מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Thesis for the degree
Master of Science

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

עבודת גמר (תזה) לתואר
מוסמך למדעים

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

By
Aviv Rotman

מאת
אביב רוטמן

לעשות יותר תמורת פחות: כיצד תאים ממזערים את עלות התרגום?

Doing More with Less: How do Cells Minimize the Cost of Translation?

Advisor:
Prof. Yitzhak Pilpel

מנחה:
פרופ' יצחק פלפל

September 2015

תשרי ה'תשע"ו

Acknowledgments

I would first like to express my gratitude to my mentor, Prof. Yitzhak Pilpel, for his dedication, caring, and support. His kindness, teaching, and passion for science are what made this project possible in a difficult time, and for this I am eternally grateful.

I would like to thank my lab members for the discussions, and late night snacks. I can't imagine a better environment to do science in and it's all because of you. A special thanks to Orna Dahan and Ernest Morderet for the special attention and support they gave throughout the project.

I would also like to thank my project partner, Idan Frumkin, you are an inseparable part of this project, a great scientist in the making, a wonderful person, and I don't know how I would have done this without you.

I would also like to thank my family and partner, for their unfaltering support in the face of great adversity. I am lucky to have you, and to have found you, and I am grateful for that.

Finally, I would like to dedicate this work to my Father, who I'd like to hope would be proud of me. I cherish his memory and miss him greatly.

"

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference.

"

ROBERT FROST
The Road Not Taken | 1920

1| Abstract

Protein translation is one of life's most costly processes as it extensively consumes production resources, raw material consumables and energy. Thus, organisms have evolved ways to synthesize new proteins in efficient and regulated manners. Although translation is a well studied process, we still do not fully understand its economy, and more specifically how cells minimize the costs of genes' translation while maintaining the desired expression level. Towards that aim, we decided to utilize an integrated approach of synthetic biology, lab-evolution and deep sequencing. Recently, a large synthetic gene library with ~14K variants was created and used to study the effects of various regulatory 5' region elements over the expression level of a GFP gene in Escherichia coli. This library allows the studying of different features, such as transcription levels, translation initiation rates, codon biases and mRNA secondary structures on synthesis cost, directly from cell's fitness, as the GFP has no function in the cell. We utilized this powerful tool to learn about the effects of sequence variation on the fitness of the cell, by growing and evolving the library for 28 days in six parallel evolutionary lines and sequencing samples at regular intervals of ~30 generations. Since all designs are competing against each other, we can compute the relative fitness of each of them. A comparison between fitness and GFP expression level revealed a negative correlation, but, interestingly only above a certain threshold. Comparing the expected and observed fitness showed large span even when accounting for protein expression level. By analyzing this "fitness residual" we found that ribosome attenuation at the early elongation phase allows minimizing cost of translation. Interestingly, this attenuation is reached by three independent mechanisms: First is a tight mRNA secondary structure. Second, is a high affinity of the mRNA to the ribosome anti- Shine Dalgarno binding site. Third, a ribosomal-flow model, which relies on the translation speed of each codon, suggests that positive residual designs tend to show deeper and earlier ribosomal bottle-necks than negative designs. In addition to production resources, consumables also affected fitness. In particular, we found that incorporation of amino acids with a low energetic cost is associated with a positive fitness residual. In conclusion, our study provides a comprehensive data set and analysis that help decipher the economy of translation and identify in particular factors that allow minimizing production costs at a desired expression level.

2| Introduction

Translation is a fundamental cellular process that is performed by all life forms. Since proteins are the molecules that perform most of life's functions, cells use translation as a regulatory mechanism for central cellular functions such as response to stress[1,2], cell cycle[3,4], and differentiation[5–7]. Thus, an efficient and timely translation process is at the heart of cell biology. Optimizing translation is also cardinal for heterologous gene expression, which has become a prominent tool in the pharmaceutical, agricultural, medical, and chemical industries[8,9]. In the past, it was believed that transcription was the dominant mechanism that dictates expression levels[10]. Yet, today we realize the complexity of the picture, as signal transduction, transcription, translation, and degradation were all shown to affect the concentration of proteins in the cell[11–13]. Strikingly, although translation is a thoroughly studied process, the mechanisms of its regulation are far from being fully understood.

Translation machinery is one of the largest and most complicated complexes in the cell, comprising of the ribosome, tRNA, tRNA-synthetases and amino-acids. Ribosomal RNA comprises ~60% of the cellular RNA[14], while ribosomal protein make up ~10–20% of the proteome[15,16]. tRNA is also a highly abundant species of RNA, making up ~15% of the cellular RNA[14]. Together tRNA and rRNA comprise the vast majority of the cell's RNA content[17]. Notably, amino acids are a central metabolic resource in the cell, 65% of its consumed energy is devoted to their synthesis[18] . Interestingly, it is estimated that ~50% of the cellular proteome is related to the translation machinery[19], leading to the conclusion that around half of the translation machinery requirement is for self-maintenance.

Although translation of new proteins is mandatory to sustain life, it also bears a significant cost. This cost can be the result of either the result of the translation process itself, or the energetic cost of the resources that are utilized during translation[20,21]. It is estimated that ~75% of the cell's energy is devoted to the synthesis of new proteins[22]. Thus, mechanisms to regulate the translation process itself have evolved to ensure that it is performed efficiently. These mechanisms

assist the cell to minimize mis-translation events[23], correctly allocate ribosomes along the mRNA[24] and regulate initiation rates[25,26].

Until recently, it was believed that the only translation regulating mechanism was on initiation rates. The initiation rate, which is the rate at which translation events start, was shown to be majorly affected by the Shine-Dalgarno[26] (SD) motif in prokaryotes[27–29]. In addition to the SD motif, it was recently revealed that the secondary structure of the Ribosome Binding Site (RBS) region also participates in regulation over the initiation[30,31]. While the regulatory mechanisms of translation initiation were being studied[32], it was not entirely apparent that the elongation phase of translation also poses regulatory qualities. At the time, synonymous changes were thought to have no effect on translation and thus dubbed "silent" mutations[33,34]. Once sequencing became a mainstay in biological research and actual sequence data was obtained for the first time, it became clear that not all codons were created equal. Consistently, a phenomenon termed “codon usage bias” was revealed, in which the distribution of codons in the genome is non-random[35–37]. Notably, codon bias was more pronounced in highly expressed genes, which utilize a very specific set of codons, that differs between organisms[38,39]. Additionally, it was becoming clear that the codon usage of genes was correlated with tRNA availability[40,41]. Some hypothesized that the adaptation to the tRNA pool was what enabled the highly expressed genes to be translated efficiently, with abundant tRNAs more available in the cytoplasm and thus more rapidly translated[42]. Yet, a different approach suggested that the codon bias was better explained by the GC content of the area surrounding the gene in the genome, as genomic GC content contributes to structure and heat resistance[43].

In recent years, due to advances in high-throughput sequencing[44,45] and DNA synthesis technologies[46] alongside with the emergence of the ribosome profiling technology[47], it has become clearer that translation elongation is a non-uniform process. By sequencing all ribosome occupied RNA in the cell, it is possible to estimate elongation rate from ribosome density for every position on a transcript. Using this method exposed that translation is a highly diverse process, with yeast

translation elongation rates varying by 3 orders of magnitude[48], and different environments, such as nutrient starvation[52] or meiosis[53]. Recently, a number of ribosome profiling studies of closely related yeast strains, and their hybrids, has explored the effect of translational regulation on expression[50,51]. These studies suggest a major role for translation as a regulatory mechanism on gene expression. Conflicting data are undecided on whether this observation is a major means of regulation[50], or a minor one[51], but what is clear is that regulation of expression levels evolves on both the transcriptional and translation levels.

Many models for measuring codon usage have been proposed in the past, in hopes that the correct model would be an accurate predictor for translation speed. The first widely used model, which was suggested to model translation speed of a given codon, relied on the assumption that highly expressed genes utilize the most adapted codons. A gene with a codon bias that resembles the codon bias in highly expressed genes should be highly expressed in itself. Thus the Codon Adaptation Index (CAI)[52] was proposed, by comparing the codon usage in a gene to the codon usage of a reference set of highly expressed genes. Further findings suggested that translation speed was correlated more directly to tRNA availability[53]. This notion was somewhat confirmed by experiments heterologously over expressing tRNA genes, in which rare tRNA was expressed and changes in expression levels were observed[54]. This meant that by changing the tRNA pool, one could affect translation efficiency. Conversely, the tRNA pool of an organism does not entirely correlate with the most abundant codons in the genome, with some codons lacking a cognate (fully matching) anti codon tRNA. Thus, the tRNA Adaptation Index (tAI) was defined[55], a parameter which takes the organism's tRNA levels and wobble pairing rules into consideration when calculating a codon adaptation. tRNA levels were difficult to obtain (and still are to date) so the tAI was based on genomic data, using tRNA gene copy number as a proxy for cellular concentrations of tRNAs. When analyzing ribosome profiling data some claim that there is correlation between tAI and ribosome pausing[56,57].

tRNA availability is not the only parameter thought to explain codon bias and elongation speed. For instance, a sequence parameter not taken into account when calculating tRNA adaptation is the local mRNA secondary structure of the transcript. As mentioned earlier, secondary structure is crucial to translation initiation, but recent works show that it can have a strong effect on elongation by inducing translational pauses[58]. An additional non-tRNA related parameter is Shine-Dalgarno affinity, which can stall ribosomes by binding to the anti Shine-Dalgarno site on the riobosome[59]. This mechanism is particularly interesting since it is not codon dependent and can affect translation speed out of frame. Finally, it has been shown that the peptide sequence can also affect translation speed. By reanalyzing ribosome profiling data, Charneski *et al.*[60] deduced that peptide charge was in correlation with ribosome density; even to a great extend compared to secondary structure. In particular, these authors deduced that positively charged amino acids likely greatly retard ribosomes downstream from where they are encoded, presumably because they interact with the negatively charged ribosomal exit tunnel. This observation was further validated by Lu *et al.*[61] by synthesizing a library of peptides with increasing lengths of charged amino acids that would localize to known regions of the ribosomal exit tunnel at the time of translation. By measuring peptide lengths, the authors found that positively charged stretches promote ribosomal arrest, whereas negative or neutral stretches are translated faster. In contrast, *Artieri et al.*[62] proposed that these results were due to sequencing bias and sparse coverage of certain regions. Specifically, *Artieri et al* suggest that proline slows the ribosome when it is coded in a contiguous stretch[62–64], and this mechanism can explain most ribosomal attenuation.

Considering all the diverse mechanisms discussed above that regulate translation, it is interesting to reveal the evolutionary mechanisms and forces that led to their appearance. The two most accepted theories are selection for translation accuracy and translation efficiency. Translation efficiency has also been studied thoroughly, though the connection between translation speed and translational efficiency is still far from being understood. Initially it was assumed that genes that were better adapted to the tRNA pool, preferring to use common codons, are translated faster

and thus more efficiently[65]. This hypothesis was validated by studies which attempted to optimize the codon bias of a sequence to its host organism using codon bias data and secondary structure, leading to increased heterologous gene expression in many cases[8,66–68]. Conversely a similar study found that the use of codons less affected by amino acid starvation was the main predictor of heterologous expression levels[69]. Recently, a study on ribosome profiling data has shown that codons with less abundant cognate tRNAs are decoded slower, thus resulting in decreased elongation rates[70]. The group reanalyzed various organisms' ribosome profiling data and calculated the Mean of the Typical Decoding Rates (MTDR), and went on to show that this measure correlates well with genomic expression levels[71]. Alternatively, many studies have shown that translation efficiency relies on other parameters, such as mRNAs secondary structure[56], or the lack of Shine-Dalgarno-like sequences[59]. Finally, a recent study has shown that codon bias, and secondary structure act in a compensatory fashion. By analyzing genomic data, and reanalyzing expression data from a synthetic library, Gorochowski *et al.* found that tRNA adaptation and mRNA secondary structure are usually high when the other is low. The study explains that this phenomenon may smooth out translation elongation rates lowering the likelihood of potentially deleterious pauses or speed-ups.

The solution for this supposedly paradoxical situation may lie in the fact that the transcript has a sequential nature, and codon context is an essential variable to consider when predicting elongation efficiency. A group recently developed a model of tRNA adaptation which takes into account the location of the codon, and reported significant correlation to expression[72]. Specifically, it has been shown that 5' region of the transcript is a crucial area for regulation efficient translation[73]. This sector, the border between initiation and elongation, seems to be a highly influential area with many overlapping signals. Some groups maintain that the parameter of most significance on the 5' end is the codon adaptation, but counter intuitively the codons should be non-adaptive. The non-adaptive codons create an area of very slow translation early on in the peptide, forming a "translation ramp", which has been found in highly expressed genes both in genomic data[24] and analysis of synthetic

libraries[74]. This ramp is hypnotized to slow down translation in early elongation, but leads to overall efficient translation by preventing ribosomal jams and translation abortion by optimally spacing ribosomes. Conversely, studies have found that the determining parameter on the 5' end of the coding sequence is loose mRNA secondary[72,73,75,76]. This should allow for high initiation rate and smooth flow of the ribosome afterwards. Many of these studies have used synthetic libraries to examine this subject, by recoding and measuring the expression of genes with different codon biases and secondary structures[73,76,77]. Further studies based on multi genomic analysis claim that the codon bias observed in the 5' region is not a translational ramp, but a byproduct low GC content which seems too correlated with low tRNA adaptation[78]. Contrarily, a study has explained that the previous findings are due to the use of synthetic genes with uncommonly strong mRNA secondary structure and when considering natural secondary structure levels codon usage bias has a significant contribution to translation speed[79]. Finally, numerous groups have tried to reconcile the various hypotheses by proposing that all options can influence translation efficiency. Analysis of ribosome profiling data has shown that multiple parameters have an effect on translation efficienc[57,70,80], including both codon usage bias and mRNA secondary structure, as well as amino acid charge and proline content.

