencing due to differences in the efficiency of cDNA synthesis which may vary depending on RNA structure (discussed in 4.1.2.3). Recently, several studies utilised a novel reverse transcriptase enzyme called TGIRT (inGex), a thermostable group II intron reverse transcriptase, for the synthesis of particularly highly structured RNAs, such as tRNAs (Mohr et al., 2013; Qin et al., 2016). I therefore prepared two sequencing libraries from the same RNA sample, either using SuperScriptIII or TGIRT, to directly compare RNA levels obtained with each enzyme. The correlations between read counts obtained from both sequencing runs are shown in figure 39. Despite data overall correlating (R^2=0.7572), some variation is visible, suggesting that at least some of the discrepancy between high- and low-throughput experiments could indeed be caused by sequence biases. Since the genomic DNA sequencing of this particular set of RNA sequencing experiments was not available at the time of writing this thesis, results obtained using TGIRT could not be directly compared with qPCR results in order to address the question of sequence bias further.

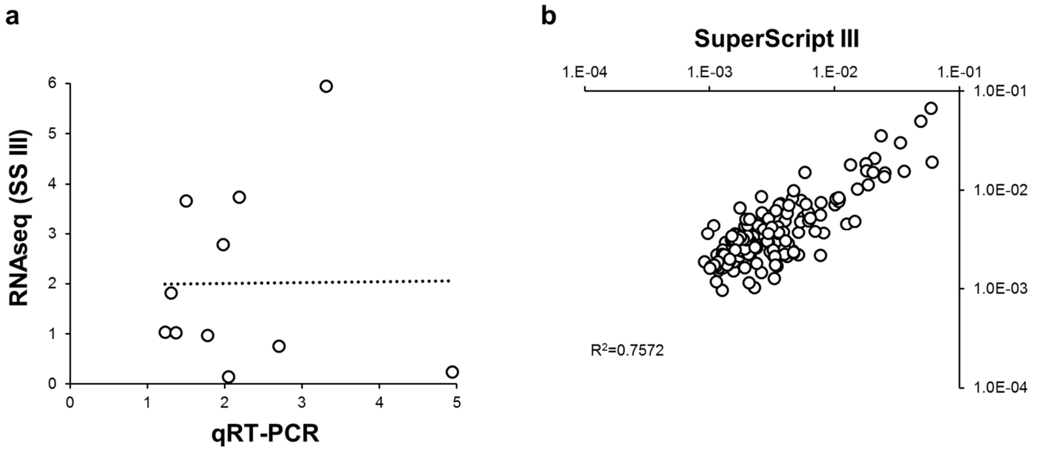


Figure 39. RNA levels measured by RNA-seq
a, RNA levels of 12 GFP variants as determined by RNA-seq of Hek293 pool cell lines using SuperScriptIII and normalised to gDNA, plotted against qRT-PCR results of transient transfected Hek293 Flp-in cells (R^2=0.0001, *p*=0.9189). **b**, Normalised read counts of RNA-seq data of HeLa GFP pool cells, prepared using either SuperScript III or TGIRT (R^2=0.7275).

Since RNA levels vary between GFP variants, it is likely that this is, at least in part, caused by differential RNA stability. To measure whether this is the case, I induced GFP expression for 24h before performing a transcription-block time course experiment using 500mM Triptolide. Unlike other more commonly used agents, such as Actinomycin D, which preferably intercalates into GC-rich DNA (Bailey et al., 1993), Triptolide is sequence-independent. Triptolide acts by inhibiting XPB, a subunit of TFIIH, which makes it highly selective for the inhibition of RNA polymerase I and II transcription initiation, allowing normalisation of any mRNA to RNA transcribed from RNA polymerase III. The exponential decay of c-Myc normalised to 7sk, a RNA polymerase III transcribed RNA unaffected by Triptolide treatment, is shown as a control for a successful block of RNA polymerase II transcription (Figure 39b). To be able to normalise GFP read counts between time points, equal amounts of GFP_000 and GFP_001 RNA (5% each) were spiked into every sample before cDNA synthesis.

As can be seen in figure 39a and c, GFP expression levels decrease with prolonged treatment with Triptolide in both Hek293 and HeLa cells. Results between HeLa cell lines are directly comparable and highlight the much higher, initial amount of total GFP (0h time point) when expressed with an intron. In order to analyse the sequencing results, the data needed to be normalised between time points for which I utilised the GFP_000 and 001 spike-ins. However, the ratio between these two variants varied up to 10-fold between samples, causing the normalisation between time points to be unreliable and data from this experiment not useable at the time of writing this thesis.

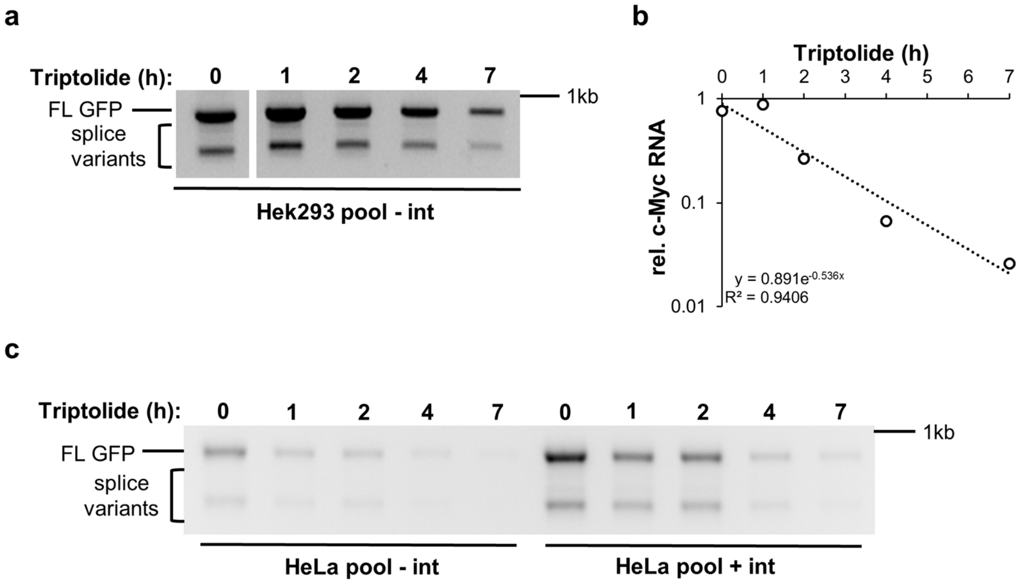


Figure 40. Transcription block time course using Triptolide to measure GFP half-life.
a, Hek293 GFP pool cells and **c,** HeLa GFP pool cells with or without intron were induced for 24h before addition of 500mM Triptolide to block transcription. Cells were harvested at indicated time points and RNA extracted. 500ng total RNA was used in cDNA synthesis and equal volumes used in subsequent PCR amplification. **b,** Exponential decay of c-Myc ($t_{1/2}$=1.29h). c-myc levels were measured using qPCR and data normalised to 7sk levels, a RNA Polymerase III transcribed RNA unaffected by Triptolide treatment.

4.3. Discussion

In this chapter, I describe the methods used to establish a human cell line-based system to monitor the phenotypic effects of codon usage on gene expression on many coding variants of GFP simultaneously. I discuss important considerations that had to be taken into account during the experimental design and present data of various control and validation experiments to establish this experimental system as a valid method that could be applied in similar phenotypic screens in the future.

4.3.1. A human cell line system for the phenotypic profiling of fluorescent reporter genes

I utilised the Flp-in system to establish a pool of cell lines in which each cell contains a stable integration of one of 217 GFP synonymous variants in the same genomic locus under the control of a tetracycline/doxycycline-inducible CMV promoter. I established Hek293 as well as HeLa GFP pool cell lines to be able to compare and contrast possible cell-specific effects seen across different cell lines, and additionally established cell pools expressing GFP with the same chimeric intron in the 5'UTR as previously used in single GFP transfection experiments presented in chapter 3. Inducibility of GFP expression of all pool cell lines was measured by FACS before the integration frequency of each GFP variant was confirmed by high-throughput sequencing of genomic DNA. However, not all GFP variants were detected in all cell line pools. One possible explanation is the poor transfectability of Flp-in Hek293 and Flp-in HeLa cell lines which seems to be different from ordinary Hek293 and HeLa cells respectively. Both lines are very sensitive to the transfection procedure as substantial cell death can be seen after transfection, as well as during chemical selection. For this reason, I prepared multiple batches of pool cell lines in the process (Table 2) after optimising the transfection protocol and upscaling the number of cells to be transfected, to be able to recover a high enough number of transfectants and to ensure the distribution of GFP variants is more representative of the frequency within the respective vector pool. The cell lines presented in this chapter are those that were used in all further experiments.

Since the preparation of high-throughput sequencing libraries can be subject to bias, I optimised all protocols to reduce noise that could be introduced due to the varying sequence features. Most polymerases do not cope well with extremely AT- or GC-rich templates, often leading to large differences in sequence coverage. This might not necessarily be a limitation in studies in which expression of particular genes is compared across various conditions, but in our case, could lead to the underestimation of transcript abundance of variants with more extreme sequence composition. To decrease such sequence biases, I selected a few GFP variants representative of the full range of possible GC to AT ratios present in the total variant pool, for optimisation experiments. In agreement with other published studies that directly compared various DNA polymerases and their effects on library compositions (e.g. %GC, average library size, Aird et al., 2011; Dabney and Meyer, 2012), I show that all tested enzymes exhibit different degrees of amplication efficiencies for different templates (Figure 27) and thus, chose the most unbiased polymerase, Accuprime Pfx, for all future experiments. Another possible source of error lies within the cDNA synthesis step. Using *in vitro* transcribed RNA of variants covering a broad range of GC content as template, as well as introducing an RNase H digest, I was able to reduce the amplification bias notably (Figure 28). Despite optimisation, the results of total RNA sequencing of Hek293 GFP pool cells divert strongly from qRT-PCR measurements of single GFP constructs in transfected cells. One possible reason could be that during library preparation, PCR products are size-selected to ensure only full length GFP amplicons will be sequenced. Furthermore, as discussed in more detail in chapter 3.8, some GFP variants are subject to cryptic splicing, which can also be seen in figure 40a and c. This may lead to further errors if RNA quantification is performed by qRT-PCR, as it is not possible to measure expression of such splice isoforms separately, since their sequences are unknown. To validate RNA-seq results, it would therefore be helpful to quantify RNA levels of either only full-length transcripts by semi-quantitative PCR, or utilise Northern blotting by using probes specific for 3' or 5' UTR to quantify RNA of the correct length only. This may lead to much better correlations, as well as provide a better picture of the possible transcript isoforms formed from each GFP variants. Likewise, it would be interesting to perform RNA-seq without size-selection to sequence splice isoforms regardless of their length. By mapping these back to the parent GFP, it would be possible to further investigate the sequence-features leading to this aberrant splicing phenotype.

Difficult sequence features, as well as the lack of 3' to 5' proof-reading ability of commonly used RTases, might lead to additional error being introduced during reverse transcription. Recently, a new class of enzymes, thermostable group II intron reverse transcriptases (TGIRTs), has become available (Mohr et al., 2013). It was shown that TGIRT can operate at higher temperatures and has higher processivity and fidelity than conventional RT enzymes, such as SuperScript III used in the described experiments above, and therefore gives a more uniform 5' to 3' coverage, as well as better coverage of small and highly structured RNA species, such as tRNAs (Mohr et al., 2013; Nottingham et al., 2016; Qin et al., 2016). Since it is likely that some GFP variants will have strong secondary structures due to their sequence composition, I made additional sequencing libraries from total RNA of the HeLa GFP pool cells using TGIRT to be able to directly compare differences in variant frequencies. Although the required gDNA results for RNA level normalisation was not available at the time of writing this thesis, the direct comparison of the total RNA sequencing results prepared with either SuperScript III or TGIRT shows that despite data overall correlating well (Figure 39b, R^2=0.7572), variation is clearly visible, confirming that both enzymes indeed have varying processing activities. Whether TGIRT provides a less biased view of RNA levels, will require data normalisation (once possible) which can then be compared to low-throughput measurements as described above. Recently, the Ellington group synthetically re-engineered a proof-reading DNA transcriptase, 'reverse transcriptase xenopolymerase' (RTX), to accept DNA as well as RNA as templates, abolishing the need for separate RT and PCR reactions altogether (Ellefson et al., 2016). However, since this is a very novel publication, no direct comparisons between RNA-seq results obtained with RTX, TGIRT and other standard enzymes have been published as yet.

4.3.2. Flow-seq data validation

I show that the here optimised and used Flow-seq approach for measuring the fluorescence of many GFP variants is highly reproducible by sequencing several technical replicates from the same day, as well as biological replicates from different days, which are all highly correlated (Figure 35b+c, R^2=0.9843 and R^2=0.8583 respectively). The obtained average fluorescence scores for each variant are also correlated to fluorescence measurements independently acquired on a plate reader (Figure 35e, R^2=0.7965), further confirming this method as a valid approach to measuring fluorescence levels. Due to the good exponential fit between plate reader and Flow-seq data, I applied an exponential transformation to all Flow-seq data using the obtained exponential coefficient, to recalibrate and fit the data to a linear scale. The resulting values represent the protein expression scores for every individual GFP variant used for any further data analysis.

The original Flow-seq method published by Kosuri et al. (2013) utilises mCherry expressed from a bidirectional promoter as an internal control to further reduce experimental noise. Fluorescence bins were defined by the mCherry-to-GFP ratio, similarly as in other studies (Dean and Grayhack, 2012; Noderer et al., 2014). The advantage of having a second reporter gene is that quantification can be independent of the copy number of expressed genes and allows to filter out cells in which expression is e.g. not fully initiated. In the cell line system used here, every cell contains only one integration site, restricting gene copy number to one, which should in theory abolish the need to correct for varying expression levels using a second reporter. In Hek293 cells, the fluorescence range covered by single GFP variants is relatively narrow already, whereas in HeLa cells, expression is much broader (Figure 33). In a study by Gamble et al. (Gamble et al., 2016), normalisation was also performed using the ratio of GFP to mCherry to reduce transcriptional noise, despite only a single gene copy being present in each cell. The correlation between Flow-seq data of different batches of HeLa cells which were induced on different days is very high (Figure 35c, R^2=0.8583), suggesting that the normalisation to another fluorescent marker is not required, but may still be useful.

When comparing the total number of reads of each variant to its respective GC3-content, no correlation was found (R^2=0.04), confirming that our protocol for library preparation indeed prevents the introduction of a systematic AT-skew. The distribution range of fluorescence varies about 100-fold for both Hek293 and HeLa cell lines, with the majority of variants exhibiting intermediate fluorescence levels with peaks in well-defined neighbouring bins (Figure 34c). During FACS-sorting, it was noticed that cells falling into the highest fluorescence bin were being sorted with a markedly lower sorting efficiency than all other bins (about 15% lower) which could not be attributed to any technical reasons (e.g. position of selected sorting stream). One common reason for low sorting efficiency are morphological differences within the cell population, either leading to issues with droplet formation and separation important for single cell analysis, or too many cells being discarded due to user-set size-gating. I speculated that high protein levels within the cells may lead to bulging of cells due to protein aggregations, however, neither cell size nor cell granularity were notably different (data not shown) suggesting that cell bulging is not the main cause for lower sorting efficiency. This observation could be followed up by a more qualitative assessment of differences in cell morphology by live cell imaging as more subtle changes in cell shape may not have been possible to filter efficiently by FACS due to the population mix.

Some variants exhibit a broad fluorescence distribution across all collected bins, more so in HeLa than Hek293, and a small number of variants display a U-shaped distribution (Figure 34d). The relatively broad fluorescence range of some variants could be the sign of partial gene silencing. Since all GFP variants are integrated into the same genomic context, this observation is unlikely to be caused by positional effects similar to what can be seen when transgenes are integrated into random genomic contexts (such as transcriptionally silenced heterochromatin). Partial silencing can also often be attributed to contact-inhibition. This mechanism acts to prevent adherent cells growing in monolayers, such as the here used HeLa and Hek293 cell lines, to overgrow. As a result, if cells become too densely grown, cell proliferation, and thus gene expression in general, will be inhibited. Throughout the experiments presented here, cells were grown at sub-confluency to avoid such effects leading to heterogeneous expression patterns. Since Flow-seq data obtained for independent batches of cells is highly correlated (Figure 35c), it is unlikely that this is a major issue in the experimental set-up. The unusual U-shaped fluorescence phenotypes of some GFP variants could however also hint towards a more permanent silencing state. DNA methylation is a highly dynamic process which has been shown to vary even within homogenous populations of cells (Smallwood et al., 2014). In the context of gene therapy, this is one of the major hurdles that needs overcoming for successful and prolonged transgene expression. A study by Bauer et al. investigated the role of CpG dinucleotide content and methylation on transgene expression, also using GFP synonymous sequence variants. By expressing CpG variants of GFP in Hek293 Flp-in and CHO Flp-in cells, it was shown that under selective pressure, expression of GFP remained constant for more than one year (Bauer et al., 2010). However, if the selection pressure was removed, expression would gradually decrease after 50 days. This was attributed to increasing CpG methylation. It was further suggested that due to the close proximity of the hygromycin resistance gene to the transgene-driving promoter, the constant and high expression of hygromycin under selective pressure keeps the chromatin directly downstream unmethylated and thus, in an open and accessible state. The loss of transgene expression after prolonged abolishment of selective pressure was shown to be linked to differential nucleosome positioning and a decrease in transcriptional activity (Bauer et al., 2010). In case of the Flow-seq experiments presented here, the passage number of GFP pool cell lines was kept as low as possible and cells were constantly grown under selective pressure. This suggests that the mechanisms leading to the very broad or even U-shaped distribution patterns are either different from the mechanism previously proposed, or are acting despite high expression of hygromycin, suggesting some sequences are more prone to methylation-mediated gene silencing than others. To test whether these observed phenotypes are indeed caused by gradual DNA methylation, the Flow-seq experiment could be repeated using cells passaged for a prolonged period of time (weeks to months) with or without selective pressure, similar as presented by Bauer et al. Results from these experiments could be directly compared to each other as well as to the results presented here. This would allow us to identify variants which are silenced more readily than others, as well as identify those that maintain high expression in both conditions. Those two groups could be further compared to each other in terms of their methylation status, e.g. through bisulphite sequencing, or within each other, as similar molecular behaviour could be caused by common sequence features between variants. Those variants could be characterised further to determine the causes for slow vs fast gene silencing. The results of such experiments would greatly benefit the design of transgenes for which high expression is desirable over a prolonged period of time (e.g. gene therapy).

It was further suggested that transgene expression driven by non-endogenous promoters, such as the here used CMV promoter, are more prone to methylation-induced silencing than when endogenous housekeeping promoters are used, such as A_2UCOE (Ubiquitous Chromatin Opening Element derived from the human HNRPA2B1-CBX3 locus) which allows efficient expression even in the absence of selective pressure or when integrated into heterochromatin (Zhang et al., 2010). In section 4.1.2.2 I demonstrate that there is no visible PCR Polymerase amplification bias between two GFP variants varying strongly in their GC content *in vitro* (Figure 17) and also do not exhibit any significant differences in nuclear transcript abundance (Figure 19a). However, this does not exclude the possibility that on a larger scale, some variants could be limited by their transcription rate. Performing transcription run-on assays to monitor this directly is technically challenging for the number of GFPs to be compared, as well as due to common sequence fragments making high-throughput approaches with this set of variants difficult to analyse. Hence, by utilising an A_2UCOE-driven expression system, it might be possible to circumvent this limitation as the comparison to previous experiments could allow us to discover a subset of variants that is inhibited either due to general low-levels of transcription, or due to the accumulation of methylation, eventually leading to complete transcriptional silencing. Additionally, this would allow the more reliable dissection of transcriptional noise from all other experimentally determined parameters and improve fluorescence measurements. Alternatively, it could also be possible to repeat similar experiments in DNMT3 knock-out cell lines. DNMT3 DNA methyltransferases are essential for *de novo* methylation and their inactivation should allow efficient expression of transgenes without the influence of methylation-induced gene silencing (Okano et al., 1999). This could be selectively tested only on those variants which are likely being silenced due to DNA methylation.

A recent paper suggests that transcription factors also bind to protein-coding regions of the human genome, with a potential preference for highly expressed genes (Stergachis et al., 2013). However, whether and how potential TF-binding affects expression, remains controversial. Data from Flow-seq experiments could therefore also be analysed in regards to the potential of particular GFP variants to interact with TFs. By performing DNA-footprinting assays and correlating expression with TF binding, this approach could shed further light on the involvement of coding-sequence TF-binding in modulating gene expression and how his might interplay with other regulatory codes (Weatheritt and Babu, 2013). It would be in particular interesting to measure correlations between GC-content, TF-binding and expression, as it was previously suggested that coding-sequencing binding TF exhibit a preference for GC-rich sequences. This was however disputed by another study claiming this result may be due to the types of computational methods applied, which might introduce sequence-bias (Agoglia and Fraser, 2016). So far, no other experimental studies have followed up on this issue. The methods applied here could therefore directly address this question further.

4.3.3. Monitoring changes in translational states using polysome profiling

To study the translational dynamics of all GFP variants in our system, I performed polysome profiling on cytoplasmic lysates of Hek293 GFP pool cells (Figure 36). This method allows to measure the abundance of transcripts in ribosome-associated fractions after sucrose gradient centrifugation. Even though it is not possible to reliably infer translational efficiency from this technique, as ribosome association is generally poorly correlated to protein yield, it can nonetheless provide an idea of the translational state of a particular transcript, e.g. is a transcript associated to ribosomes or never at all (Arava et al., 2003; Newman et al., 2016).

In the experiment presented here, I utilised cycloheximide as a ribosome inhibitor. More specifically, it inhibits ribosomal translocation, freezing actively elongating ribosomes to the DNA template. As can be seen in figure 36, I collected 18 different fractions from the sucrose gradient and isolated RNA from each. By qRT-PCR, I initially observed the distribution of GFP across the samples to firstly confirm that the method works and secondly, to observe how it relates to a highly expressed gene such as GAPDH. The results show that the GFP RNA distribution is very broad, with many transcripts sedimenting at a faster rate than 80S monosomes, suggesting that not all transcripts are actively being translated at a time. However, the majority is found in light and heavy polysomal fractions. To distinguish the translational states of GFP variants, I prepared sequencing libraries from different fractions. Since this was an initial test experiment, I pooled RNA from multiple fractions to overall obtain four different sequencing libraries: unassociated free mRNPs, monosomes, light polysomes (2-4) and heavy polysomes (5+). In later experiments, it could be considered to change the concentration of the sucrose gradient to obtain a finer separation of polysomes which would also allow us to sub-divide fraction pools further to obtain a better resolution picture for each variant. Overall, the results shown here confirm that this method can be used to measure differences in the translational state between genes, demonstrated by differences in distribution obtained for a control gene, GAPDH, and the total GFP pool. A study by Presnyak et al. utilised polysome profiling to measure the effects of codon optimality on ribosome translocation in yeast (Presnyak et al., 2015). By comparing the polysomal retention of optimal and non-optimal transcripts, it was demonstrated how codon optimality modulates translational elongation rates. The results from this experiment here can in first instance be used to analyse the effects of codon optimality on ribosome association. In addition, a similar experiment utilising e.g. harringtonine to block translation initiation followed by ribosomal run-off could be used to assess whether differences in elongation rates can be observed, similarly as shown in yeast (Presnyak et al., 2015).

Recently, more sophisticated protocols have been developed to study translational dynamics more globally with techniques such as Ribosome profiling (Ribo-seq) revolutionising the field (Ingolia et al., 2009, 2012). Ribosome profiling is a method which creates a footprint view of ribosome protected fragments (RPFs) on a nucleotide levels. By freezing ribosomes on the RNA, ribosome isolation and the partial digest of any associated RNA, followed by high-throughput sequencing, this method can be used to infer ribosome movements along genes in a global manner by comparing the density of associated ribosomes associated, as well as their distribution along the CDS (Ingolia et al., 2009). Since Ribo-seq relies on the assignment of RPFs to their respective RNA template, gene sequences need to be diverse enough to be able to accurately map the reads. In case of the GFP library used in the experiments here, some variants contain common sequence fragments or differ only by a few nucleotides. Therefore, back-mapping of the footprint reads to the respective GFP variant might not be reliable enough for meaningful analyses. With a different, i.e. larger and more complex library design, this current limitation could be circumvented and Ribo-seq could further our understanding of differences in ribosome movements along particular variants, by revealing differences in decoding speeds in individual codons (Tuller et al., 2011) or translational pause sites (Ingolia et al., 2011). However, comparing translational efficiency between variants is only possible when translation elongation rates are similar. Another possible limitation could be the short length of RPFs which might complicate data analysis if several distinct mRNA subpopulations are present, such as alternative splice forms. Thus, Ribo-seq combined with Polysome profiling may be required to build a more comprehensive and complete view of codon content and its effect on translation.

4.3.4. Investigating RNA phenotypes

Several studies have implicated codon usage as an important determinant of RNA export and subcellular localisation. I therefore performed nuclear-cytoplasmic fractionation of both Hek293 and HeLa GFP pool lines to measure potential differences in RNA localisation (Figure 38). To do so, I calculated the relative cytoplasmic concentration (RCC) of each GFP variant which indicates whether a particular transcript is either more enriched in the nucleus or cytoplasm, or whether it is roughly equally distributed. Furthermore, this experiment can also be used to study the effects of splicing on RNA localisation.

The calculated RCC values for all variants in the pool differ but correlate well between replicates. For HeLa, it is also noticeable that none of the variants expressed without intron have an RCC above 0.6, whereas on introduction of an intron, the histogram is shifted to the right with some variants now becoming more cytoplasmic (>0.6). This is an expected result since splicing is known to increase transcript stability as well as to facilitate nuclear-cytoplasmic export by mediating interactions between transcript and export machinery (Choi et al., 1991; Valencia et al., 2008). It is however unclear whether changes in the localisation ratio are exclusively caused by differences in mRNA export since splicing may also increase transcript stability (Gupta et al., 2013).

