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
Omer Asraf

מאת
עומר אסרף

הרעבה לחומצות אמינו משפיעה על העלות של ביטוי חלבונים בחיידק אי קולי.

Starvation to amino acids modifies the cost of gene expression in *E. coli*.

Advisor:
Prof. Yitzhak Pilpel

מנחה:
פרופ' יצחק פלפל

June 10, 2017

ט"ז בסיון, תשע"ז

Contents

Abstract

Proteins expression is mandatory as they take part in all functions of the cell. As such, a major portion of cellular resources is devoted to protein synthesis, and optimizing the cost of expression is crucial for the organism. Studies have explored the outcomes of amino acids starvation on the ability to express proteins; however, what are the implications of amino acids starvation on the economy of gene expression remains unclear. In the presented study we sought to answer that question by introducing a synthetic GFP library to *E. coli* and examining the fitness of the different variants under phenylalanine and leucine starvation. Furthermore, we asked how would evolutionary solutions to such starvations will alter the basal cost of expression when the cells are starved or in a nutrient-rich environment. Our results show that contrary to expected, starvation to an amino acid would provide fitness advantage to cells hosting variants that encode the depleted nutrient; moreover, evolution under starvation to an amino acid can benefit with cells that harbor variants encoding an amino acid that is metabolically-related to the depleted amino acid. Two theories were proposed to explain the findings: one from a supply-and-demand point of view, and the other by consideration of translation efficiency. We examined the mutation dynamics and the most prominent observation was the contribution of mutations at conserved sites of the promoter to the variants' fitness. Our data has also revealed that variants with higher translation initiation rate had higher mRNA levels, a relation that is likely to be mediated by degradation prevention.

Introduction

Protein expression is an essential for all life forms, as proteins are involved in every cellular process. Unsurprisingly, a larger portion of available cellular resources is devoted to protein expression[1,2]; therefore, efficient and cost-effective expression of proteins is a significant selective force during evolution. Whereas highly-expressed genes were optimized for the use of common, quickly-translated codons, genes that have lower expression levels are characterized by a usage of rare codons[3] due to smaller selective force. A recent study[4] conducted in our lab (by Idan Frumkin, Dvir Schirman and Aviv Rotman) sought to discover the genetic features that optimize the protein expression process in *Escherichia coli* by utilizing a synthetic GFP library. The library contained 14,234 variants that had either strong or low promoter, the translation was driven by either high, medium or weak ribosome binding site (RBS), and the start codon was followed by a peptide sequence composed of 10 amino acids that were taken from a pool of 137 native, highly-expressed *E. coli* genes that were each recoded in 13 different nucleotide sequences. A pooled competition was conducted among the library's variants, and the fitness of each variant was evaluated. Since the GFP does not provide any benefit to the cell, the variants' fitness reflected the cost associated with expression. By examining which genetic features are associated with high or low expression-adjusted fitness, fundamentals of cost-effective gene expression were exposed.

Among several factors, two interesting determinants of the fitness of each variant were the amino acid and nucleotide composition of their library variant. Purines enrichment in the coding sequence was associated with high expression-adjusted fitness variants. This finding is surprising, since the synthesis cost of purines is higher than that of pyrimidines; however, purines have been found to encode cheaper amino acids[5], suggesting that a major component of cost is in amino acid synthesis.

With respect to features of the peptide sequence, cost-effective variants were found to mainly consist of metabolically-cheap amino acids, and hydrophobic amino acids led to an increase in expression cost, possibly due to protein aggregation which exerts toxicity to the cell. Here I looked for ways to further verify the relations between amino acid synthesis cost and the representation of amino acids among the library variants and their fitness.

At times of stress, such as starvation, the integrity of protein expression is compromised. Normally, 0.8 of the isoacceptor tRNA molecules of an amino acids are charged[6]; at times of starvation, a differential charging of the tRNA isoacceptors is achieved because of differences in the charging kinetics of tRNA-aminoacyl-synthetases[7,8]. Furthermore, the cell's proteome is changed; genes that promote proliferation are

downregulated, and expression of stress-response genes is induced[9,10]. As mentioned earlier, rarely expressed genes (among them, stress-associated) are enriched with rare codons. Thus, the selective charging of tRNA molecules can serve as a mechanism that allows robust expression of required genes despite shortage in resources.

Nature's way of solving difficulties encountered by living beings is evolution. Specifically, at times of stress, an individual cell that would manage to evade growth-arrest and restore its capacity to reproduce is expected to take over the population. An evolution under constant starvation is expected to give rise to solutions that will relieve the stress inflicted by nutrient insufficiency and promote normal growth. It is conceivable that such solutions will alter the economy of gene expression: changing the composition of the tRNA pool or rewiring precursor allocation to favor the synthesis of an amino acid over the other are two among many constitute feasible solutions.

In the present study, we aimed at elucidating the consequences of amino acid starvation on the cost of expressing proteins with different genetic features, and how would evolution in absence of an amino acid would alter the cellular economy. To do so, we introduced the aforementioned GFP library to either WT *E. coli*, or to strains that evolved in the absence of either phenylalanine or leucine; then, we carried 84-generations pooled competitions for each strain, either on media complemented for all amino acids or a depleted media. By inferring the variants' fitness and comparing them across conditions, we sought to discover the rules that govern efficiency of protein expression at times of nutrient depletion, and how can the course of evolution change these rules.

Results

Amino acids starvation as an environmental perturbation that promotes trans-adaptation

The rationale behind using amino acid depletion as the environmental perturbation can be explained in several manners. Such perturbation is exogenous to the cells, therefore it cannot be modified by the cells over time. The consequences of amino acid starvation are relatively well studied and understood, providing us with better insights regarding the cellular components that might be involved in trans-acting adaptations and hence can allow better interpretation of the experimental results. And, since amino acid depletion affects virtually every gene in the genome by affecting the rate and accuracy of translation, such perturbation promotes trans- but not cis- acting adaption.

The candidate amino acids to be starved had to fulfill two conditions. First, in order to elicit a large-scale and complex response to the depletion, the candidate amino acids had to be relatively common among the library's variants, and to have a large variance in their level of occurrence across the different variants. Second, to exert maximal selective pressure, the candidate amino acids had to be as expensive as possible in terms of synthesis costs for the cell.

With respect to these two terms, phenylalanine (F) and leucine (L) were selected as the amino acids for depletion (fig. 1), the former due to its high cellular synthesis cost and the latter because of its variation in the library, in which some variants do not contain leucine at all, and at the other extreme, others have as many as 5 leucine residues in the variable region.

Amino acids depletion impairs growth of WT strain

We sought to assess the burden inflicted by amino acids depletion on the fitness of the WT strain (which was pre-evolved prior to starvation on the MOPS complemented medium; see next section for details). For that we have compared the growth of the WT strain in presence or absence of phenylalanine or leucine. As can be seen in figure 2, inoculation in both depleted media led to growth deficits compared to the complemented medium. Growth in absence of phenylalanine caused relatively minor decline in growth, and differences in the growth curve are seen mainly as a lower yield. In contrast, growth in the absence of leucine led to a substantially long lag phase, although the growth rate at exponential phase and the final yield resemble that of the WT on complemented media.