A basic question arises from the aforementioned findings, how do we measure translation efficiency? As shown above, the majority of recent research on the subject uses expression level as the measure of translational efficiency. This is an imperfect definition since not all protein production is required to sustain maximum capacity, and more importantly, it disregards the other variable that defines efficiency, cost. As discussed above, efficient expression is not maximal expression; rather it is expression at a desired level with minimal cost[81]. When ignoring the effect of cost on efficiency the evolutionary question is lost, and we cannot surmise from the data what strategy maximizes fitness. Recently, *Ellis et al.*[82] have tried to answer this question by measuring constitutive expression of a reporter gene, on the background of various synthetic constructs that varied in gene copy number, transcription level, initiation rate, and codon optimization. By comparing reporter

gene output level among all designs, *Ellis et al.* were able to measure the effect of these different parameters on the translation capacity of the cell. The authors found that optimal designs, which maximized both cellular capacity and protein production of the synthetic construct, demonstrated higher transcription levels and lower initiation rates. While interesting, this work only measured translational cost of a construct on a reporter gene, but did not address the link to cellular fitness.

In our project we try to pour light onto this unanswered question, how does the cell optimize expression for fitness? We do this by using a synthetic library developed in one of the previously mentioned studies. In a recent study Goodman *et al.*[76] addressed the question of maximizing protein production by examining expression of a synthetic library. The group synthesized a library combining 2 promoters with different transcription levels, 3 RBSs with varying initiation rates, and 137 initiator peptides taken from the highest expressed proteins in the E. coli genome. These 11 amino acid peptides were each recorded 13 times to represent different codon biases and secondary structures (See methods and figure 1). The library was expressed in E. coli, and the expression level of each design was analyzed by combining flow cytometry and next generation sequencing. The obvious result was that expression level correlates well with the different transcription and translation initiation rates, but these alone were not sufficient to explain all the variance in expression levels. The group concluded that the main parameter during early elongation dictating efficient expression was the secondary structure, with little correlation to codon bias.

While that experiment helped clarify which sequence features maximize expression, it did not provide an idea on which features affect the cost of production. While cost increases with production level, and fitness decreases with cost, an interesting question is which sequence features minimize fitness cost at a given expression level. In our experiment we used the Goodman *et al.* library to examine the relation between sequence variation and expression cost. We evolved the library in the lab for ~270 generations in 6 replicates, and sequenced the population at different time points, and different lineages. This procedure allowed us to examine the relative

fitness of all the designs and find which sequence parameters effect fitness. Predictably, the population changes drastically over time, with some designs disappearing completely and others taking over more than 80% of the population. Additionally, we observed a significant correlation between protein level and fitness, but only above a certain threshold of expression. This indicates our set up can detect the cost of expression, but only when it is strong enough to overcome inherent cellular noise. We calculated a "fitness residual', being the difference between expected fitness at a given expression level and the observed fitness at that level. We found that several parameters that promote ribosome attenuation in early elongation are correlated with fitness residual. We also found that protein energetic cost has a strong influence on fitness as well.

3 | Goals

In this work we aimed to shed light on a relatively untouched aspect of translation- its cost on the cell. We set out to understand the mechanisms in which the cell optimizes expression, namely minimizing cellular cost while maintaining a desired expression level. Particularly, we focused on the importance of sequence features at the 5' end of the gene, and how they affect the relative fitness of the cell.

Goal I

We aimed to monitor the effect of translation on cellular fitness by utilizing an existing, synthetic library. This goal was achieved by subjecting the library to a lab-evolution experiment followed by sequencing samples to track the frequency of each design according to time.

Goal II

Next, we aimed to elucidate sequence parameters that minimize translational cost. By linking protein expression to relative fitness in the population, we hoped to find parameters that explain the deviations from expected fitness at a given expression level. To this end, we analyzed various nucleotide and amino acid sequence parameters and compared them to the deviation from expected fitness. Thus, we were able to uncover parameters that led to higher than expected fitness.

4 | Methods

4.1 Library architecture

The library (Fig. 1) reported in Goodman *et al.*[76] was synthesized by Agilent Technologies using an oligo synthesis process[83]. All oligos were then ligated to the plasmid (pGERC) directly upstream to a GFP reporter. Each design in the library is composed of a promoter, a Ribosome Binding Sites (RBS) and a linker of 11 amino acids. The library as a whole includes: two promoters with either Low or High transcription levels. Three synthetic RBSs with Strong, Mid, or Low translation initiation rates, as well as 137 different genomic RBSs that were defined as the 20bps upstream to the ORF of the 137 most highly expressed genes in E. coli genome (WT). Finally, 137 coding sequences (CDS) consisting of the first 11 amino acids from the same highly expressed genes. Notably, Each CDS appears in the library in 13 different synonymous forms: the WT sequence, the most common codons based on the frequencies in the E. coli genome, the rarest codons, and additional 10 designs with increasing mRNA secondary structure folding energies. All combinations amounted in 14,234 distinct designs.

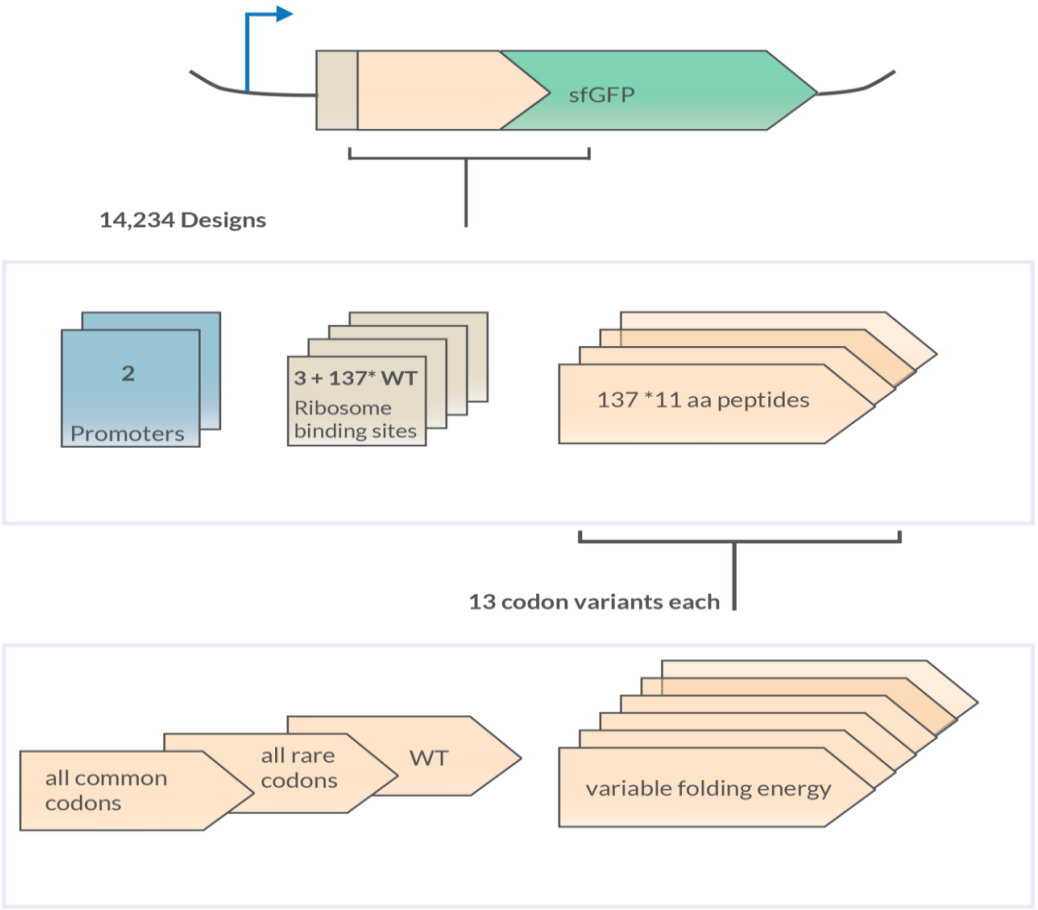


Figure 1 – Goodman *et al.* library design
The library is composed of 14,234 designs linked to a sfGFP reporter gene. Each design includes a promoter (High and Low), an RBS (Strong, Mid, Weak, and WT), and a stretch of 11 aa taken from the 137 highest expressed E. coli genes. Each CDS was recoded synonymously 13 times to represent 10 levels for mRNA secondary structure, the WT sequence, the rarest and most common codons.

4.2 Evolution

Lab-evolution experiment was carried out by serial dilution. The library was grown on 1.2 ml of LB + 50µg/ml kanamycin at 30°C in six parallel lineages and was diluted daily by a factor of 1:128 into fresh media (results in ~6.9 generations per dilution). This procedure was repeated for 28 days and every four days (~27 generations) samples were taken from each lineage, mixed with glycerol and kept at -80°C. Thus, each of the six independent lineages has seven samples at days 4, 8, 12, 16, 20, 24 & 28.

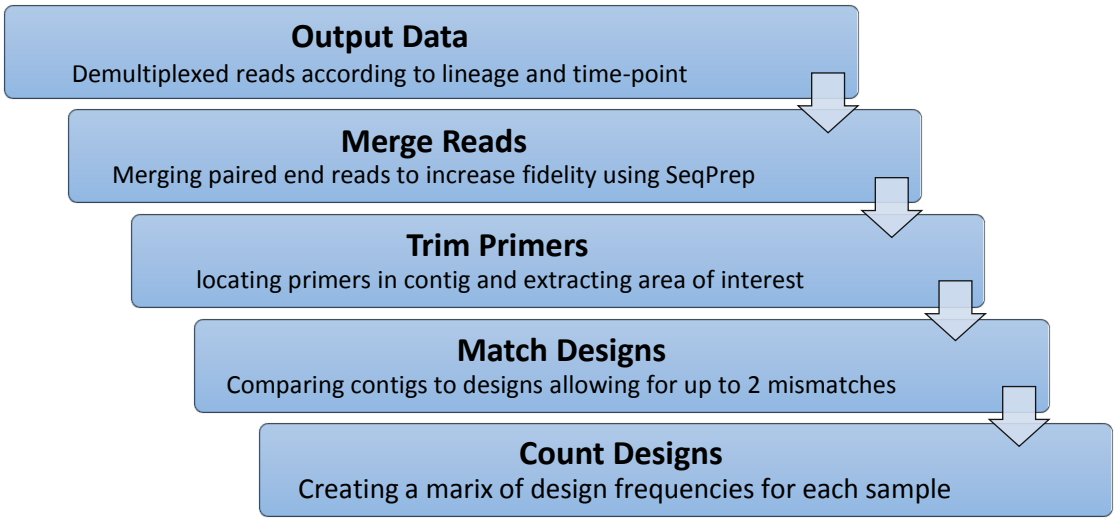
4.3 Library preparation and sequencing

Plasmids from time zero (library "ancestor") and 42 evolution samples were purified with a QIAgene mini-prep kit and used as templates for PCR to amplify specifically the variable region of all designs in the population. To minimize PCR and sampling biases, we used a large amount of template, ~500ng of DNA, and a relatively short PCR of 26 rounds. The forward primer (sequence CAGCTCTTCGCCTTTACGCATATG) was paired with 5 different reverse primers (sequences: R1: GACAATGAAAAGCTTAGTCATGGCG, R2: ACAATGAAAAGCTTAGTCATGGCG, R3: CAATGAAAAGCTTAGTCATGGCG, R4: AATGAAAAGCTTAGTCATGGCG, R5: ATGAAAAGCTTAGTCATGGCG) that are one bp shifted from each other to insure that library complexity was high enough for Illumina sequencing. PCR products were then run on BluePipin to capture the correct amplicon size of ~140 bps and remove any un-specific amplicons. Then, DNA buffer was exchanged using Agencourt AmPure SPRI bead cleanup protocol. Hiseq library was prepared next using the sequencing library module from Blecher-Gonen, R. *et al.* 2013[84]. In short, blunt ends were repaired, Adenine bases were added to the 3' end of the fragments, barcode adapters containing a T overhang were ligated, and finally the adapted fragments were amplified. The process was repeated for each sample with a different Illumina DNA barcode for multiplexing, and then all samples were pooled in equal amounts and sequenced. We preformed a 125 bp paired end high output run on the HiSeq 2500 PE Cluster Kit v4. Base calling is performed by RTA v. 1.18.64, and de-multiplexing is carried out with Casava v. 1.8.2, outputting results in FASTQ format.

4.4 Data processing

De-multiplexed data was received in the form of FASTQ files split into samples. First, SeqPrep[85] was used to merge paired reads into a single contig, to increase sequence fidelity over regions of dual coverage, as was done in the Goodman and Kosuri *et al.* studies[76,86]. The size of each contig was then compared to the theoretical combined length of the forward primer, the reverse primer and the variable region of the designs. Next, the forward and reverse primers were found on each contig (allowing

for 2 mismatches) and trimmed out. This step was performed for both the forward and reverse complement sequences of the contig, to account for non-directional ligation of the adaptors during library preparation. Next, the reverse primer was searched at the last 5 nucleotides of the contig to account for different primer lengths, as explained in section 4.2. Once primers were trimmed, the contig was tested again for its length to ensure no indels had occurred. The percentage of contigs that were filtered out at this step was not dependent on the evolutionary time of the sample, suggesting that indels do not play a major role in the evolutionary competition. Then, contigs were discarded if they included ambiguous bases anywhere along their sequence. Contigs were then compared sequentially to the entire library, comparing the sequence each contig to the sequence of each design. Any contig without a matching design within two mismatches or less was discarded. Contigs with more than a single matching design with the same reliability were also discarded due to ambiguity. Mismatches were calculated using a simple base comparison between the contig and the sequence of each design. Each contig that passed these filters was counted in key value data structure, connecting all designs in the library to their frequency in each sample. All contigs discarded from any of the steps were recorded with their sequence quality data and reason of failure. This data was then used for all downstream analyses.



Flow Chart 1 - FitSeq Data Processing Flow

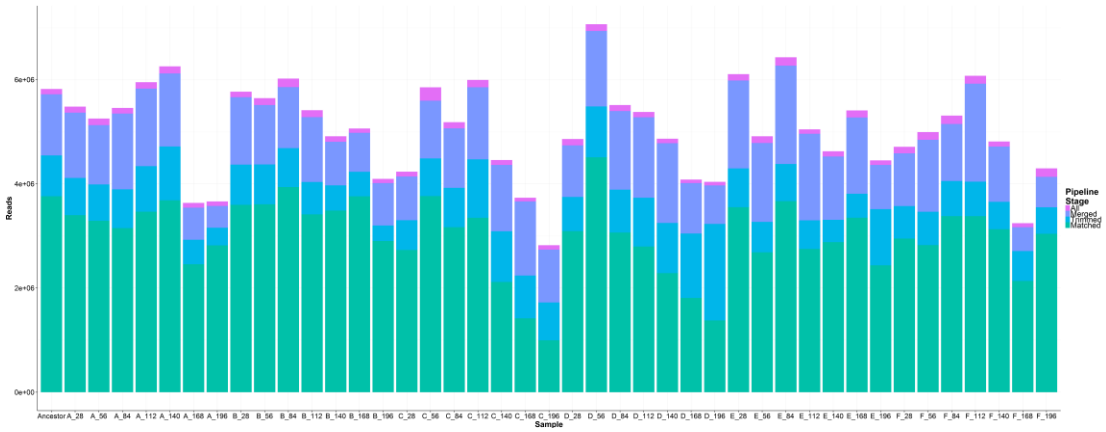


Figure 2 – Number of reads passing the various pipeline stages per sample

Each column represents a sample, in the form of <lineage>_<time-point>, with each of the colors matching each of the pipeline stages: pink = all reads of a given sample, purple= successfully merged reads, blue= successfully trimmed reads, turquoise = reads that were matches to a single design in the library. Notably, the variability in number of reads passing each stage is dependent on the total number of reads in the sample, and not on the time or lineage of the experiment.