Recently published studies in yeast and zebrafish demonstrate a strong link between codon usage and RNA stability (Mishima and Tomari, 2016; Presnyak et al., 2015). To study effects on transcript stability, I performed a transcription block time course experiment with both Hek293 and HeLa GFP pool cell lines in order to be able to measure differences in RNA half-lives between GFP variants (Figure 40). As is already evident from semi-quantitative RT-PCR results of such a time course (Figure 40c), GFP levels are overall elevated for those constructs expressing GFP with an intron compared to those without. This suggests higher total RNA expression either due increased RNA export and/or stability. To be able to reduce sequencing noise caused by strongly varying transcript populations within samples from each time point, I used two reference variants, GFP_000 and GFP_001, as spike-ins for each sample to be able to normalise read counts across the entire time course. However, in the sequencing results of this experiment, the ratio between the reference variants changed between time points by about 10-fold, suggesting that this approach by itself might not be sufficient for normalisation and further adjustments have to be made.

5. Sequence determinants of gene expression in human cells

In the previous chapter I described the development and validation of a method to assess the effects of codon usage on the expression of several hundred synonymous variants of GFP by measuring multiple experimental phenotypes. Here, I present the results of previously described experiments. I compare and contrast results from two tested cell lines, highlight the relationship between various experimentally derived parameters, as well as relate results to calculated and predicted sequence features.

5.1. GFP expression varies between cell lines

To investigate the effects of synonymous codon usage on GFP expression, I used an approach based on Flow-Seq to estimate protein expression by GFP fluorescence (described in detail in chapter 4). I established GFP pool cell lines expressing 217 variants of GFP in two different host cell line backgrounds to test whether this approach is a valid method to measure general effects of codon usage on gene expression, as well as cell type specific differences. The Flow-seq experiments were conducted with Hek293 and HeLa GFP pool cells. From the resulting read count distribution across all fluorescence bins, the mean bin localisation of each GFP variant was calculated and data calibrated to previously obtained data from single GFP fluorescence measurements (described in chapter 3). Since not all GFP variants were equally well represented in both cell line pools, some variants were filtered out due to low read counts across all bins (<1000). Data representing measurements in HeLa cells in this and in the following sections are the average of 3 replicate experiments (R^2 = 0.82–0.96 between experiments); Hek293 data represents results from one experiment. After data filtering, 169 GFP variants were observed in both Hek293 and HeLa data sets (Figure 41). GFP expression correlates between both cell lines ($R^2 = 0.267$; $p = 6.4e^{-13}$) but large variation is also visible.

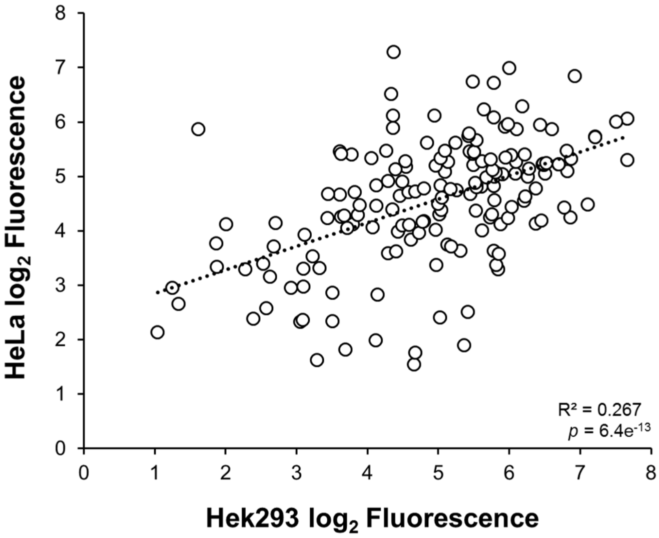


Figure 41. Comparison of GFP fluorescence levels between Hek293 and HeLa pool cells. Shown are the mean bin scores for 169 GFP variants observed in both HeLa and Hek293 data sets. HeLa data points represent the mean of 3 experiments, Hek293 data points are representative of 1 experiment. All cell lines were induced for 24h prior to FACS-sorting into 8 fluorescence bins.

5.2. GC3 as a determinant of GFP expression

Previously published studies showed a strong correlation between GC3 content and protein expression (Kudla et al., 2006). In chapter 3 of this thesis, I demonstrate that this can also be observed across a larger set of GFP variants and I expected to observe similar correlations in Flow-seq experiments. Since protein measurements between Hek293 and HeLa data show significant variation (Figure 41), this was done separately for each cell line to assess whether variation related to GC content is more pronounced in a particular cell type (Figure 42).

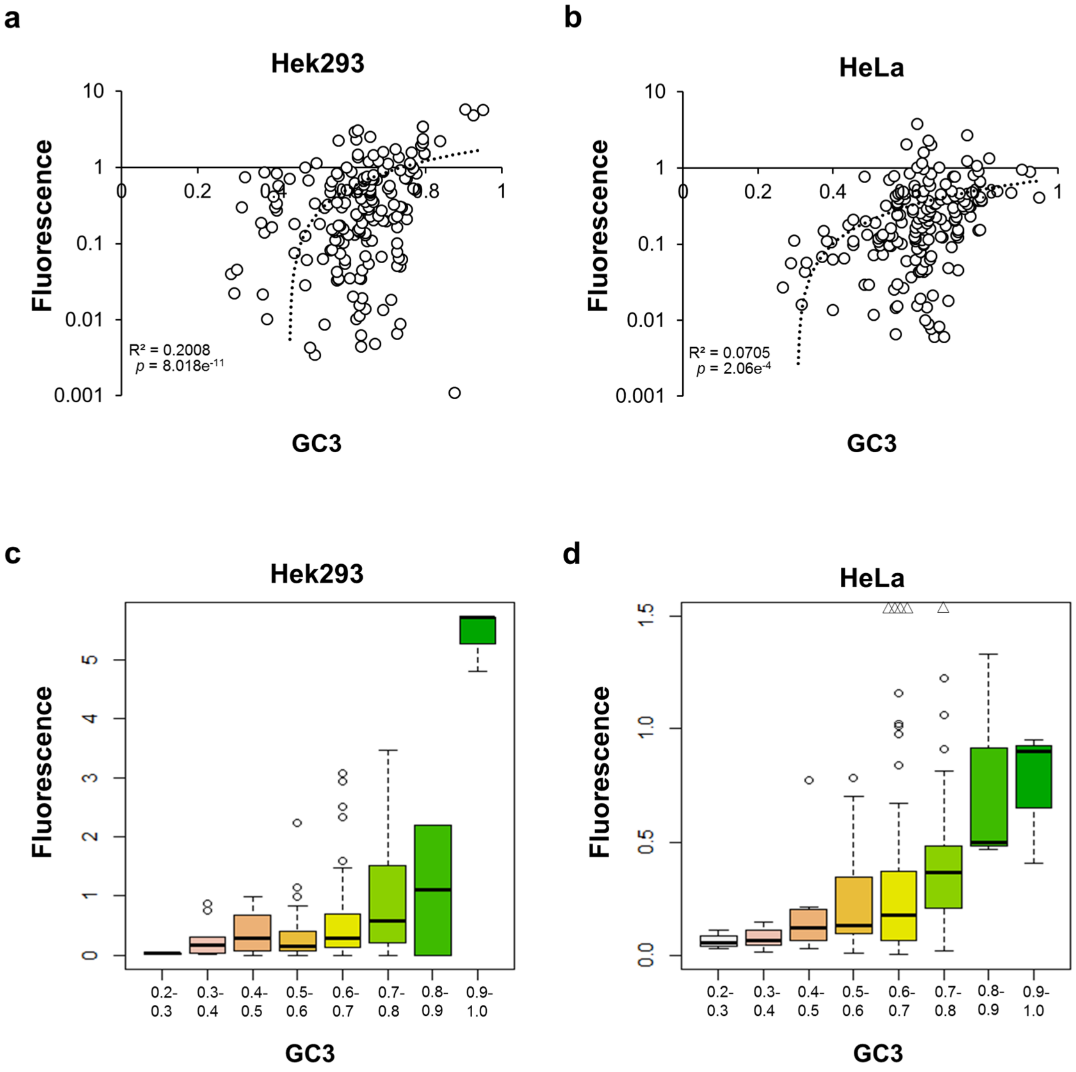


Figure 42. Comparison of GFP fluorescence with GC3 content.
Shown are the average bin localisations for 169 GFP variants observed in both **a**, Hek293 and **b**, HeLa data sets. Hek293 data is representative of one experiment, HeLa data points represent the mean of 2 experiments. All cell lines were induced for 24h prior to FACS analysis. **c + d**, Fluorescence distribution within defined GC3 bins for variants measured in Hek293 (**c**) and HeLa (**d**). Triangles indicate outliers.

A correlation can be seen in Hek293 (R^2=0.2008, p=8.018e^{-11}) and also, but to a much lesser extent, in HeLa (R^2=0.0705, p=2.06e^{-4}). This is surprising since I previously observed very strong correlations in single GFP transfection experiments conducted in HeLa cells (R^2=0.599, p=3.4e^{-7} for 36 variants; section 3.3.). Besides fundamental differences in the experimental methods, another possible explanation for the lower than expected correlation could be differences in sequence composition of some of the GFP sequences (discussed in the following section).

5.2.1. GFP expression is influenced by GC distribution

The majority of GFP variants used in the cell line pools were assembled from three distinct parts with an approximate length of 1/3 each (assembly described in Kudla et al., 2009, see Supplementary Figures 1-3). These 'thirds' can vary substantially in their GC content, leading to an overall uneven/heterogeneous GC distribution along the coding sequence ('uneven GC variants', Figure 43a). For this reason, 23 further sequences were designed to cover a very broad range of GC3 (0.26-0.95) with a more even/homogeneous distribution of GC along their sequences and were acquired as synthetic gene fragments (gBlock Gene Fragments, IDT; 'even GC variants', Figure 43a). As illustrated in Figure 43b, the fluorescence range covered by this set of variants is representative of the overall range covered by the entire library. As the sequences of all variants are known, it is possible to investigate whether differences in sequence composition can have an effect on expression. I therefore treated each group of GFP variants separately to assess if GC distribution is a relevant factor in determining protein levels.

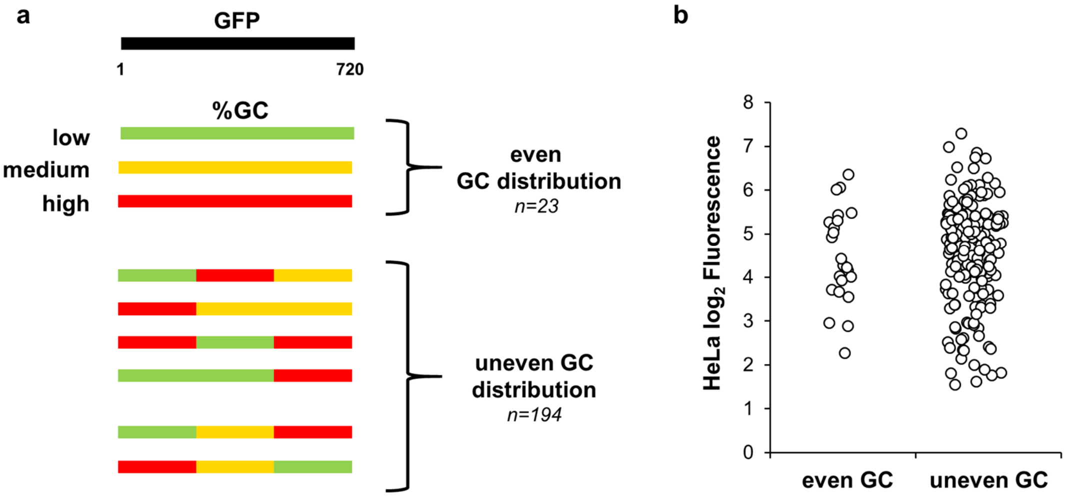


Figure 43. GC distribution varies across all GFP sequences.
a, 195 of 217 GFP variants present in the pool are made up of three thirds with varying GC composition, leading to an overall heterogeneous distribution of GC along their sequences. In contrast, 23 variant sequences were designed to cover a very broad GC3 range and to be homogenously distributed. **b**, GFPs with homogenous GC distribution are representative of the total fluorescence range covered by all other GFP variants in HeLa cells. Data shown are the means of 2 experiment conducted on HeLa cells.

When plotting protein levels against GC3 separately for each variant group, significant correlation can be observed for both groups (Figure 44a-d), however the correlation is stronger for variants with even GC distribution, both in HeLa and Hek293 cells. This observation indicates that the distribution of GC along the coding sequence is important for protein expression. For example, if high GC in a specific part of the gene was an important determinant of expression, then one would expect a high correlation of total GC with expression among even GC but not uneven GC variants. Alternatively, the discrepancy between even and uneven GC variants could be the result of experimental and/or computational biases, as discussed below.

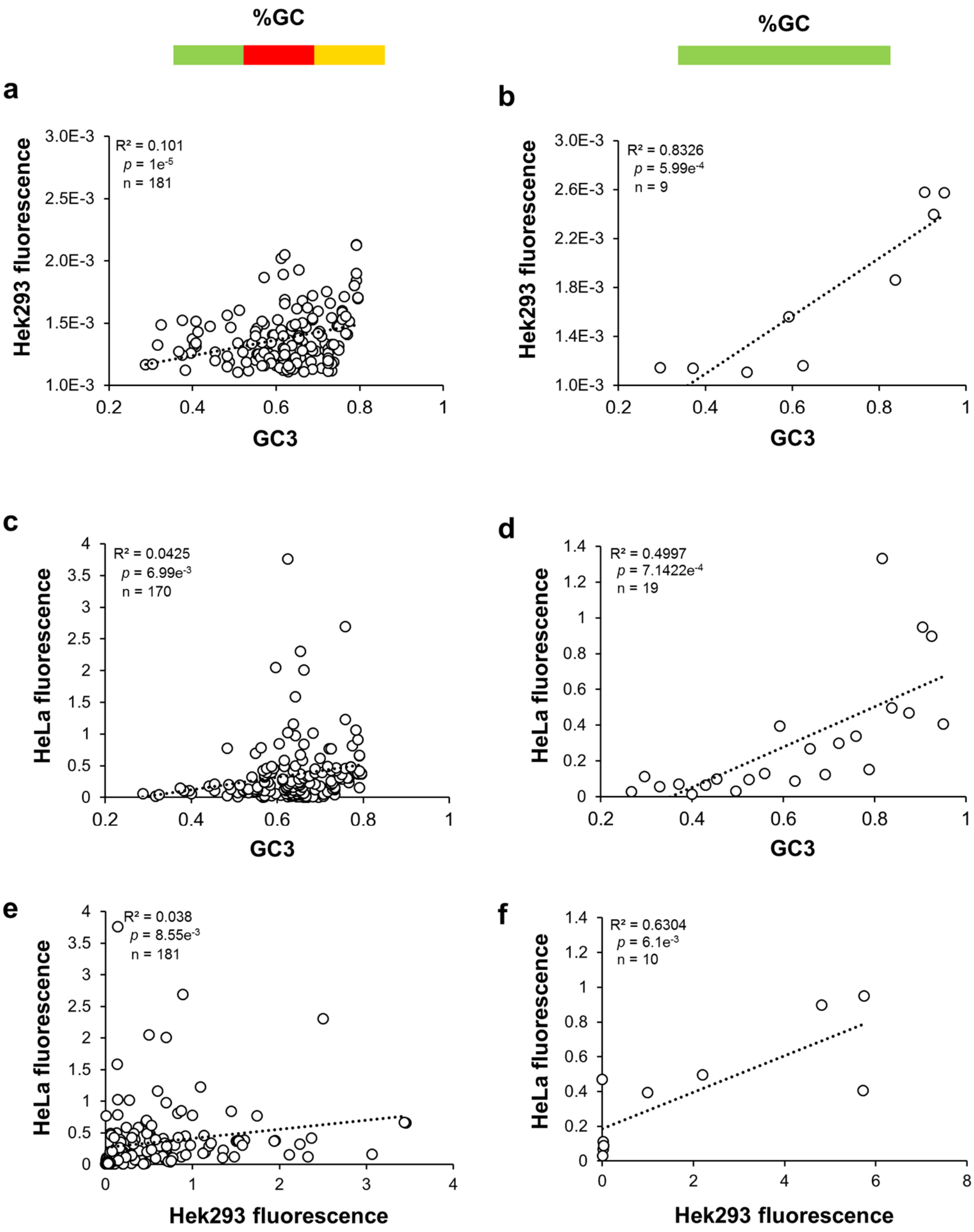


Figure 44. The effect of GC sequence composition on protein levels
GFP variants were separated into two groups according to their GC distribution: even (left column) or uneven (right column). **a-d**, Protein levels for each group are plotted individually against their total GC3 content for both Hek293 (**a+b**) and HeLa cells (**c+d**). **e+f**, Comparison of GFP expression between cell lines for different variant groups individually.

Noticeably, the number of data points for each group is very different. Variants with a homogenous GC make-up are underrepresented in the Hek293 data set, which leads to many of those being omitted from this analysis due to low read counts. Additionally, the total GC3 range covered by either group varies distinctly as the higher extremes of GC3 are absent in the heterogeneous GC3 group, whereas variants of the other group were specifically designed to cover a very broad range. When comparing expression between both cell lines individually for each variant group, data correlates between variants with even distribution (R^2=0.6304), but only weakly for those variants with uneven GC (R^2=0.038; Figure 44e+f). Overall, these results suggest that the GC distribution pattern is a factor contributing to GFP expression.

5.2.2. High GC at the beginning of the GFP gene increases protein expression

Since the relationship between high GC3 content and efficient protein expression potentially also depends on the overall distribution of GC along the sequence, I tested whether the position of GC rich sequences was important for efficient expression by correlating protein levels with the GC3 of each individual sequence third, separately for each GC distribution group. In Hek293 cells, the correlation is the highest with the first third (Figure 45, left column) whereas in HeLa cells, the correlation is highest with the second third (Figure 45, right column).

To further test the importance of GC3 of individual sequence thirds, I used multiple regression analysis to assess the relative contribution of each individual third to protein expression (Figure 46). In Hek293 cells, the GC3 content of the first third contributes the most to the overall variation seen on protein levels (R^2=0.2187, p=5.594e^{-10}), whereas the contributions of the other thirds become decreasingly insignificant. In HeLa cells, no clear difference in the contributions of different thirds can be seen as only the GC3 of the second third results in a marginally significant *p*-value (p=0.0438). The overall *p*-value however is significant (p=3.385e^{-3}), suggesting that the total GC3 is important, but might be independent of the position within the GFP sequence.

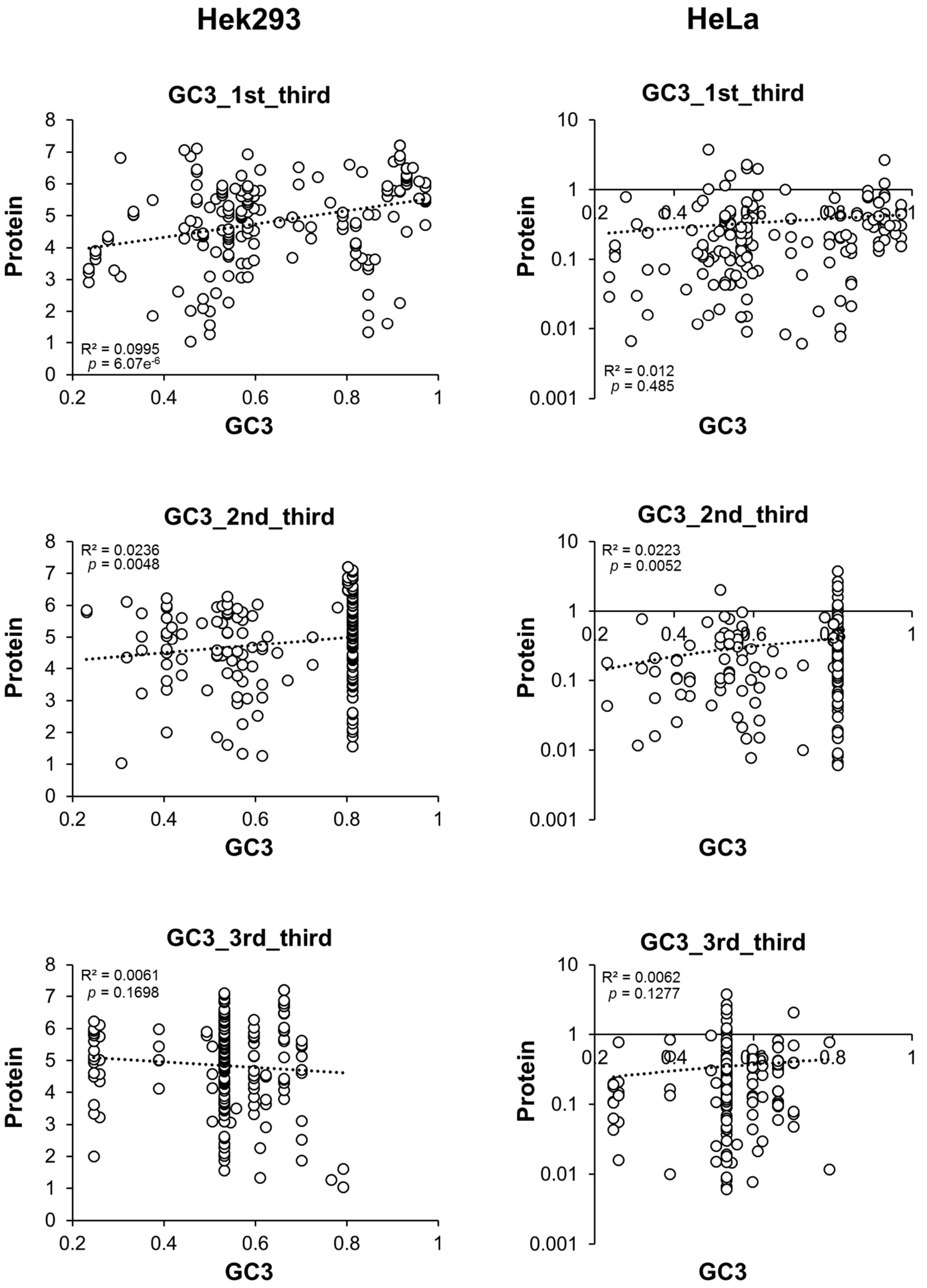


Figure 45. The influence of GC3 of three sequence thirds on GFP protein levels. Plotted are the protein levels as measured by Flow-seq in Hek293 (left column) and HeLa (right column) against the GC3 calculated for each sequence third individually.

a

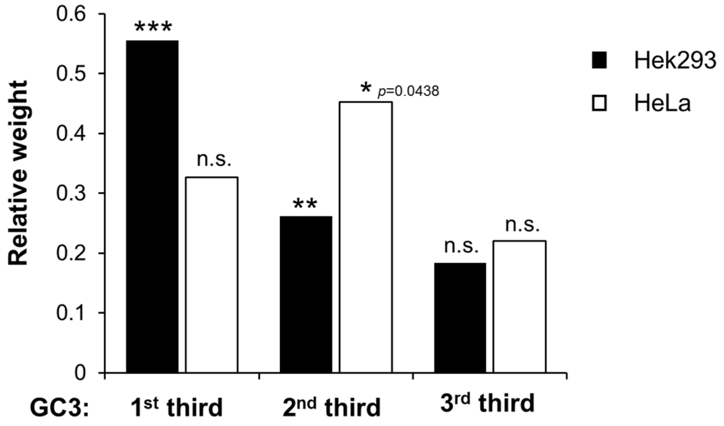


Figure 46. The relative contribution of GC3 of different sequence parts on protein expression.
Bar plots represent the relative weights of the GC3 content for each sequence third of GFP sequences. The sum of weights for all three thirds equal to 100% of the total variation observed in a particular cell line. Hek293 (black bars): R^2=0.2187, *p*=5.594e^{-10}; HeLa (white bars): R^2=0.0707, *p*=3.385e^{-3}.

To test the hypothesis that high GC3 at the beginning of the gene is of greater importance for protein expression, we created various fusion constructs composed of one GC-poor and one GC-rich genes in various orientations (Figure 47a). The following described experiment was conducted by Jeanne Bazile, a visiting MSc student who was working on this project under my daily supervision. We chose two variants of GFP that are very different in GC-content, GFP_000 (GC=0.40) and GFP_001 (GC=0.62). These are the same variants used in chapter 3 and were also used in other published studies investigating the role of GC content on gene expression (Kudla et al., 2006). In addition, we chose a second fluorescent reporter gene, mKate2, and used two similarly GC poor (GC=0.39) and GC rich (GC=0.58) variants of this protein to create translational gene fusions between one GFP and one mKate2 variant in all possible combinations (8 in total, Figure 47). The objective was to quantitatively assess differences in expression levels between these constructs and to verify that these could be explained by the orientation of the fused genes. The fusion constructs were cloned into CMV-driven mammalian expression vectors and reverse transfected into HeLa cells for 24h. Expression of the fusion proteins was estimated by measuring red fluorescence (shown in Figure 47b) and green fluorescence (similar results, data not shown).