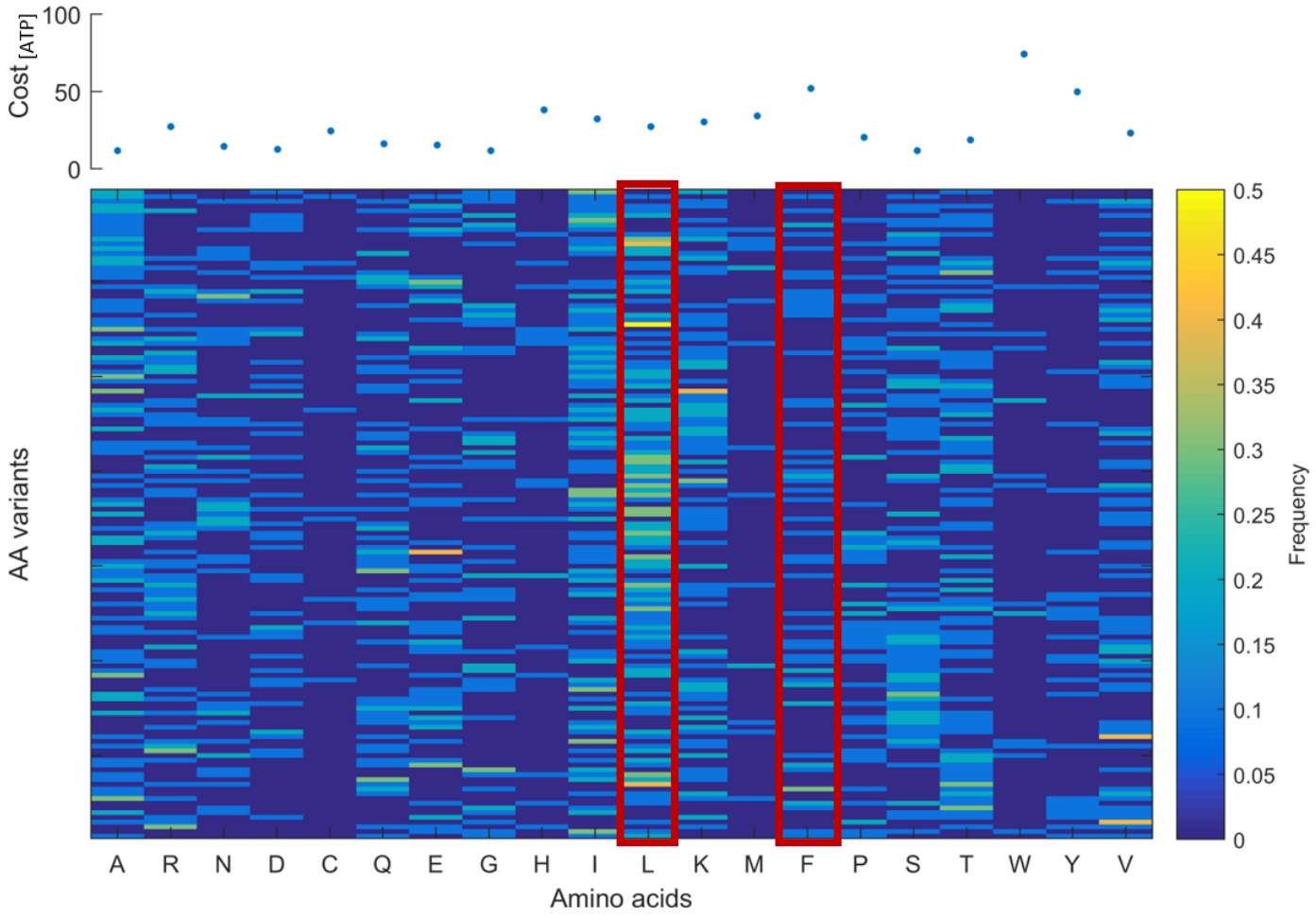


Fig. 1. Normalized distribution of amino acids in the library variants. Top panel: synthesis costs of the amino acids. Bottom panel: amino acids' frequency in each of the peptide variants in the library. Each row indicated a peptide sequence, for a total of 137 peptides represented in the library. Columns indicates amino acids, and the heatmap represents the frequency of each amino acid in each peptide. Selected amino acids, phenylalanine (F) and leucine (L), are marked in red.

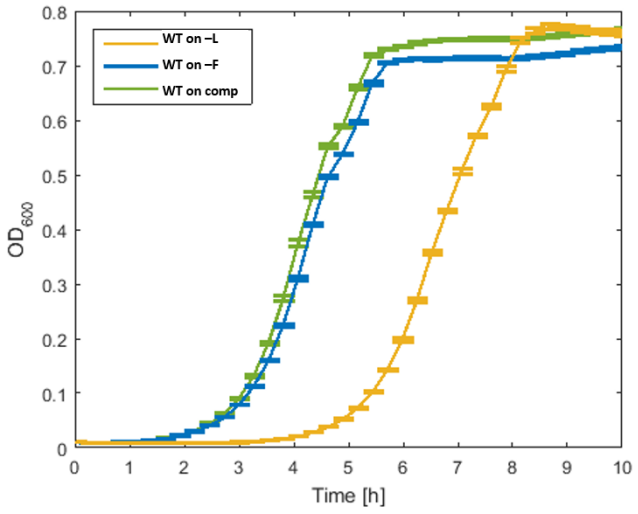


Figure 2. Amino acids depletion inflicts differential growth defect. Growth curves of the ancestral (WT) strain on complemented media, F depleted media and L depleted media (green, blue and yellow, respectively).

Experimental procedure

In the presented study, we used a synthetic GFP library composed from 14,234 variants. Each variant had either strong or weak promoter, a high, medium or low RBS, and a short peptide sequence composed from the start codon and 10 amino acids that were taken from a pool of 137 native, highly-expressed *E. coli* genes. Each peptide sequence was recoded using either WT, rare or common codons, as well as coding sequences that spanned a range of high to low mRNA folding energy; in total, each peptide sequence was recoded 13 times. In addition to the 3 constant RBS configurations, each peptide sequence was also introduced with the RBS of the gene from which it had been derived.

Since the wild type *E. coli* cells are accustomed to LB media in routine use, we wanted to reduce the possibility that variant-harboring cells will gain a fitness advantage during the competition phase by becoming better adapted to the experimental media. Thus, we began by adapting the wild-type *E. coli* to MOPS medium complemented with all amino acids by employing a lab-evolution technique in a serial dilution regime, in which cells were introduced to a fresh media, allowed to reach stationary phase by being grown for 24 hours in constant temperature, moisture and shaking conditions, and then diluted 1:120 (to a total volume of 1210 µl) to a fresh media, for a total of ~120 generations (3 weeks).

Following this evolution step, we measured the fitness of several clones originating from the evolved population. A representative clone with the highest fitness (termed WT) was evolved for ~500 generations or ~420 generations on either phenylalanine- or leucine- depleted MOPS media, respectively, by conducting a serial dilution assay as described above. The fittest clone from each of the evolutions (evo^F and evo^L for phenylalanine depletion and leucine depletion, respectively) was selected and served as an ancestor for the competition assays. The selected clones, as well as the WT clone, were modified to accept the GFP library plasmids via electroporation, and following introduction of the library, 12 days (~84 generations) of pooled competition assays were conducted on either complemented medium or the depleted medium to which the strain has evolved employing the aforementioned scheme. Sequencing libraries were constructed from samples of generation 0 and generation 84 by two PCR reactions. The libraries were deep-sequenced and following data processing, the fitness of each variant was evaluated as the binary logarithm of the ratio between generation 84 and generation 0 (fig. 3).

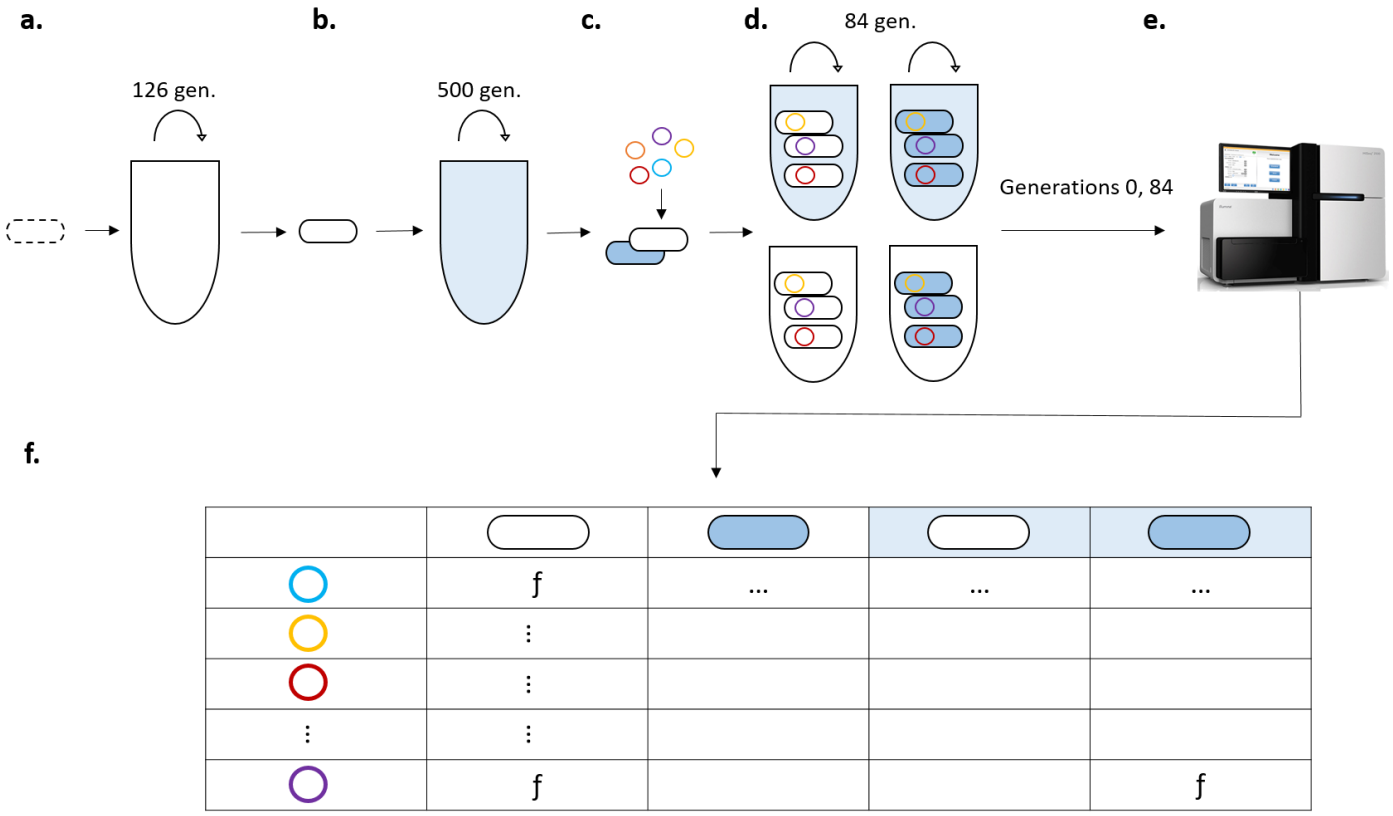


Figure 3. Schematic illustration of the experiment. The displayed scheme describes the experimental procedure per single (either phenylalanine or leucine) amino acid. A. Wild-type *E. coli* (dashed ellipsoid) was adapted to complemented media (white tube). B. Adapted strain (termed WT in the present work, solid elipsoid) was evolved in absence of an amino acid (light blue tube). C. Library plasmids (colored circles) were introduced to both evolved ($evo^{F\backslash L}$, blue elipsoid) and ancestral (WT) strains. D. 84-generations competitions were conducted in presence and absence of the amino acid. E. samples from generations 0, 84 were sequenced. F. Fitness of each variant was evaluated for each condition.