4.5 Sequence features

Data set from Goodman *et al.*

The data-set from Goodman *et al.* was obtained and provided the framework for our data-set. In short, the Goodman *et al*. data included the sequence of each design, DNA count (plasmid copy number of each design), RNA level normalized to DNA, raw FlowSeq data, protein levels calculated from FlowSeq data, tRNA adaptation data such as tAI and CAI, secondary structure (ΔG) and GC content. More detailed explanations on this data and its derivation can be found in the supplementary material of Goodman *et al.* 2013.

Calculating translation initiation rate per design

We estimated the translation initiation rate of each design with the "RBS calculator"[87,88] which simulates initiation rates given a UTR and a coding sequence. This calculation is achieved by using a biomechanic model combining the affinity to the anti Shine-Dalgarno sequence at the ribosome binding site, mRNA secondary structure of the UTR and coding sequence, and steric interference of the ribosome and the mRNA.

TASEP simulation

To model the effects of sequence on translation we collaborated with the lab of Prof. Tamir Tuller at Tel Aviv University to simulate ribosome flow on each of the designs. To this end, we used the Totally Asymmetric Simple Exclusion Process (TASEP)[89] model of particle flow and translation speed taken from the MTDR[70] work of the Tuller lab. The MTDR data is derived from ribosome sequencing data by calculating the ribosome profile distribution of each codon. This measurement symbolizes the translation speed of each codon, and it correlates significantly with tRNA availability. In the TASEP simulation, ribosomes are injected to the transcript according to an initiation rate (taken from the RBS calculator data), and they translate at a per-codon speed that is described by the MTDR value of each codon. The simulation is run until steady state is reached and average density of ribosomes is recorded for each codon for a given sequence. We used the TASEP simulation to obtain the mean ribosome number occupying a single mRNA, the mean ribosomal density per codon, the ribosomal density at the first codon, and the location and depth of the ribosome-flow bottleneck, being the codon with the highest ribosome density

Peptide properties

Peptide properties were calculated using the "Peptides" R package[90]. We calculated different amino acid composition metrics, aliphatic index[91], Boman index[92], charge[93], hydrophobic index[94], hydrophobicity[95], instability index[96], molecular weight, and pI. Amino-acid cost was derived using the "aacost" table from the "seqinr" package[97], which holds for each amino acid the amount of energy consumed for its production in high energy ATP or GTP bonds[98]. Cost was either evaluated per amino acid or summed for the whole peptide.

Shine-Dalgarno affinity

The Shine-Dalgarno affinity was calculated identically to Li *et al.* Nature 2012[98]. In short, for each position we calculated the affinity occurring 8-11 bp upstream, the distance between the ribosome A site and the anti Shine-Dalgarno site. The affinity was calculated for all 10 bp sequences in the area, and the maximum value was returned. Affinities were pre calculated for all possible 10mers, using the RNA annealing function from the Vienna package. We then calculated the affinity for the entire variable sequence and recorded for each design the affinity of the first

position, the number of positions with a non-negligible affinity (ΔG< 0), the median, mean, and standard deviation for the non-negligible positions, and the position with the maximum affinity in the design. We repeated this analysis for the constant region, taking into account that the affinity of the first 11 bases of the GFP depended on the sequence of the variable region.

4.6 Statistical analysis of FitSeq data and library parameters

Linear regression of fitness based on protein levels

A one parameter linear regression was calculated between protein levels and fitness on a log2*log2 scale. As a proxy for fitness, the design frequency at generation 84 (i.e. normalized read count of the design out of the total sample read count) was divided by the initial frequency of the design in the ancestor population. We filtered out designs that had a log 2 protein level above 17.5, since this is the point of saturation in Goodman et al. measurements. Additionally, designs with log 2 protein levels of less than 14 were discarded as their fitness didn't show significant correlation with protein level (see Fig. 9). Notably, only designs with High promoters were included in the analysis, since almost all Low promoter designs did not pass the protein level filter. This decision was essential as there are very few low-promoter designs that pass the threshold, and these designs demonstrate unique features such as a very low GC-content that could mask other signals.

Calculation of fitness residuals and classifying designs into positive and negative fitness residual groups

We defined the "fitness residual" of a design as the difference between the fitness predicted by the linear model and its expression, and the observed fitness that was calculated as explained above. We then split the designs into two groups- positive fitness residual designs and negative fitness residual designs. To account for random noise during evolution and sampling biases we only included designs that showed an identical fitness residual sign in five or six lineages. The set of all the above filters resulted in 613 designs in the negative group and 951 in the positive.

Parameter comparison between two fitness residual groups

A one-sided Wilcoxon rank-sum test was used to compare the distributions of different sequence parameters between the positive and negative fitness residual groups[99], and resulting p-values were corrected for multiple hypotheses using the Holm–Bonferroni method[100]. This correction was performed separately for either peptides or nucleotide sequence parameters as these two parameter types were analyzed independently. Next, we tested the effect size of each parameter using the Hodges–Lehmann estimator[101] and set a cutoff of 5% for parameters we defined as important. We choose this threshold since it demonstrated the sharpest reduction in effect size when all parameters were ranked from highest to lowest effect size. Here too, the analysis was performed separately for amino acid and nucleotide sequences. Notably, the 5% cutoff was adequate in both cases. All effect sizes, p values, and q values for both amino acid and nucleotide sequences can be found in appendix A.

Amino acid enrichment and enrichment ratio calculation

To calculate the frequency of the various amino acids in the collective proteome in either the positive or the negative fitness residual group, we counted the occurrences of each amino acid in each design. We then summed this number for each amino acid across all designs in each group and divided the sum by the number of designs in each group multiplied by 11. To quantify the relationship between amino acid enrichment and energetic-cost we calculated the frequency ratio of each amino acid by dividing the amino acid frequency of the positive fitness residual group by the frequency of the negative group. If this enrichment value is larger than 1 for a given amino acid, it is enriched in the positive group. If the value is smaller than 1, then the amino acid is enriched in the negative fitness residual group. We then calculated the Pearson correlation between the amino acid enrichment ratio and the amino acid energetic-cost (see 4.5).

5 | Results

5.1 Library designs show growth rate differences even with a short duration of growth

To study the relation between sequence parameters in the 5' region of a gene to the output expression level, *Goodman et al.* synthesized a synthetic library with ~14K different designs expressing a GFP gene with an upstream variable region[76] (see methods and figure 1). Then, the GFP expression level of each design was measured and correlated with different sequence parameters (see introduction). To check whether different designs demonstrate a range of fitness values, we decided to analyze the DNA coverage of each design in *Goodman et al*. published data, which we hypothesized to serve as a proxy for cellular fitness. If expression of the GFP gene indeed serves as a burden on the cell's ability to grow, expression levels should correlate negatively with fitness. Since the library was grown for an estimated duration of ~50-70 generations at the time of *Goodman et al.* experiment, we hypothesized that major growth differences could already be detected.

Indeed, we found that not all designs are represented equally in the population, with design frequencies spanning three orders of magnitudes (Fig. 3). Notably, when splitting the library into design groups according to the promoter-RBS classes, we observed an association between the median frequency of each class and the class identity, which in turn are associated with GFP expression level (see introduction). Notably, designs with the High promoter showed significantly less DNA coverage than designs with the Low promoter, simply indicating that high transcription rate was selected against during the process of library preparation. A similar trend was found among the different RBSs as the group with the strongest RBS demonstrated the lowest DNA coverage. These observations suggest that the various designs in the library not only range in GFP expression levels but also in fitness, making it ideal for posing questions about the efficiency of the translation process and how cells minimize its cost.

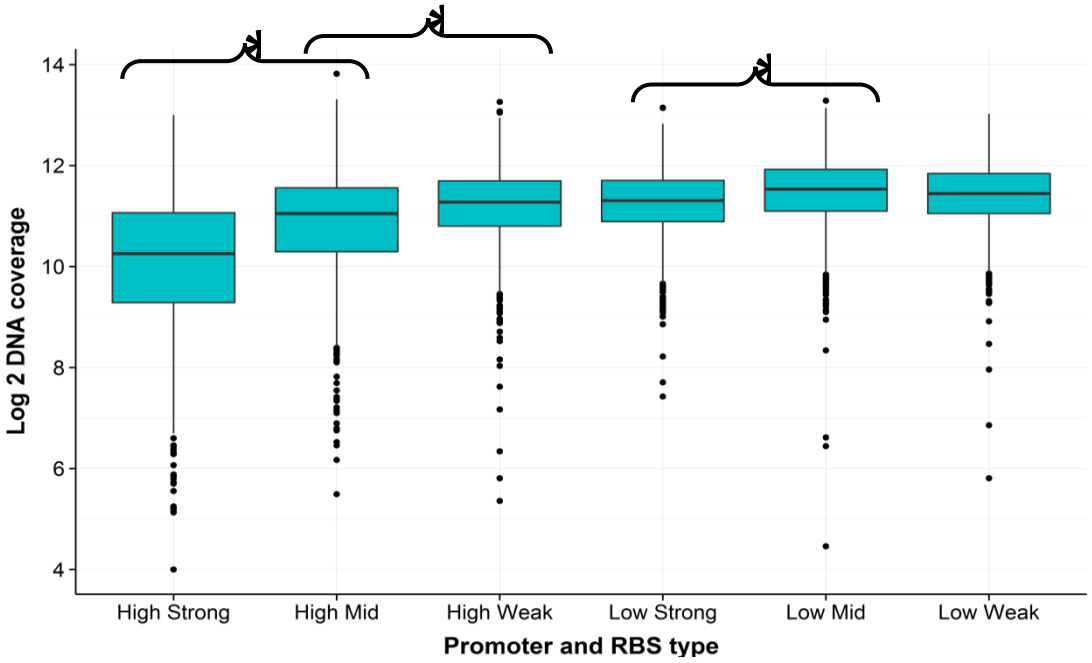


Figure 3 – DNA coverage data from Goodman *et al.* shows high level of variability and association with expression level

The DNA sequencing data was collected to estimate the plasmid copy number of each design in the population for normalization purposes. Notably, the variability of the DNA coverage value is very high with a values ranging between 15 and 14,463 with a mean of 2,535 and a standard deviation of 1,342. Interestingly, the DNA coverage of the groups defined by promoter and RBS are different, in respective to their expression level. The differences between the promoters are most pronounced, but the RBS show the same pattern as well. P. value < 2.2e-16 is marked by an asterisk (*).

5.2 FitSeq experiment further elucidates fitness differences in the library

To take full advantage of the potential of the library and reveal more subtle growth differences among the designs, we designed a new approach, "FitSeq" (Fig. 4 and see methods). In FitSeq, we performed a lab-evolution experiment with the library as the ancestor source. The experiment was run in six parallel lineages for ~200 generations, with samples taken every ~30 generations. The ancestor and all samples were then sequenced and analyzed for dynamics and sequence parameters that minimize translation cost.

5.2 Figure 03

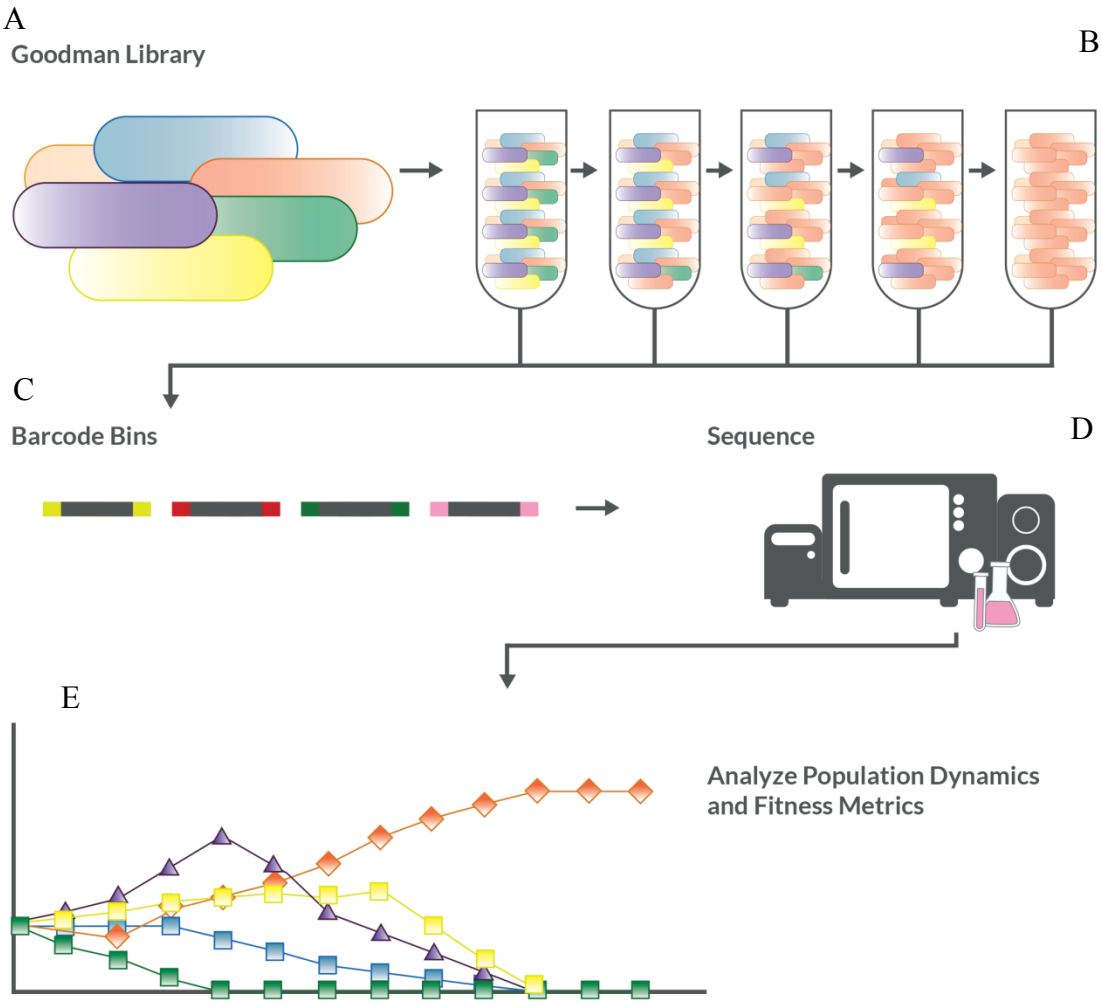


Figure 4 – FitSeq – a method to compare the relative fitness of a synthetic library
A) The library (in this case from Goodman *et al.*) is evolved in a serial dilution set up, in multiple lineages. B) The evolution is sampled at constant time intervals. C) The sample DNA is extracted, amplified, and barcoded. D) Barcoded fragments are sent to deep sequencing. E) Sequencing results are processed and analyzed according to time point and lineage.