Due to the good spectral separation, no mKate2 signal can be detected for either of the two GFP variants when expressed as individual coding sequences (Figure 47b). The GC-poor variant of mKate2 exhibits no detectable fluorescence on its own, whereas the GC-rich variant of mKate2 shows over 60-fold high fluorescence. This confirms that differences in expression levels caused by differences in GC can also be seen for mKate2, a fluorophore unrelated to GFP. Overall, the expression levels of all fusion constructs vary strongly. The main observation is that GC-poor mKate2, which by itself does not show any measureable expression, can, however, be detected if fused to the 3' end of GC-rich GFP, but not if fused to the 5' end. The opposite can be observed for GC-rich mKate2 – when fused to the 3'end of GC-poor GFP, expression is decreased compared to mKate2 alone or when fused to GC-rich GFP. To test whether similar effects can be seen on the RNA level, HeLa cells were transfected with some of these constructs and RNA isolated after 24h for qRT-PCR analysis. As can be seen in figure 47c, similar effects can be seen on the RNA level: Expression is decreased when GC-poor mKate is fused to the 5' end of GC-rich GFP, but not when in reverse order. Taken together, these results suggest that high GC at the beginning of a gene is important for efficient protein expression.

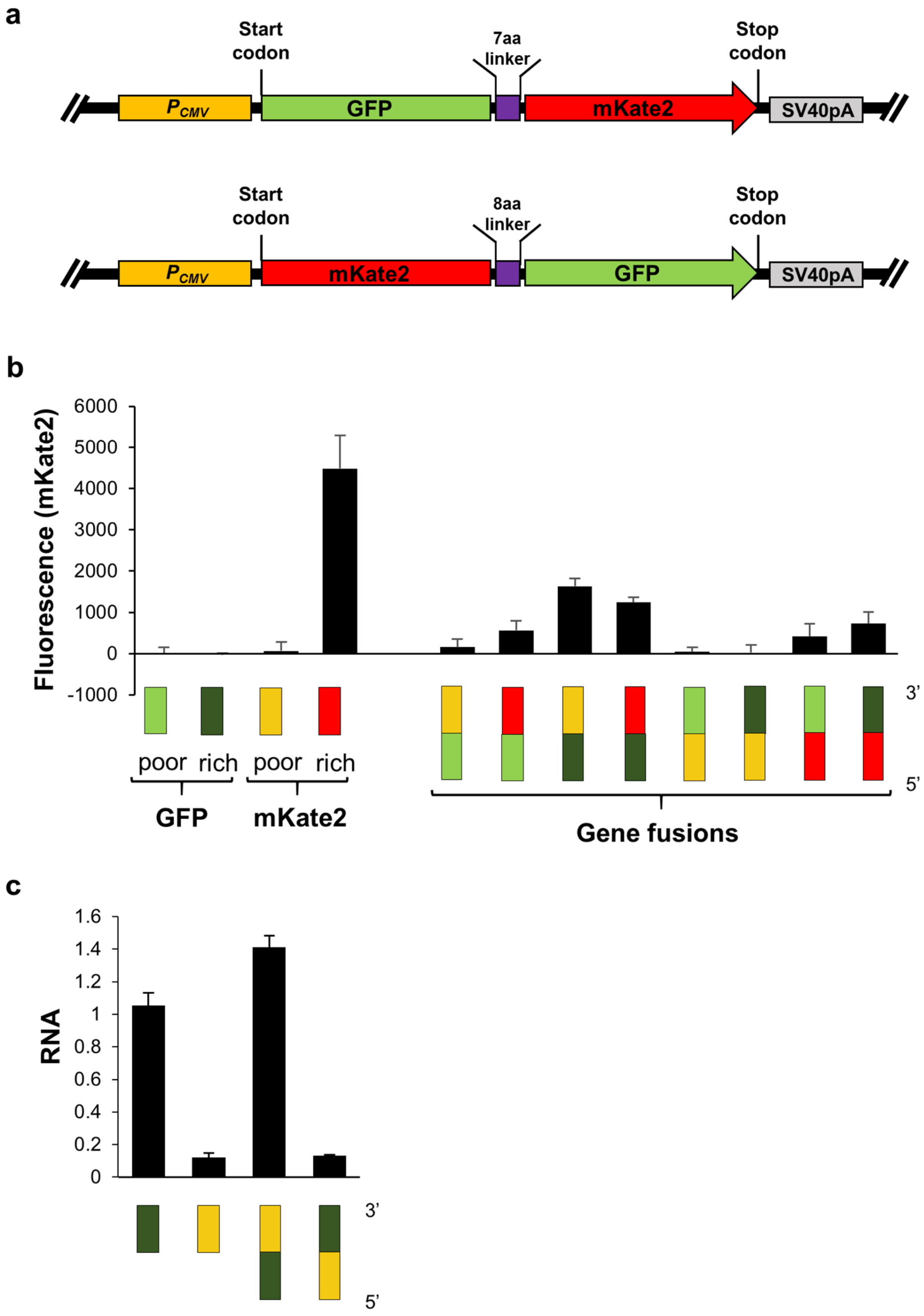


Figure 47. Expression levels of mKate2 and GFP fusion constructs.
a, Translational fusion constructs were created between GC-poor or GC-rich variants of GFP (%GC=0.4 and 0.62) and mKate2 (%GC=0.39 and 0.58). Variants were fused in all 8 possible combinations: constructs with an N' terminal GFP are linked via a 7 amino acid linker, constructs with an N' terminal mKate2 are linked via an 8 amino acid linker. **b**, Constructs were transfected into 3 wells of a 96-well plate and fluorescence measured 24h later using a plate reader as previously described (chapter 3). Bar plots represent the averages of 3 wells transfected with the same plasmid preparation. Error bars denote the standard deviation. Data shown are representative of 3 independent experiments. **c,** Fusion constructs were expressed in HeLa cells and RNA isolated 24h post-transfection followed by qRT-PCR analysis. Shown are the relative RNA levels of the fusion RNA constructs normalised to Gapdh. Error bars denote the standard deviation of 3 technical replicates.

5.2.3. Codon usage affects GFP translation

Since codon usage is usually studied in the context of translation, I utilised polysome profiling to assess the translation dynamics of GFP coding-sequence variants in Hek293 cells. As described in 4.2.5, I used the translation elongation inhibitor cycloheximide to block actively elongating ribosomes from translocating, and thus freeze them at their particular location. Four fraction pools were sequenced: (A) Free Ribonucleoprotein (RNP) complexes, (B) monosomes, (C) light polysomes (2-4) and (D) heavy polysomes (5+) (Figure 48a).

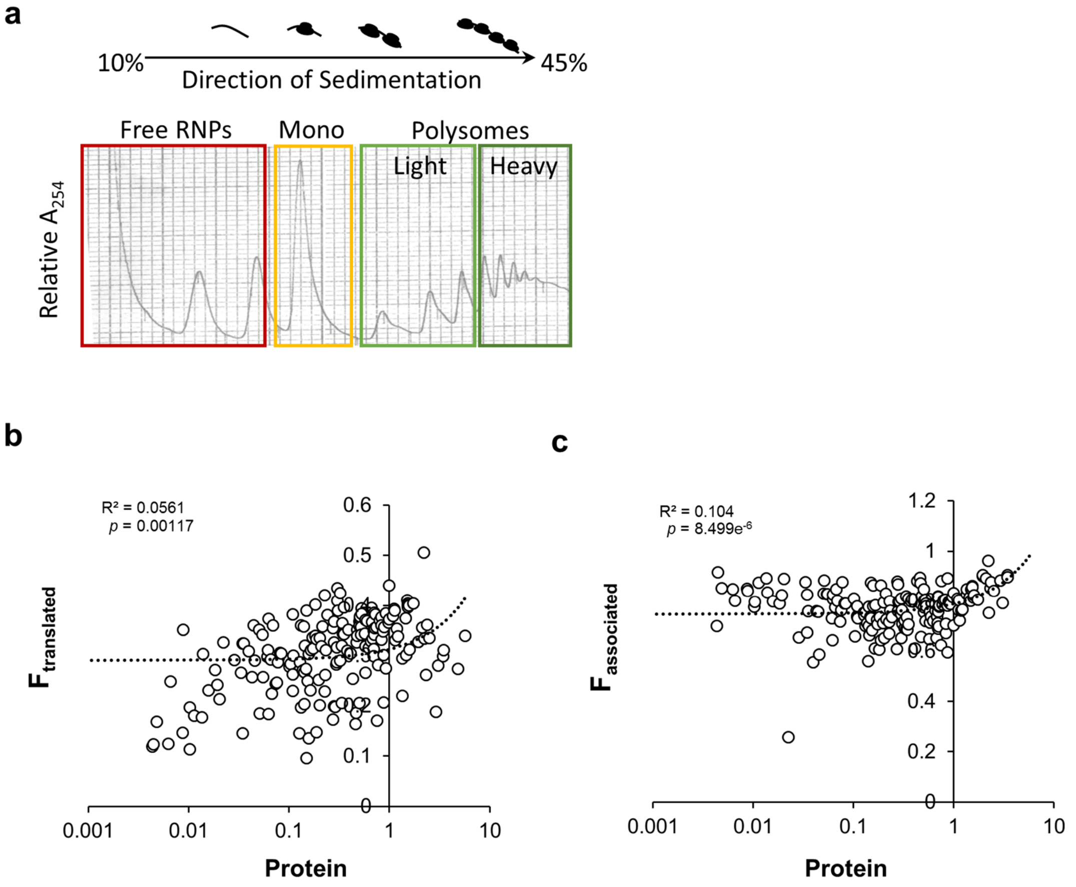


Figure 48. Polysome profiling of Hek293 GFP pool cell lines.
a, UV profile of polysomal separation after sucrose gradient centrifugation. 18 fractions in total were collected and pooled as indicated by the coloured boxes into 4 samples and prepared for high-throughput sequencing: Free RNPs (red), monosomes (yellow), light (light green) and heavy (dark green) polysomes. **b**, Correlations between Protein levels and $F_{translated}$ or **c**, $F_{associated}$ as described below. *n*=1.

To study GFP translation, I defined two different scores:

(a) $F_{\text{associated}} = \frac{(mono+light+heavy)}{free\ RNA}$

- which is the fraction of transcripts that is associated with fully assembled 80S ribosomes and therefore theoretically translatable, and

(b) $F_{\text{translated}} = \frac{heavy}{light+\ mono+heavy}$

- which scores the translational state of a transcript, i.e. what is the fraction of transcripts that are actively being translated.

When correlating both scores to protein measurements obtained by Flow-seq, some of the variation in the data can be explained ($F_{\text{translated}}$: R^2= 0.0561 and $F_{\text{associated}}$: R^2=0.104), but both are not very strong predictors for GFP protein levels (Figure 48b).

Since I had previously observed differences in phenotypic behaviour for different groups of variants depending on their GC distribution, I correlated GC3 content to $F_{\text{associated}}$ and $F_{\text{translated}}$ separately for each group (Figure 49). For both variant groups, GC3 correlates with $F_{\text{associated}}$ (uneven GC: R^2= 0.4554, p=1.136e^{-26}; even GC: R^2= 0.8557, p=0.02436) and also with $F_{\text{translated}}$, however only significantly for those with uneven GC distribution (R^2= 0.0231, p=0.0394). This suggests that the GC distribution pattern has a stronger influence on the amount of transcripts that are associated to fully initiated 80S ribosomes, but to a lesser extent, if at all, influences the translational state of a given variant.

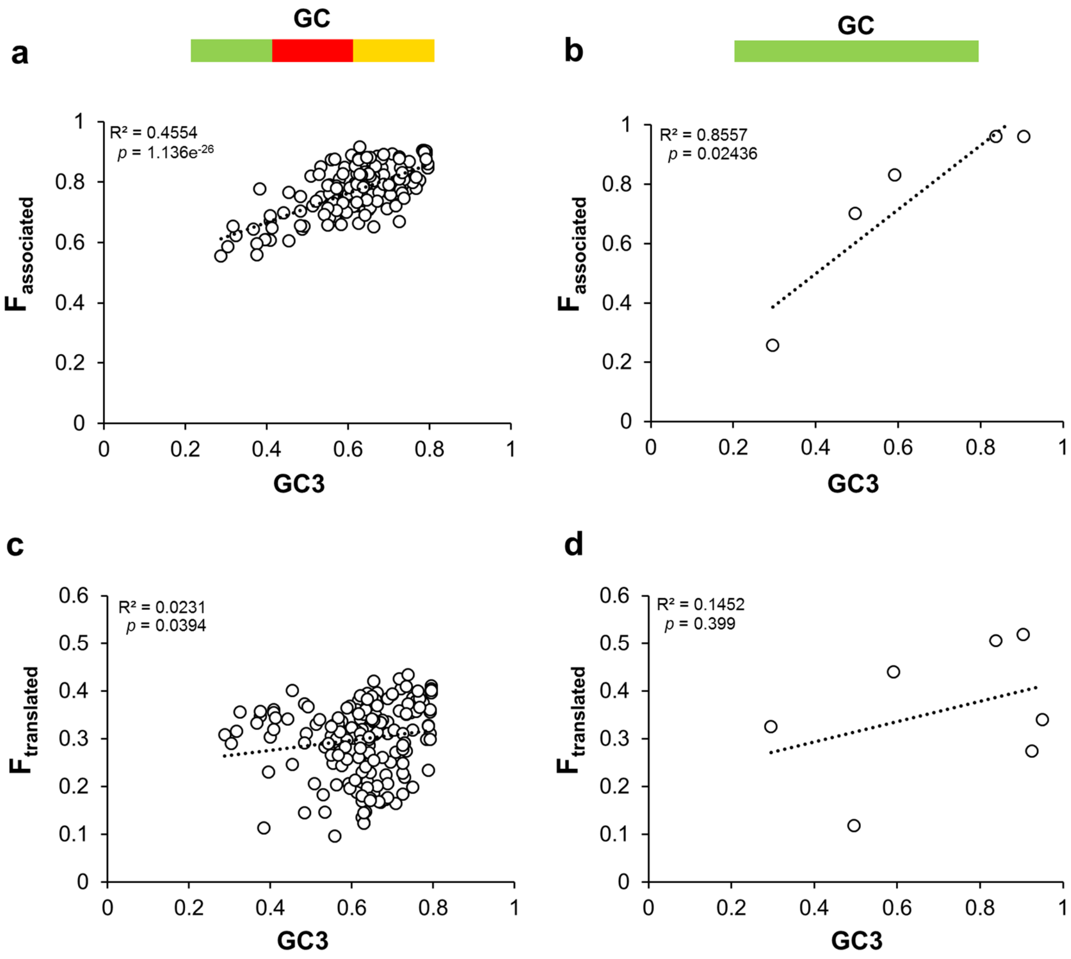


Figure 49. The effect of GC distribution on GFP translation in Hek293 pool cells. For each variant group (left column: uneven GC; right column: even GC), $F_{associated}$ (**a + b**) and $F_{translated}$ (**c + d**) are plotted against GC3.

5.2.4. GC-content affects mRNA subcellular localisation of GFP

Some reports have implicated codon usage as a factor contributing to nucleo-cytoplasmic RNA export efficiency (Kotsopoulou et al., 2000; Malim et al., 1989; Nguyen et al., 2004). Using two GFP variants with very different GC content, I previously demonstrated that high GC-content leads to an increase of cytoplasmic RNA levels (see 3.6). To test whether similar effects can be seen across many variants, I performed cellular fractionation followed by RNA extraction of Hek293 and HeLa GFP pool cell lines after 24h of GFP induction. For each GFP variant, I calculated the Relative Cytoplasmic Concentration (RCC) from normalised read counts in each fraction (described in 4.2.4). Variants with an RCC below 0.4 are considered as mostly nuclear, whereas those with an RCC above 0.5 are regarded as more cytoplasmic (Figure 50+51). In Hek293 cells, the average GC3 of more cytoplasmic variants is significantly higher compared to more nuclear variants for both variants group (Figure 50a+c) which is also reflected in the correlations between RCC and GC3 (Figure 50b+d, even GC: R^2=0.8108, *p*=2.0707e^{-5} ; uneven GC: R^2=0.1606, *p*=7.732e^{-7}).

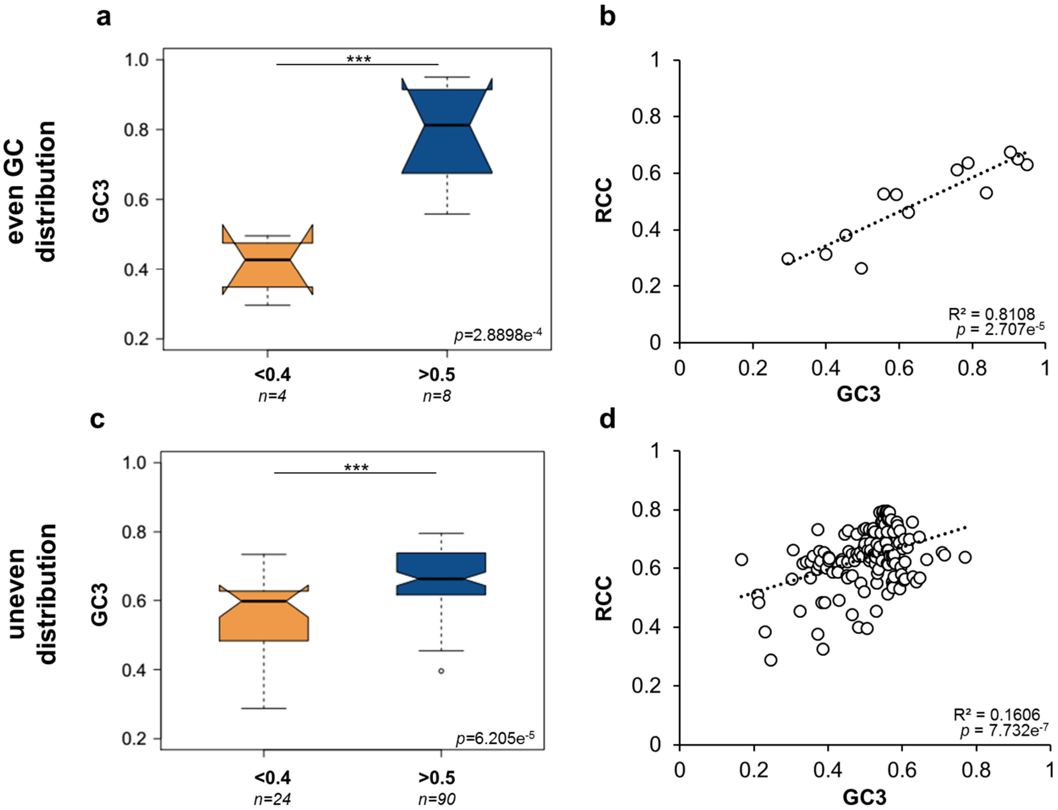


Figure 50. The effect of GC3 on the RCC of GFP RNA in Hek293 pool cells.
GFP variants were divided into more nuclear (RCC<0.4) and more cytoplasmic (RCC>0.5). **a**, Boxplots of variants with even GC distribution separated by localisation (*p*=2.9e^{-4}). **b,** Scatterplot of the RCC for all variants with even GC distribution plotted against their respective GC3. **c**, Boxplots of variants with uneven GC distribution separated by localisation (*p*=6.205e^{-5}). **d,** Scatterplot of the RCC for all variants with uneven GC distribution plotted against their respective GC3. *n*=1.

Similar observations can be made from data from HeLa GFP pool cells (Figure 51, left column). I also performed the same experiment with HeLa expressing GFP with an intron in their 5'UTR (Figure 51, right column). Interestingly, the difference in mean GC3 between more nuclear and more cytoplasmic variants loses its significance. Additionally, the GC3 scores of cytoplasmic variants with even GC distribution cover the full range of possible GC3 values (Figure 51, top right).

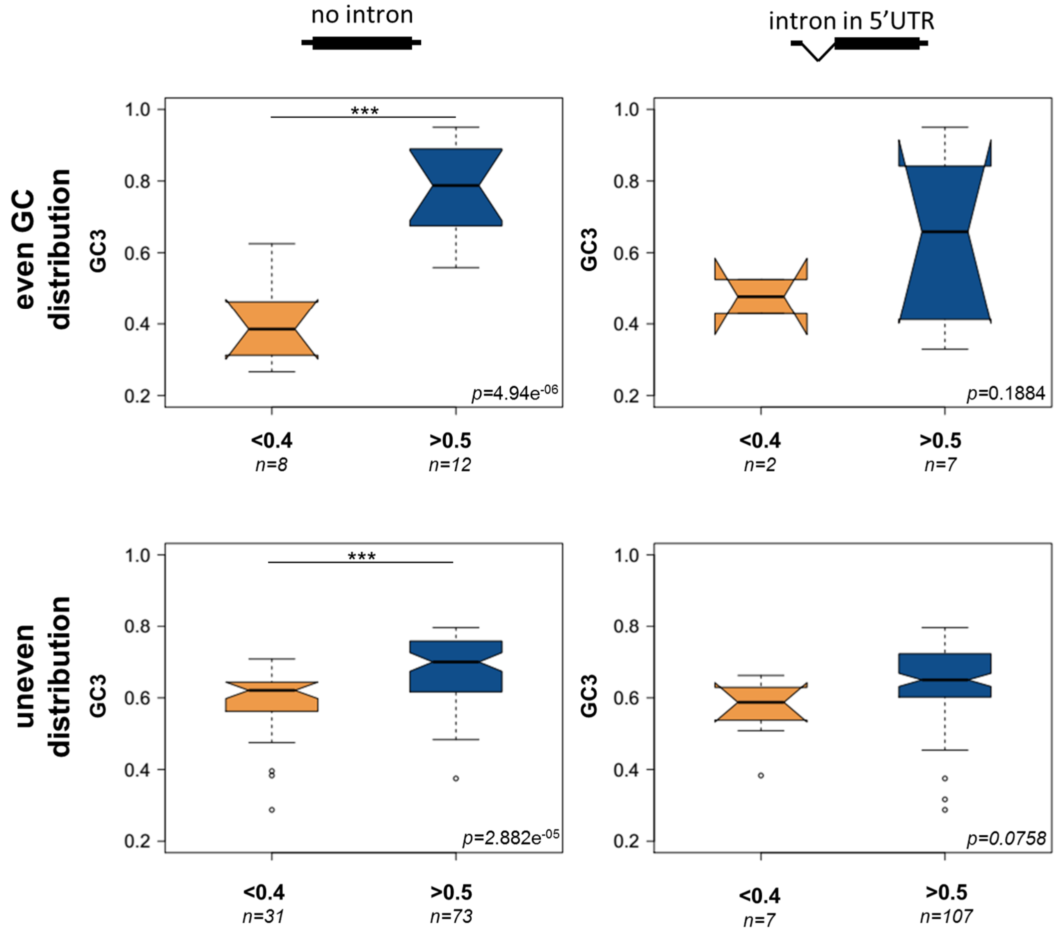


Figure 51. The effect of GC3 on the RCC of GFP RNA in HeLa pool cells.
Boxplot representations of RCC scores and their link to GC3 content of GFP variants grouped by even (top row) and uneven (bottom row) GC distribution. GFP variants were expressed in the absence (left column) and presence (right column) of a 5'UTR intron. *n*=2.

One of the expectations for the comparison of RNA levels between variants expressed with and without intron would be that splicing enhances the expression of at least a subset of variants due to its stabilising effect on transcripts. Indeed, the presence of a 5'UTR intron causes a significant increase in the mean RCC score for GFP variants with low GC content (%GC<0.4; compare Figure 52a+b). On the other hand, splicing does not increase the RCC for variants with high GC content any further (%GC>0.7).

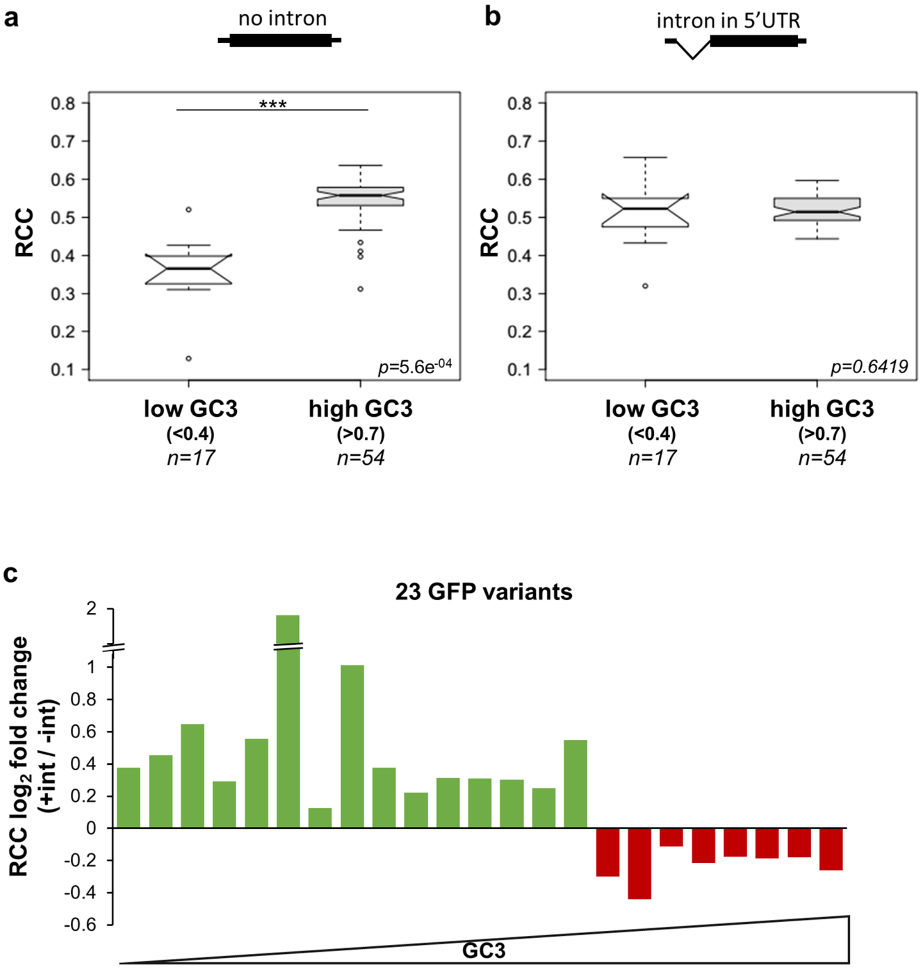


Figure 52. The effect of splicing on the RCC of GFP variants.
Boxplot representation of the distribution of RCC scores for **a**, unspliced or **b**, spliced GFP variants with low GC3 (<0.4; white) or high GC3 (>0.7; grey) in HeLa cells. **c,** Log2 fold-change of 23 GFP variants with significant differences in RCC scores (p<0.05) when expressed with intron (+int) compared to no intron (-int). Variants are arranged by increasing GC3 content (GC3 range = 0.29 – 0.95).