Growth deficits due to amino acids starvation are alleviated during evolution

To assess how did the evolution under starvation mended the growth impairments, we plotted the growth curves of the evolved and WT strains when grown on amino-acids depleted media. Growth curves of evo^{F} and WT strain in absence of F demonstrate reduction in the lag phase, higher growth rate and higher yield obtained during evolution (fig. 4.a). Such evolutionary modifications are observed to a larger extant with leucine starvation when evo^{L} is plotted against WT on L depleted media (fig.4.b), where all the above parameters improve dramatically. This can be explained since the average fitness of the WT strain is remarkably low when grown on leucine-depleted medium compared to a complemented medium, and therefore evolutionary solutions can (in theory, and as observed) have a significant beneficial contribution to fitness.

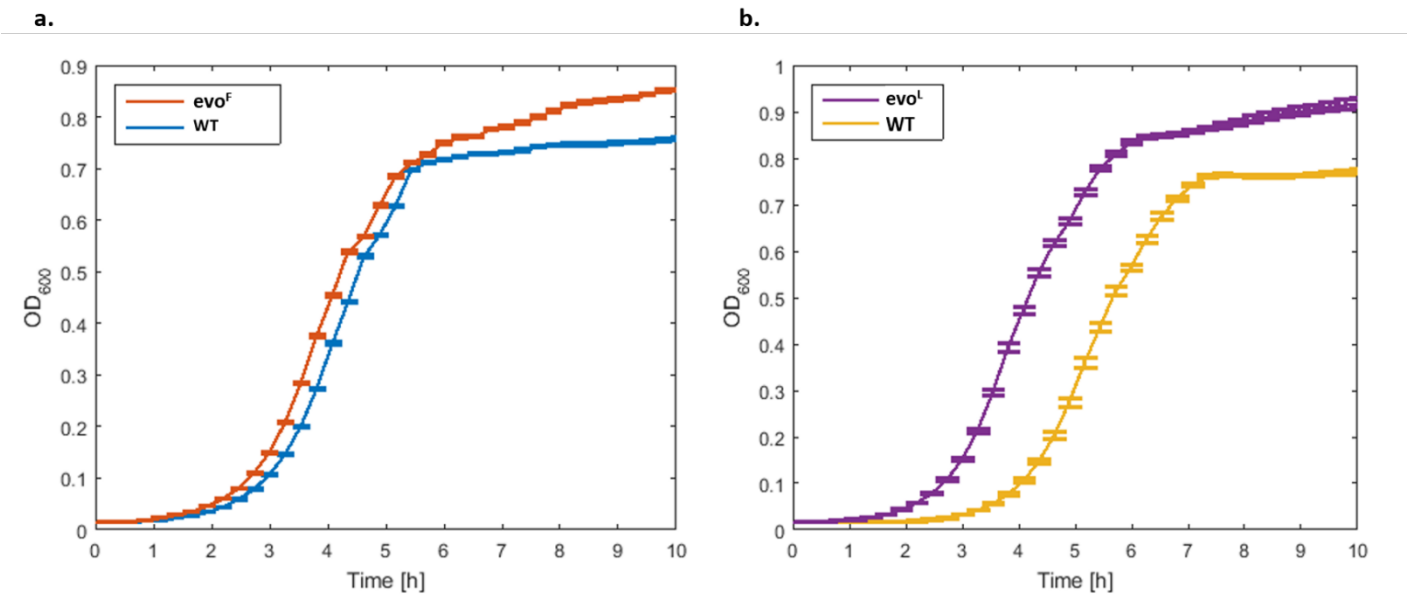


Figure 4. Amino acids depletion inflicts a growth defect that was mended during evolution. a) Growth curves of WT (blue) and evo^{F} (orange) strains on F depleted media. b) Growth curves of WT (yellow) and evo^{L} (purple) strains on L depleted media.

Library preparation

Since our ancestral culture has already evolved during synthesis of the library, there was a 2 orders of magnitude difference between the most common and the rarest variants. We therefore estimated that we need coverage of ~10^6 cells for the beginning of the competition phase. We evaluated the electroporation efficiency to be ~4-5*10^5 cells out of 1.4*10^9 cells, by plating the electroporated cells and counting the number of colonies. Therefore, for each ancestral sample (WT, evo^{F} and evo^{L}), we conducted 10 electroporation reactions and grew 2 plates per electrotoration reaction, to prevent competition on resources on the plates, to a total of 20 plates. That gave us an estimated coverage of 10^6 colonies. We grew the plates overnight, pooled the colonies to a single tube, and grew a large portion of that tube overnight. The stationary cultures served as the ancestral samples.

Following the competition, samples from generation 84 of each condition, as well as the ancestral samples, were grown overnight and library plasmids were extracted. The sequencing libraries were prepared by conducting two PCR reactions, the first has amplified the variable region and the second introduced Illumina adapters to the amplicons. Final product size was ~270 bp (fig. 5). Sequencing libraries were cleaned and sent for sequencing at the G-INCPM unit.

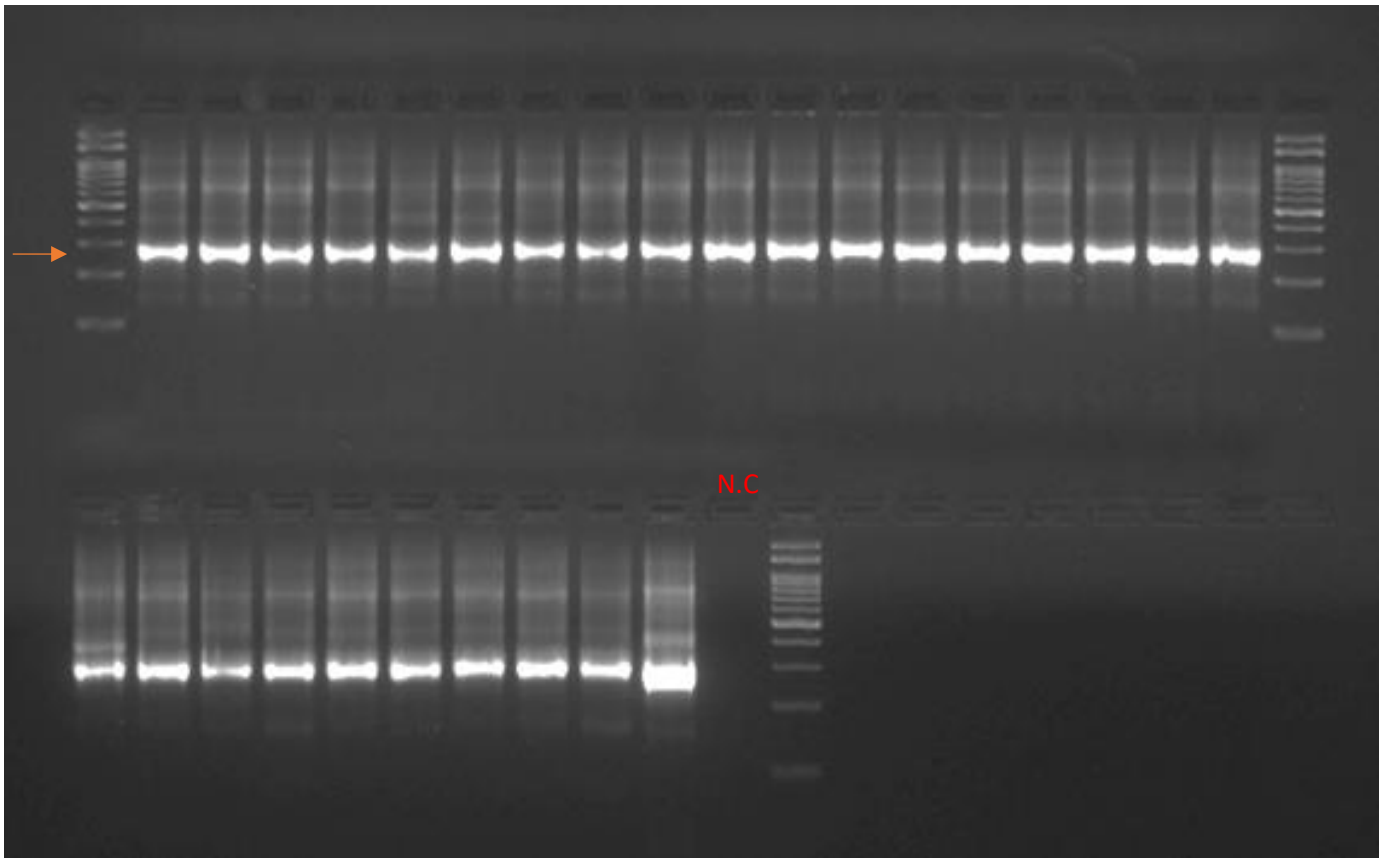


Figure 5. Final products of sequencing libraries. The samples' sequencing libraries, with product size of ~270 bp (marked with an arrow), run on a 2% agarose gel. 100bp ladders were used.