5.3 FitSeq data reveals evolutionary dynamics in which most variants go extinct while others fixate in the population

To learn about the population dynamics in our lab-evolution experiment, we first looked at design extinction over time by calculating the percentage of designs covered with at least a single read over time (Fig. 5). Consistently, the number of designs covered by sequencing declines over time as some are out-competed by their opponents. Interestingly, design coverage remains similar among independent lineages until generation ~112, while differences in emerge after that time point. The major collapse in design coverage occurs at generation ~168 and by generation ~200

coverage ranges between ~87-93%, signifying in the extinction of around ~1400 designs.

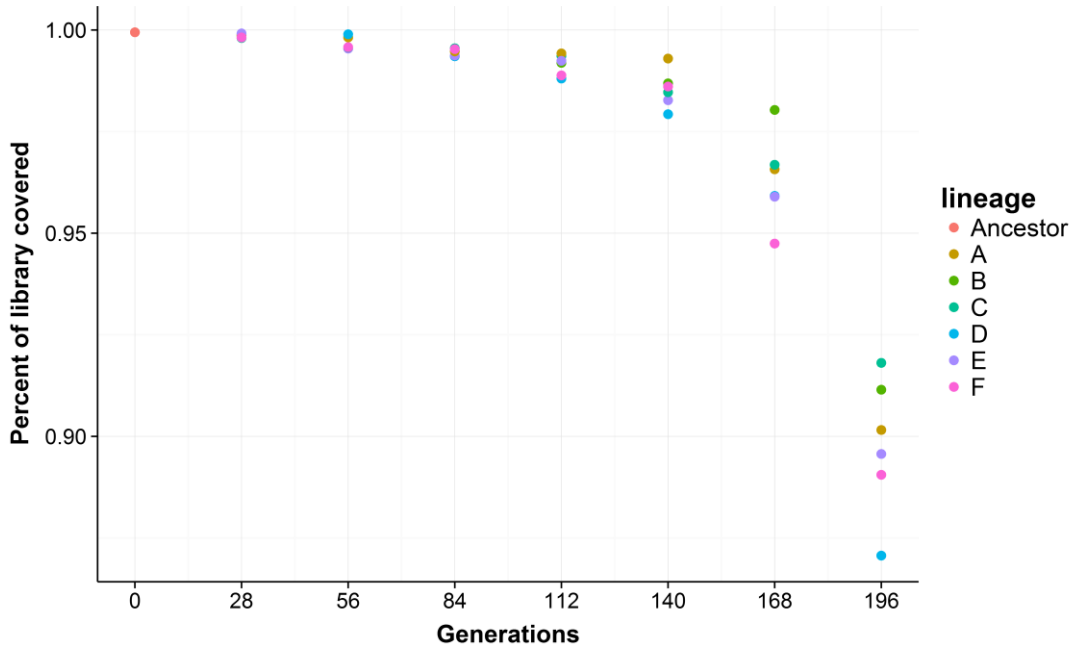


Figure 5 – Percent of library designs covered over time
Percent of designs in population covered by at least 1 read. Lineages start diverging in at generation 140.

To quantitatively compare the evolutionary dynamic among all samples, we calculated the Gini index[102], which reflects the inequality in a population, for each time point per lineage. Here, we calculated the inequality in designs' frequencies where a Gini score of 0 means that all library designs share the same frequency, and a score of 1 means that only a single design exists in the population. Consistent with our previous finding, the initial Gini score of the ancestor is above 0. Examining the Gini index over time (Fig. 6) revealed that inequality rises throughout all lineages, with the most rapid escalation happening after generation 112. The final Gini scores are close to 1, indicating that few designs fixate in the population.

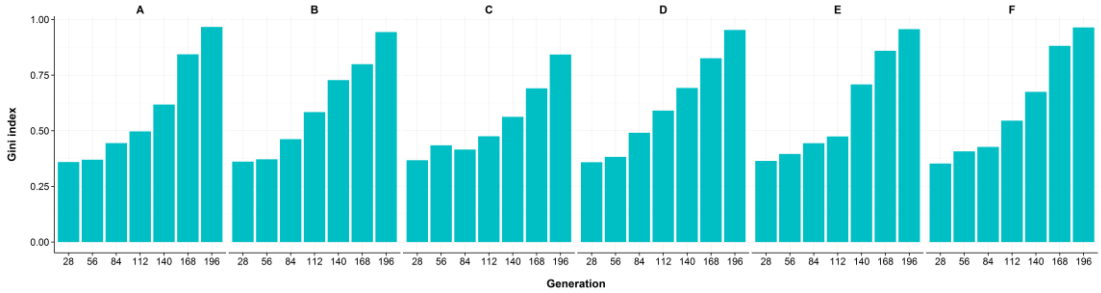


Figure 6 – Gini index score over time by lineage
The Gini score represents the inequality in frequency distribution in the population. A score of 0 represents equal frequencies among designs in the population, and a score of 1 means that a single design exists. The Gini scores starts climbing in all lineages at around generation 112. The final Gini score of all lineages is close to 1, meaning most of the population is composed of very few designs.

Then, we turned to elucidate the population dynamics over time when splitting the library to six groups according to the promoter-RBS classification (Fig. 7). While frequency distributions of all groups are similar at first, expression-based differences emerge starting at generation 84. Interestingly, towards the end of the experiment the entire population collapses, consistent with the Gini index score discussed above.

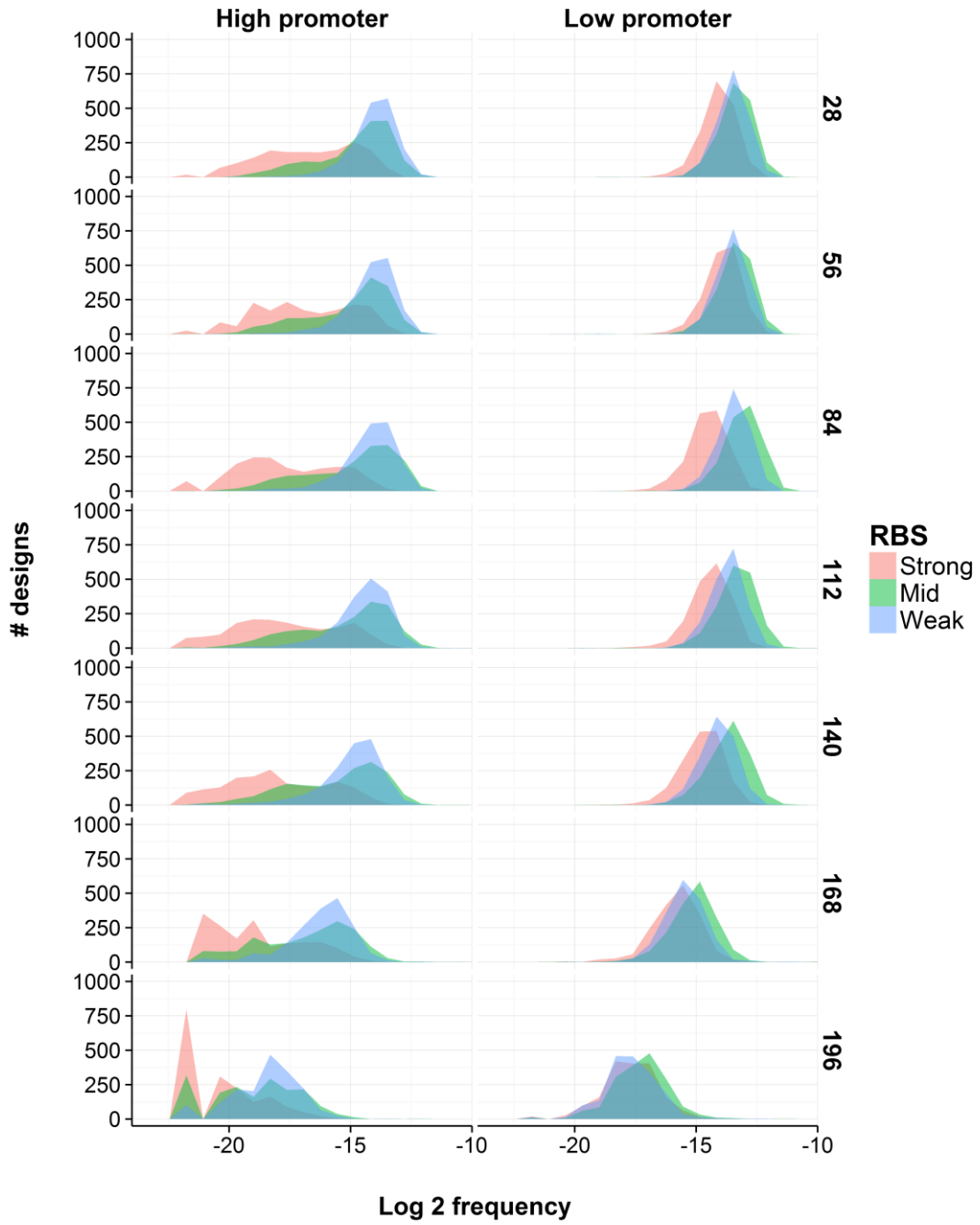


Figure 7 – Frequency distribution divided by generation, promoter, and RBS
The distribution of frequency in lineage A, representing all lineages, is initially similar among the three RBS and only slightly different between the promoters. At generation 84, the high promoter designs decline, especially those with the strong RBS. At generation 140 the population means start declining and reach a very low frequency by generation 196. This is probably due to single designs fixating in the population. The same dynamics are reproduced for all lineages.

We then further explored the similarities among samples by clustering them (UPGMA) according to either Pearson or Spearman correlation of design frequencies. With 7 time points, 6 repeats and 1 ancestor we had a total of 43 samples to compare to one another. We assessed similarity between each pair of samples,

based on read counts of all ~14,000 designs through two means, the Pearson, or Spearman correlation. The Pearson dendrogram (Fig. 8A) shows that in early time points the samples are clustered according to the generation of the sample and not by the independent lineage. However, starting around generation 112, samples from the same lineage tend to cluster together. This observation suggests that the population dynamics in our experiment is split into two phases. First, a deterministic phase in which the designs compete among themselves and the dynamic is governed by the fitness of each design compared to the mean fitness in the population. This phase is less subjected to random events and thus the dynamics in each lineage is similar to others. At a certain point, depending on factors like dilution ratio or population size, beneficial mutations emerge either on the variable region of each design or in the genome itself. Since these mutations occur randomly in cells with different background, a stochastic element is affecting population dynamics and thus few, arbitrary designs fixate in the population of each lineage. Interestingly, the Spearman-based clustering revealed that samples were clustered together according to generation, rather than lineage, throughout the experiment (Fig. 8B). This difference between Person and Spearman may be the result of the non-parametric nature of the Spearman correlation, which is less affected by extreme outliers. Thus, in our case, the Spearman-based clustering demonstrates the deterministic forces in the experiment (governed by fitness) rather than the stochastic ones (governed by random mutations and drift).

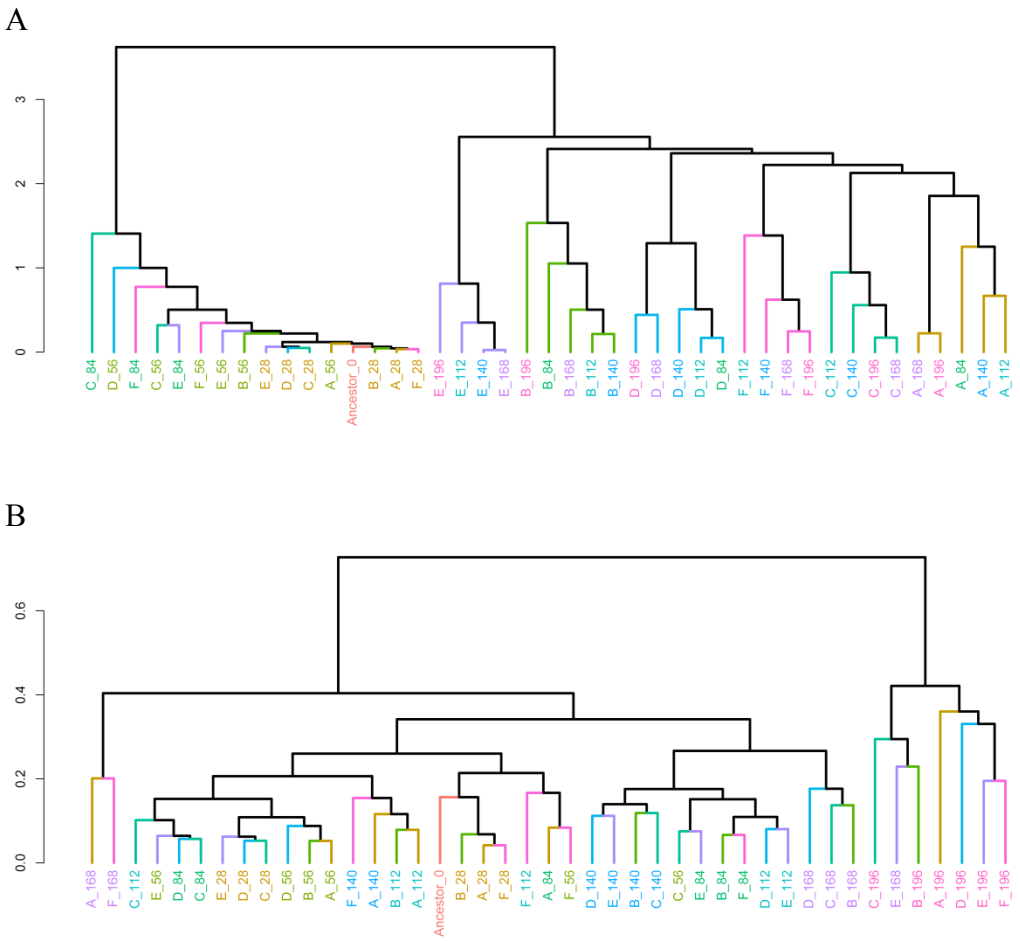


Figure 8 – Sample clustering according to Pearson or Spearman correlation
The UPGMA dendrograms were created using the correlation of design frequencies between each pair of samples as a distance matrix. A) Pearson correlation dendrogram, the samples initially cluster by generation, and later cluster by lineage. B) Spearman correlation dendrogram in which samples cluster mostly by generation throughout the entire experiment. Labels are colored according to generation, whereas branches according to lineage.