Since these box plot representations only provide an overview across different variant populations, I was also interested in the effects on a single-variant level. To do so, I selected only those variants for which the fold change in RCC changed significantly when variants were expressed with an intron (cut-off p<0.05; n=23) and plotted those in order of increasing GC3 (Figure 52c). This analysis indicates that splicing primarily increases the cytoplasmic localisation of GC-poor variants, whereas it leads to an increase in nuclear levels in case of GC-rich transcripts.

I previously demonstrated by qRT-PCR that increasing GC3 content also increases total RNA levels (chapter 3.4, Figure 14). I expected that variants with overall high RNA levels should also have high RCC scores as prolonged nuclear retention will lead to transcript degradation. I therefore correlated the RCC scores against qRT-PCR data for variants with even GC distribution as these represented the majority of variants previously tested by qRT-PCR (Figure 53). Variants expressed without intron show a significant correlation between RNA levels and cytoplasmic localisation (R^2=0.3502, *p*=0.0258; figure 53a) whereas when these variants are expressed with an intron, the correlation loses significance (*p*=0.0558), suggesting that splicing strongly reduces the total variation in RCC (compare to Figure 52b).

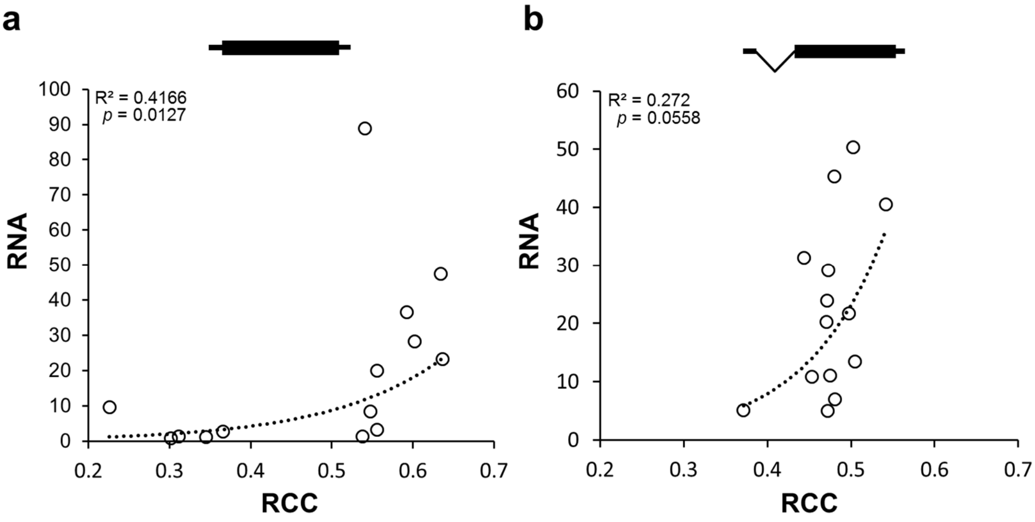


Figure 53. Correlation between RCC and RNA levels for variants with even GC distribution.
The RCC for variants with even GC distribution expressed in HeLa cells **a,** without intron and **b,** with intron are plotted against RNA levels as measured by qRT-PCR (see 3.4).

5.2.5. Determining the relative importance of experimental phenotypes

As an overview of the overall correlations between experimentally measured parameters as well as GC content, all data from Hek293 experiments are shown in a correlation matrix below (Figure 54). As I demonstrated in previous sections that GC distribution is a factor contributing to expression phenotypes, I also created such matrices separately for variants with even and uneven GC distribution (Figure 54, middle and bottom row). Similar matrices are also shown for data obtained from HeLa pool cells.

Multiple aspects that were described in previous sections are reflected in the overall results, such as the correlations between total GC3 and all measured parameters. For variants with uneven GC distribution, overall GC3 and the GC3 of the first third are correlated with protein levels in Hek293, whereas in HeLa only the overall GC3 is significantly correlated. In Hek293, $F_{associated}$ strongly correlates with GC3 (*r*=0.67), in particular with the 2nd and 3rd thirds (*r*=0.59 and 0.52 respectively). The RCC correlates similarly well with GC3 in both cell lines (Hek293 *r*=0.28; HeLa *r*=0.3), with the GC3 of the first third showing the highest correlation. Quite strikingly different is the picture for variants with even GC distribution. All measured parameters are strongly correlated to each other, as well as to GC3 content, in both Hek293 and HeLa cells. In Hek293, some of the parameters are not significant despite high *r* values. This is most likely caused by the low number of useable data points due to the underrepresentation of this GFP variant group in this particular cell line pool. The correlations between $F_{associated}$ and RCC still persist, whereas some previously positive correlations become insignificant (e.g. RCC with $F_{translated}$). Overall, protein levels correlate strongly with translational scores, RCC as well as GC3.

These correlations are based on simple linear regression and therefore do not take into account the possibility that some variables might be highly correlated with each other (multicollinearity). In order to dissect the individual contributions of each experimentally derived parameter, I used multiple regression analysis in order to obtain the relative contribution of each to overall GFP protein levels in Hek293 cells (Figure 55). By fitting both translation scores and RCC into the model with protein as the response variable, an overall 32.33% of variation can be explained (p=5.029e^{-12}). Both $F_{associated}$ and $F_{translated}$ are almost equally weighted (41.93% and 42.6% respectively), however, when RCC is added to the model, its relative weight is not significant (p=0.183).

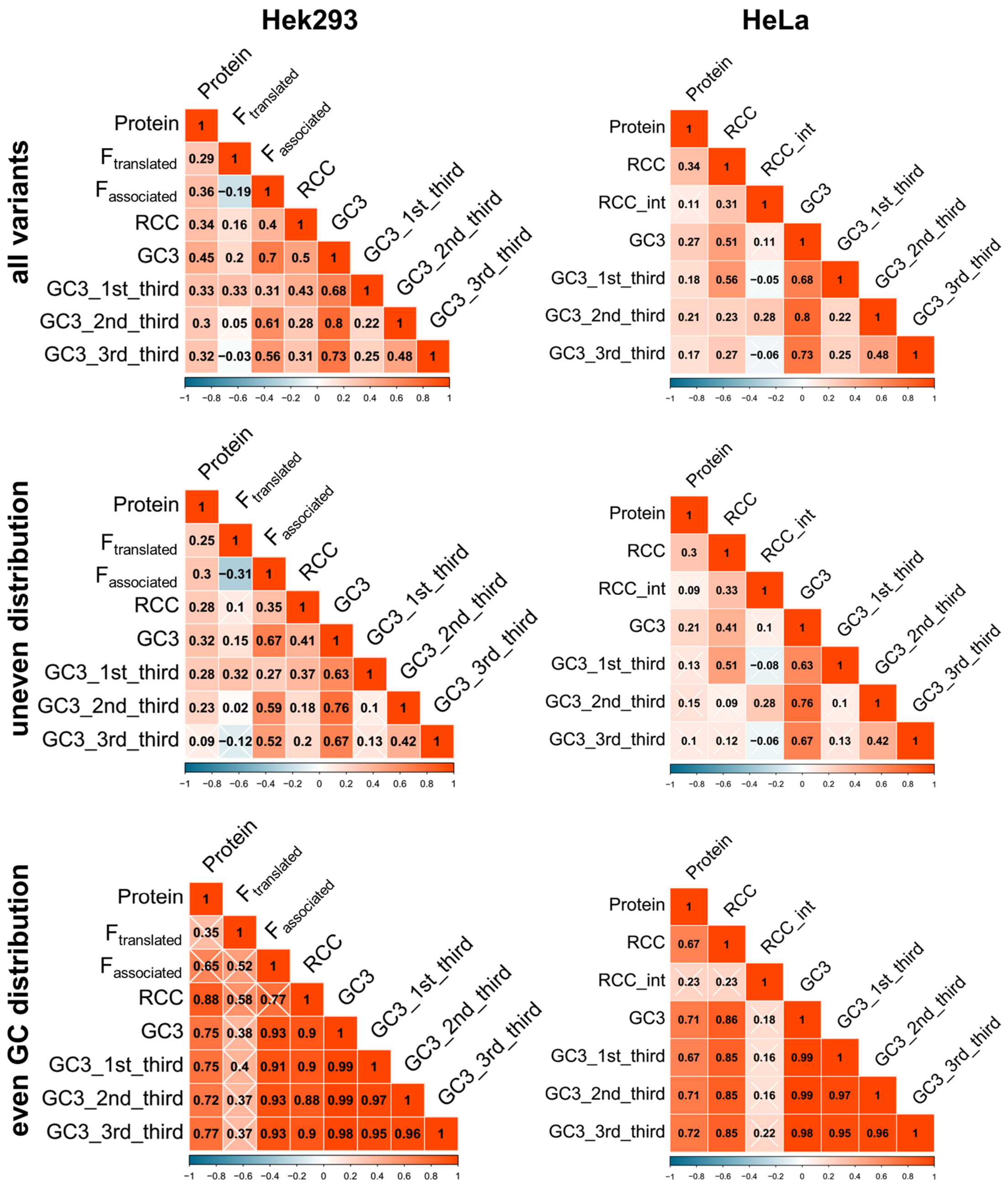


Figure 54. Pearson correlation matrix of experimentally measured parameters and GC3 of GFP.

Shown are various experimental data and calculated features of GFP variants measured in Hek293 and HeLa cells. Each entry in the matrix shows the correlation between measurements of a pair of samples for all variants together (top row), variants with uneven GC distribution (middle row) or even GC distribution (bottom row). White crosses indicate *p*>0.05.

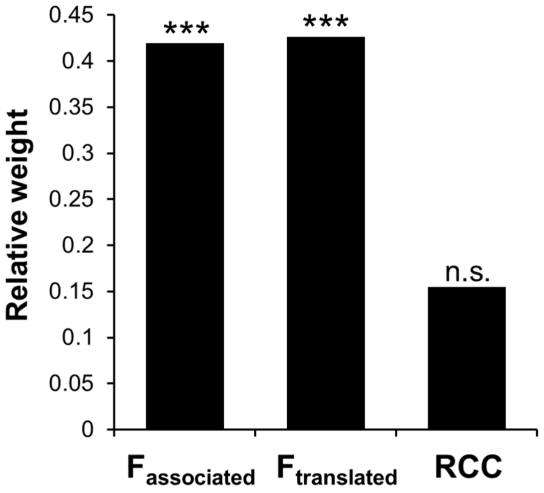


Figure 55. The relative contribution of translational scores and RNA localisation on GFP protein levels.

Shown are the relative weights of $F_{asociated}$, $F_{translated}$ and RCC with GFP protein levels measured in Hek293 cells as response variable. Total variation explained: 32.33%, *p*=5.029e^{-12} based on 146 observations.

5.3. The relative contribution of calculated sequence features in GFP expression

In the previous section, I quantified the relative contribution of experimentally obtained measurements to overall protein levels. In this section, I will attempt to explain the observed variation in terms of sequence features thought to influence expression levels. A common area of application for such calculations is gene sequence optimisation or ‘codon optimisation’. Since codon usage varies strongly across genes, the expression of transgenes often results in poor efficiency due to unfavourable sequence features. It might therefore be beneficial to replace codons with synonymous codons to better match the host cell’s codon preferences. Many computational tools for coding sequence optimization are now available. These use a gene sequence as input, scan it for undesirable sequence features and re-design the sequence in order to optimise the gene for high protein expression in a particular species. One such tool offered by GeneArt/ThermoFisher, called GeneOptimizer, generates from one input sequence thousands of variants optimised for high expression in human cells. It takes various sequence properties into account, such as cryptic splice sites, destabilising RNA elements or rare codons. Based on my experiments, I would like to calculate or predict sequence features that can explain the variation that I observe in GFP expression data. We therefore established a formal collaboration with GeneArt to be able to use their knowledge of undesirable sequence features to individually score each GFP variant, and quantitatively assess the contribution of such features to overall expression. GeneArt provided us with a list of 21 sequence annotations and their motifs. All parameters with their descriptions and motifs are listed in table 3. Since the sequences of all GFP variants are known, G. Kudla used the provided information to score the occurrence of such features in all GFP sequences. Additionally, we also calculated other sequence properties, such as the codon optimisation index (CAI), tRNA adaption index (tAI), the effective number of codons (NC), as well as folding energy (Gibbs free energy) and the frequency of CpG dinucleotides. The Pearson correlation matrix below contains all 33 parameters and illustrates their relationships (Figure 56). Most features cluster with other similar features, e.g. parameters scoring splice donor motifs, or features that score related parameters such as GC3 and CpG. As would be expected, parameters describing very opposing features, such as GC related parameters and AU-centric ones (e.g. poly(A) or ARE motifs), are negatively correlated with each other. Since many of these parameters score similar or same sequence features (e.g. motif count vs motif maxScore), some of these are highly correlated. For this reason, I omitted some of these parameters that are arranged in highly correlated clusters in later data analyses, but retained at least one parameter for each sequence feature in addition to those that are of particular interest for this study, such as the GC3 of individual sequence thirds.

Table 3. Description of sequence parameters

Feature name	Feature definition	Note	Reference
GeneArt_ARE_1_count	ATTTA	(a)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10826
GeneArt_ARE_1_maxScore	ATTTA	(b)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10827
GeneArt_ARE_2_count	ATTTTA	(a)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10828
GeneArt_ARE_2_maxScore	ATTTTA	(b)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10829
GeneArt_AT_stretch_count	[AT]{9}	(a)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10830
GeneArt_AT_stretch_maxScore	[AT]{9}	(b)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10831
GeneArt_GC_stretch_count	[GC]{9}	(a)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10832
GeneArt_GC_stretch_maxScore	[GC]{9}	(b)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10833
GeneArt_PolyA_ANRU_PSSM_count	Position-specific scoring matrix	(a)(c)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10834
GeneArt_PolyA_ANRU_PSSM_maxScore	Position-specific scoring matrix	(b)(c)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10835
GeneArt_polyPurine_PSSM_count	[AG]{22}	(a)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10836
GeneArt_polyPurine_PSSM_maxScore	[AG]{22}	(b)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10837
GeneArt_Splice_acceptor_PSSM_count	Position-specific scoring matrix	(a)(d)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10838
GeneArt_Splice_acceptor_PSSM_maxScore	Position-specific scoring matrix	(b)(d)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10839
GeneArt_Splice_donor_consensus_count	RGGTNNGT	(a)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10840
GeneArt_Splice_donor_consensus_maxScore	RGGTNNGT	(b)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10841
GeneArt_Splice_donor_cryptic_count	RSGTNNHT	(a)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10842
GeneArt_Splice_donor_cryptic_maxScore	RSGTNNHT	(b)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10843
GeneArt_Splice_donor_PSSM_count	Position-specific scoring matrix	(a)(e)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10844
GeneArt_Splice_donor_PSSM_maxScore	Position-specific scoring matrix	(b)(e)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10845
GeneArt_PolyA_generic	Regular expression	(f)	Graf M. et al (2000), J Virol 2000 Nov; 74(22): 10822–10826
GC3	GC content in third position of codons		

Feature name	Feature definition	Note	Reference
CpG	number of CpG dinucleotides		
CpG_1st_third	number of CpG dinucleotides (nt 1-216)		
CpG_2nd_third	number of CpG dinucleotides (nt 217-489)		
CpG_3rd_third	number of CpG dinucleotides (nt 490-720)		
GC3_1st_third	GC content in third position of codons (nt 1-216)		
GC3_2nd_third	GC content in third position of codons (nt 217-489)		
GC3_3rd_third	GC content in third position of codons (nt 490-720)		
dG_-4_38	mRNA folding energy (nt -4 - 38)		Markham NR, Zuker M, Methods Mol. Biol. 453 (2008) 3–31.
dG_-40_40	mRNA folding energy (nt -40 - 40)		Markham NR, Zuker M, Methods Mol. Biol. 453 (2008) 3–31.
HsCAI	Codon Adaptation Index (H.sapiens)		Sharp and Li (1987), Nucl Acids Res 15(3): 1281-1295
HsAllGenesCAI	Codon Adaptation Index (H.sapiens)		Sharp and Li (1987), Nucl Acids Res 15(3): 1281-1295
Nc	Effective number of codons		Wright F (1990), Gene. 1990 Mar 1;87(1):23-9
tAI	tRNA adaptation index		dos Reis et al. (2003) Nuc. Acids Res. 31:6976

(a), number of times motif was identified in each GFP sequence, calculated by FIMO (http://meme-suite.org/)

(b), top score of motif match in each GFP sequence, calculated by FIMO (http://meme-suite.org/)

(c), ((47,3,0,50)(18,6,9,67)(53,12,12,23)(59,6,0,35)(70,6,6,18))

(d), ((11,37,10,39);(13,31,10,41);(3,19,15,57);(5,23,13,56);(13,35,9,42);(9,40,10,41);(17,35,17,31);(17,41,5,37);(13,41,3,44);(10,37,6,48);(26,33,14,27);(6,68,0,25);(100,0,0,0);(0,0,100,0);(21,8,62,8))

(e), ((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)(16,17,21,46))

(f), ATGAAA|ACTAAA|ATTAAA|AACCAA|ATATAA|AATTAA|AATACA|AAAATA|AATATA|ATACTA|ATTGTA|ATAAA|AATAA|ATTTA|ATTAAT|ATACAT|AAGCAT|ATATATTT|ATTGTT|GTTAAA|TTTGCA|TACATA|TATATA|TTTATA|TTTGTA|TTTTTATA|TAGTAGTA

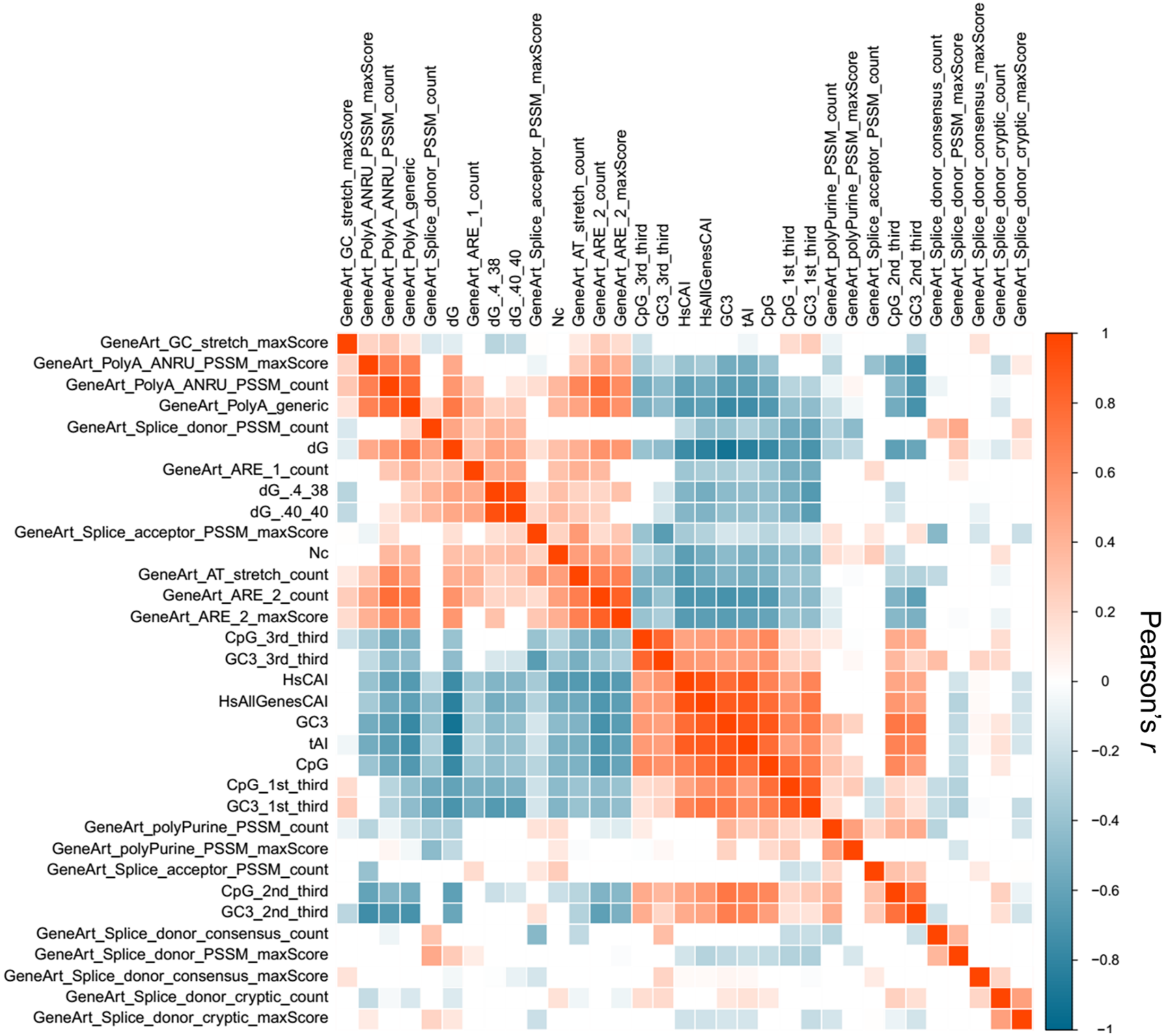


Figure 56. Pearson correlation matrix of 33 sequence features for 217 GFP variants.

5.3.1. Comparisons between HeLa and Hek293

To study the sequence determinants of protein abundance across the GFP library, I correlated 25 parameters of the list above to protein levels measured in both Hek293 (Figure 57 a,c,e) and HeLa (b,d,f), either for all variants together, or individually for the two variant groups with differing GC distribution pattern.

In Hek293 as well as HeLa data, GC3 is highly correlated with protein levels across all variants (*r*=0.21–0.75). As expected, other sequence features related to GC content, such as CpG content and GC stretch count, also strongly correlate with protein levels. Parameters measuring features with high A or AT, such as AREs or AT-stretches, are negatively correlated with expression, which is expected due to their role in transcript destabilisation. Commonly used codon optimality measures, such as the CAI and tAI, are positively correlated with expression, whereas the Nc is expectedly negatively correlated, as the Nc reaches its minimal value when codon usage is the most biased, i.e. not random or constraint, and therefore decreases with e.g. increasing GC3 content. For variants with uneven GC distribution, overall results are similar, however, correlation strengths are generally weaker and often result in insignificant *p*-values. Variants with homogenous GC distribution on the other hand, exhibit more extreme correlation patterns compared to those with uneven GC patterns. Similarly as with correlations to experimental data, a possible explanation could be the lower number of variants measured for this variant group (Hek293 - 9; HeLa – 23) compared to all others (He293 – 181; HeLa – 170). Overall, most patterns are very similar between both cell lines and trends tend to be the same. Some groups of sequence features stand out more than others (marked by black boxes; groups 1 - 3) and will be discussed in more detail in the following sections.

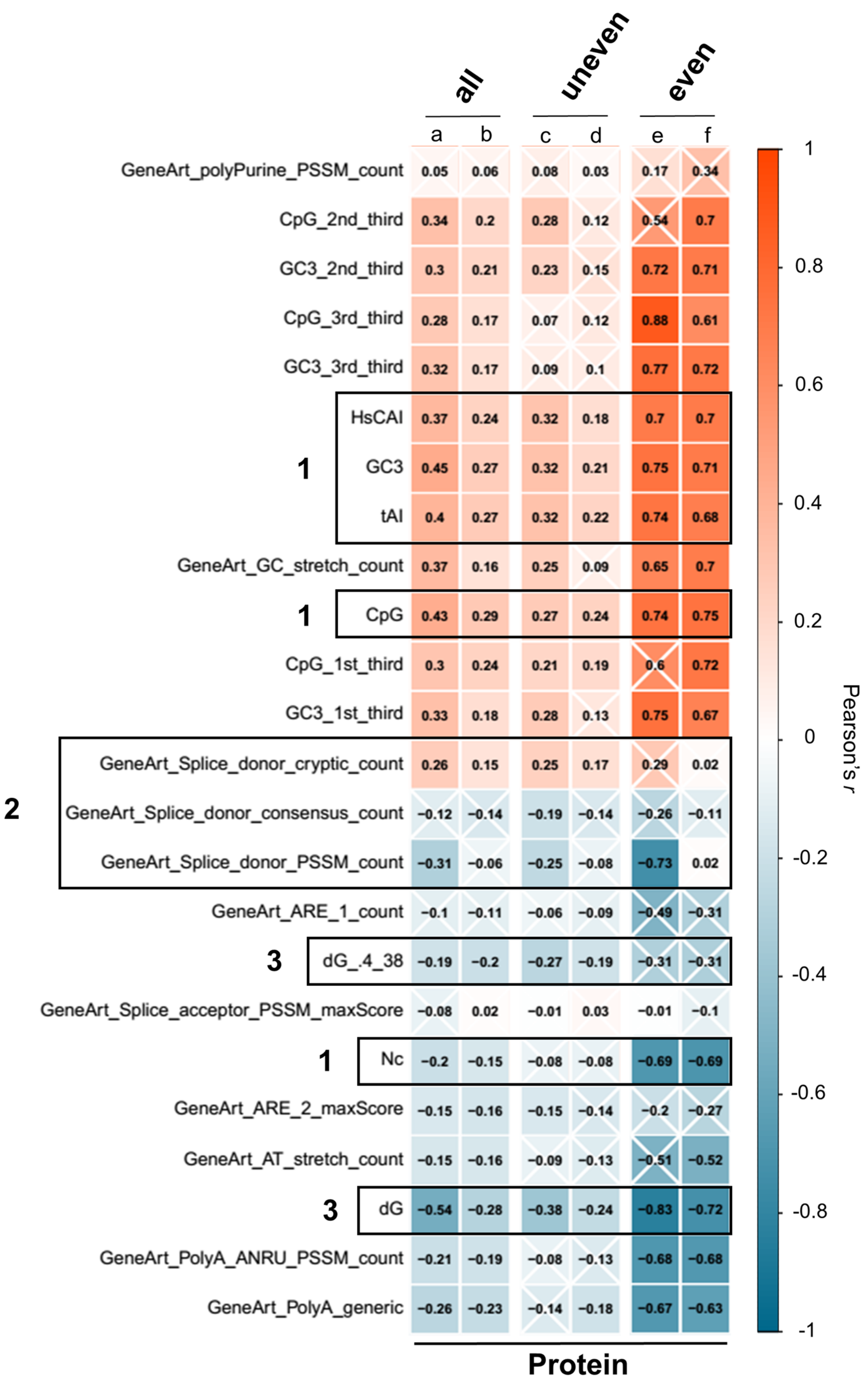


Figure 57. Pearson correlations between calculated and predicted parameters and GFP protein levels in Hek293 and HeLa cells.