Library characterization

Fundamentals of population genetics state that the course of change in the composition of a certain population over time, the GFP library in our case, is dependent on both the fitness and initial frequency of each clone. Therefore, we started by examining the frequencies at generations 0 and 84 for each competition assay (fig. 6.a). The frequencies distributions span ~3 orders of magnitude, and are narrower for ancestral samples. Importantly, an increase in the portion of variants with low frequency is observed at generation 84 (data now shown). The fitness distributions clarify the picture (fig. 6.b); generally, the fitness distribution is centered around 0 (no change) and negatively skewed due to variants that show a significant decrease in frequency. However, there are exceptions: most variants of the WT strain on L depleted media decrease in frequency at least in one order of magnitude, and evo^{F} strain shows greater fluctuations in fitness, with many variants demonstrating a decrease in frequency.

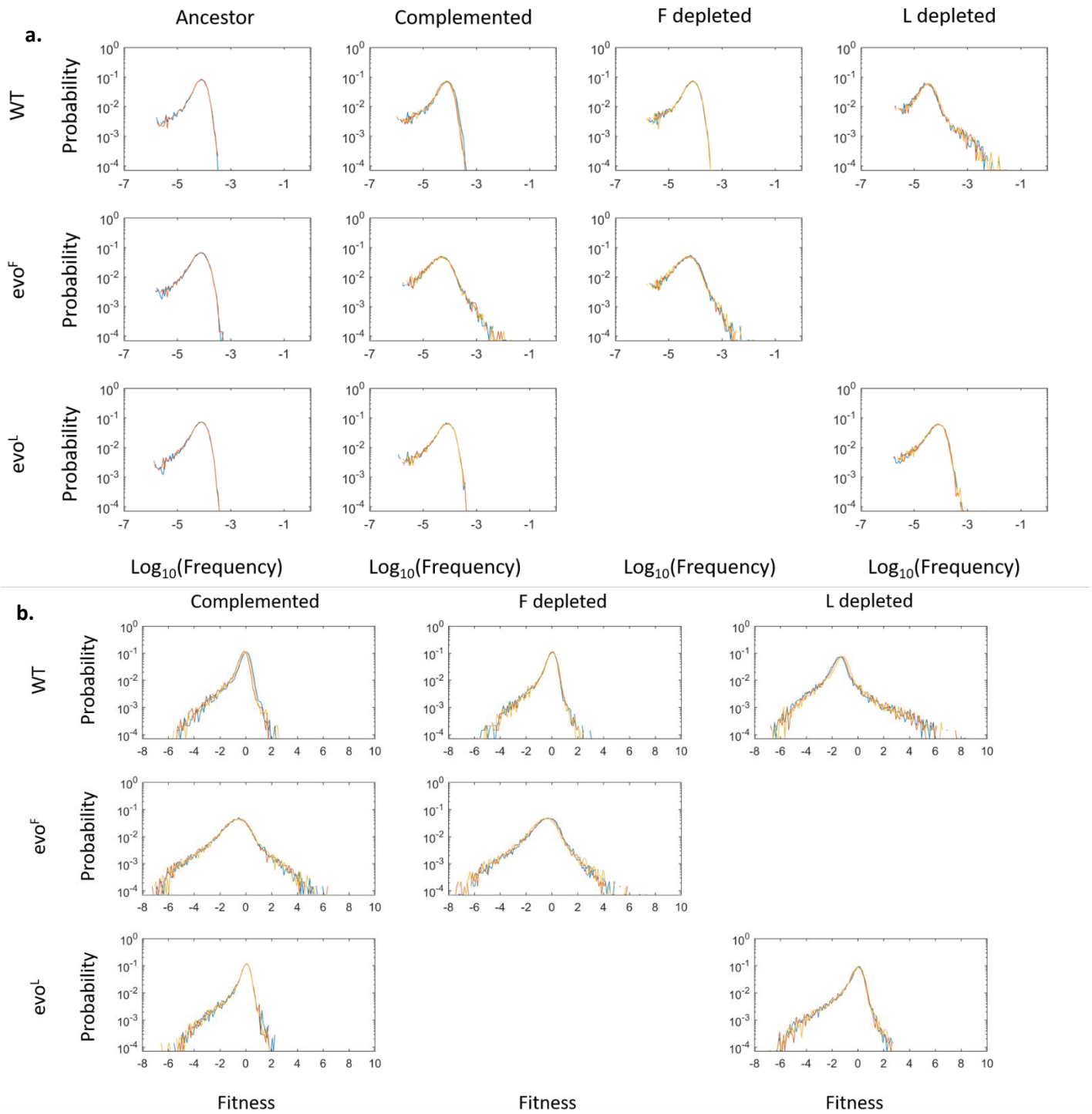


Figure 6. Distributions of frequency and fitness values. a) Frequencies' distribution at the generation 0 (first column) and generation 84 (columns 2 to 4) for the different strains and competition media. Each subplot contains 3 biological repetitions for generation 84 and 2 technical repetitions for generation 0.
b) distribution of fitness values ($\log_2$ of frequencies' ratios) across different strains and conditions.

Samples' frequency cluster together according to strain, with the exception of WT on L depleted media, and tend to be sub-clustered according to the media used for the competition (fig. 7). The correlations between the samples are high ($r > .6$ for the least similar clusters). Notably, the ancestral samples capture much of the competed samples composition. It may be due to a competition occurring during preparation of the ancestors or because of small selective pressures during the competition, leading to insignificant change in the population. As expected, the high-expression variants (identified by high promoter and strong \ medium RBS) are associated with low frequency (fig. 7). Surprisingly, variants that are designed to provide the lowest GFP expression levels (low promoter; weak and WT RBS) are demonstrating lower frequency than variants with low promoter but medium \ strong RBS. Variants' fitness cluster in the same fashion as their frequency (fig. 8), and provides support for the main findings. The WT on L depleted media form a separate clade when clustered by fitness, indicating that the samples' have gone through a different competitional (and perhaps, evolutionary) trajectory.

Both frequency and fitness correlate moderately with GFP levels (fig. 9). An important observation is that while frequency decreases monotonically with GFP level, the fitness values diverge at high expression; this may be due to measurement error caused from low sample size: low-coverage variants are expected to demonstrate higher fluctuations in read counts due to genetic drift, sampling and sequencing biases; while the effect of this noise is small in absolute frequency values, it is more dominant when taking ratios of frequencies. We should also note that GFP expression values represent measurements done for the wild type in LB medium.