In this work, our aim was to first reveal the fitness differences among all designs in the library in order to elucidate sequence parameters that minimize expression costs. Hence, we needed to choose the most appropriate time in our lab-evolution experiment that shows growth differences on the one hand but is not affected by beneficial mutations, fixation of mutated designs or stochastic events on the other hand. Given figures 5-8, we found that generation 84 fit these criterions best.

5.4 Correlation between fitness and GFP expression levels appears only above a threshold

Since expressing the GFP deprives valuable resources from the cell, the higher the GFP expression level is the faster the design is out-competed by other designs, and is less prevalent in the population. Thus, we hypothesized that fitness and GFP expression level should correlate across all GFP expression levels. As a proxy for fitness, we used the frequency of each design at generation ~84 normalized to its initial frequency in the ancestor, as explained in section 5.3 (see methods). Fig. 9 demonstrates the fitness over GFP expression level for all designs with the high promoter. Interestingly, it seems that fitness is affected by expression only above a threshold of 14 (arbitrary units of GFP reads). This observation suggests that expression of a functionless protein can cause a burden on cell growth only above a certain expression level. Below this level, we found that the effect of gene expression on fitness is negligible, presumably because resources are not considerably more wasted compared to other processes that lead to such misuse, such as inaccurate gene expression or noise.

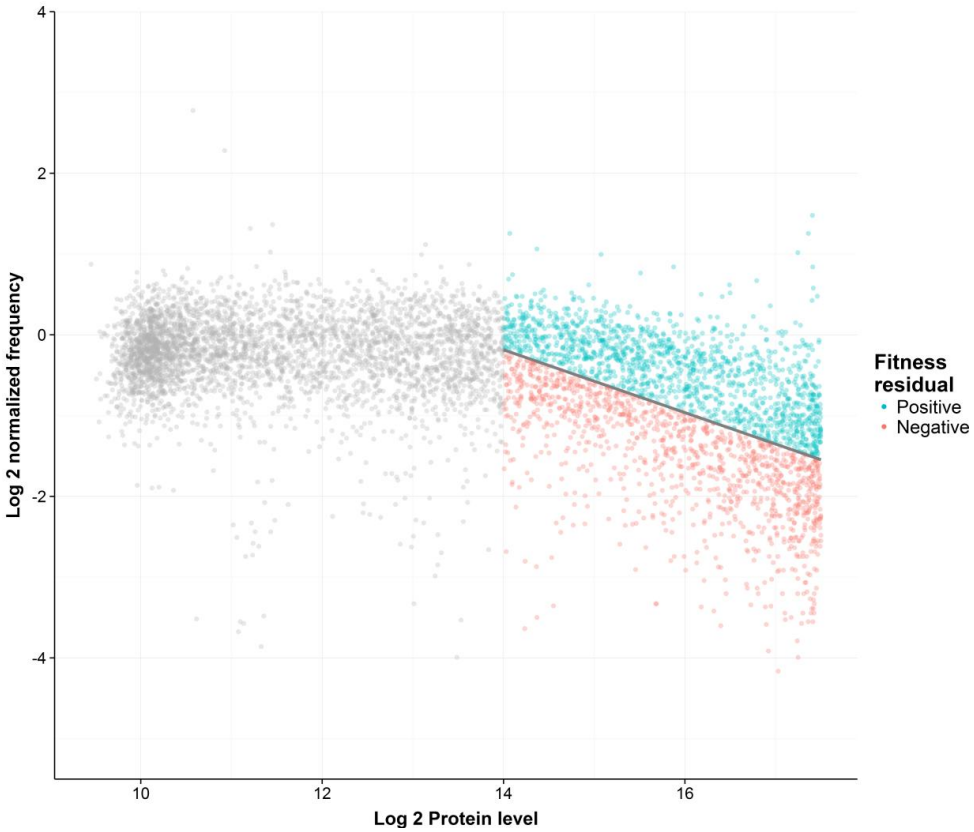


Figure 9 – Fitness correlated to protein level

The log 2 protein levels from Goodman *et al.* correlate (slope = -0.39 p-value = 2.70e-160 RMSD = 0.72) with the frequency of the designs normalized to their initial frequency (fitness), but only above a protein level of 14. Fitness residual is defined as the distance on the Y axis between any point and the linear model. The blue points are the positive fitness residual group, whereas the red points are the negative fitness residual group. The grey points are all points below the expression threshold that do not show correlation with fitness and are not included in the linear model or the fitness residual groups.

5.5 Ribosome attenuation by multiple mechanisms at 5' of coding sequence is advantageous to the cell

By quantifying the effect of protein expression on fitness, we could now characterize mechanisms that allow designs to minimize expression cost. To achieve this goal, we first performed a linear regression between fitness and expression for all designs with a protein level between 14 and 17.5 (see methods). Then, we defined the "fitness residual" of a design as the difference between the expected fitness, as predicted by this linear regression, to the observed fitness in the lab-evolution experiment. Then, we split the designs into two groups of either positive (designs that are doing better than expected, blue points in Figure 9) or negative (designs doing worse than expected, red points in Figure 9). Since the observed fitness

residual is prone to noise due to drift and sampling errors, we only classified designs whose fitness residual sign was identical in five or six lineages. We then compared various parameters between the two fitness residual groups, and found three advantageous mechanisms that are associated with positive fitness residual designs.

Strong secondary structure at the 5' of the coding sequence is associated with higher than expected fitness

Goodman *et al.* showed that loose mRNA secondary structures are correlated with high GFP expression levels. Interestingly, our work revealed that a desired expression level could be reached with a smaller translational cost if the transcript demonstrates a strong secondary structure at its 5’ end. Figure 10A shows the GC-content distributions of the positive and negative fitness residual groups while figure 10B shows the distributions of the simulated free energy of the mRNA secondary structure. Evidently, positive fitness residual designs demonstrate higher GC-content (Effect size = 7.15%, q. value = 1.67E-22, Wilcoxon rank sum) and stronger secondary structures (Effect size = 6.73, q. value = 2.65E-13, Wilcoxon rank sum) compared to the negative fitness residual designs. Importantly, the position of strong secondary structured areas in the 5' region of the transcript is of relevance to the fitness residual sign. Specifically, when the ΔG is calculated for the entire variable region of the transcript (UTR+variable region) there is no significant difference between the positive and negative fitness residual groups (Fig 10C, Effect size = 2.59%, q. value = 1.34E-03, Wilcoxon rank sum). This observation suggests that positive fitness residual is specifically associated with strong secondary structures immediately after the start codon at the 5' of the ORF and not with up-stream regulatory elements such as the RBS. This conclusion further supports the need to attenuate ribosome at the early phase of elongation to increase the fitness residual.

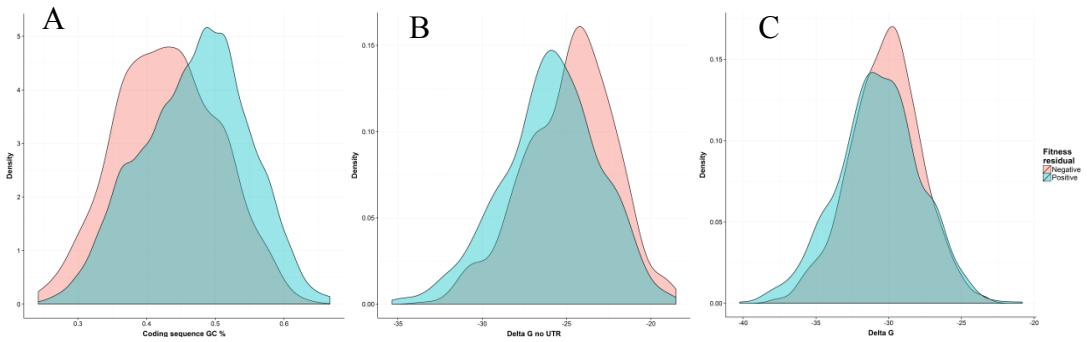


Figure 10 – Comparison of ΔG and GC content between positive and negative fitness residual groups
Distributions of negative (pink) and positive (blue) fitness residuals. A) GC-content of the coding sequence, positive higher than negative (Effect size = 7.15%, q. value = 1.67E-22). B) Folding energy of coding sequence, positive higher than negative (Effect size = 6.73%, q. value = 2.65E-13). C) Folding energy of coding sequence and UTR, insignificant (Effect size = 2.59%, q. value = 1.34E-03).

A deep ribosome-flow bottleneck at the beginning of translation elongation is associated with positive fitness residual

The adaptation of a transcript to the cellular tRNA pool has been widely studied in recent years, and demonstrated as a regulatory mechanism for elongation. Specifically, a ramp model of slowly translated codons at the 5' of the transcript has been suggested to support translation of genes (see introduction). Thus, we compared the tAI and cAI indexes, being a proxy for translation speed of a given transcript, of the positive and negative fitness residual groups. We hypothesized that small tAI or cAI scores, which mean stalled ribosome, should associate with positive fitness residual. Yet, our data showed no significant difference between the two groups (Fig. 11A+B, cAI: Effect size = 4.03%, q. value = 5.68E-05, tAI: Effect size = 3.83%, q. value = 3.84E-05, Wilcoxon rank sum).

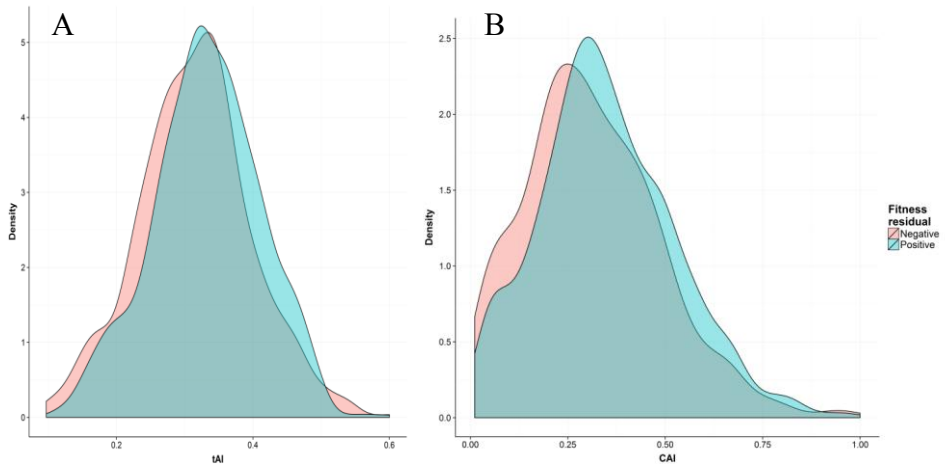


Figure 11 – tAI and CAI show no association with fitness residual
Distributions of both tAI (A Effect size = 3.83%, q. value = 3.84E-05) and cAI (B, Effect size = 4.03%, q. value = 5.68E-05) do not demonstrate significant difference between positive or negative fitness residual design.

The tAI and cAI indexes provide a crude estimation of translation rate for an entire sequence. This characteristic may obscure delicate differences regarding ribosomal flow among different sequences. Thus, to examine more subtle effects of codon adaptation on ribosomal flow and link them to translational dynamics, we collaborated with the Tuller lab at Tel Aviv University to simulate the ribosomal density profile of each design. To this end, we used the TASEP model that is based on both simulated initiation rates and MTDR translation rates, a measure of codon translation time derived empirically from ribosome profiling data (see methods). By analyzing the simulation data at steady-state we obtained the ribosomal bottleneck for each design, being the position with the highest ribosomal density on the transcript. The distribution of bottleneck positions shows that deeper bottlenecks are associated with positive fitness residual designs (Fig 12, Effect size = 9.72% q. value = 1.42E-13, Wilcoxon rank sum). This observation serves as an additional support that an early ribosomal attenuation reduces translational costs at a given expression level and further shows that the speed at which codons are translated could be utilized by cells as a mechanism of translation elongation cost regulation. Interestingly, splitting the designs into the three RBS types shows a modest, yet potentially significant difference: Strong RBS shows the most statistically significant difference between the distributions of bottleneck depth of the negative and positive fitness residuals (Effect size of 5.99%, p-value = 7.04e-04) compared the Mid

(Effect size = 1.94%, p value = 1.39e-01) and Weak RBSs (Effect size = 6.82%, p value = 2.32e-02). This is consistent with the ramp theory which predicts that bottleneck is beneficial only at high initiation rate (data not shown).