Correlations were calculated for GFP variants in groups with differential GC distribution, even (right) or uneven (middle), or all together (left). Shown are correlations for protein data obtained from Hek293 (a, c and e) and HeLa cells (b, d and f). Parameters of interest were grouped into *(1)* measures of codon optimality and GC, *(2)* splicing and *(3)* folding energy. Observations used in correlations: HeLa – 23 even/170 uneven; Hek293 – 9 even/181 uneven. White crosses indicate *p*>0.05.

5.3.1.1. Measures of codon optimality and GC content (group 1)

The codon adaptation index (CAI), tRNA adaptation index (tAI) and effective number of codons (Nc) are very commonly used measures of codon usage and often applied to predict or optimise a gene's expression level. To see how these measures relate to one another, to GC content and to my experimental data, they are visualised in the correlation matrices below (Figure 58a+b). All are highly correlated and as expected, Nc is negatively correlated to the others, as unlike CAI and tAI, this measure is based on the coding sequence only, without taking other parameters other than codon usage into account, and will become lower with increasing codon bias (e.g. because of increasing GC content). All three measures are also strongly correlated to both CpG and GC3. For CAI, this could be explained by how it is calculated. The CAI is determined relative to a set of highly expressed reference genes (dos Reis et al., 2004). To analyse the origin of the correlation between CAI and GC3, I compared the range of GC3 in the reference genes to the GC3 distribution of all human coding genes (Figure 58c). Since there is no significant difference in the mean GC3 for both gene sets, the correlation between CAI and GC3 cannot be explained simply by the nucleotide composition of the used reference genes. In summary, all measures of codon optimality used here are strongly positively correlated with RCC, translational state, as well as protein levels of GFP, in both Hek293 and HeLa cells. Interestingly, when GFPs are expressed with an intron (RCC_int), all previous positive correlations are lost, suggesting that the contribution of these parameters to overall protein expression is strongly decreased in the presence of a splicing event.

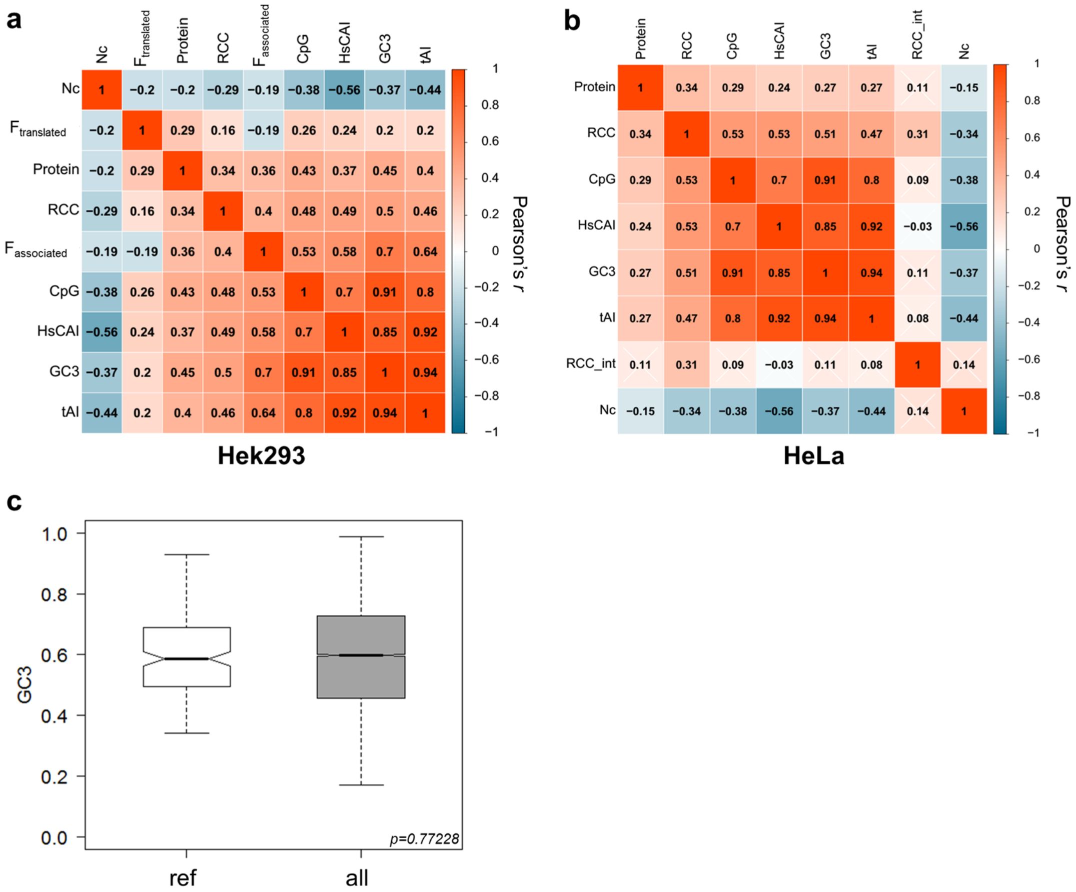


Figure 58. Pearson correlation matrix of experimentally derived parameters for GFP variants compared to measures of codon optimality.
Correlations were calculated with all GFP variants measured in **a,** Hek293 or **b,** HeLa cells. **c**, The GC3 distribution of highly expressed human genes used as reference for CAI calculations (‘ref’, n=192) compared to the GC3 distribution of all human genes (‘all’, n=30,455).

5.3.1.2. Splice site prediction (group 2)

In figure 57 described in 5.3.1, all splice donor related parameters are correlating to varying extents with protein expression in Hek293 (denoted as group 2). I previously observed a cryptic splicing phenotype for some GFP variants (chapter 3.8, Figure 21) and noticed that often the first third of the GFP sequence is spliced out. This led me to speculate that sequence variation primarily within this region must cause this particular sequence to be falsely recognised as intron. Some of the GFP variants used in these experiments only vary in the first third of their sequence and are identical otherwise (n=51). I correlated the same parameters as above to the expression of this particular subset of variants with the assumption that any significant correlations would primarily arise due to differences in sequence composition within the first third of the GFP sequence (Figure 59a). Only two parameters correlate significantly with expression: 'Splice donor consensus count' in Hek293 (r= -0.3; p=3.316e^{-2}, Figure 59b) and 'ARE2 maxScore' in HeLa (r=-0.31; p=3.067e^{-2}, Figure 59c). 'Splice donor consensus count' is negatively correlated with GC3 (r=-0.36, p=0.0320) and CpG (r=-0.40, p=4.851e^{-3}). These observations prompted me to do a multiple regression analysis to test whether 'Splice donor consensus count' influences expression independently of GC3 and CpG. The calculated beta coefficient for GC3 becomes negative when CpG is added to the regression model, although it is established that the relationship between GC and protein is positive. Additionally, the overall p-value of this analysis is not significant (p=0.5088; R^2=0.0277). This result indicates that due to the very high correlation between these two parameters, it is not possible to reliably dissect their individual contributions using this approach for this particular set of variants.

Since I have previously demonstrated that the first sequence third of GFP is of higher importance for efficient protein expression, I was interested in how this might reflect in the cellular localisation of RNA of this particular subset of GFP variants (Figure 59d). The results show that 'Splice donor consensus count' is negatively correlated with RCC in Hek293. All GC-linked parameters are positively correlated with RCC in HeLa (e.g. GC3, CpG) but not significant in Hek293, although trends are similar. When GFP variants are expressed with an intron in the 5'UTR, none of the parameters is significantly correlated with RCC.

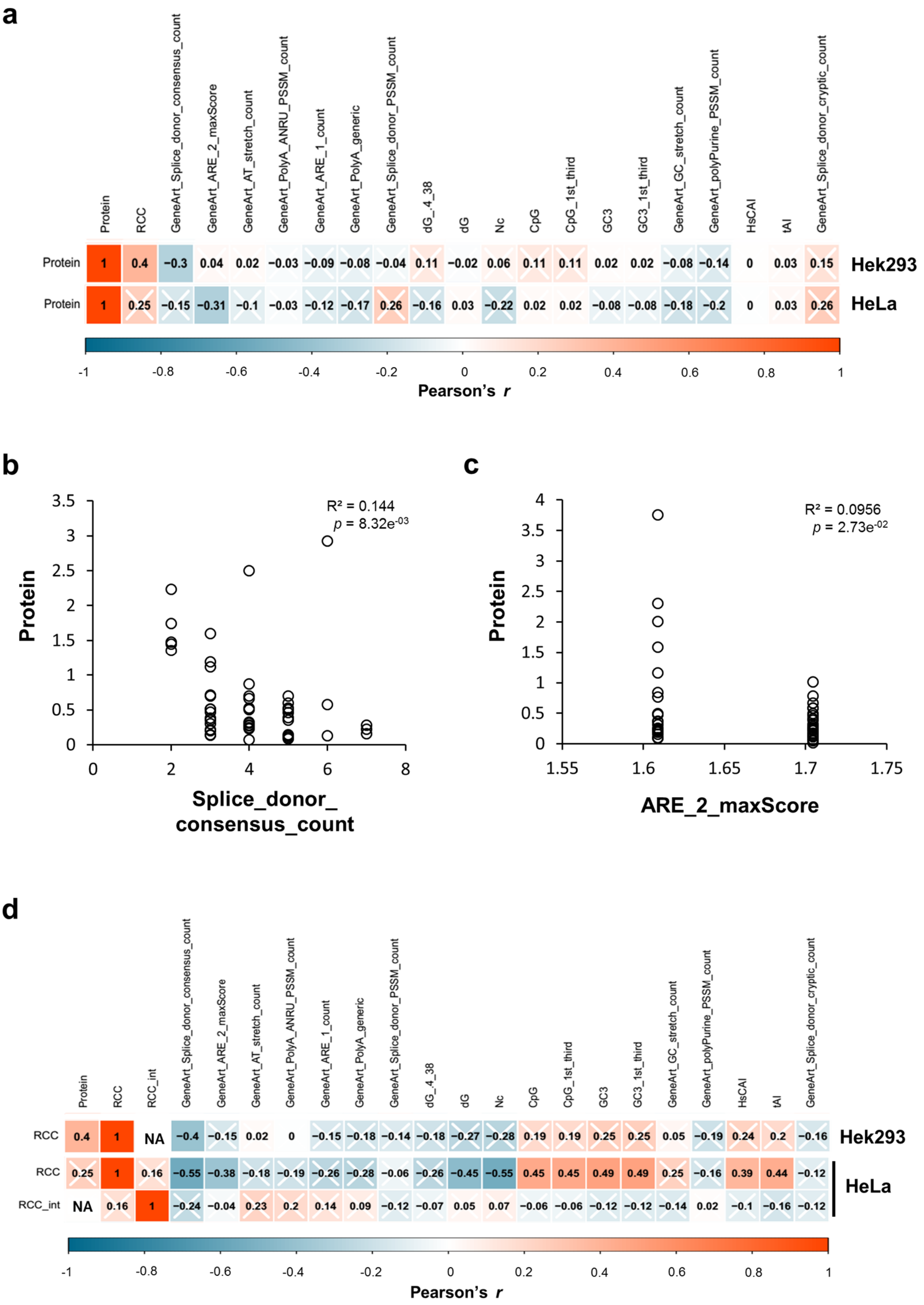


Figure 59. Pearson correlations between protein levels of GFP variants with identical 2nd and 3rd thirds and sequence features.

a, Correlations of protein levels with sequence features in Hek293 (top) and HeLa (bottom). **b + c**, Correlations of Splice_donor_consensus_count and ARE_2_maxScore with protein levels in HeLa and Hek293 respectively. **d**, Correlations of RCC with sequence features in Hek293 (top) and HeLa (middle and bottom). 'RCC' = GFP expressed without intron, 'RCC_int' = GFP expressed with intron in 5'UTR.

5.3.1.3. High folding energy at the beginning of GFP correlates with high expression (group 3)

Gibbs free energy (dG) is commonly used as a measure of folding energy of RNA molecules and strongly connected to sequence composition due to the fact that 3 hydrogen bonds are required for forming G:C base pairs whereas only 2 are needed for A:T pairs. It is therefore expected that dG is strongly correlated to GC3 (boxes denoted with '3' in figure 57). Previous published studies demonstrated that strong RNA folding near the ribosomal start site inhibit protein expression in *E.coli* (Kudla et al., 2009). However, the data presented here suggests the opposite is the case in human cells, since folding energy correlates not only with GC3 content (*r*= -0.91, *p*=5.47e^{-78}) but also positively with protein expression (Hek293: *r*= -0.54 *p*=1.47e^{-15}, HeLa: *r*= -0.28, *p*=7.47^{-5}). I have previously shown that GC3 at the beginning of the gene is of greater importance (see 5.2.2) and therefore expected to see similar effects for folding energy. Using multiple regression analysis with protein level as predictor variable, the amount of variation that can be explained by folding energy differs between cell types (Hek293: R^2=0.1287; HeLa: R^2=0.0595) but in both cases, folding energy of the first two thirds have a greater relative weight than the last third in determining overall GFP protein levels (Figure 60).

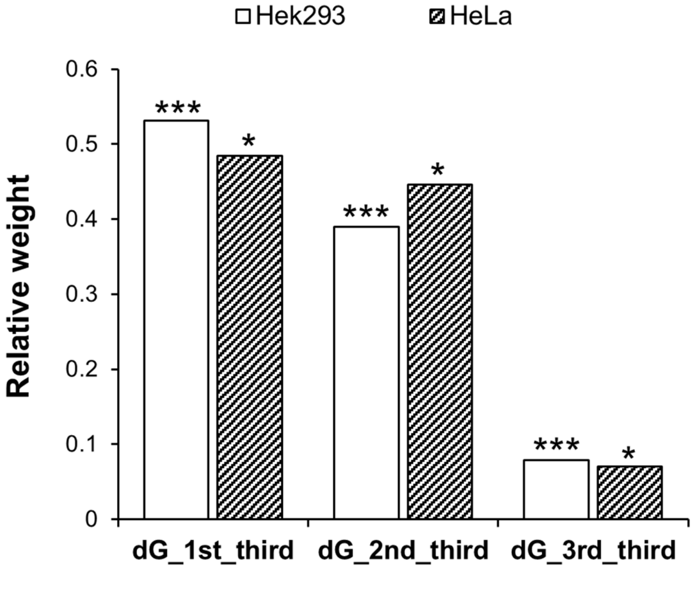


Figure 60. The relative importance of folding energy on protein expression.
Shown are the relative weights of folding energy individually for all three thirds with protein expression as response variable. Results are shown for both Hek293 (R^2=0.1287, p=2.014e^{-5}, 181 observations) and HeLa (R^2=0.0595, *p*=0.01826, 167 observations).

5.3.2. The link between RCC and protein levels

To compare whether the same GFP variants show similar phenotypic behaviour in both Hek293 and HeLa cells, I divided the number of variants for which I obtained RCC scores into thirds and compared the top third (preferential cytoplasmic localisation) and the bottom third (more nuclear) between both cell lines (Figure 61a+b). Out of 60 variants in HeLa, and 52 in Hek293, 25 variants with high RCC and 25 with low RCC are common in both cell lines (Figure 61a+b). Since for this particular analysis, I did not separate GFP variants by their GC distribution pattern, I was interested whether there are general differences in the average GC3 for those variants found in the top and bottom third within each cell line, irrespective of their absolute RCC value. If GC3 generally increases the nuclear to cytoplasmic ratio, it would be expected that results of this analysis should reflect previously shown data in which I utilised defined RCC cut-offs for comparison between nuclear and cytoplasmic variants (Figure 50+51, 5.2.4). As expected, in both cell lines, the average GC3 of variants with high RCC scores is overall higher (Hek293: *p*=9.25e^{-7}; HeLa: *p*=1.65e^{-12}).

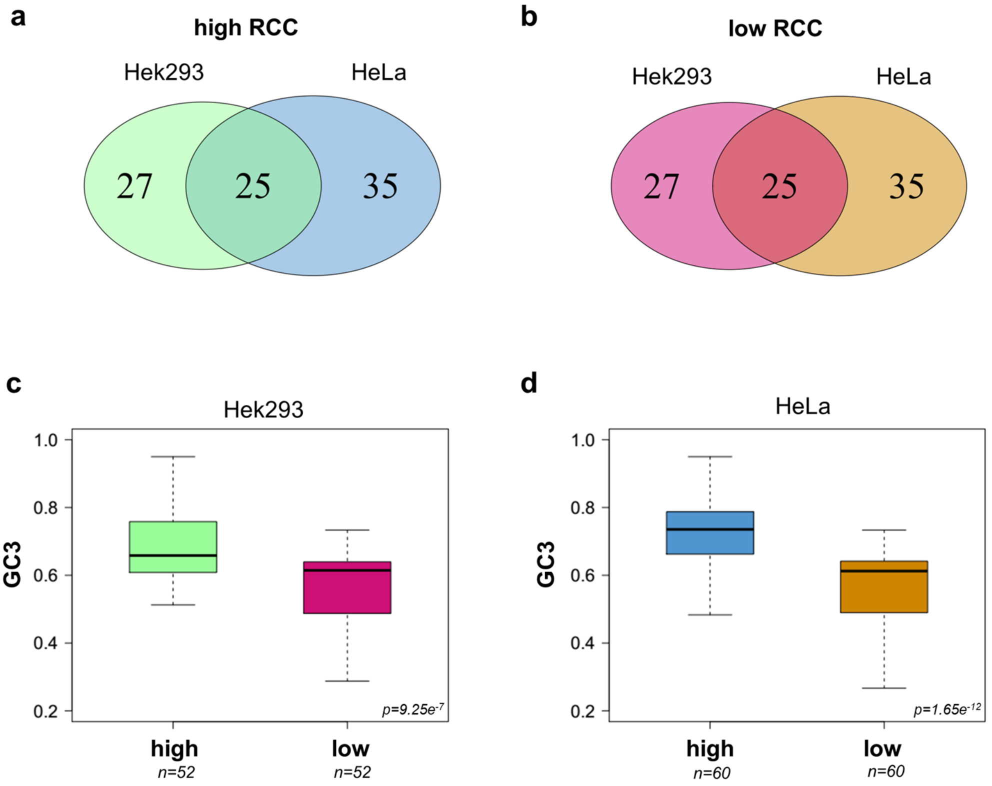


Figure 61. Comparison of GFP variants with highest or lowest RCC between cell lines. Venn diagrams show the intersect between the top (**a**) or bottom (**b**) third of variants between both cell lines (high RCC: *p*=0.0027; low RCC: *p*=0.0027; Fisher's exact test). The mean GC3 for variants in the top third is significantly higher compared to the bottom third in both Hek293 (**c,** *p*=9.25e^{-7}) and HeLa (**d,** *p*=1.65e^{-12}).

As some variants are common in both cell lines, and therefore behave phenotypically similar in terms of their nucleo-cytoplasmic distribution, I probed whether these two groups differ fundamentally in sequence parameters other than GC3 (Figure 62). The groups differ significantly in most sequence features, in particular in those related to AT/GC content. Those with low RCC values score higher in parameters which are generally associated with transcript destabilisation, such as 'AT_stretch_count' (*p*=2.23e^{-4}), 'PolyA_generic' (*p*=7.33e^{-3}), 'ARE_1_count' (*p*=1.83e^{-3}) and 'ARE_2_maxScore' (*p*=1.03e^{-3}). Additionally, the low RCC group scores on average higher in two splicing related parameters; 'Splice_donor_PSSM_count' (*p*=8.97e^{-4}) and 'Splice_acceptor_PSSM_count' (*p*=3.15e^{-2}).

Following on from this result, I performed a similar analysis as above focussing on the highest and poorest expressed variants in regards to protein levels (Figure 63a): 36 highly expressed variants are common between both cell types, as well as 32 poorly expressed variants. Within the poorly expressed variants, 12 variants were previously also found to have consistently low RCC scores (intersect in Figure 61b). Within the 36 highest expressed variants, 8 were also found within those variants with the highest RCC scores. To investigate how sequence features might contribute to the differences in protein levels, I compared each variant group in terms of their sequence features as before (Figure 63c). As expected, all GC-related features are significantly increased in highly expressed variants. Only two other features are significantly different between the two groups, 'Splice_donor_consensus_count' (*p*=2.72e^{-3}) and 'Splice_donor_PSSM_count' (*p*=2.77e^{-5}), which are both increased in poorly expressed variants, once more suggesting cryptic splicing as a major factor influencing GFP protein levels across all variants.

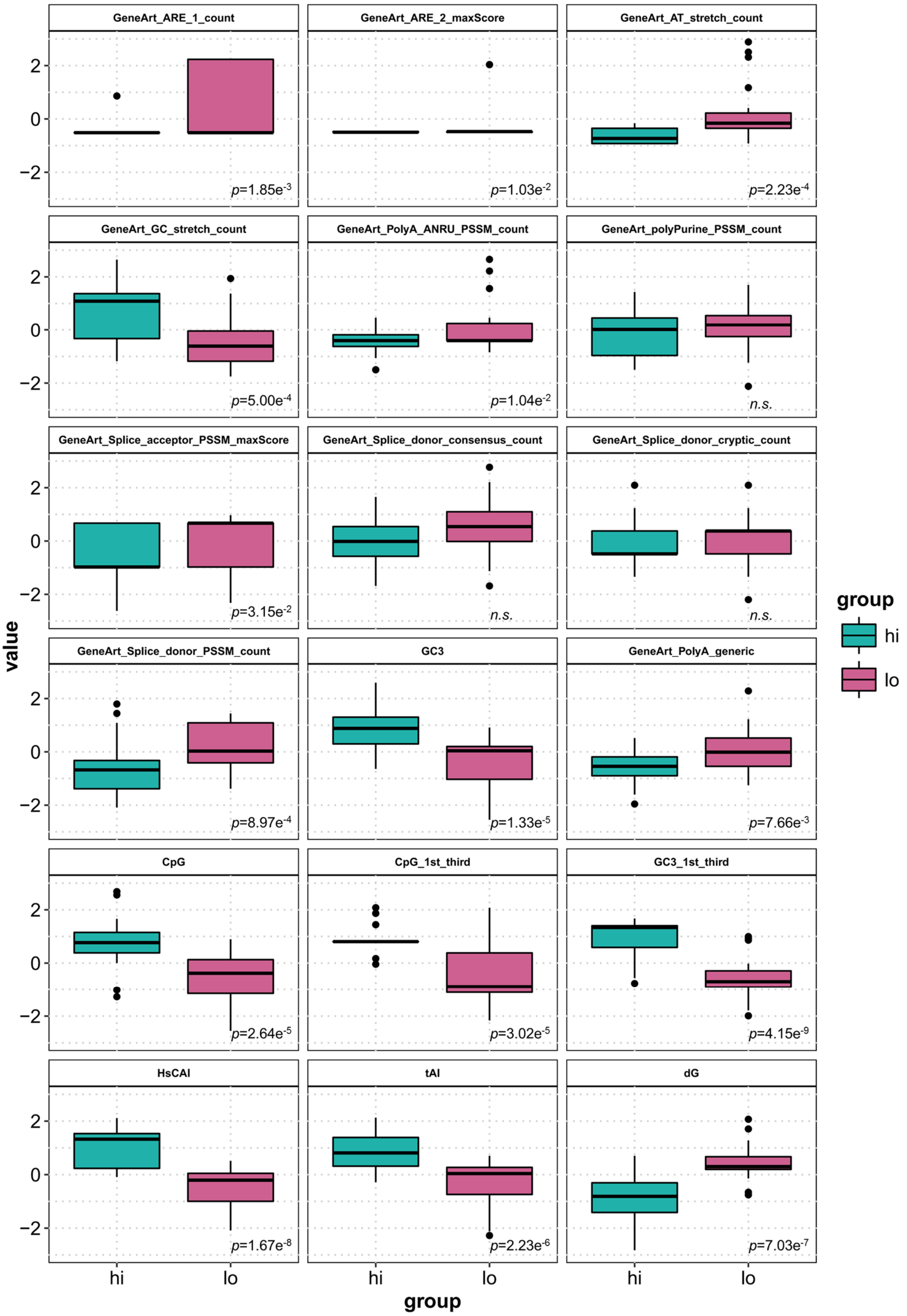


Figure 62. Comparison of sequence features between GFP variants with highest and lowest RCC in HeLa and Hek293 cells.

For variants expressed in each cell line, variants with the 1/3 highest and 1/3 lowest RCC scores were intersected (see Figure 61 a+b; n=25 each) and sequence properties compared. Data presented in boxplots represents normalised values. n(high)=25; n(low)=25.