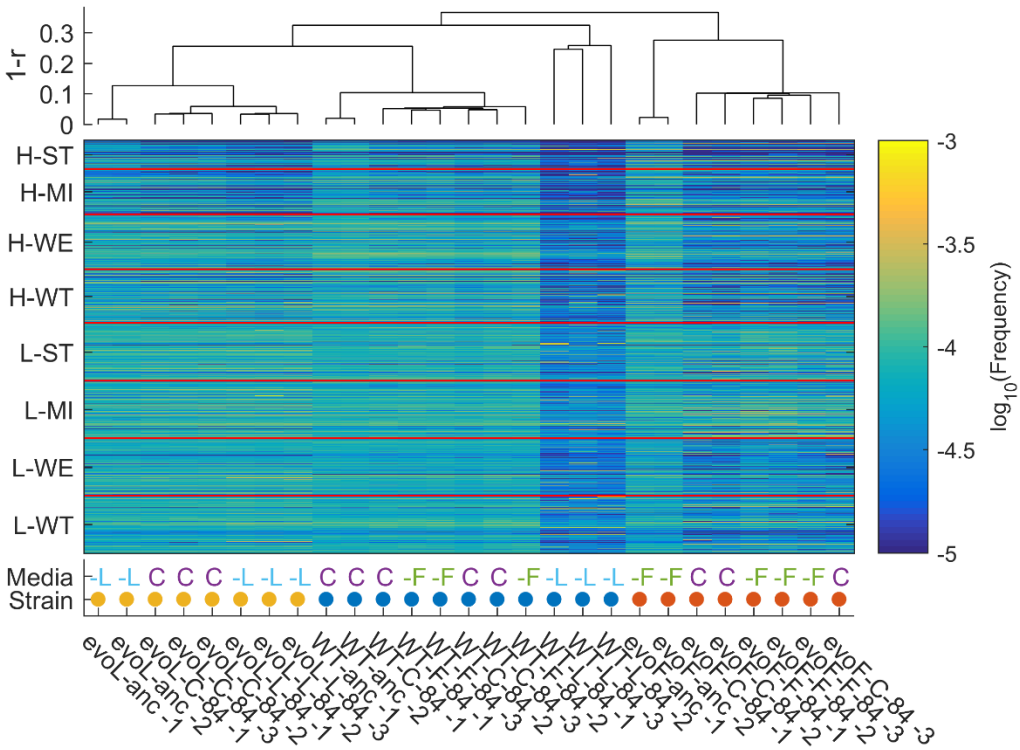


Figure 7. Samples' frequencies clusters according to strains and then sub-cluster according to media. heatmap showing varaints' frequencies. Rows are ordered by promoter-RBS composition, columns are ordered by clustering. The dendrogram shows the distance between the clusters, as 1-r (pearson's coefficient). Panel legend: strains: yellow – evo^{L}, blue – WT, orange – evo^{F}.

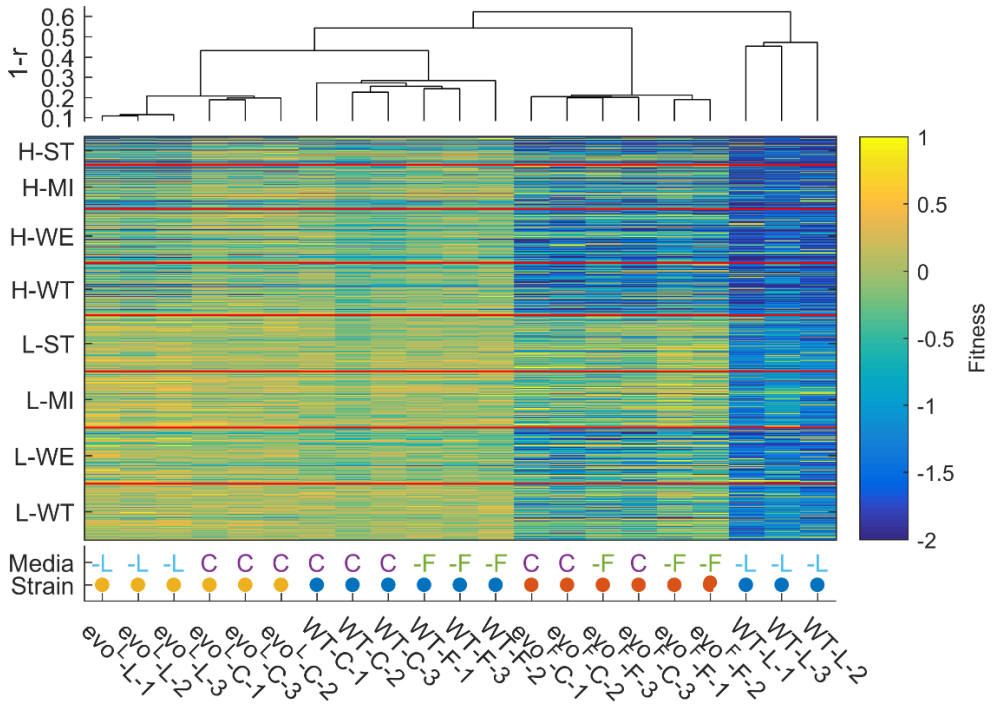


Figure 8. Samples' fitness clusters according to strains and then sub-cluster according to media. heatmap showing varaints' fitness. Rows are ordered by promoter-RBS composition, columns are ordered according to clustering order. The dendrogram shows the distance between the clusters, as 1-r (pearson's coefficient). Panel legend: strains: yellow – evo^{L},
blue – WT, orange – evo^{F}.

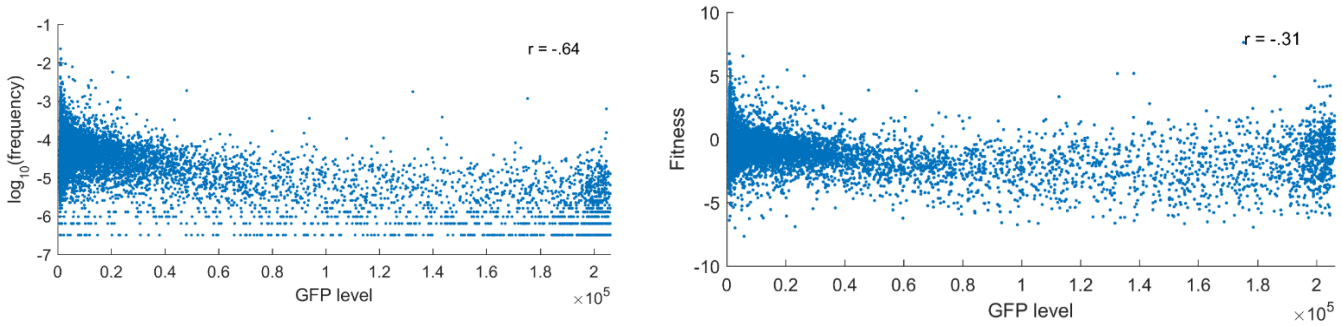
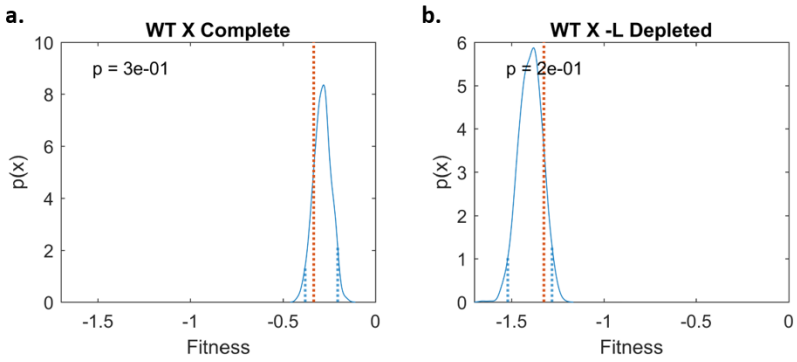


Figure 9. Correlation between variants' frequencies, fitness and GFP expression levels. Both frequency (left) and fitness values (right) correlate with GFP expression levels, although fitness values tend to vary towards the high end of the GFP distribution.

Effect of amino acids composition on variants' fitness

Next, we sought to discover how the variants' amino acid composition has affected its fitness under starvation for each amino acid. More specifically, we wanted to assess the fitness deficit (or benefit) associated with high phenylalanine (leucine) content in a variant; how is that deficit modulated during starvation to the inquired amino acid; and whether evolving the ancestral (WT) strain in absence of the amino acids in the medium alleviates the observed deficit, whether on complemented or depleted media. We started by inspecting the fitness of variants that have a high content of phenylalanine ($n \geq 2/10$) and leucine ($n \geq 4/10$) on the different genetic backgrounds and media. L-enriched variants have lower-than-average fitness when harbored by WT strain on complemented media (fig. 10.a), and slightly higher fitness than population mean when the WT strain is inoculated in L depleted media (fig. 10.b). This is consistent with the moderate cost of synthesis of this amino acid (see fig. 1). Similar trend, albeit less noticeable, is evident on the evo^L strain (fig. 10. c,d). Phenylalanine-enriched variants demonstrate a similar phenomenon: these strains have a significantly lower-than-average fitness on WT and evo^F strains under complemented media (fig. 11. a,c), consistent with the high cost of this amino acid. Surprisingly, these strains experience an increase in fitness compared to the population when provided F-depleted media (fig. 11. b,d).

In order to compare the fitness of F- enriched and L-enriched variants between the different conditions we first calculated the distance of the inquired sample (e.g. F-enriched variants) from the population mean, in units of standard deviation, for each condition; then, the difference in the distances was calculated for every two conditions. As a control, same procedure was carried out for 1000 random samples with the same size of the inquired sample, generating a control distribution of differences. The significance of the observed difference in distances was evaluated with respect to that distribution using a Z-test. p-values were corrected for testing of multiple hypotheses employing a false detection rate of .05.