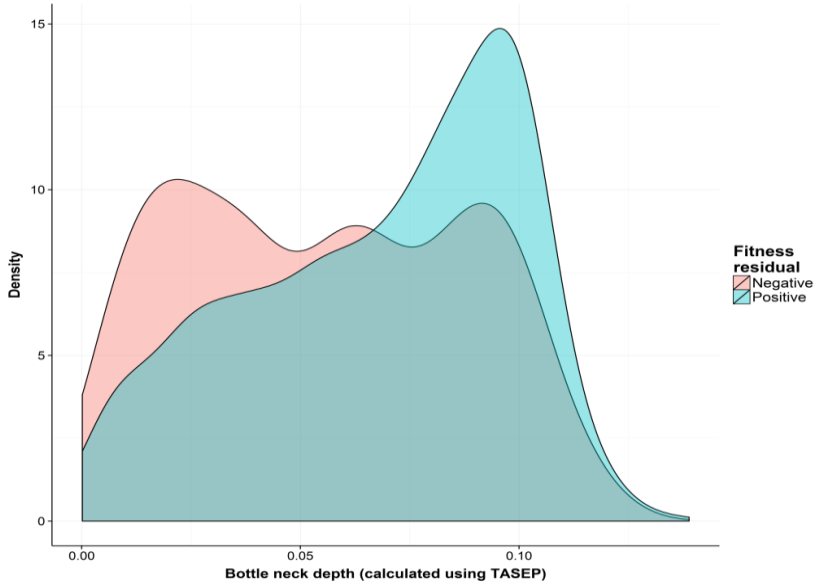


Figure 12 –Simulated ribosomal bottleneck depth is associated with fitness residual
A) Positive (blue) fitness residual designs demonstrate deeper bottlenecks (Effect size = 9.72% q. value = 1.42E-13).

Affinity to anti Shine-Dalgarno motif in early elongation leads to better fitness

A recently discovered mechanism for ribosome attenuation suggests that the affinity of sequences to the ribosome anti Shine-Dalgarno motif leads to ribosome pausing[59]. We calculated the Shine-Dalgarno affinities along the sequence of each design (see methods) and found that strong affinity early in the gene coincides with positive fitness residual designs. This observation is based on the number of positions along the sequence with a non-negligible affinity to the ribosome anti Shine-Dalgarno sequence, which was found to be higher for positive fitness residual designs compared to the negative group (Fig. 13, Effect size = 7.41% q. value = 1.36E-12). Thus, we suggest that Shine-Dalgarno motif associated with ribosomal attenuation is an additional mechanism in the cellular toolbox to maximize fitness.

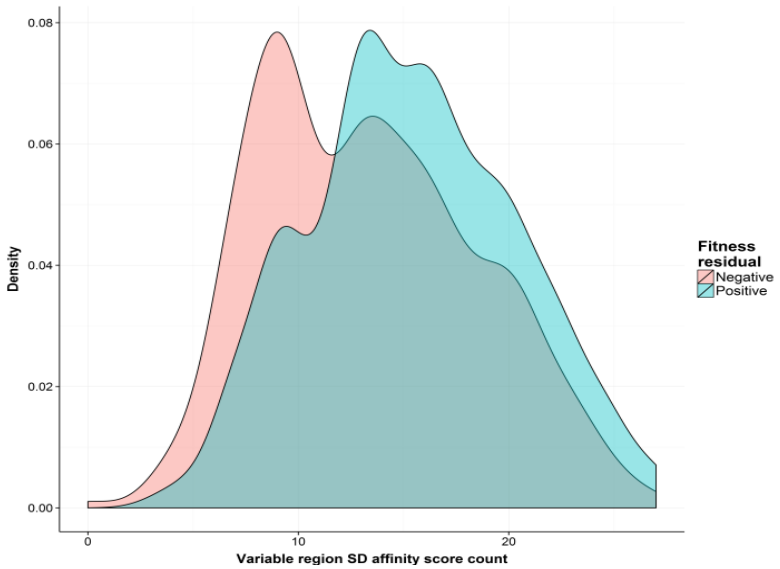


Figure 13 – High affinity to anti Shine Dalgarno site is associated with positive fitness residual
High Shine-Dalgarno affinity is associated with positive (blue) fitness residual designs as seen by counting the number of positions with a non-negligible affinity to the anti Shine-Dalgarno motif (Effect size = 7.41% q. value = 1.36E-12)

5.6 Amino acid properties affect fitness residual

The building blocks of all proteins are amino acids, making them the most consumed resource in translation. Each amino acid has its own unique chemical properties and may interact differently with the translation machinery. We thus hypothesized that the amino acid composition of the nascent polypeptide may affect translational cost. Indeed, we found that different amino acid parameters differ between the positive and negative fitness residual groups. Interestingly, positive fitness residual designs tend to have a more hydrophilic peptide (Fig. 14A, Effect size = 13.50%, q. value = 2.76E-34). Consistently, a polar amino acid content was also found to associate with a positive fitness residual (Fig. 14B, Effect size = 14.30%, q. value = 4.37E-32). These findings, along with other, similar parameters such as aliphatic residue composition, Boman index, charged residue composition, and aliphatic index (See appendix A.1), suggest that hydrophilic peptides lead to a higher fitness residual. We hypothesize that these peptide parameters are associated with the tendency of the protein to form aggregates[103] (see discussion).

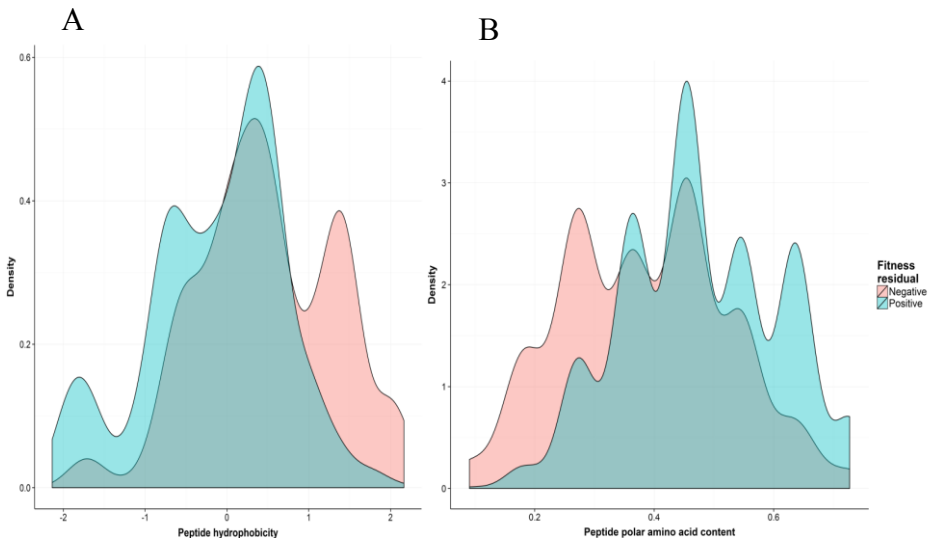


Figure 14 – Comparison of peptide properties between positive and negative fitness residual groups Hydrophilic peptides (A, p Effect size = 13.50%, q. value = 2.76E-34), and polarity (B, Effect size = 14.30%, q. value = 4.37E-32) are associated with positive fitness residual designs.

Energetic-cost of amino-acid production is correlated with fitness residual

Amino-acids do not differ only in their chemical nature, but also by the energy-rich ATP or GTP bonds that the cell consumes in their metabolic production[98]. Hence, we hypothesized that usage of energetically-expensive amino-acids may provide a heavier burden compared to cheaper ones. Indeed, lower-cost peptides were found to associate with positive fitness residual designs (fig. 15A, Effect size = 11.20%, q. value =8.90E-48). Additionally, we examined the frequency of each amino acid in the positive and negative fitness residual groups (see methods). Figure 15B demonstrates the enrichment of each amino-acid in either the positive or the negative design group. Strikingly, amino-acids with a low energetic-cost (Glu, Gln & Ala) were found to be more frequent in the positive fitness residual group, while high-cost amino acid (Phe, Ile & Tyr) are enriched in the negative fitness residual group. To reveal how translation rate and energetic-cost of amino-acid are linked, we calculated the enrichment ratio of each amino acid by dividing the amino-acid frequency in the positive group by the frequency in the negative group (see methods). Remarkably, this enrichment-ratio was found to correlate with the metabolic cost of each amino-acid (fig. 15C, Pearson correlation: 0.52, p-value: 0.018), demonstrating the significance of this parameter to determining the fitness residual of a design. Indeed, the strongest correlation between enrichment-ratio to

cost was found for designs with the strong RBS, while the mid RBS designs show a weaker correlation that disappears completely for the weak RBS with a slow initiation rate (Fig 16). These observations suggest that energetically-expensive amino-acid do not only burden cells during their costly production but also while they are being utilized by the ribosome during the translation process, presumably due to a feedback that increases their synthesis upon consumption in translation

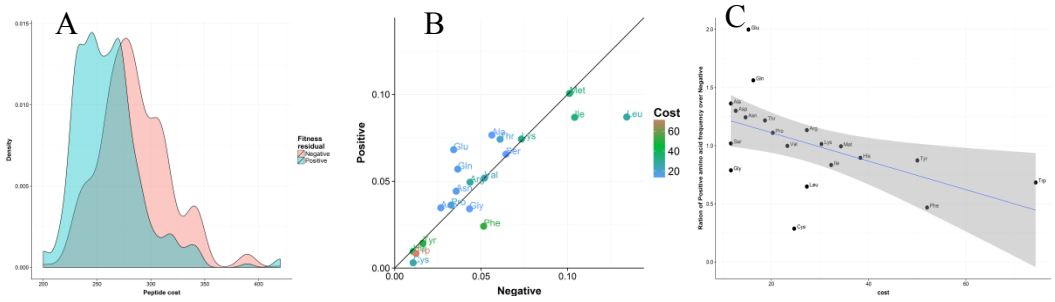


Figure 15 – Energetic cost of amino acid is correlated with fitness residual
A) Energetically-cheap peptides are associated with positive (blue) fitness residual designs (fig. 15A, Effect size =11.20%, q. value =8.90E-48). B) High cost amino acids (Phe, Ile, Tyr) are more abundant amongst negative fitness residual designs and cheap amino acids (Glu, Gln, Ala) are more frequent in positive fitness residual designs. C) The ratio between amino acid ratio and energetic cost is correlated (Pearson correlation: -0.52, p-value: 0.018).

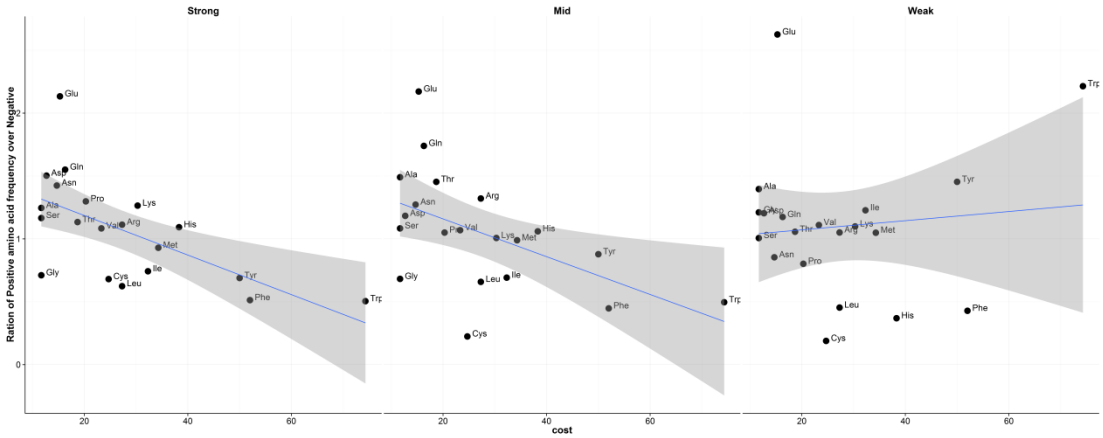


Figure 16 – Correlation between amino acid ratio and cost is more significant for strong RBS
The correlation between amino acid ratio and energetic cost of amino acid is strongest for designs with the strong RBS (Pearson correlation: -0.63, p-value: 0.003), weaker for the mid RBS (correlation: -0.53, p-value: 0.015), and is non-significant for the weak amino acid (correlation: 0.1, p-value: 0.66).

5.7 Secondary structure, ribosomal flow, Shine-Dalgarno affinity, hydrophility and amino-acid energetic-cost contribute to fitness residual independently

In this work, we revealed different mechanisms that minimize translational cost and increase the fitness of the cell. Although these mechanisms are different by nature they may be inter-connected, namely that designs that score high on one of these parameters may tend to score highly on others. For example, Shine-Dalgarno affinity could correlate with secondary structure as both parameters are influenced by GC-content. To check this possibility, we computed the correlation among the different parameters (fig. 17). Reassuringly, no strong, significant correlation was found between any two parameters either when correlating all designs in the library (fig. 17A) or only the designs with a consistent fitness residual sign (fig. 17B). This analysis strengthens our observations for each of the mentioned mechanisms and their independent contributions to fitness.

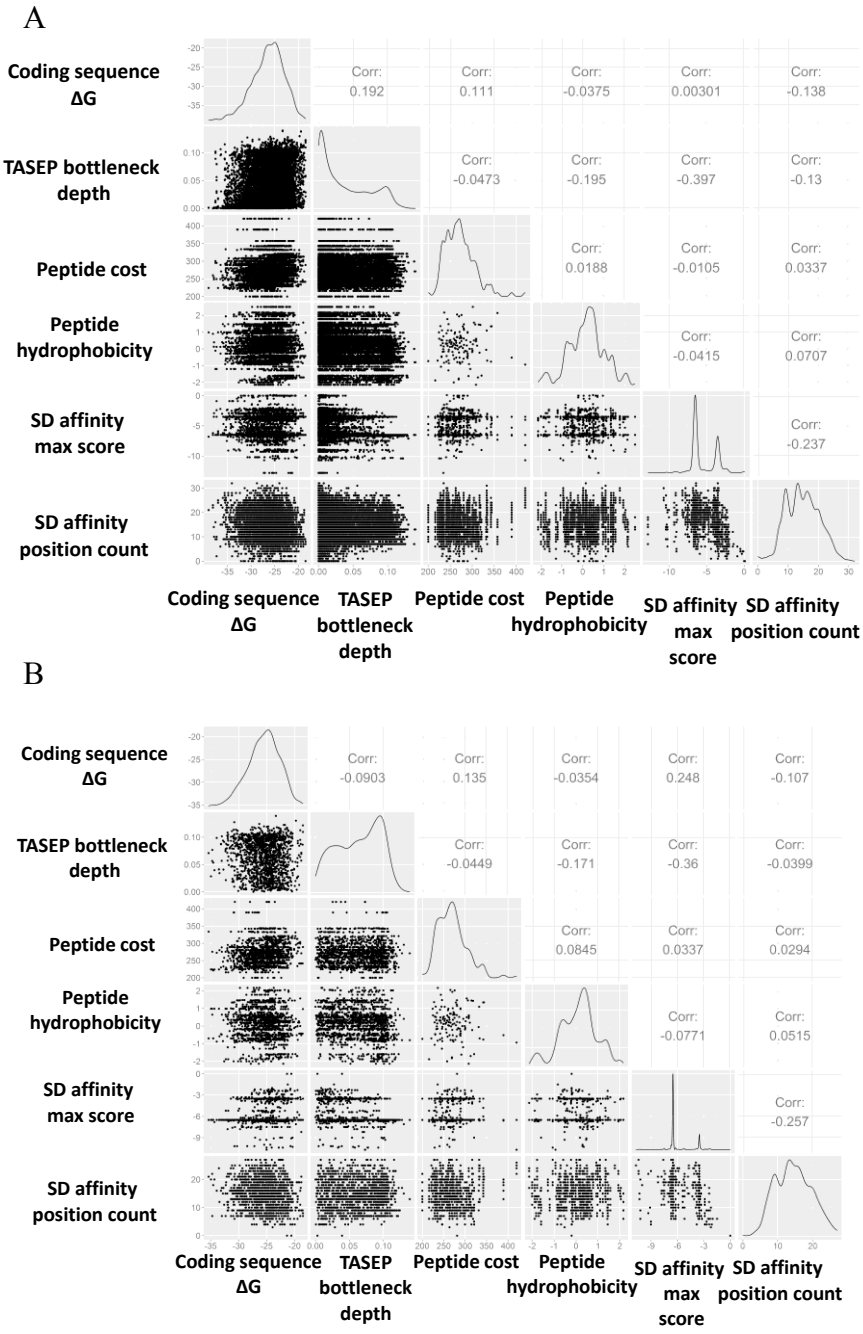


Figure 17 – Comparison of sequence parameters between positive and negative fitness residual group

No correlation exists among secondary structure, bottleneck strength, peptide cost, peptide hydrophobicity, and Shine-Dalgarno affinity, either when considering all designs of the library (A) or only designs with identical fitness residuals in 5 or 6 lineages (B).