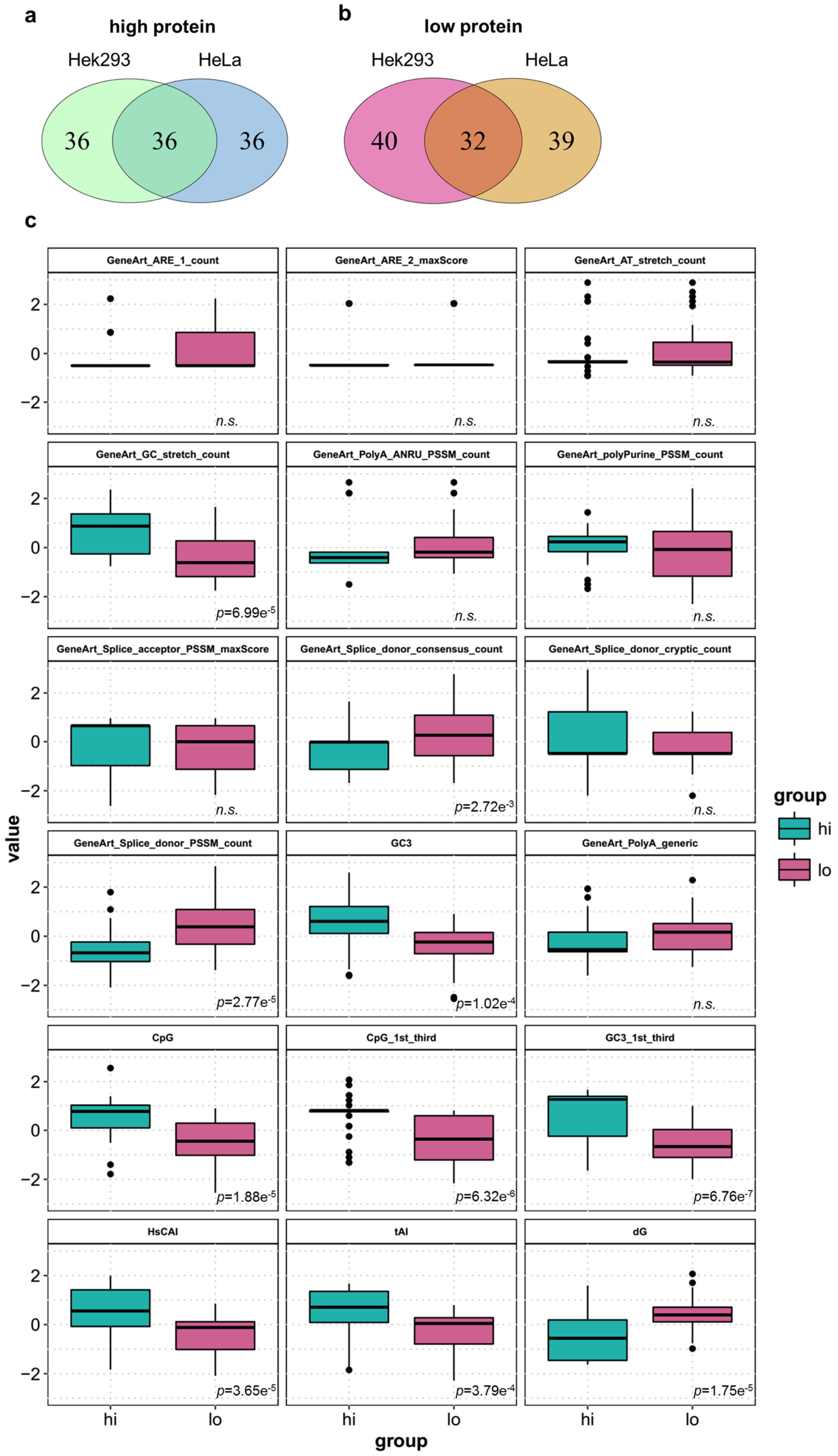


Figure 63. Comparison between highest and lowest expressed GFP variants in HeLa and Hek293 cells.

For variants expressed in each cell line, the 1/3 highest (**a**) and 1/3 lowest (**b**) expressed variants were intersected (high protein: *p*=4e^{-4}; low protein: *p*=1e^{-4}; Fisher's exact test) and for common variants sequence properties compared (**c**). Data presented in boxplots represents normalised values. n(high)=36; n(low)=32.

5.4. Discussion

In this chapter I present experimental data from various experiments conducted on Hek293 and HeLa cell lines expressing 217 GFP variants with the aim to study the relationship between codon usage, RNA localisation and protein levels, as described in chapter 4. I correlate experimentally derived parameters with various calculated and predicted sequence features, in order to explain variation seen in the data. I further establish GC content as a strong determinant of GFP expression and demonstrate that high GC at the beginning of GFP is of greater importance for high protein levels. Directly linked to this observation is also the finding that, unlike in bacteria and yeast, high folding energy is positively correlated with GFP protein expression in human cells. Additionally, I show evidence that a cryptic splice event is likely a major factor contributing to poor GFP expression.

5.4.1. Comparative analysis of GFP expression in Hek293 and HeLa cells

I quantified protein levels of 217 GFP variants using a method based on Flow-seq (as described in detail in chapter 4). I conducted this experiment using two different cell lines, Hek293 and HeLa, in order to validate this method as a novel approach for the phenotypic profiling of fluorescent reporter genes in human cell lines. The fluorescence profiles of both cell lines are correlated (Figure 41, R^2=0.267, p=6.4e^{-13}), but some GFP variants also show strikingly different patterns of expression. For example, GFP_126 reproducibly identifies as highly expressed in HeLa, but poorly in Hek293, whereas GFP_253 shows an opposite pattern. The comparative analysis of tRNA expression levels in different tissues has shown significant differences in the abundance of tRNA species (Dittmar et al., 2006), suggesting that tRNA levels may play a role in regulating protein synthesis. Some more recent studies have suggested that codon usage is coupled to functional differences in subsets of mammalian genes through adaptation to the tRNA pool (Gingold et al., 2014; Plotkin et al., 2004). However, a follow-up study has found that such differences are most likely driven by the underlying genomic context, such as GC content, and that tRNA pools of any cell type are equally efficient at translating any transcript population (Rudolph et al., 2016; Sémon et al., 2006). Since the expression of GFP variants in the experiments shown here is driven from the same locus within each cell line, it is unlikely that differences in genomic context, such as GC content around the insertion sites, could account for opposing effects on particular variants when comparing expression patterns between cell lines. This would therefore suggest that differences in the tRNA pools between Hek293 and HeLa cells might be a likely cause for the observed intercellular variation in GFP expression.

5.4.2. GC content and even distribution increase GFP expression

GFP protein levels correlate with high GC3 content in plate reader and FACS experiments, and the distribution of GC content along the sequence influences this correlation (Figure 44). This is also reflected in the comparison of protein levels between cell lines, as the correlation becomes better when only comparing variants with even GC distribution, but not between those with an uneven distribution (Figure 44e+f). This suggests that for variants with more varying GC distribution along their sequence, factors important for efficient expression either vary or act to different degrees in different cell lines. This led to the question whether a specific region within the CDS was of greater importance for determining protein levels. I used multiple regression analysis to address this question, as well as to get an estimate of the relative importance of any particular sequence segments (Figure 46). For Hek293, the results suggest that high GC3 within the first third of the sequence is of greater importance than in the 2nd or 3rd third and that GC3 can overall explain 21.87% of the variation in protein levels. In contrast, results for HeLa cells do not seem to suggest a particular segment to be of bigger importance, but rather the overall GC content, which contributes to 7.07% of variation in expression. We confirmed the greater importance of high GC in the first part of the gene as a more general effect by obtaining similar results using translational gene fusion of GC-poor and a GC-rich gene variants of GFP and mKate2 described in 5.2.2. (Figure 47).

In several studies it has been shown that changing codon usage can decrease protein functionality, likely by affecting the maturation process in which e.g. translational pauses, caused by varying decoding speeds of codons, allow chaperones to assist peptide segments to fold into their correct secondary structure (Fu et al., 2016; Kimchi-Sarfaty et al., 2007; Zhou et al., 2015). Removal of such pause sites could therefore result in protein misfolding. If this was the case in our experiments, this could lead to a decrease in fluorescence signal due to a potential loss of GFP functionality. In the gene fusion experiments described in 5.2.2, some highly fluorescent speckles in cells expressing protein fusions of GC-rich variants of GFP and mKate2 could be observed (Jeanne Bazile, unpublished). This indicates that proteins are aggregating due to a failure to adopt their correct conformational structure, which likely also explains their decreased expression compared to single gene controls (Figure 47b). We did not make such observations for any other expressed fusion constructs. Since qRT-PCR analysis showed that RNA levels are decreased as well, the results we obtain for constructs with a GC-poor gene at the beginning of the fusion protein, are in line with my previous findings showing that GC-poor GFP variants generally have lower RNA expression (see 3.4), and further underlines the finding that high GC near the 5'end of a gene leads to increased expression. It will be interesting to see whether similar effects can be observed for RNA levels across the GFP library, in particular those variants with uneven GC distribution.

A possible explanation for low RNA levels could be a decrease in transcript stability. Recent studies have reported codon usage as a major determinant of RNA stability in yeast (Presnyak et al., 2015), zebrafish (Mishima and Tomari, 2016) and across other species (Bazzini et al., 2016), mediated by the DEAD-box helicase Dhh1 (DDX6) accumulating along the mRNA transcript when ribosomes are progressing slowly or are stalled. This is followed by transcript deadenylation through the CCR4-NOT complex (Radhakrishnan et al., 2016). I conducted a stability time course on HeLa GFP pool cell lines (see 4.2.5) which should be able to provide clues on whether high GC near the beginning of GFP also exerts a positive effect on overall stability or whether other mechanisms are involved. Differences in transcript stability between variants can be seen, however, due to problems with control spike-ins, the data could not be normalised and therefore not further analysed (4.2.5).

Transcript stability can also vary between different cellular compartments as the major 3'-5' decay machinery, the RNA exosome complex, associates with different cofactors and adaptors in different cell compartments (Kilchert et al., 2016). It could therefore be interesting to additionally measure transcript half-lives separately in the nuclear and cytoplasmic fraction of cells to study if particular sequence features are also involved in the differential recognition by exosomal subunits, mediated for example through differential binding of particular mRNA binding proteins, such as DDX6 (Radhakrishnan et al., 2016). A knock-down of DDX6 could provide further insight into its role in codon-dependent transcript stabilisation in human cells, as so far, this has only been studied in yeast (Radhakrishnan et al., 2016). This experiment could be complemented with a series of knock-down experiments, e.g. targeting the exosomal subunit RRP6, to see whether poor GFP expression of some variants can be rescued. A related project aiming to identify further proteins specifically involved in the regulation of GC-poor and GC-rich genes using a siRNA knock-down approach, is currently ongoing and may shed further light onto the roles of DDX6 and RRP6, as well as other key regulators of RNA metabolism (Miriam Pedron, Christine Mordstein, in collaboration with the Dziembowski lab, Warsaw, unpublished).

Another possible explanation for sequence-specific differences in molecular phenotypes could potentially be attributed to other properties that are directly linked to GC content. A study by

Bauer et al. investigated the influence of intragenic CpG dinucleotide distribution on gene expression (Bauer et al., 2010). By comparing the expression profiles of synonymous gene variants of GFP, one containing no CpGs (GC=55%), one containing 60 CpGs (GC=61%), the authors demonstrate that protein expression is about 29-fold reduced in stable Hek293 Flp-in cells and suggest that this is not caused by reduced mRNA stability or export, but rather by differences in transcriptional activation. However, the authors used Actinomycin D as reagent to block RNA polymerase II mediated transcription which is known to act by intercalating into GpC dinucleotides (Lo et al., 2013), hence, this particular drug is not suitable for comparing variants specifically designed to differ in this particular feature. I addressed a similar issue in section 3.7 in which I show that using a sequence-independent reagent a marked decrease in stability can be observed between two sequence-variants of GFP which only differ in their GC-content, rectifying a previous study which used Actinomycin D and reported no difference (Kudla et al., 2006). It is therefore very likely that differences in transcript stability are also contributing to the overall expression levels. Bauer et al. however also demonstrate differences in nucleosome positioning *in vitro* which might also occur *in vivo* (Bauer et al., 2010). Comparable conclusions were drawn from another study using a similar approach but with cytokines as model genes (Krinner, 2012). Interestingly, the question of positional importance of CpG dinucleotides was addressed using gene variants with enrichments in 5', 3' or central regions of the coding sequence, comparable to the GFP sequence thirds used here to investigate the effect of GC-content variation along the gene. The results suggest that close proximity of CpG dinucleotides to the transcription start site increase gene expression (Krinner, 2012), similar as to what I can observe in my data set as well (Figure 45+46). From the finding that CpGs are enriched around the TSS of the top 5% highest expressed genes compared to bottom 5% lowest expressed genes, the authors concluded that high CpG density close to the TSS is a general feature of highly expressed genes. However, no clear distinction between CpG and GC3 was actually made. Overall, my findings that high GC3 and high CpG correlate with high GFP expression agree with this study. I attempted to address the question of whether GC3 or CpG content is of greater importance by conducting a multiple regression analysis in order to calculate their relative contributions to observed protein expression. However, since the results of this analysis are not significant (p=0.5088; R^2=0.0277), I conclude that it is not possible to dissect the individual contributions of GC3 and CpG by using the current set of GFP variants as both parameters are highly correlated (r=0.9, p=1.53e^{-82}). In the future, this question could be addressed by designing a different set GFP constructs specifically enriched or depleted for either parameter, as well as with different distribution patterns along the coding sequence. A similar phenotypic screen as presented in this thesis could then be used to study the effects of both GC3 and CpG individually, and to further dissect and investigate the underlying mechanisms.

Directly linked to GC-content is also the strength of the folding energy of molecules. I demonstrate that higher folding energy at the beginning of a gene is also correlated to and important for high protein expression. This is in direct contrast to studies performed in *E.coli* showing that strong folding structures near the translation start site inhibit protein production (Goodman et al., 2013; Kudla et al., 2009). In bacteria, ribosomes directly recognise the ribosomal binding site at which translation initiates, whereas in eukaryotes, ribosomes associate to the 5'UTR and start scanning the transcript for the start codon. Therefore, it is not unexpected that structured RNA can be permissive for efficient translation initiation in eukaryotes, but not in bacteria. In addition, considering the importance of correct 5'end processing on several RNA processing steps, including the nucleo-cytoplasmic translocation of mRNA in eukaryotes, it could be speculated that strong folding structures may serve as binding sites important for the recruitment of additional factors, either required for e.g. the interaction with the 5' cap complex in promoting efficient transcript expression, or for transcript stabilisation.

5.4.2.1. GC content affects RNA localisation

To study differences in RNA localisation between GFP variants, I performed cellular fractionation on Hek293 pool cells, followed by high-throughput sequencing of RNA isolated from both nuclear and cytoplasmic compartments. To study the effects of GC3 on GFP mRNA localisation, I compared the average GC3 content of variants that are more nuclear (RCC<0.4) with those that are more cytoplasmic (RCC>0.5) (Figure 50a+b). The results suggest that high GC leads to an increase in cytoplasmic RNA levels, which can also be seen when correlating GC with the corresponding RCC in each variant group: GFPs with even GC distribution are highly correlated with cytoplasmic RNA abundance, whereas all others show a more subtle, but still significant correlation (Figure 50+51). This further suggests that high GC content, in addition to even GC distribution, correlates with high cytoplasmic RNA levels. Whether this can be attributed to increased RNA export or differences in RNA stability, cannot be concluded from this experiment. The results from an RNA stability time course in combination with cellular fractionation experiments could therefore provide a better insight into the role of codon usage in determining transcript stability. Another potential method to test the involvement of RNA export could be to exploit an expression system used by viruses. Studies on HIV genes have shown that the expression of certain late viral transcripts depends on the early viral protein Rev (Malim et al., 1989). Rev recognises the Rev-response element (RRE), an RNA sequence found on several viral mRNAs encoding structural proteins. Rev binds the viral mRNA in the nucleus and promotes efficient nucleo-cytoplasmic export. It was shown that increasing the GC-content of these viral genes circumvents the requirement for Rev (Kotsopoulou et al., 2000). We therefore hypothesise that low expression of particularly GC-poor GFPs could be due to inefficient RNA export. To test this, the RRE could be cloned into the 3' UTR of GFP constructs and expression measured in the presence or absence of Rev (e.g. by co-transfection of a Rev-expression plasmid). Comparing RNA levels between both conditions should reveal a subset of variants limited primarily by RNA export. As the outcome of this experiment would strongly depend on the successful introduction of the Rev plasmid into cells, it might be useful to express another fluorescent reporter gene on the Rev-expression plasmid, which could be used to FACS sort cells according to the expression of this reporter. This experiment could be followed either by cellular fractionation, to detect a subset of variants limited by RNA export which should be reflected in an increased RCC, and/or followed by Flow-seq, as this could further help to identify variants which are limited by RNA export, and/or translation. Since Rev-dependent RNA export has been shown to be mediated via CRM1, it could also be possible to block CRM1 mediated nuclear export using Leptomycin B. Leptomycin B is a drug which inhibits nuclear export of signal-containing proteins, such as Rev, by directly binding to CRM1 (Wolff et al., 1997). Since it is likely that different GFP variants adopt very different RNA structures, each facilitating the interaction with different sets of RNA-binding proteins (RBPs), treatment with Leptomycin B could be used to identify those variants, which depend on adaptor proteins to mediate CRM1-dependent RNA export, similar as e.g. snRNAs require PHAX to target CRM1 to the Cap-binding complex (CBC) (Müller-McNicoll and Neugebauer, 2013). CRM1 has also been shown to be involved in the export of some ARE-containing mRNAs (Gallouzi and Steitz, 2001). Such sequences could be targets for particular adaptor proteins, such as HuR (López de Silanes et al., 2004), which can ultimately link transcripts to the CRM1 export pathway. ARE-like sequences are negatively correlated with RCC scores in HeLa cells (Figure 59d, *r*= -0.38), suggesting that their presence affects the cytoplasmic-to-nuclear RNA ratio negatively. Whether this is through transcript destabilisation or through the interaction with particular RBP inhibiting efficient export, is, however, not clear.

In HeLa, more parameters are significantly correlated with calculated RCC scores than with protein levels (compare figure 59d with 59a). GC3 as well as CpG are highly correlated with high RCC scores when variants are expressed without an intron. This contrasts protein measurements for the same set of GFPs, for which GC3 and CpG are not significantly correlated with expression, suggesting that high GC3/CpG content may be primarily required for efficient cytoplasmic export. Trends in Hek293 are similar, though not significant, possibly due to the lack of replicate experiments (n=1 vs n=2 in HeLa). Quite dramatically different are the results when an intron is introduced into the 5'UTR. None of the calculated parameters are significantly correlated with RCC, indicating that the occurrence of a controlled splicing event may overwrite the influence of other sequence features. This is expected since splicing is known to facilitate efficient expression by mediating various steps in gene expression, as well as enhancing transcript stability (Choi et al., 1991; Nott et al., 2003). This is supported by the finding that the range of RCC covered by variants with even GC distribution is much tighter compared to variants expressed without an intron, which is mostly caused by an increase of RCC of those variants, with otherwise low RCC scores (Figure 53). This finding is also reflected in the increase of GC3 range covered by variants with more cytoplasmic localisation in the presence of an intron (Figure 51), indicating that GC-poor variants are more likely to benefit from a controlled splicing event, whereas with increasing GC3, RCC scores may not improve any further and are more likely to decrease instead (Figure 52c). This is in agreement with a previous finding in stable Hek293 and HeLa, in which the introduction of an intron markedly decreases protein levels of a very GC3-rich variant (GFP_001; GC3=0.97; Figure 13, section 3.3), further demonstrating that splicing does not necessarily improve the protein output. For this particular variant, protein expression is already thought to be at its maximum and it would therefore not be expected to significantly improve due to splicing. To put this result in relation to my previous findings on single GFP variants showing that RNA levels do increase with increasing GC3, even in the presence of an intron (Figure 14e, chapter 3.4), this indicates that here, the primary effect of splicing is on transcript stabilisation in both cellular compartments, leading to an overall unchanged cytoplasmic-to-nuclear RNA ratio. However, I previously also observed that visibly more variation in transcript populations occurs in the presence of an intron as assessed by qRT-PCR using primers specific for either 3'UTR or 5'UTR (Figure 14b+c, chapter 3.4), which suggests that some transcripts are indeed truncated. Since the sequencing library preparation requires the presence of both UTR sequences for PCR amplification, truncated mRNAs are therefore selected against and the experiments presented here will only take full-length GFP transcripts into account. The potential effect of splicing on transcript stability should be further studied by Northern Blotting to validate the high-throughput RNA measurements, but also to confirm that splicing indeed does not rescue the expression of particularly GC-rich variants. This could be complemented with Flow-seq experiments using cell lines expressing GFP with an intron to further elucidate the role of underlying sequence features and splicing. I would expect that for normally highly expressed variants, the presence of an 5'UTR intron might further contribute to a cryptic splice phenotype as the presence of a strong splice donor site might also lead to the utilisation of other, alternative acceptor sites, which otherwise might not have an effect at all. This could potentially by mediated by differences in GC-content (Amit et al., 2012) leading to the removal of essential coding fragments and thus, a decrease in protein levels.

5.4.2.2. High GC increases ribosome association with GFP mRNA

A vast amount of studies investigating the functional consequences of codon usage focusses on effects on translation, but the correlations between transcript abundance and protein levels are often poor (Ingolia, 2014). To study translational dynamics, I performed polysome profiling on Hek293 GFP pool cells (Figure 48). In this experiment I used Cycloheximide, which is an agent that blocks ribosomes from translocating, effectively freezing them at their current location on the transcript. I was interested in the fraction of transcripts that are associated with fully assembled 80S ribosomes, termed $F_{associated}$, as this can indicate whether a particular variant could in theory, due its association with 80S, be translated. I found that increasing GC3 leads to an increase in this fraction (Figure 49a, R^2=0.454, p=1.136e^{-26}). Additionally, I assessed the change in the translational state by scoring the abundance of a transcript occurring in those fractions, that could be considered as actively translating, termed $F_{translated}$, which, however, does not correlate with GC3 very well (Figure 49c+d). Both measures correlate to protein levels to varying degrees ($F_{associated}$: R^2=0.104; p=0.0394; $F_{translated}$: R^2=0.0561, p=0.00117) but not particularly strongly. This is expected since it was previously shown that ribosome density is highly variable on endogenous genes in yeast (Ingolia et al., 2009) and inferring translational efficiency might therefore not be meaningful. The results shown here confirm that for a pool of sequence variants, translational dynamics vary strongly and the interpretation of a transcripts individual translational state might be difficult to directly relate to protein yield. The relationship between GC3 and $F_{associated}$ suggests that with increasing GC-content, more ribosomes associate to transcripts. Whether this is due to an increase in transcript abundance, e.g. through higher cytoplasmic localisation, allowing an increase in ribosome association, or whether an increase in ribosome binding leads to higher transcript stability, is not clear. For variants with uneven GC distribution, high $F_{translated}$ scores correlate with protein levels (r=0.25, p=0.0384). No correlation can be observed with even GC variants, suggesting that despite an increase of monosome-associated transcripts, no other general shift in the translational state can be observed.

If, for any given GFP transcript, translation is inhibited during elongation, the abundance of RNA molecules would be expected to be shifted to heavier fractions, as slowly or inefficiently decoding ribosomes would be more abundant (Presnyak et al., 2015). It is, however, difficult to infer this from the data presented here due to the lack of resolution. In the future, it could be considered to increase such by subdividing polysomal fractions further. In the case that expression is limited by translation initiation, lower amounts of ribosome association would be expected, shifting RNA molecules rather towards lighter fractions (Bulmer, 1991). Since my data shows a stronger correlation with ribosome association, it would be interesting to repeat this experiment using a reagent that specifically blocks translation at the initiation stage, such as Harringtonine (Fresno et al., 1977; Ingolia et al., 2012). Comparing the results from both experiments could further elucidate the role of GC3 in modulating the translation dynamics by revealing variants, that are either limited at translation initiation or elongation. Overall, these findings suggest that GC3 is a determinant for ribosome association, but not necessarily of the translational state of a transcript.

Studies utilising ribosome-profiling to investigate the relationship between codon usage and translation are not always conclusive. On the one hand, no correlations between frequently used codons and high translation elongation were found (Ingolia, 2014), and more frequent codons were reported to be translated with the same speed as rare codons (Pop et al., 2014; Qian et al., 2012). On the other hand, some studies suggest that rare codons are decoded slower due to lower levels of cognate tRNAs (Dana and Tuller, 2014; Gardin et al., 2014) or wobble pairing (Stadler and Fire, 2011). Therefore, the role of codon usage in translation modulation is still unclear. On the basis that elongation speed does vary between codons, an analysis of the distribution patterns of rare and frequent codons revealed a ramp sequence immediately downstream of the ATG consisting of stretches of rare codons (Tuller et al., 2010a, 2010b). It was suggested, that slow elongation caused by low folding energy at the beginning would help maximise protein yield in yeast and bacteria (Goodman et al., 2013; Shah et al., 2013; Tuller et al., 2010b). It would therefore be interesting to investigate whether similar effects can be seen across the GFP library, although so far, the data suggests that strong mRNA folding is beneficial for high expression, at least on the mRNA level. To address this question properly, more variants would have to be designed to specifically vary within the region of the predicted ramp (first ~50nt). Another observation that was made, was the overproportional co-occurrence of certain codons. It was suggested that codons, that are recognised by the same tRNA, are more likely to occur in clusters, facilitating tRNA recycling in order to increase translation speed (Cannarozzi et al., 2010). This could also be tested by specifically designing variants, in which codons are arranged to either minimise or maximise tRNA recycling.