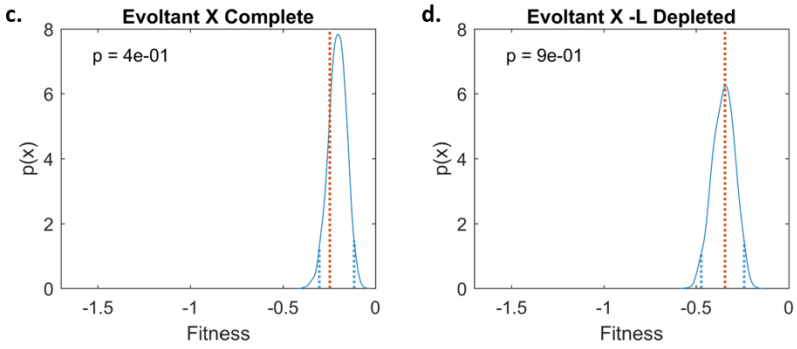
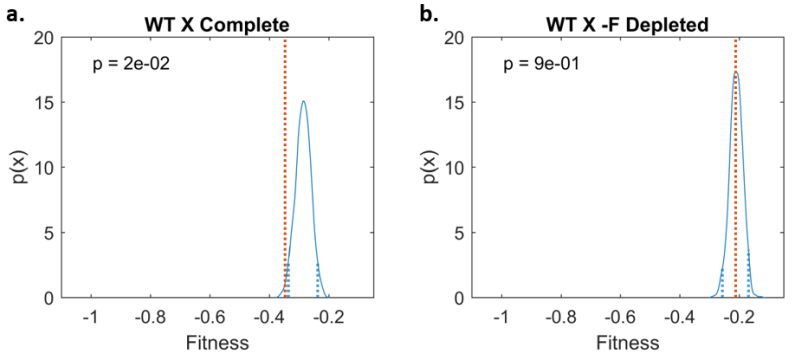


Figure 10. Fitness of leucine-enriched variants compared to the population fitness under different conditions. Mean fitness of leucine-enriched variants (orange, n_{leu} >= 4 out of 10) plotted against the bootstrapped distribution of the population mean (solid blue) in different conditions. Alpha values of .025 and .975 (significance of .05, two sided) are plotted in dashed blue. a) WT strain on complemented media. b) WT strain on L depleted media. c) evo^{L} on complemented media. d) evo^{L} on L depleted media.



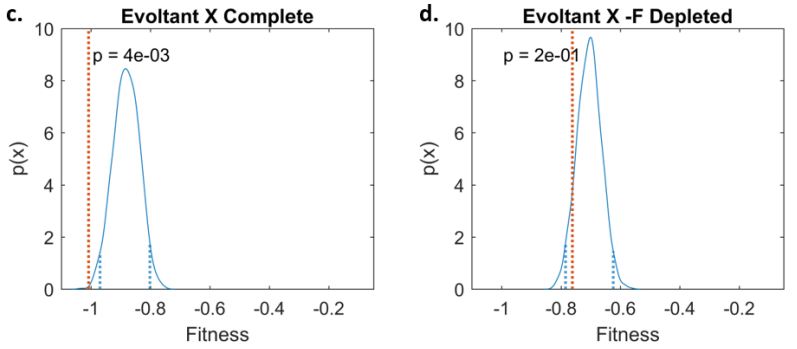


Figure 11. Fitness of phenylalanine-enriched variants compared to the population fitness under different conditions. Mean fitness of phenylalanine-enriched variants (orange, n_{phe} >= 2 out of 10) plotted against the bootstrapped distribution of the population mean (solid blue) in different conditions. Alpha values of .025 and .975 (significance of .05, two sided) are plotted in dashed blue. a) WT strain on complemented media. b) WT strain on F depleted media. c) evo^{F} on complemented media. d) evo^{F} on F depleted media.

This statistical procedure reveals that variants with high phenylalanine content experienced a significant improvement in fitness when the hosting WT strain was starved for phenylalanine, compared to complemented media (fig. 12). Such observation is evident, despite not significant, for the evolved strain as well. This was a surprising observation that opposed our expectations. Intuitively, one would argue that a depletion of an amino acids from the media would be harmful to variants that contain many occurrences of the depleted amino acid, for several reasons: the variants will experience the depletion stronger than variants that do not contain the depleted amino acid and therefore do not waste it on a un-needed protein; due to an increase in the synthesis rate of the depleted amino acid, which reflects energetic and metabolic cost; and because stalled ribosomes, which are expected to increase in frequency when amino acids are missing, trigger the starvation-induced stress response.

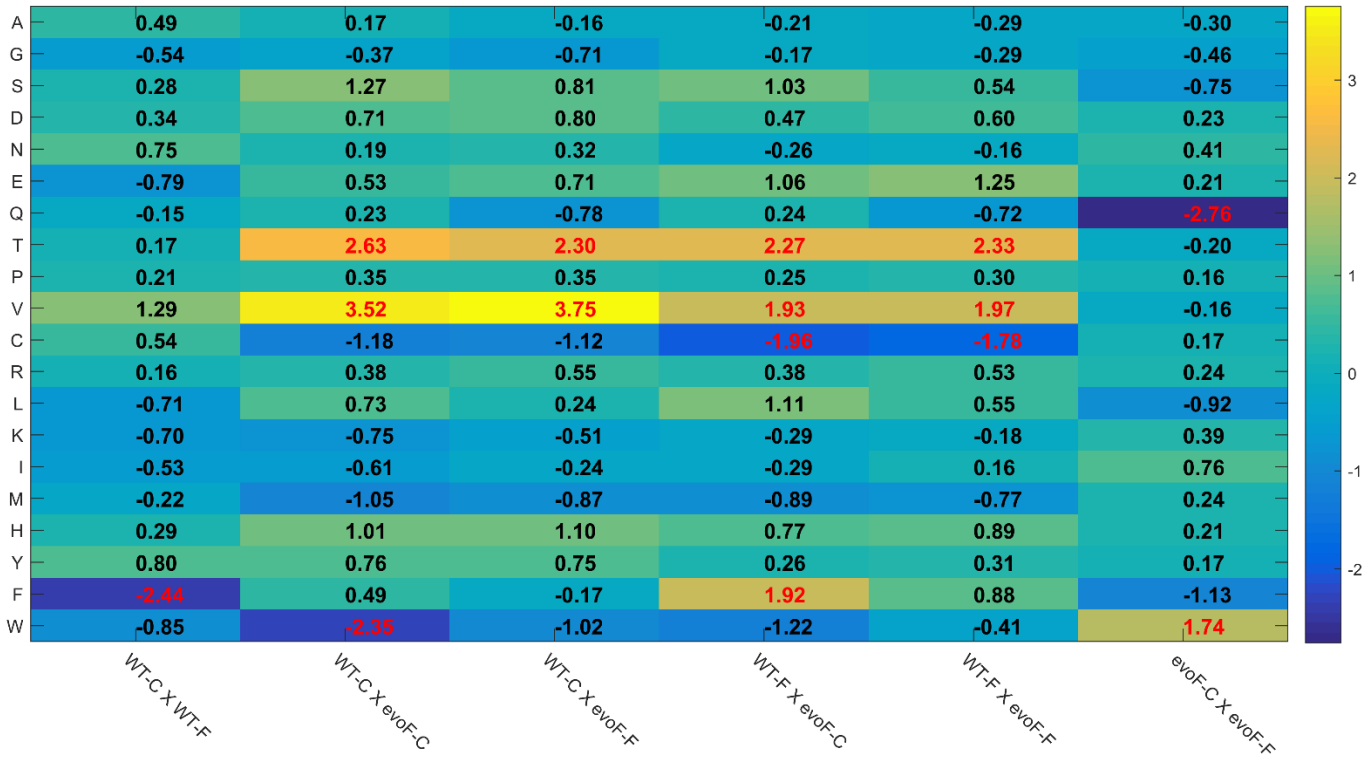


Figure 12. Comparison of variants' fitness in phenylalanine-related conditions pairs, as a function of their amino acid content. Each row represents variants that are enriched for the indicated amino acid. Columns indicate comparisons between the labeled conditions. Heatmap shows $-\log_{10}$(FDR-corrected p-value) of a statistical test used to evaluate the fitness difference of the enriched variants from the remaining population under the two conditions (see text). Sign indicates directionality: positive (negative) value indicated that the variants had higher fitness on the former (latter) condition. Red color indicates significance.

Our data allow us to examine how starvation to leucine and phenylalanine affects the fitness to strains that feature a high representation to each of the other amino acids as well. Figure 12 shows the change in fitness of variants rich in each of the 20 amino acids between pairs of strains grown on the different media. One interesting finding is that tryptophan- (W) enriched variants had a significant fitness enhancement when hosted by evo^F, compared to WT, and competed on complemented media. Put differently, this surprising observation indicates that one way to improve the fitness of a strain that is about to express a foreign protein rich with tryptophan is to evolve the strain in absence of phenylalanine before inducing the forced expression. In contrast such prior evolution to depletion of phenylalanine does not seem to improve the fitness of variants that are rich in this amino acid.