6 | Discussion

Translation is a cardinal, basic, and a wide spread cellular process. Hence, its efficiency is of high importance to the cell. One way to rationalize efficiency is defined as the benefit-cost ratio[104]. The cost of expressing a protein can affect the fitness of a cell by wasting valuable resources and burdening the production of other essential proteins[81]. This possibility may lead to a lower fitness of a given cell and thus its extinction from the population. Thus, selection may act to minimize the cost of translation via diverse molecular mechanisms that are only partly elucidated. Although translational efficiency has been highly studied in the past, most works have been dedicated to researching regulatory mechanisms that maximize protein expression[66,72,73,75–77,105]. In this work, we focused on translational cost, cellular fitness and the molecular mechanisms that link the two.

To this end, we developed a method, termed FitSeq, to infer the cost of protein expression at various levels. Our findings suggest that there is a significant correlation between protein expression and fitness, although this correlation is not as simple as we predicted. Interestingly, the correlation is apparent only above a certain threshold of protein expression. We propose that this lack of correlation below the threshold is the result of other cellular processes, which alter translational cost, such as variance in gene expression, that mask the cost of expression. If any of these factors has a larger impact on the cell than the cost of expressing the GFP, then it would be difficult for us to detect the cost above the random noise.

We then defined the "fitness residual" of each design as the difference between the fitness predicted by the correlation with expression, and the actual fitness observed. We split the designs into two groups, those with positive fitness residuals which were doing better than expected, and those with negative fitness residuals, which were doing worse. Interestingly, many designs had very noisy fitness residuals, demonstrating both positive and negative signs across the independent lineages of our experiment. Only hundreds designs showed positive or negative fitness residual consistently. This observation could be due to high impact of stochastic events throughout the experiment on fitness residual. To take this noise into account, we only considered designs with fitness residual signs that were identical in 5 or 6 lineages, thus strengthening the significance of our observations.

Our lab previously hypothesized that a translational ramp in the early elongation region of the transcript, generated by codons that are translated by tRNA with low cellular concentrations, would lead to high translational efficiency[24,74]. Conversely, other works have shown that loose mRNA secondary structure is correlated with high expressions levels[76,77]. Remarkably, our work suggests that several independent mechanisms can govern this translational ramp and increase cellular fitness. Indeed, we revealed that three parameters, which lead to slow translation speed, are associated with positive fitness residuals: strong secondary structures, high occurrence of significant Shine-Dalgarno affinities, and low concentrations of the corresponding tRNA. From these observations we conclude that ribosome attenuation in the early phase of translation elongation is advantageous to cellular fitness, regardless of the stalling mechanism.

In addition to nucleotide parameters that were revealed to affect the translational cost, we next examined peptide parameters. Notably, we found that peptide hydropholicity is associated with positive fitness residuals. Since amino acid hydrophobicity is highly correlated with measures of protein aggregation propensity[103], we speculate that the negative fitness residuals observed among the more hydrophobic peptides might represent a tendency of the gene product to aggregate. If true, this hypothesis adds an additional component of cost, protein toxicity.

Additionally, we witnessed a correlation between positive fitness residuals and amino acid energetic-cost. Previous works found genomic evidence that highly expressed genes demonstrate higher frequencies of low-cost amino-acids[98,106]. The authors suggest that this observation is the outcome of metabolic cost on translation and natural selection. This idea was further substantiated by a study fusing the cellular transcriptome, ribosome profile data and flux balance analyses[21]. The work of Hu *et al.* showed that amino-acid metabolism positively correlates with translation efficiency and ribosome density. Thus, the consumption of raw materials by translation results in production of these materials to compensate for their usage. However, in measurements performed by Stoebel *et al.* no cost for amino acid consumption was found, and all cost was attributed to transcription and translation[20]. Notably, our results confirm the theoretical possibility that selection

acts on the amino-acid sequence to minimize its metabolic cost. Although the chemical nature of the amino acids probably affects their usage by cells, there are evolutionary scenarios in which evolution could influence the choice of amino acid on the basis of frugal metabolic cost. Examples for this could be choosing a the cheaper of the three basic amino acids when a positive charge is needed, or preferring to choose non polar amino acids of low cost in transmembrane domains.

In conclusion, the goal of my thesis was to examine the mechanisms by which the cell minimizes translational cost. We developed a method to derive the relative fitness of a synthetic library, relaying on lab-evolution and high-throughput sequencing. This approach enabled us to reveal the link between protein cost and fitness. We were then able to ascertain fitness parameters that allow designs to perform better than expected according to their expression levels. In the future, our system may be applied to study more sequence parameters and how they affect translational cost, or even the cost of other stages in expression such as transcription, degradation, or even splicing. By combining forward engineering of sequences with controlled lab-evolution environments, we could discover more mechanisms that minimize expression cost in the cell.

7 | Future plans

In light of our findings, we have several future plans we wish to pursue. First, we aim to understand the topic of cost in different environments, as these environments will surely change cellular priorities and translation regulation in turn. We plan to perform additional FitSeq experiments with the library we utilized in this work under different conditions, such as amino acid starvation or temperature induced stress. Such conditions are expected to expose other mechanisms that affect translational cost because starvation is likely to change translation speed due to depletion of charged tRNA and stress will lead to vast changes of the proteome and thus translational speed will probably be one of the regulatory mechanisms of this change.

Since the library created by *Goodman et al.* only allowed us to explore the 5' of the transcript, we would aim to construct additional synthetic libraries that would allow us to explore sequence parameters on other areas of the transcript such as translation termination and mRNA degradation. Furthermore, this approach will enable us to examine the effect of ribosomal pausing in all regions of the transcript, by modifying areas mid translation. These studies could shed more light on the subject of translational cost.

Separately, we wish to validate our findings in the genomic context of prokaryotes. Hence, we plan to examine the presence of the mechanisms we found to halt ribosomal flow at the 5' of the transcript in natural genes. Additionally, we would like to explore the proteomic trends we found in this work. Our results suggest that hydrophylic residues close to the start codon are beneficial, possibly due lower aggregation tendencies, and we wish to learn if natural genes demonstrate this phenomenon. Moreover, we showed that a low amino-acid energetic-cost minimizes translational cost. Interestingly, recent studies showed that costly amino-acids appear in smaller frequencies in highly expressed genes[98,106]. We thus plan uncover additional traces of cost-based evolution. Specifically, we would expect to see non-functional regions comprised mostly by cheap amino acids compared to functional positions of proteins.

Finally this study has interesting biotechnological implications, in the field of heterologous gene expression. In this field people express foreign genes within cells and hope to maximize production. While most current efforts are geared towards maximization of protein produced, the well-being of the host cell is likely to be very crucial too. The set of fitness-residual minimizing properties suggested here may thus help in better design of heterologously expressed genes in such a way that their production within a host cell will minimize the reduction in costs. Since often the amino acid sequence in such heterologous expression systems is predetermined, it is mainly the ribosome allocation properties that would prove themselves as relevant to such expression optimization. We thus would aim to develop a fitness residual calculator to assist in the forward design of heterologous genes, which would maximize the fitness of the cell while maintaining high production values.

8| Appendix A – Parameter lists

8.1 List of nucleotide sequence parameters and statistical test results

Parameter	Description	Direction	p value	q values	Effect size
GC1 content	GC% at the 1st position of the codon	Positive > Negative	1.18E-15	2.71E-14	11.10%
GC3 content	GC% at the 3rd position of the codon	Positive > Negative	4.80E-10	8.64E-09	11.10%
Average ribosome number on mRNA	Number of ribosome sequestered to the transcript as simulated by TASEP	Positive > Negative	3.59E-16	9.33E-15	10.70%
Average ribosome density	Average density of ribosomes for entire transcript as simulated by TASEP	Positive > Negative	3.61E-16	9.33E-15	10.70%
Ribosomal bottleneck depth	Density of position with highest density in transcript as simulated by TASEP	Positive > Negative	1.71E-15	3.76E-14	9.72%
Ribosomal density at start codon	Density of starting codon as simulated by TASEP	Positive > Negative	1.29E-16	3.48E-15	8.92%
Average translation rate	Rate of protein production as simulated by TASEP	Positive > Negative	3.61E-16	9.33E-15	8.55%
Shine-Dalgarno affinity position count	Number of positions that show a non-negligible SD efficiency	Positive > Negative	6.79E-14	1.36E-12	7.41%
Coding sequence GC%	GC content of coding sequence	Positive > Negative	5.58E-24	1.67E-22	7.15%
Transcript GC%	GC content of transcript, including coding sequence and UTR	Positive > Negative	9.95E-20	2.89E-18	7.04%
Coding sequence ΔG	ΔG of mRNA secondary structure, without UTR	Positive < Negative	1.26E-14	2.65E-13	6.73%
CAI	Codon Adaptation Index	Positive > Negative	4.06E-06	5.68E-05	4.03%
tAI	tRNA Adaptation Index	Positive > Negative	2.56E-06	3.84E-05	3.83%
Shine-Dalgarno affinity standard deviation	Standard deviation of SD affinity for all non-negligible location	Positive > Negative	1.66E-05	1.83E-04	3.68%
Transcript ΔG	ΔG of transcript, including UTR	Positive < Negative	1.49E-04	1.34E-03	2.59%
RBS initiation rate	Initiation rate calculated by Ribosome Binding Site calculator	Positive > Negative	1.61E-17	4.51E-16	2.03%
Shine-Dalgarno affinity mean	Mean value of SD affinity for all non-negligible positions	Positive > Negative	8.04E-04	6.43E-03	1.36%
Shine-Dalgarno affinity median	Median value of SD affinity for all non-negligible positions	Positive > Negative	9.83E-04	6.88E-03	1.23%
GFP Shine-Dalgarno affinity median	Median value of SD affinity for all non-negligible positions in GFP	Positive < Negative	4.14E-01	4.14E-01	0.06%
GFP Shine-Dalgarno affinity mean	Mean value of SD affinity for all non-negligible positions in GFP	Positive < Negative	1.46E-01	3.89E-01	0.02%
GFP area Shine-Dalgarno affinity standard deviation	Standard deviation value of SD affinity for all non-negligible positions in GFP	Positive < Negative	5.33E-02	2.78E-01	0.01%
GC2 content	GC% at the 2nd position in the codon	Positive > Negative	5.21E-06	6.77E-05	0.01%
GFP Shine-Dalgarno maximal affinity score	Max value of SD affinity for all non negligible positions in GFP	Positive > Negative	4.64E-02	2.78E-01	0.00%
Shine-Dalgarno maximal affinity score	Max value of SD affinity for all non-negligible positions	Positive < Negative	1.13E-05	1.36E-04	0.00%
RBS Shine-Dalgarno affinity	Value of SD affinity of the Ribosome Binding Site	Positive < Negative	8.54E-09	1.45E-07	0.00%
GFP Shine-Dalgarno affinity position count	Number of positions that have non-negligible SD efficiency in GFP	Positive > Negative	1.91E-06	3.06E-05	0.00%
Ribosomal bottleneck position	Position with highest density in transcript as simulated by TASEP	Positive < Negative	1.00E-10	1.90E-09	0.00%
Shine-Dalgarno maximum affinity position	Position with maximum value of SD affinity for coding sequence	Positive < Negative	9.72E-02	3.89E-01	0.00%
GFP Shine-Dalgarno maximum affinity position	Position with maximum value of SD affinity in GFP	Positive > Negative	1.25E-01	3.89E-01	0.00%

8.2 List of peptide sequence parameters and statistical test results

Parameter	Description	Direction	p value	q value	Effect Size
Aliphatic residue composition	Percent of initiator peptide that is aliphatic	Positive < Negative	6.47E-30	3.30E-27	16.70%
Charged residue composition	Percent of initiator peptide that is charged	Positive > Negative	5.21E-18	1.87E-15	16.70%
Non polar residue composition	Percent of initiator peptide that is non polar	Positive < Negative	7.13E-35	4.37E-32	14.30%
Polar residue composition	Percent of initiator peptide that is polar	Positive > Negative	7.13E-35	4.37E-32	14.30%
Peptide Hydrophobicity	Calculated as in Kyte *et al.*[95]	Positive < Negative	3.82E-37	2.76E-34	13.50%
Peptide aliphatic index	Calculated as in Ikai *et al.*[91]	Positive < Negative	6.27E-25	2.90E-22	11.90%
Peptide Boman index	Calculated as in Boman *et al*[92].	Positive > Negative	1.92E-36	1.28E-33	11.80%
Peptide energetic cost	Calculated as in Akashi *et al.*[98]	Positive < Negative	1.14E-50	8.90E-48	11.20%
Peptide instability index	Calculated as in Guruprasad *et al.*[96]	Positive > Negative	1.18E-14	3.07E-12	7.91%
pI	Calculated as in Bjellqvist *et al.*[107]	Positive < Negative	3.10E-04	3.69E-02	2.34%
Peptide Molecular Weight	Calculated as in Wilkins *et al.*[108]	Positive < Negative	7.28E-03	4.37E-01	1.94%
Hydrophobic moment	Calculated as in Eisenberg *et al.*[94]	Positive < Negative	1.71E-01	1.00E+00	0.84%
Acidic residue composition	Percent of initiator peptide that is acidic	Positive > Negative	4.18E-23	3.60E-02	0.02%
Basic residue composition	Percent of initiator peptide that is basic	Positive > Negative	1.12E-01	1.74E-20	0.01%
Aromatic residue composition	Percent of initiator peptide that is aromatic	Positive < Negative	6.23E-17	1.00E+00	0.00%
Tiny residue composition	Percent of initiator peptide that is tiny	Positive > Negative	1.81E-03	1.85E-14	0.00%
Small residue composition	Percent of initiator peptide that is small	Positive > Negative	2.89E-10	1.54E-01	0.00%

9| Bibliography

1. Spriggs, K. A., Bushell, M. & Willis, A. E. Translational regulation of gene expression during conditions of cell stress. *Mol. Cell* **40,** 228–37 (2010).