5.4.3. Low GFP expression is partially caused by cryptic splicing

To explain more variation seen in my data, I utilised a set of sequence features, either calculated or predicted, from a list of negative sequence motifs obtained from GeneArt, in order to correlate those to the expression patterns observed across all GFP variants. By doing so, I show that poor RCC scores of variants which only differ in the first sequence third can partially be explained by the higher occurrence of splice donor consensus sequences and ARE-like motifs (Figure 59d). Splicing donor consensus motifs are also amongst the most enriched when comparing the highest with poorest expressed variants across the entire GFP pool (Figure 62 and 63). This is in line with my previous observation, that for some GFP variants, a cryptic splice phenotype can be seen, in which often the first sequence third is removed (see section 3.8, figure 21). This is likely caused by conserved sites that weakly resemble splice acceptor/donor sites. I suspect that certain differences in sequence composition within this region greatly contribute to this segment being falsely recognised as an intronic sequence, followed by splicing. Differences in intron vs exon GC-content have been shown to play a major role in correct splice-site recognition, and changes in GC-composition of exons may result in altered splicing patterns, with lower GC leading to increased splicing (Amit et al., 2012). I therefore speculate that the observed cryptic splicing phenotype might directly be linked to the decreased GC content within this particular region. To test this, it would be required to specifically pick variants that are either very GC-poor, or very GC-rich within this region, but do not differ otherwise. If similar mechanisms are acting here, it would be expected that those variants with low GC in this region, would more likely lead to the false removal of this segment. However, this experiment does not exclude the possibility that other short motifs, such as splice enhancer or silencer motifs, might be present in this region as well, which could further complicate the splicing pattern, possibly leading to even more splice isoforms. Variants specifically lacking such sequences could be designed to further elucidate the role of codon composition on splice site recognition.

5.4.4. Quantifying the influence of sequence features on expression

Most of the analyses presented here tried to elucidate the relationship between individual transcript properties and total protein levels. A common approach for dissecting the contributions of individual variables to an observation is by applying multiple regression analysis (e.g. Qian et al., 2012; Tuller et al., 2010). The general approach of this analysis is to feed multiple predictor variables (e.g. GC-content, CAI) into a model, in order to explain the observed variation in a particular response variable (e.g. protein levels) as the relative contribution of each predictor to the overall variation (R^2). By doing so, I demonstrate the higher importance of strong folding energy (Figure 60) and high GC3 (Figure 46) within the first GFP sequence third. However, this approach may not work well depending on how closely certain predictor variables are related. When applying multiple regression analysis to obtain the relative weights of translation scores and RCC in predicting protein levels, overall 32.33% of the variation could be explained by $F_{associated}$ and $F_{translated}$ (p=5.029e^{-12}), however RCC lost its significance completely. This is partially expected since cytoplasmic abundance is inherently coupled to translation, as RNA binding proteins, acting as e.g. translational enhancers, usually do so indirectly by stabilising the transcripts and thus, differences in translation dynamics would also be expected to influence the translational yield. On the other hand, it would also be expected that poor cytoplasmic localisation should, at least in part, determine protein levels due to the depletion of the available and translatable RNA pool. This demonstrates the issues in identifying causative factors when predictor variables are strongly correlated to each other, such as in case of many calculated sequence features used in this study (Figure 56). When all parameters are included in a regression model, previously strongly positive factors, such as GC content, either become negative predictors, or result in insignificant *p*-values (data not shown). A study by Neymotin et al. (2015) tried to overcome such hurdles by two similar approaches: either by pre-selecting predictors in pair-wise correlations to only include significant ones into a multiple regression model, by step-wise deletion of those, with the highest *p*-value to retain only significant predictors, or, by building a model with predictors with the lowest *p*-values in pair-wise correlations, and with each new predictor added, the model is re-run to monitor changes in significance, to only retain those that explain the most variation (Neymotin et al., 2015). Relying on the statistical significance of relative weights is a popular approach, but has also been argued to not be meaningful due to issues in partitioning shared variance of heavily correlated predictors (Tonidandel et al., 2009). It was therefore suggested that the significance of relative weights should be tested via bootstrapping, i.e. the repeated sampling with replacement from the existing data set to create a larger data set with which a confidence interval can be built around each relative weight (Tonidandel et al., 2009). On applying this strategy to the data presented here, the issue of insignificance of the majority of predictors could not be resolved. Several reasons could account for this, such as the relatively small number of variants tested (in comparison to genome/transcriptome-wide data sets), as well as the required parameter standardisation, which might lead to data skewing. In the future, the method of data normalisation needs to be reconsidered to accommodate the very different scales between parameters. Additionally, alternative modelling approaches, which may not need extensive normalisation, such as decision based modelling (e.g. Random Forests), could be explored as well.

6. Conclusion and outcome of this thesis

Codon usage has been extensively studied, in particular in bacteria and yeast, but only a few studies attempted to quantitatively measure the contribution of synonymous codon usage on gene expression in human cells. By expressing a library of a reporter containing random synonymous substitutions, I systematically studied the effects of coding-sequence variation on expression of a protein-coding gene.

I expressed several synonymous coding variants of GFP in transient and stable expressing human cell lines and could demonstrate the positive correlation of GC3 with protein expression. Using a transcription inhibitor acting independent of GC-content, Triptolide, I could show a link between increased transcript stability and higher protein levels, a finding supplementing previous published data from our lab (Kudla et al., 2006). In order to study phenotypic effects of codon usage variants on a larger scale, I established and validated the use of stable Flp-in cell lines for the expression of several hundred GFP variants from the same genetic locus, a system which allows the simultaneous measurements of several molecular phenotypes via high-throughput sequencing, without the influence of positional effects or differences in gene regulatory regions (same promoter, same UTRs). Highly reproducible data from Flow-seq experiments confirm GC content as a determinant of gene expression and additionally reveal the positional importance of GC distribution as a novel factor influencing protein levels, with high GC3 within the first ~200nt leading to higher expression, regardless of the downstream sequence composition. This finding is in agreement with other studies suggesting high CpG near the 5'end is important for more efficient expression (Bauer et al., 2010). This result is reproducible within two tested cell lines, Hek293 and HeLa, however, clear dissimilarities in expression patterns can also be observed, possibly caused by tissue-specific differences in codon adaptation. The positional effect of high GC near the 5' end is also directly linked to the observation that strong secondary structure formation near the translation start site is highly permissible for efficient gene expression in human cells. This result directly opposes results obtained in bacteria, for which high folding energy is non-permissive for translation initiation due to ribosome occlusion (Kudla et al., 2009).

Several sequence features are correlated with GC content, however the underlying mechanisms are not clear. In yeast, it was recently shown that codon adaptation is strongly correlated with RNA stability, mediated through the binding of the RNA helicase Dhh1 on transcripts with low ribosome density, leading to transcript destabilisation and reduced expression (Presnyak et al., 2015; Radhakrishnan et al., 2016). I demonstrate that low GC content is correlated with poor gene expression, however it is unclear whether this is caused by reduced RNA stability, export or translation, and whether trans-acting factors, similar as to Dhh1 in yeast, are involved. A systematic siRNA screen of >200 candidate factors involved in RNA metabolism is currently underway and may shed light on the differential regulation of GC-poor and GC-rich genes (in collaboration with the Andrzej Dziembowski lab, Warsaw). This screen could be further expanded to a genome-wide scale using e.g. genome-wide siRNA or shRNA libraries. Alternatively, any potential trans-activating regulators, such as the human orthologue of Dhh1 (DDX6), could be directly tested in knock-down experiments using the established GFP pool cell lines followed by Flow-seq measurements.

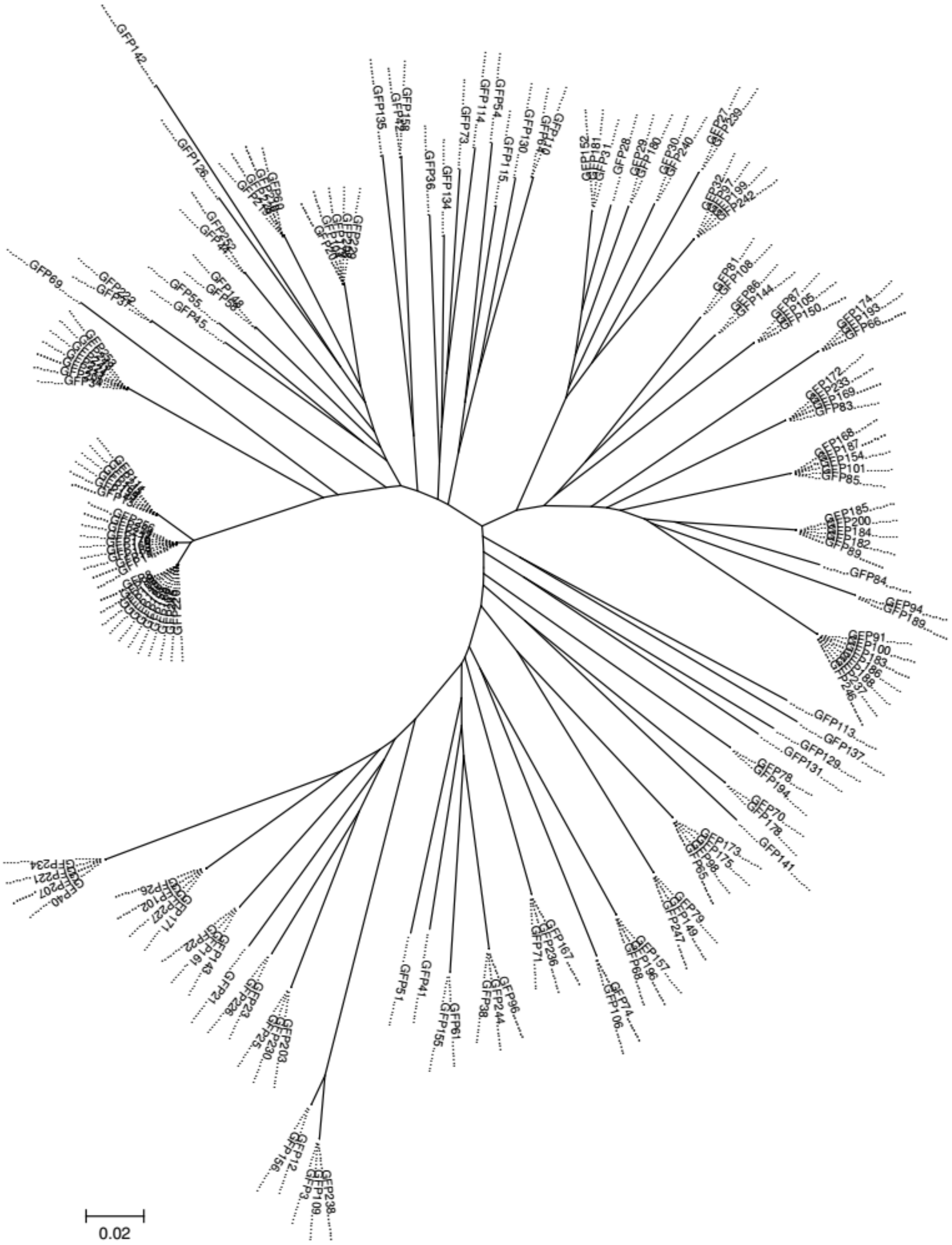
The most studied phenotypic consequences of synonymous mutations in human disease are related to splicing, and the generality of this was recently highlighted in a series of high-throughput studies (Julien et al., 2016; Rosenberg et al., 2015). It remain, however unclear, how splicing affects other directly linked RNA processes, such as e.g. RNA stability and export, and how these effects may be modulated through synonymous changes. I observed the occurrence of a cryptic intron within many of the GFP variants, but the frequency of splicing varies for each. Using similar high-throughput sequencing approaches as described in this thesis, the probability of splicing could be quantified and correlated to the nucleotide sequence. A further related questions that I addressed, is, whether gene expression regulation varies between genes that are unspliced or spliced. By placing an intron in the 5'UTR of the GFP expression constructs, I demonstrated that splicing increases the cytoplasmic localisation of particularly GC-poor variants, but not of GC-rich variants, which generally already have high transcript levels. To be able to study and quantify the differences in regulation on multiple steps in gene expression further, the high-through put methods established in this thesis will aid uncover the sequence properties of variants for which expression depends e.g. on splicing.

Recent published data shows that synonymous substitution rates are lower in exonic splice enhancer elements (ESEs) due to stronger purifying selection (Cáceres and Hurst, 2013; Savisaar and Hurst, 2016). Such elements however, are often not removed in the design of heterologous gene expression systems, in which intronless versions of genes are often preferred. A collaboration between the Kudla and Hurst group aims to further understand the regulation of splicing mediated by ESEs in combination with the underlying sequence composition, in order to develop new algorithms for codon optimisation of genes. To establish the generalisability of novel findings, experiments will be conducted using variants of genes relevant in biomedicine.

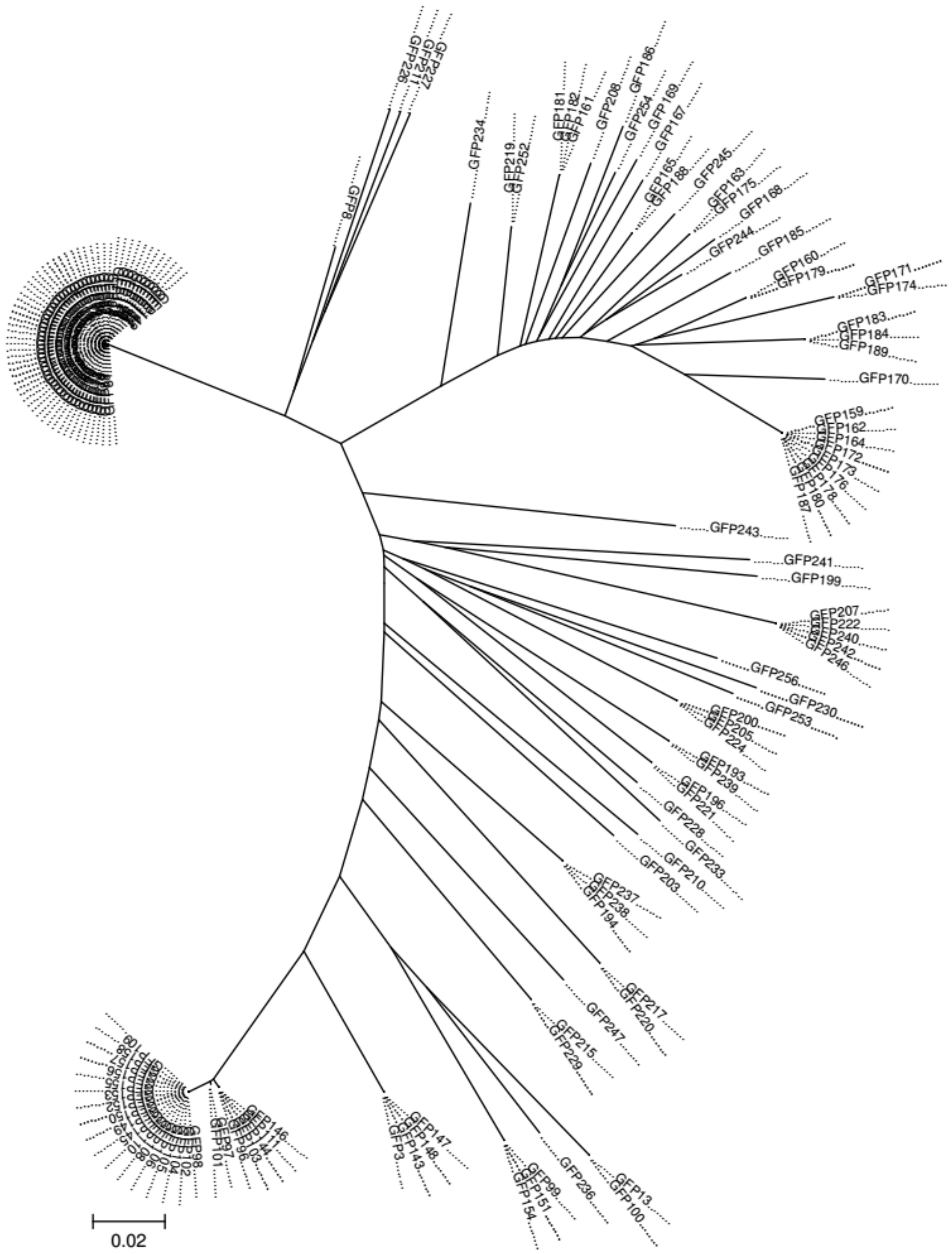
By using a list of sequence parameters utilised by GeneArt's commercially available codon optimisation tool, I further demonstrated that cryptic splice site recognition is a major factor leading to decreased GFP levels of especially GC-poor variants. Whether this is caused by the introduction and/or disruption of splicing regulatory sequences, or directly linked to changes in sequence composition potentially enhancing the usage of weak splice sites, is, as yet, unclear and will require further investigation. In an attempt to quantitatively measure the contribution of several sequence features thought to be beneficial for expression, such as tAI, GC3 and CpG, by correlating experimentally derived data using multiple regression analysis, the issue of multicollinearity between several parameters led to inconclusive results, leaving the requirement for the exploration of alternative, e.g. machine-learning approaches, to decipher more complicated relationships between variables. A collaboration between the Kudla lab and GeneArt will further investigate and quantify the involvement of common codon optimisation features on gene expression by utilising a much larger GFP variant library of (~50,000 variants) in Flow-seq experiments. Results from this screen could be integrated into already existing algorithms for codon optimisation and tested in proof-of-concept experiments using biomedically relevant genes.

Taken together, the methods established and the results presented in this thesis add to the current understanding of how synonymous substitutions affect gene expression by exploring the effects of codon usage on RNA export, stability, splicing and translation, as well as protein yield. Knowledge gained from this project, as well as ongoing work, will greatly benefit biomedical research by improving heterologous gene design for gene therapy, therapeutic protein production, DNA/RNA vaccines and synthetic biology.

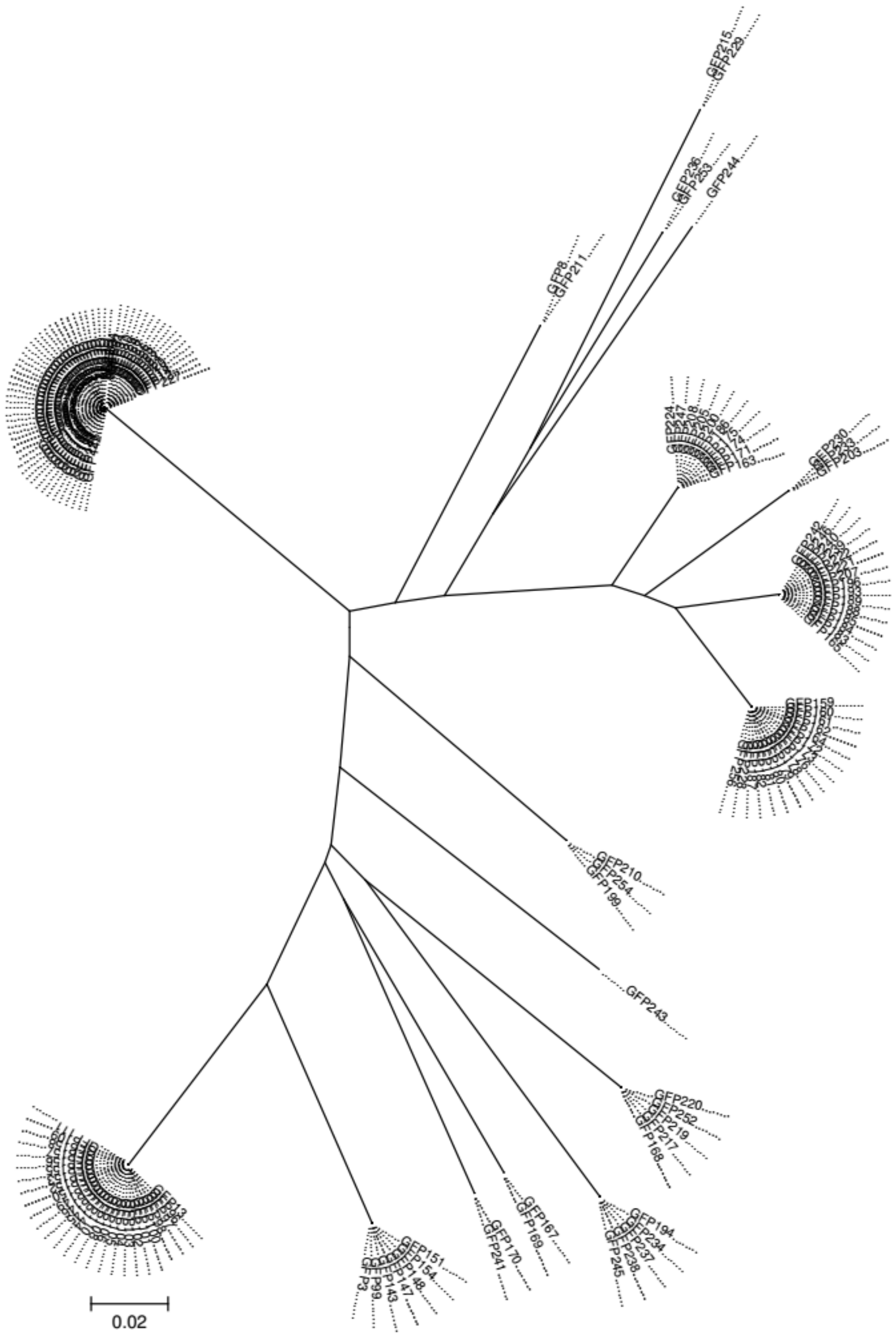
7. Appendix



Supplementary Figure 1. Distance tree of the first sequence third of synthetic GFP genes. An un-rooted tree generated by neighbour-joining, based on the pairwise hamming distance among 168 synthetic GFP genes. Trees generated for nucleotides 1-216 (of 720).



Supplementary Figure 2. Distance tree of the second sequence third of synthetic GFP genes. An un-rooted tree generated by neighbour-joining, based on the pairwise hamming distance among 168 synthetic GFP genes. Trees generated for nucleotides 217-489 (of 720).



Supplementary Figure 3. Distance tree of the third sequence third of synthetic GFP genes. An un-rooted tree generated by neighbour-joining, based on the pairwise hamming distance among 168 synthetic GFP genes. Trees generated for nucleotides 490-720 (of 720).

8. References

Abaza, I., and Gebauer, F. (2008). Trading translation with RNA-binding proteins. RNA 14, 404–409.

Agoglia, R.M., and Fraser, H.B. (2016). Disentangling sources of selection on exonic transcriptional enhancers. Mol. Biol. Evol. 33, 585–590.

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 12, R18.

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: Natural selection and translational accuracy. Genetics 136, 927–935.

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., et al. (2012). Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep. 1, 543–556.

Arava, Y., Wang, Y., Storey, J.D., Liu, C.L., Brown, P.O., and Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S. A. 100, 3889–3894.

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.

Bailey, S. a, Graves, D.E., Rill, R., and Marsch, G. (1993). Influence of DNA base sequence on the binding energetics of actinomycin D. Biochemistry 32, 5881–5887.

Bakheet, T., Frevel, M., Williams, B.R., Greer, W., and Khabar, K.S. (2001). ARED: human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins. Nucleic Acids Res. 29, 246–254.

Barreau, C., Paillard, L., and Osborne, H.B. (2005). AU-rich elements and associated factors: Are there unifying principles? Nucleic Acids Res. 33, 7138–7150.

Bartel, D.P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. Cell 136, 215–233.

Bauer, A.P., Leikam, D., Krinner, S., Notka, F., Ludwig, C., Längst, 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, 1721–1843.

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

Berg, O.G., and Kurland, C.. (1997). Growth rate-optimised tRNA abundance and codon usage. J. Mol. Biol. 270, 544–550.

Bernardi, G. (1993). The Vertebrate Genome : Isochores and Evolution. Mol. Biol. Evol. 10, 186–204.

Bernardi, G. (2012). Isochores. eLS 1–6.

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 8, 1499–1504.

Biro, J.C. (2008). Correlation between nucleotide composition and folding energy of coding sequences with special attention to wobble bases. Theor. Biol. Med. Model. 5, 14.

Brennan, C.M., Gallouzi, I.E., and Steitz, J.A. (2000). Protein ligands to HuR modulate its interaction with target mRNAs in vivo. J. Cell Biol. 151, 1–13.

Brest, P., Lapaquette, P., Souidi, M., Lebrigand, K., Cesaro, A., Vouret-Craviari, V., Mari, B., Barbry, P., Mosnier, J.-F., Hébuterne, X., et al. (2011). A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nat. Genet. 43, 242–245.

Buchan, J.R., Aucott, L.S., and Stansfield, I. (2006). tRNA properties help shape codon pair preferences in open reading frames. Nucleic Acids Res. 34, 1015–1027.

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. Genetics 129, 897–907.

Cáceres, E.F., and Hurst, L.D. (2013). The evolution, impact and properties of exonic splice enhancers. Genome Biol. 14, R143.

Cannarozzi, G., Cannarrozzi, G., Schraudolph, N.N., Faty, M., von Rohr, P., Friberg, M.T., Roth, A.C., Gonnet, P., Gonnet, G., and Barral, Y. (2010). A role for codon order in translation dynamics. Cell 141, 355–367.

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W., and Prasher, D. (1994). Green fluorescent protein as a marker for gene expression. Science (80-.). 263, 802–805.

Chamary, J. V, and Hurst, L.D. (2005). Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. 6, R75.

Chen, C.Y.A., and Shyu, A. Bin (1995). AU-rich elements: characterization and importance in mRNA degradation. Trends Biochem. Sci. 20, 465–470.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L., and McAdams, H.H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci. U. S. A. 101, 3480–3485.

Chevalier-Mariette, C., Henry, I., Montfort, L., Capgras, S., Forlani, S., Muschler, J., and Nicolas, J.-F. (2003). CpG content affects gene silencing in mice: evidence from novel transgenes. Genome Biol. 4, R53.

Choi, T., Huang, M., Gorman, C., and Jaenisch, R. (1991). A Generic Intron Increases Gene-Expression in Transgenic Mice. Mol. Cell. Biol. 11, 3070–3074.