Notably, threonine (T) and valine (V) -rich variants performed better on WT than on evo^F, regardless of media; cysteine (C) demonstrate an opposite (although not significant) trend. The three amino acids are involved in the same metabolic network: valine acts as an activator to the enzyme that catalyzes the conversion of threonine to a isoleucine precursor, in a committing step; isoleucine and valine compete for the same precursor. Furthermore, isoleucine acts as an inhibitor for the aforementioned enzyme; cysteine is catalyzed to pyruvate, which, among its many other roles, serves as a precursor for valine. We cannot decipher the similarity in the response of variants containing these amino acids to the hosting genetic background (either WT or evo^F).

Same analysis for leucine depletion resulted in much less significant observations of change in strains fitness that can be associated to their amino acid content (fig. 13); we believe it is due to the evolution that took place in WT on L depleted media, in which the genetic properties of the GFP variant expressed by the cell had a secondary contribution to the cell's fitness.

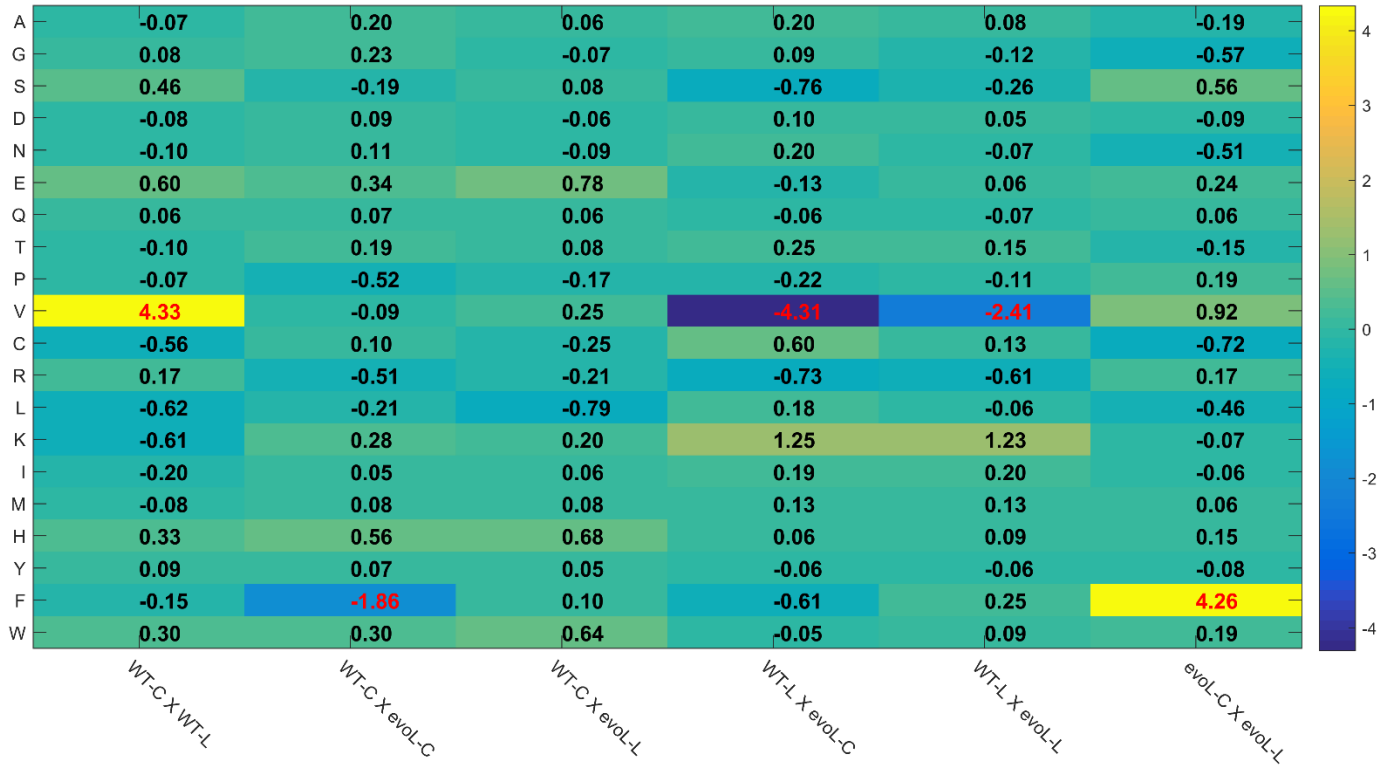


Figure 13. Comparison of variants' fitness in leucine-related conditions pairs, as a function of their amino acid content. Each row represents variants that are enriched for the indicated amino acid. Columns indicate comparisons between the labeled conditions. Heatmap shows -$\log_{10}$(FDR-corrected p-value) of a statistical test used to evaluate the fitness difference of the enriched variants from the remaining population under the two conditions (see text). Sign indicates directionality: positive (negative) value indicated that the variants had higher fitness on the former (latter) condition. Red color indicates significance.

<u>High translation initiation rate is associated with increased mRNA levels, likely due to prevention of degradation</u>

During initial analysis of the data we've observed that variants with stronger RBS had higher mRNA levels (fig. 14). Conventionally, mRNA levels are considered to depend on the core promoter and other transcription regulatory elements, and on mRNA degradation rates, but they are considered separate from translation related features such as RBS strength. We had two competing hypotheses addressing the directionality of information flow from the RBS component to mRNA levels. The first model has to do with a degradation-protecting effect that translating ribosomes might exert of mRNAs. According to this explanation a strong RBS, that is characterized by low amount of mRNA structure, recruits the ribosome to translation more efficiently, and in return the translating ribosomes protect the mRNA from degradation more effectively.

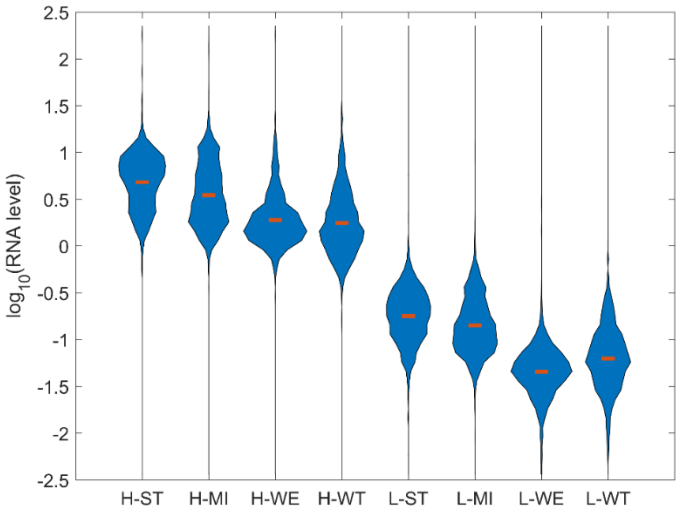


Figure 14. mRNA levels increase with promoter and RBS strength. Distributions of mRNA levels are shown for each composition of promoter and RBS. High promoter has higher basal mRNA levels than the low promoter, and stronger RBSs are associated with higher mRNA levels regardless of the promoter to which they are fused.

Low GC content in that region, a determinant of low RNA secondary structure content, might thus contribute to RNA stability through a translation-mediated effect. The alternative explanation suggests that the region that we consider as an RBS, also exerts its effect on transcription initiation or elongation. Here too low GC content would be associated with high RNA level, but this time due to an effect originating from the DNA and not the RNA melting. According to this model, more easily melting DNA duplex will allow higher transcription and hence higher mRNA level. Distinguishing between the two models is essential, since one of them clearly implicates translation with RNA degradation and the other would suggest a modified view of genes promoters that is affected by the RBS too. The challenge is that both models generate same prediction – that low GC content at the RBS would lead to high abundance of mRNA level, either due to translation-coupled or transcription-couple mechanism. We looked for a differential prediction of the two models. We decided to compute the DNA melting temperature (Tm), and also the RNA secondary structure free energy of each variant. As expected the two measures display high (negative) correlation (fig. 15 a,b,c) – variants with high DNA melting temperature tend to show high secondary structure of the RNA ($r = -.59$, $p = 0$). This correlation is to be expected because high GC content variants will show both tight mRNA structure and high DNA melting temperature. Yet the correlation is far from perfect (fig. 15 a-c): strains with same DNA melting temperature may show a wide RNA folding energies. We decided to ask if RNA levels vary more strongly with RNA or DNA melting. We found that the mRNA levels correlate only weakly with Tm while controlling for ΔG ($r = .16$, $p < 10^{-80}$, fig 15.e). In contrast mRNA levels correlate more strongly with ΔG at a given Tm ($r = -.5$, $p = 0$, fig. 15.d). This observation indicates that indeed the mRNA secondary structure,

and not the DNA accessibility, is the factor that modulates the mRNA levels. This analysis provides some support for the translation-governed mRNA degradation-protection hypothesis.