2. Holcik, M. & Sonenberg, N. Translational control in stress and apoptosis. *Nat. Rev. Mol. Cell Biol.* **6,** 318–27 (2005).

3. Stumpf, C. R., Moreno, M. V, Olshen, A. B., Taylor, B. S. & Ruggero, D. The translational landscape of the mammalian cell cycle. *Mol. Cell* **52,** 574–82 (2013).

4. Sachs, A. B. Cell Cycle–Dependent Translation Initiation. *Cell* **101,** 243–245 (2000).

5. Gingold, H. *et al.* A dual program for translation regulation in cellular proliferation and differentiation. *Cell* **158,** 1281–92 (2014).

6. Van der Velden, A. W. & Thomas, A. A. . The role of the 5' untranslated region of an mRNA in translation regulation during development. *Int. J. Biochem. Cell Biol.* **31,** 87–106 (1999).

7. De Moor, C. H., Meijer, H. & Lissenden, S. Mechanisms of translational control by the 3' UTR in development and differentiation. *Semin. Cell Dev. Biol.* **16,** 49–58 (2005).

8. Gustafsson, C., Govindarajan, S. & Minshull, J. Codon bias and heterologous protein expression. *Trends Biotechnol.* **22,** 346–53 (2004).

9. Elena, C., Ravasi, P., Castelli, M. E., Peirú, S. & Menzella, H. G. Expression of codon optimized genes in microbial systems: current industrial applications and perspectives. *Front. Microbiol.* **5,** 21 (2014).

10. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25,** 117–24 (2007).

11. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13,** 227–32 (2012).

12. De Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. *Mol. Biosyst.* **5,** 1512–26 (2009).

13. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–42 (2011).

14. Warner, J. R. The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.* **24,** 437–440 (1999).

15. Liebermeister, W. *et al.* Visual account of protein investment in cellular functions. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 8488–93 (2014).

16. Valgepea, K., Adamberg, K., Seiman, A. & Vilu, R. Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. *Mol. Biosyst.* **9,** 2344–58 (2013).

17. Rosenow, C., Saxena, R. M., Durst, M. & Gingeras, T. R. Prokaryotic RNA preparation methods useful for high density array analysis: comparison of two approaches. *Nucleic Acids Res.* **29,** E112 (2001).

18. Stouthamer, A. H. A theoretical study on the amount of ATP required for synthesis of microbial cell material. *Antonie Van Leeuwenhoek* **39,** 545–565 (1973).

19. Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z. & Hwa, T. Interdependence of cell growth and gene expression: origins and consequences. *Science* **330,** 1099–102 (2010).

20. Stoebel, D. M., Dean, A. M. & Dykhuizen, D. E. The cost of expression of Escherichia coli lac operon proteins is in the process, not in the products. *Genetics* **178,** 1653–60 (2008).

21. Hu, X.-P., Yang, Y. & Ma, B.-G. Amino Acid Flux from Metabolic Network Benefits Protein Translation: the Role of Resource Availability. *Sci. Rep.* **5,** 11113 (2015).

22. Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467,** 929–34 (2010).

23. Farabaugh, P. J. & Björk, G. R. How translational accuracy influences reading frame maintenance. *EMBO J.* **18,** 1427–34 (1999).

24. Tuller, T. *et al.* An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141,** 344–354 (2010).

25. De Smit, M. H. & van Duin, J. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc. Natl. Acad. Sci.* **87,** 7668–7672 (1990).

26. Shine, J. & Dalgarno, L. Determinant of cistron specificity in bacterial ribosomes. *Nature* **254,** 34–8 (1975).

27. Steitz, J. A. & Jakes, K. How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **72,** 4734–8 (1975).

28. Chen, H., Bjerknes, M., Kumar, R. & Jay, E. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. *Nucleic Acids Res.* **22,** 4953–7 (1994).

29. Johnson, G., Widner, W., Xin, W. N. & Feiss, M. Interference with phage lambda development by the small subunit of the phage 21 terminase, gp1. *J. Bacteriol.* **173,** 2733–8 (1991).

30. Voges, D., Watzele, M., Nemetz, C., Wizemann, S. & Buchberger, B. Analyzing and enhancing mRNA translational efficiency in an Escherichia coli in vitro expression system. *Biochem. Biophys. Res. Commun.* **318,** 601–14 (2004).

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

32. Vellanoweth, R. L. & Rabinowitz, J. C. The influence of ribosome-binding-site elements on translational efficiency in Bacillus subtilis and Escherichia coli in vivo. *Mol. Microbiol.* **6,** 1105–1114 (1992).

33. KIMURA, M. Evolutionary Rate at the Molecular Level. *Nature* **217,** 624–626 (1968).

34. King, J. L. & Jukes, T. H. Non-Darwinian Evolution. *Science (80-.).* **164,** 788–798 (1969).

35. Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pavé, A. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* **8,** r49–r62 (1980).

36. Sharp, P. M. & Li, W.-H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24,** 28–38 (1986).

37. Hartl, D. L., Moriyama, E. N. & Sawyer, S. A. Selection intensity for codon bias. *Genetics* **138,** 227–234 (1994).

38. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* **9,** r43–74 (1981).

39. Gouy, M. & Gautier, C. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* **10,** 7055–7074 (1982).

40. Ikemura, T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2,** 13–34 (1985).

41. Bennetzen, J. L. & Hall, B. D. Codon selection in yeast. *J. Biol. Chem.* **257,** 3026–31 (1982).

42. Ikemura, T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J. Mol. Biol.* **151,** 389–409 (1981).

43. Bernardi G et al. (1985), The mosaic genome of warm-blooded vertebrates. - Xenbase Paper. at <http://www.xenbase.org/literature/article.do?method=display&articleId=29209>

44. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The next-generation sequencing revolution and its impact on genomics. *Cell* **155,** 27–38 (2013).

45. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24,** 133–41 (2008).

46. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11,** 499–507 (2014).

47. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7,** 1534–50 (2012).

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

49. Brar, G. A. *et al.* High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335,** 552–7 (2012).

50. Artieri, C. G. & Fraser, H. B. Evolution at two levels of gene expression in yeast. *Genome Res.* **24,** 411–21 (2014).

51. Albert, F. W., Muzzey, D., Weissman, J. S. & Kruglyak, L. Genetic influences on translation in yeast. *PLoS Genet.* **10,** e1004692 (2014).

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

53. Varenne, S., Buc, J., Lloubes, R. & Lazdunski, C. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J. Mol. Biol.* **180,** 549–76 (1984).

54. Zhang, G., Hubalewska, M. & Ignatova, Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Biol.* **16,** 274–80 (2009).

55. Dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32,** 5036–44 (2004).

56. Tuller, T. *et al.* Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol.* **12,** R110 (2011).

57. Dana, A. & Tuller, T. Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. *PLoS Comput. Biol.* **8,** e1002755 (2012).

58. Yang, J.-R., Chen, X. & Zhang, J. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol.* **12,** e1001910 (2014).

59. Li, G.-W., Oh, E. & Weissman, J. S. The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484,** 538–541 (2012).

60. Charneski, C. A. & Hurst, L. D. Positively charged residues are the major determinants of ribosomal velocity. *PLoS Biol.* **11,** e1001508 (2013).

61. Lu, J. & Deutsch, C. Electrostatics in the ribosomal tunnel modulate chain elongation rates. *J. Mol. Biol.* **384,** 73–86 (2008).

62. Artieri, C. G. & Fraser, H. B. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome Res.* **24,** 2011–21 (2014).

63. Peil, L. *et al.* Distinct XPPX sequence motifs induce ribosome stalling, which is rescued by the translation elongation factor EF-P. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 15265–70 (2013).

64. Chevance, F. F. V, Le Guyon, S. & Hughes, K. T. The Effects of Codon Context on In Vivo Translation Speed. *PLoS Genet.* **10,** (2014).

65. Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 3645–50 (2010).

66. Angov, E., Hillier, C. J., Kincaid, R. L. & Lyon, J. A. Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. *PLoS One* **3,** e2189 (2008).

67. Maertens, B. *et al.* Gene optimization mechanisms: a multi-gene study reveals a high success rate of full-length human proteins expressed in Escherichia coli. *Protein Sci.* **19,** 1312–26 (2010).

68. Zhou, M. *et al.* Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* **495,** 111–5 (2013).

69. Welch, M. *et al.* Design parameters to control synthetic gene expression in Escherichia coli. *PLoS One* **4,** e7002 (2009).

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

71. Dana, A. & Tuller, T. Mean of the Typical Decoding Rates : a new translation efficiency index based on the analysis of ribosome profiling data.

72. Hockenberry, A. J., Sirer, M. I., Amaral, L. A. N. & Jewett, M. C. Quantifying position-dependent codon usage bias. *Mol. Biol. Evol.* **31,** 1880–93 (2014).

73. Allert, M., Cox, J. C. & Hellinga, H. W. Multifactorial determinants of protein expression in prokaryotic open reading frames. *J. Mol. Biol.* **402,** 905–18 (2010).

74. Navon, S. & Pilpel, Y. The role of codon selection in regulation of translation efficiency deduced from synthetic libraries. *Genome Biol.* **12,** R12 (2011).

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

76. Goodman, D. B., Church, G. M. & Kosuri, S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science (80-.).* **475,** science.1241934– (2013).

77. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science (80-.).* **324,** 255–258 (2009).

78. Shah, P., Ding, Y., Niemczyk, M., Kudla, G. & Plotkin, J. B. Rate-limiting steps in yeast protein translation. *Cell* **153,** 1589–601 (2013).

79. Supek, F. & Šmuc, T. On relevance of codon usage to expression of synthetic and natural genes in Escherichia coli. *Genetics* **185,** 1129–34 (2010).

80. Gardin, J. *et al.* Measurement of average decoding rates of the 61 sense codons in vivo. *Elife* **3,** e03735 (2014).

81. Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436,** 588–92 (2005).

82. Ceroni, F., Algar, R., Stan, G.-B. & Ellis, T. Quantifying cellular capacity identifies gene expression designs with reduced burden. *Nat. Methods* **12,** 415–418 (2015).

83. LeProust, E. M. *et al.* Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38,** 2522–40 (2010).

84. Blecher-Gonen, R. *et al.* High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. *Nat. Protoc.* **8,** 539–54 (2013).

85. St. John, J. SeqPrep. at <https://github.com/jstjohn/SeqPrep>

86. Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 14024–9 (2013).

87. Salis, H. M., Mirsky, E. a & Voigt, C. a. Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* **27,** 946–950 (2009).

88. Espah Borujeni, A., Channarasappa, A. S. & Salis, H. M. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res.* **42,** 2646–2659 (2014).

89. Tracy, C. A. & Widom, H. Asymptotics in ASEP with Step Initial Condition. *Commun. Math. Phys.* **290,** 129–154 (2009).

90. Daniel Osorio, P. R.-V. and R. T. Peptides: Calculate Indices and Theoretical Properties of Protein Sequences. (2015). at <http://cran.r-project.org/package=Peptides>

91. Ikai, A. Thermostability and aliphatic index of globular proteins. *J. Biochem.* **88,** 1895–1898 (1980).

92. Boman, H. G. Antibacterial peptides: basic facts and emerging concepts. *J. Intern. Med.* **254,** 197–215 (2003).

93. Moore, D. S. Amino acid and peptide net charges: A simple calculational procedure. *Biochem. Educ.* **13,** 10–11 (1985).

94. Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. U. S. A.* **81,** 140–4 (1984).

95. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–32 (1982).

96. Guruprasad, K., Reddy, B. V & Pandit, M. W. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.* **4,** 155–61 (1990).

97. Lobry, D. C. and J. R. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. *Biological and Medical Physics, Biomedical Engineering* 207–232 (2007). at <https://cran.r-project.org/web/packages/seqinr/index.html>

98. Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 3695–700 (2002).

99. Individual Comparisons by Ranking Methods on JSTOR. at <http://www.jstor.org/stable/3001968?seq=1#page_scan_tab_contents>

100. A Simple Sequentially Rejective Multiple Test Procedure on JSTOR. at <http://www.jstor.org/stable/4615733>

101. Hodges, J. L. & Lehmann, E. L. Estimates of Location Based on Rank Tests. *Ann. Math. Stat.* **34,** 598–611 (1963).

102. Ceriani, L. & Verme, P. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *J. Econ. Inequal.* **10,** 421–443 (2011).

103. Ahmed, A. B. & Kajava, A. V. Breaking the amyloidogenicity code: methods to predict amyloids from amino acid sequence. *FEBS Lett.* **587,** 1089–95 (2013).

104. Kalisky, T., Dekel, E. & Alon, U. Cost-benefit theory and optimal design of gene regulation functions. *Phys. Biol.* **4,** 229–45 (2007).

105. Welch, M., Villalobos, A., Gustafsson, C. & Minshull, J. *Designing genes for successful protein expression*. *Methods in Enzymology* **498,** (Elsevier Inc., 2011).

106. Raiford, D. W. *et al.* Metabolic and translational efficiency in microbial organisms. *J. Mol. Evol.* **74,** 206–16 (2012).

107. Bjellqvist, B. *et al.* The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* **14,** 1023–31 (1993).

108. Wilkins, M. R. *et al.* Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.* **112,** 531–52 (1999).