Cline, J., Braman, J., and Hogrefe, H. (1996). PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Res. 24, 3546–3551.

Coleman, J.R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., and Mueller, S. (2008). Virus attenuation by genome-scale changes in codon pair bias. Science 320, 1784–1787.

Cong, L., Ann Ran, F., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L. a., et al. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. Science (80-.). 339, 819–823.

Corneo, G., Ginelli, E., Soave, C., and Bernardi, G. (1968). Isolation and characterization of mouse and guinea pig satellite deoxyribonucleic acids. Biochemistry 7, 4373–4379.

Costantini, M., and Bernardi, G. (2008). The short-sequence designs of isochores from the human genome. Proc. Natl. Acad. Sci. U. S. A. 105, 13971–13976.

Crick, F.H. (1966). Codon--anticodon pairing: the wobble hypothesis. J. Mol. Biol. 19, 548–555.

Crick, F.H., Barnett, L., Brenner, S., and Watts-Tobin, R.J. (1961). General nature of the genetic code for proteins. Nature 1227–1232.

Culjkovic-Kraljacic, B., and Borden, K.L.B. (2013). Aiding and abetting cancer: mRNA export and the nuclear pore. Trends Cell Biol. 23, 328–335.

Dabney, J., and Meyer, M. (2012). Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. Biotechniques 52, 87–94.

Dalle, B., Rubin, J.E., Alkan, O., Sukonnik, T., Pasceri, P., Yao, S., Pawliuk, R., Leboulch, P., and Ellis, J. (2005). eGFP reporter genes silence LCRbeta-globin transgene expression via CpG dinucleotides. Mol. Ther. 11, 591–599.

Dana, A., and Tuller, T. (2014). The effect of tRNA levels on decoding times of mRNA codons. Nucleic Acids Res. 42, 9171–9181.

De Conti, L., Baralle, M., and Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip. Rev. RNA 4, 49–60.

Dean, K.M., and Grayhack, E.J. (2012). RNA-ID, a highly sensitive and robust method to identify cis-regulatory sequences using superfolder GFP and a fluorescence-based assay. RNA 18, 2335–2344.

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., and Xu, T. (2005). Efficient Transposition of the piggyBac (PB) Transposon in Mammalian Cells and Mice. Cell 122, 473–483.

Dittmar, K.A., Goodenbour, J.M., and Pan, T. (2006). Tissue-Specific Differences in Human Transfer RNA Expression. PLoS Genet. 2, e221.

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., Dölken, 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. Sci. Rep. 3, 1318.

Duan, J., Wainwright, M.S., Comeron, J.M., Saitou, N., Sanders, A.R., Gelernter, J., and Gejman, P. V. (2003). Synonymous mutations in the human dopamine receptor D2

(DRD2) affect mRNA stability and synthesis of the receptor. Hum. Mol. Genet. 12, 205–216.

Duret, L. (2000). tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. 16, 287–289.

Ellefson, J.W., Gollihar, J., Shroff, R., Shivram, H., Iyer, V.R., and Ellington, A.D. (2016). Synthetic evolutionary origin of a proofreading reverse transcriptase. Science (80-.). 352, 1590–1593.

Eskesen, S.T., Eskesen, F.N., and Ruvinsky, A. (2004). Natural selection affects frequencies of AG and GT dinucleotides at the 5′ and 3′ ends of exons. Genetics 167, 543–550.

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

Fairbrother, W.G., Yeh, R.F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science (80-.). 297, 1007–1013.

Fan, X.C., Myer, V.E., and Steitz, J. a. (1997). AU-rich elements target small nuclear RNAs as well as mRNAs for rapid degradation. Genes Dev. 11, 2557–2568.

Fischer, U., Huber, J., Boelens, W.C., Mattajt, L.W., and L??hrmann, R. (1995). The HIV-1 Rev Activation Domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs. Cell 82, 475–483.

Fischer, U., Meyer, S., Teufel, M., Heckel, C., Lührmann, R., and Rautmann, G. (1994). Evidence that HIV-1 Rev directly promotes the nuclear export of unspliced RNA. EMBO J. 13, 4105–4112.

Fluman, N., Navon, S., Bibi, E., and Pilpel, Y. (2014). mRNA-programmed translation pauses in the targeting of E. coli membrane proteins. Elife 3, e03440.

Forman, J.J., and Coller, H.A. (2010). The code within the code: MicroRNAs target coding regions. Cell Cycle 9, 1533–1541.

Frenkel-Morgenstern, M., Danon, T., Christian, T., Igarashi, T., Cohen, L., Hou, Y.-M., and Jensen, L.J. (2012). Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. Mol. Syst. Biol. 8, 572.

Fresno, M., Jiménez, A., and Vázquez, D. (1977). Inhibition of translation in eukaryotic systems by harringtonine. Eur. J. Biochem. 72 VN-r, 323–330.

Fu, J., Murphy, K.A., Zhou, M., Li, Y.H., Lam, V.H., Tabuloc, C.A., Chiu, J.C., and Liu, Y. (2016). Codon usage affects the structure and function of the Drosophila circadian clock protein PERIOD. Genes Dev. 1761–1775.

Fu, Y., Foden, J.A., Khayter, C., Maeder, M.L., Reyon, D., Joung, J.K., and Sander, J.D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat. Biotechnol. 31, 822–826.

Furger, A., O'Sullivan, J.M., Binnie, A., Lee, B. a, and Proudfoot, N.J. (2002). Promoter proximal splice sites enhance transcription. Genes Dev. 16, 2792–2799.

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.

Gallouzi, I., and Steitz, J.A. (2001). Delineation of mRNA Export Pathways by the Use of Cell-Permeable Peptides. Science (80-.). 294, 1895–1902.

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159, 907–911.

Gamble, C.E., Brule, C.E., Dean, K.M., Fields, S., Grayhack, E.J., Gamble, C.E., Brule, C.E., Dean, K.M., Fields, S., and Grayhack, E.J. (2016). Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast. Cell 166, 1–12.

Gao, F., Li, Y., Decker, J.M., Peyerl, F.W., Bibollet-Ruche, F., Rodenburg, C.M., Chen, Y., Shaw, D.R., Allen, S., Musonda, R., et al. (2003). Codon usage optimization of HIV type 1 subtype C gag, pol, env, and nef genes: in vitro expression and immune responses in DNA-vaccinated mice. AIDS Res. Hum. Retroviruses 19, 817–823.

Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., and Futcher, B. (2014). Measurement of average decoding rates of the 61 sense codons in vivo. Elife 3.

Gingold, H., Tehler, D., Christoffersen, N.R., Nielsen, M.M., Asmar, F., Kooistra, S.M., Christophersen, N.S., Christensen, L.L., Borre, M., Sørensen, 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 (80-.). 342, 475–479.

Graf, M., Bojak, A., Deml, L., Bieler, K., Wolf, H., and Wagner, R. (2000). Concerted action of multiple cis-acting sequences is required for Rev dependence of late human immunodeficiency virus type 1 gene expression. J. Virol. 74, 10822–10826.

Graf, M., Ludwig, C., Kehlenbeck, S., Jungert, K., and Wagner, R. (2006). A quasi-lentiviral green fluorescent protein reporter exhibits nuclear export features of late human immunodeficiency virus type 1 transcripts. Virology 352, 295–305.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pavé, A. (1980). Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8, 197.

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.

Gupta, S.K., Carmi, S., Ben-Asher, H.W., Tkacz, I.D., Naboishchikov, I., and Michaeli, S. (2013). Basal splicing factors regulate the stability of mature mRNAs in trypanosomes. J. Biol. Chem. 288, 4991–5006.

Gutman, G.A., and Hatfield, G.W. (1989). Nonrandom utilization of codon pairs in Escherichia coli. Proc. Natl. Acad. Sci. 86, 3699–3703.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ann Ran, F., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. 31.

Hurst, L.D. (2006). Preliminary assessment of the impact of microrna-mediated regulation on coding sequence evolution in mammals. J. Mol. Evol. 63, 174–182.

Hyde, S.C., Pringle, I.A., Abdullah, S., Lawton, A.E., Davies, L. a, Varathalingam, A., Nunez-Alonso, G., Green, A.-M., Bazzani, R.P., Sumner-Jones, S.G., et al. (2008). CpG-free plasmids confer reduced inflammation and sustained pulmonary gene expression. Nat. Biotechnol. 26, 549–551.

Ingolia, N.T. (2014). Ribosome profiling: new views of translation, from single codons to genome scale. Nat. Publ. Gr. 15.

Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., and Weissman, J.S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc. 7, 1534–1550.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324, 218–223.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147, 789–802.

Itakura, K., Hirose, T., Crea, R., Riggs, A., Heyneker, H., Bolivar, F., and Boyer, H. (1977). Expression in Escherichia coli of a Chemically Synthesized Gene for the Hormone Somatostatin. Science (80-.). 198, 1056–1063.

Ivics, Z., Hackett, P.B., Plasterk, R.H., and Izsvák, Z. (1997). Molecular Reconstruction of Sleeping Beauty, a Tc1-like Transposon from Fish, and Its Transposition in Human Cells. Cell 91, 501–510.

Ivics, Z., Li, M.A., Mates, L., Boeke, J.D., Nagy, A., Bradley, A., and Izsvak, Z. (2009). Transposon-mediated genome manipulation in vertebrates. Nat. Methods 6, 415–422.

Jeanty, C., Longrois, D., Mertes, P.-M., Wagner, D.R., and Devaux, Y. (2010). An optimized protocol for microarray validation by quantitative PCR using amplified amino allyl labeled RNA. BMC Genomics 11, 542.

Jeong, S., and Stein, A. (1994). Micrococcal nuclease digestion of nuclei reveals extended nucleosome ladders having anomalous DNA lengths for chromatin assembled on non-replicating plasmids in transfected cells. Nucleic Acids Res. 22, 370–375.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. Elife 2.

Joung, J.K., and Sander, J.D. (2013). TALENs: a widely applicable technology for targeted genome editing. Nat Rev Mol Cell Biol 14, 49–55.

Julien, P., Miñana, B., Baeza-Centurion, P., Valcárcel, J., and Lehner, B. (2016). The complete local genotype–phenotype landscape for the alternative splicing of a human exon. Nat. Commun. 7, 11558.

Kilchert, C., Wittmann, S., and Vasiljeva, L. (2016). The regulation and functions of the nuclear RNA exosome complex. Nat. Rev. Mol. Cell Biol. 17, 227–239.

Kimchi-Sarfaty, C., Oh, J.M., Kim, I.-W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S. V, and Gottesman, M.M. (2007). A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315, 525–528.

Kimura, M. (1980). Journal of Molecular Evolution A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences. J. Mol. Evol 16, 111–120.

Kitabayashi, M., and Esaka, M. (2003). Improvement of reverse transcription PCR by RNase H. Biosci. Biotechnol. Biochem. 67, 2474–2476.

Kosovac, D., Wild, J., Ludwig, C., Meissner, S., Bauer, A., and Wagner, R. (2010). 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.

Kreitman, M. (1996). The neutral theory is dead. Long live the neutral theory. Bioessays 18, 678–683.

Krinner, S. (2012). The impact of intragenic CpG content on epigenetic control of transgene expression in mammalian cells.

Krinner, S., Heitzer, A.P., Diermeier, S.D., Obermeier, I., Längst, G., and Wagner, R. (2014). CpG domains downstream of TSSs promote high levels of gene expression. Nucleic Acids Res. 42, 3551–3564.

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 (80-.). 324, 255–258.

Lander, E.S., Heaford, a, Sheridan, a, Linton, L.M., Birren, B., Subramanian, a, Coulson, a, Nusbaum, C., Zody, M.C., Dunham, a, et al. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Lei, H., Dias, A.P., and Reed, R. (2011). Export and stability of naturally intronless mRNAs require specific coding region sequences and the TREX mRNA export complex. Proc. Natl. Acad. Sci. U. S. A. 108, 17985–17990.

Leuenroth, S.J., and Crews, C.M. (2008). Triptolide-induced transcriptional arrest is associated with changes in nuclear substructure. Cancer Res. 68, 5257–5266.

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.

Lo, Y.S., Tseng, W.H., Chuang, C.Y., and Hou, M.H. (2013). The structural basis of actinomycin D-binding induces nucleotide flipping out, a sharp bend and a left-handed twist in CGG triplet repeats. Nucleic Acids Res. 41, 4284–4294.

López de Silanes, I., Zhan, M., Lal, A., Yang, X., and Gorospe, M. (2004). Identification of a target RNA motif for RNA-binding protein HuR. Proc. Natl. Acad. Sci. U. S. A. 101, 2987–2992.

Louhichi, A., Fourati, A., and Rebaï, A. (2011). IGD: A resource for intronless genes in the human genome. Gene 488, 35–40.

Ma, L., Cui, P., Zhu, J., Zhang, Z., and Zhang, Z. (2014). Translational selection in human: more pronounced in housekeeping genes. Biol. Direct 9, 17.

Malim, M.H., Hauber, J., Le, S.Y., Maizel, J. V., and Cullen, B.R. (1989). The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. Nature 338, 254–257.

Mao, Z., Bozzella, M., Seluanov, A., and Gorbunova, V. (2008). Comparison of nonhomologous end joining and homologous recombination in human cells. DNA Repair (Amst). 7, 1765–1771.

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

Mitsui, M., Nishikawa, M., Zang, L., Ando, M., Hattori, K., Takahashi, Y., Watanabe, Y., and Takakura, Y. (2009). Effect of the content of unmethylated CpG dinucleotides in plasmid DNA on the sustainability of transgene expression. J. Gene Med. 11, 435–443.

Miura, M., Tanigawa, C., Fujii, Y., and Kaneko, S. (2013). Comparison of six commercially-available DNA polymerases for direct PCR. Rev Inst Med Trop Sao Paulo 55, 401–406.

Mladenova, V., Mladenov, E., and Russev, G. (2009). Organization of Plasmid DNA into Nucleosome-Like Structures after Transfection in Eukaryotic Cells. Biotechnol. Biotechnol. Equip. 23, 1044–1047.

Mohr, S., Ghanem, E., Smith, W., Sheeter, D., Qin, Y., King, O., Polioudakis, D., Iyer, V.R., Hunicke-Smith, S., Swamy, S., et al. (2013). Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. RNA 19, 958–970.

Mouchiroud, D., D'Onofrio, G., Aïssani, B., Macaya, G., Gautier, C., and Bernardi, G. (1991). The distribution of genes in the human genome. Gene 100, 181–187.

Mouchiroud, D., Fichant, G., and Bernardi, G. (1987). Compositional compartmentalization and gene composition in the genome of vertebrates. J. Mol. Evol. 26, 198–204.

Müller-McNicoll, M., and Neugebauer, K.M. (2013). How cells get the message: dynamic assembly and function of mRNA-protein complexes. Nat. Rev. Genet. 14, 275–287.

Narum, D.L., Kumar, S., Rogers, W.O., Fuhrmann, S.R., Liang, H., Oakley, M., Taye, A., Sim, B.K.L., and Hoffman, S.L. (2001). Codon optimization of gene fragments encoding Plasmodium falciparum merzoite proteins enhances DNA vaccine protein expression and immunogenicity in mice. Infect. Immun. 69, 7250–7253.

Newman, Z.R., Young, J.M., Ingolia, N.T., and Barton, G.M. (2016). Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9. Proc. Natl. Acad. Sci. U. S. A. 113, E1362-71.

Neymotin, B., Ettorre, V., and Gresham, D. (2015). Global determinants of mRNA degradation rates in Saccharomyces cerevisiae. bioRxiv 14845.

Nguyen, K.-L., Llano, M., Akari, H., Miyagi, E., Poeschla, E.M., Strebel, K., and Bour, S. (2004). Codon optimization of the HIV-1 vpu and vif genes stabilizes their mRNA and allows for highly efficient Rev-independent expression. Virology 319, 163–175.

Nierhaus, K.H. (2006). Decoding errors and the involvement of the E-site. Biochimie 88, 1013–1019.

Noderer, W.L., Flockhart, R.J., Bhaduri, A., Diaz de Arce, A.J., Zhang, J., Khavari, P.A., and Wang, C.L. (2014). Quantitative analysis of mammalian translation initiation sites by FACS-seq. Mol. Syst. Biol. 10, 748–748.

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.

Nottingham, R.M., Wu, D.C., Qin, Y., Yao, J.U.N., Hunicke-smith, S., and Lambowitz, A.M. (2016). RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. RNA 22, 597–613.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99, 247–257.

Ossareh-Nazari, B., Bachelerie, F., and Dargemont, C. (1997). Evidence for a role of CRM1 in signal-mediated nuclear protein export. Science 278, 141–144.

Oyola, S.O., Otto, T.D., Gu, Y., Maslen, G., Manske, M., Campino, S., Turner, D.J., MacInnis, B., Kwiatkowski, D.P., Swerdlow, H.P., et al. (2012). Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. BMC Genomics 13, 1.

Pagani, F., Raponi, M., and Baralle, F.E. (2005). Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc. Natl. Acad. Sci. U. S. A. 102, 6368–6372.

Pechmann, S., and Frydman, J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat. Struct. Mol. Biol. 20, 237–243.

Pechmann, S., Chartron, J.W., and Frydman, J. (2014). Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. Nat. Struct. Mol. Biol. 21, 1100–1105.

Percudani, R., Pavesi, A., and Ottonello, S. (1997). Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. J. Mol. Biol. 268, 322–330.

Pfaffl, M.W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res. 29, e45.

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. 101, 12588–12591.

Pollard, V.W., and Malim, M.H. (1998). The HIV-1 Rev Protein - Overview of the Retroviral Life Cycle. 491–532.

Pop, C., Rouskin, S., Ingolia, N.T., Han, L., Phizicky, E.M., Weissman, J.S., and Koller, D. (2014). Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. Mol. Syst. Biol. 10, 770.

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.

Puchta, O., Cseke, B., Czaja, H., Tollervey, D., Sanguinetti, G., and Kudla, G. (2016). Network of epistatic interactions within a yeast snoRNA. Science (80-.). 352, 840–844.

Qian, W., Yang, J.-R., Pearson, N.M., Maclean, C., and Zhang, J. (2012). Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genet. 8, e1002603.

Qin, Y., Yao, J., Wu, D.C., Nottingham, R.M., Mohr, S., Hunicke-Smith, S., and Lambowitz, A.M. (2016). High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. RNA 22, 111–128.

Quax, T.E.F., Claassens, N.J., Söll, D., and van der Oost, J. (2015). Codon Bias as a Means to Fine-Tune Gene Expression. Mol. Cell 59, 149–161.

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.

Reeves, R., Gormanl, C.M., and Howard2, B. (1985). Minichromosome assembly of non-integrated plasmid DNA transfected into mammalian cells. Nucleic Acids Res. 13, 3599–3615.

Robberson, B.L., Cote, G.J., and Berget, S.M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell. Biol. 10, 84–94.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. Cell 163, 698–711.

Rudolph, K.L.M., 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.

Saccone, S., Federico, C., and Bernardi, G. (2002). Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. Gene 300, 169–178.

Savisaar, R., and Hurst, L.D. (2016). Purifying selection on exonic splice enhancers in intronless genes. Mol. Biol. Evol. 1–64.

Schneider, R., Campbell, M., Nasioulas, G., Felber, B.K., and Pavlakis, G.N. (1997). Inactivation of the human immunodeficiency virus type 1 inhibitory elements allows Rev-independent expression of Gag and Gag/protease and particle formation. J. Virol. 71, 4892–4903.

Schneider-Poetsch, T., Ju, J., Eyler, D.E., Dang, Y., Bhat, S., Merrick, W.C., Green, R., Shen, B., and Liu, J.O. (2010). Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. Nat. Chem. Biol. 6, 209–217.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic Regulation of Nucleosome Positioning in the Human Genome. Cell 132, 887–898.

Schor, I.E., Rascovan, N., Pelisch, F., Alló, M., and Kornblihtt, A.R. (2009). Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. Proc. Natl. Acad. Sci. U. S. A. 106, 4325–4330.

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., and Ragg, T. (2006). BMC Molecular Biology The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol. Biol. 7.

Sémon, M., Lobry, J.R., and Duret, L. (2006). No evidence for tissue-specific adaptation of synonymous codon usage in humans. Mol. Biol. Evol. 23, 523–529.

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. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28–38.

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

Shin, Y.C., Bischof, G.F., Lauer, W. a, and Desrosiers, R.C. (2015). Importance of codon usage for the temporal regulation of viral gene expression. Proc. Natl. Acad. Sci. U. S. A. 112, 14030–14035.

Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat. Methods 11, 817–820.

Sokolowski, M., Tan, W., Jellne, M., and Schwartz, S. (1998). mRNA Instability Elements in the Human Papillomavirus Type 16 L2 Coding Region. J. Virol. 72, 1504–1515.

Stadler, M., and Fire, A. (2011). Wobble base-pairing slows in vivo translation elongation in metazoans. RNA 17, 2063–2073.

Stergachis, A.B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., Raubitschek, A., Ziegler, S., LeProust, E.M., Akey, J.M., et al. (2013). Exonic transcription factor binding directs codon choice and affects protein evolution. Science 342, 1367–1372.

Tan, W.E.I., Felber, B.K., Zolotukhin, A.S., Pavlakis, G.N., and Schwartz, S. (1995). Efficient Expression of the Human Papillomavirus Type 16 L1 Protein in Epithelial Cells by Using Rev and the Rev-Responsive Element of Human Immunodeficiency Virus or the cis - Acting Transactivation Element of Simian Retrovirus Type 1. J. Virol. 69, 5607–5620.

Taniguchi, I., Mabuchi, N., and Ohno, M. (2014). HIV-1 Rev protein specifies the viral RNA export pathway by suppressing TAP/NXF1 recruitment. Nucleic Acids Res. 42, 6645–6658.

Tillo, D., and Hughes, T.R. (2009). G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics 10, 442.

Titov, D. V, Gilman, B., He, Q.-L., Bhat, S., Low, W.-K., Dang, Y., Smeaton, M., Demain, A.L., Miller, P.S., Kugel, J.F., et al. (2011). XPB, a subunit of TFIIH, is a target of the natural product triptolide. Nat. Chem. Biol. 7, 182–188.

Tonidandel, S., Lebreton, J.M., and Johnson, J.W. (2009). Determining the statistical significance of relative weights. Psychol. Methods 14, 387–399.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010a). An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. Cell 141, 344–354.

Tuller, T., Veksler-Lublinsky, I., Gazit, N., Kupiec, M., Ruppin, E., and Ziv-Ukelson, M. (2011). Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol. 12, R110.

Tuller, T., Waldman, Y.Y., Kupiec, M., and Ruppin, E. (2010b). Translation efficiency is determined by both codon bias and folding energy. Proc. Natl. Acad. Sci. U. S. A. 107, 3645–3650.

Urnov, F.D., Rebar, E.J., Holmes, M.C., Zhang, H.S., and Gregory, P.D. (2010). Genome editing with engineered zinc finger nucleases. Nat Rev Genet 11, 636–646.

Valencia, P., Dias, A.P., and Reed, R. (2008). Splicing promotes rapid and efficient mRNA export in mammalian cells. Proc. Natl. Acad. Sci. 105, 3386–3391.

van der Kuyl, A.C. (2012). HIV infection and HERV expression: a review. Retrovirology 9, 6.

Varathalingam, A., Lawton, A., Munkonge, F., Chan, M., Pringle, I., Greisenbach, U., Alton, E., Gill, D., and Hyde, S. (2005). Novel CPG-Depleted and Codon-Optimised CFTR CDNAs Maintain the Structure and Function of CFTR Protein. | UK CF Gene Therapy Consortium. p.

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

Wang, Y., Lu, J., He, L., and Yu, Q. (2011). Triptolide (TPL) inhibits global transcription by inducing proteasome-dependent degradation of RNA polymerase II (Pol II). PLoS One 6, e23993.

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.

Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell 119, 831–845.

Watson, J.D., and Crick, F.H. (1953). The structure of DNA. Cold Spring Harb. Symp. Quant. Biol. 18, 123–131.

Weatheritt, R.J., and Babu, M.M. (2013). The hidden codes that shape protein evolution. Science 342, 1325–1326.

Wolff, B., Sanglier, J.-J., and Wang, Y. (1997). Leptomycin B is an inhibitor of nuclear export : inhibition of translocation of the human immunodeficiency virus type 1 (HIV-l) Rev protein and Rev-dependent mRNA Jean-Jacques. Chem. Biol. 4, 139–147.

Xing, K., and He, X. (2015). Reassessing the "Duon" hypothesis of protein evolution. Mol. Biol. Evol. 32, 1056–1062.

Xu, Y., Ma, P., Shah, P., Rokas, A., Liu, Y., and Johnson, C.H. (2013). Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. Nature 495, 116–120.

Yant, S.R., Wu, X., Huang, Y., Garrison, B., Burgess, S.M., and Kay, M.A. (2005). High-Resolution Genome-Wide Mapping of Transposon Integration in Mammals. Mol. Cell. Biol. 25, 2085–2094.

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.

Zhang, F., Frost, A.R., Blundell, M.P., Bales, O., Antoniou, M.N., and Thrasher, A.J. (2010). A ubiquitous chromatin opening element (UCOE) confers resistance to DNA methylation-mediated silencing of lentiviral vectors. Mol. Ther. 18, 1640–1649.

Zhang, G., Hubalewska, M., and Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat. Struct. Mol. Biol. 16, 274–280.

Zhou, M., Wang, T., Fu, J., Xiao, G., and Liu, Y. (2015). Nonoptimal codon usage influences protein structure in intrinsically disordered regions. Mol. Microbiol. 97, 974–987.

Zhou, M., Guo, J., Cha, J., Chae, M., Chen, S., Barral, J.M., Sachs, M.S., and Liu, Y. (2013). Non-optimal codon usage affects expression, structure and function of clock protein FRQ. Nature 495, 111–115.

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.