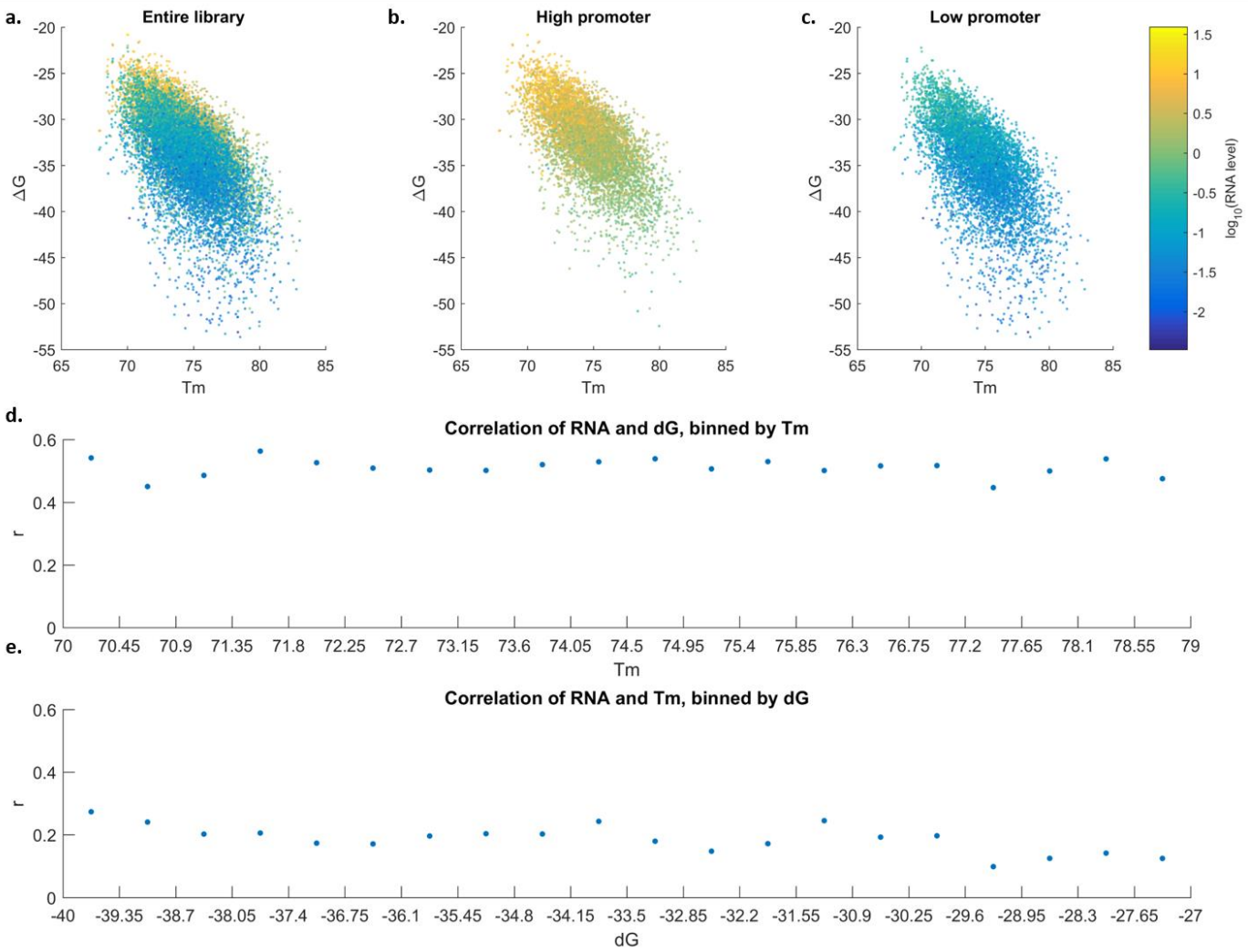


Figure 15. mRNA levels depend on mRNA structure more than on DNA melting temperature. Variants' Tm values were plotted against corresponding ΔG values, and were colored by RNA level for a) entire library. b) High-promoter variants. c) Low-promoter variants. d,e) Variants were binned according to Tm values and correlations between ΔG and mRNA levels were plotted, or vice versa.

Characterization of mutations dynamics during the competition

My study also provides a unique opportunity to reveal mutation dynamics along the lab evolution experiment. We decided to examine the notion that mutations might not be spreading randomly in across variants.

We noticed that under several conditions, many of the variants had a tremendous increase in frequency, and wanted to examine whether cis-mutations (i.e. mutations in the variant itself) or trans-mutations (i.e. mutations elsewhere in the genome (not sequenced so far)) may have caused the fitness improvements of certain variants. By plotting the fraction of mutated bases observed for each variant against the variants' fitness or frequency in the pool (fig. 16 a and b, respectively), we have observed a negative correlation with fitness ($r = -.33$, $p < 10^{-150}$) and frequency ($r = -.58$, $p = 0$). Such observation could be explained simply as a statistical artifact; if mutations are uniformly distributed, rare and random, low-coverage variants will show higher mutation rate by definition – a mismatch that occurred by chance in a single-read of a 100bp variant will have a frequency of 0.01 (1/variant length), and if it had occurred in a 5 reads variant it'll have a five-fold lower frequency. We have conducted simulations that operated under the assumption that all mismatched are derived from PCR and sequencing biases (error rate = $4*10^{-3}$, fig. 16. c); the simulation's results predict an order-of-magnitude lower mutation frequency, and a unimodal distribution of the mutation versus the bimodal distributions observed.

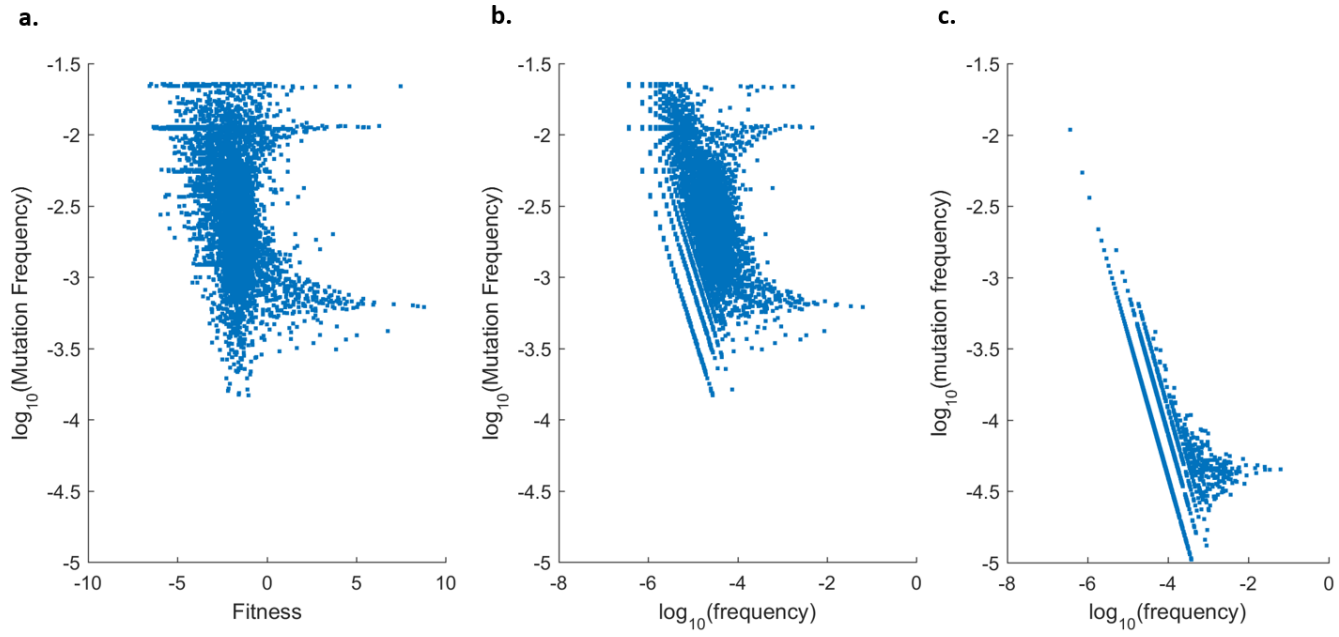


Figure 16. Mutation fraction as a function of variants' fitness. mutations frequency as a function of a) fitness, b) frequency, c) frequency, as obtained from a simulation.

Why are certain variants more likely than others to show mutations? Mutations in genetics and evolution can be either fitness affecting (beneficial or harmful), but they can also be neutral. We wanted to understand which of the mutations detected here represent adaptive and which represent neutral events. Accumulation of neutral mutations at certain variants in the library could hint that such variants simply increase the rate of mutation irrespective of their fitness effect.

To gain first insights on the spread of mutations over the different biological components of the variants, we split the variants by promoter and RBS and plotted, for variants with the same promoter and RBS, the frequency of mismatches observed per position in each variant (fig. 17, shown for variants with high promoter and mid RBS, a representative sample). Our data indeed show an accumulation of mutations at positions with relevance to the protein expression process, such as the -35 and -10 (TATA box) of the promoter (marked in orange). Surprisingly, the start codon was not a hotspot for mutations, despite the potential fitness advantage that can be obtained by eliminating translation.

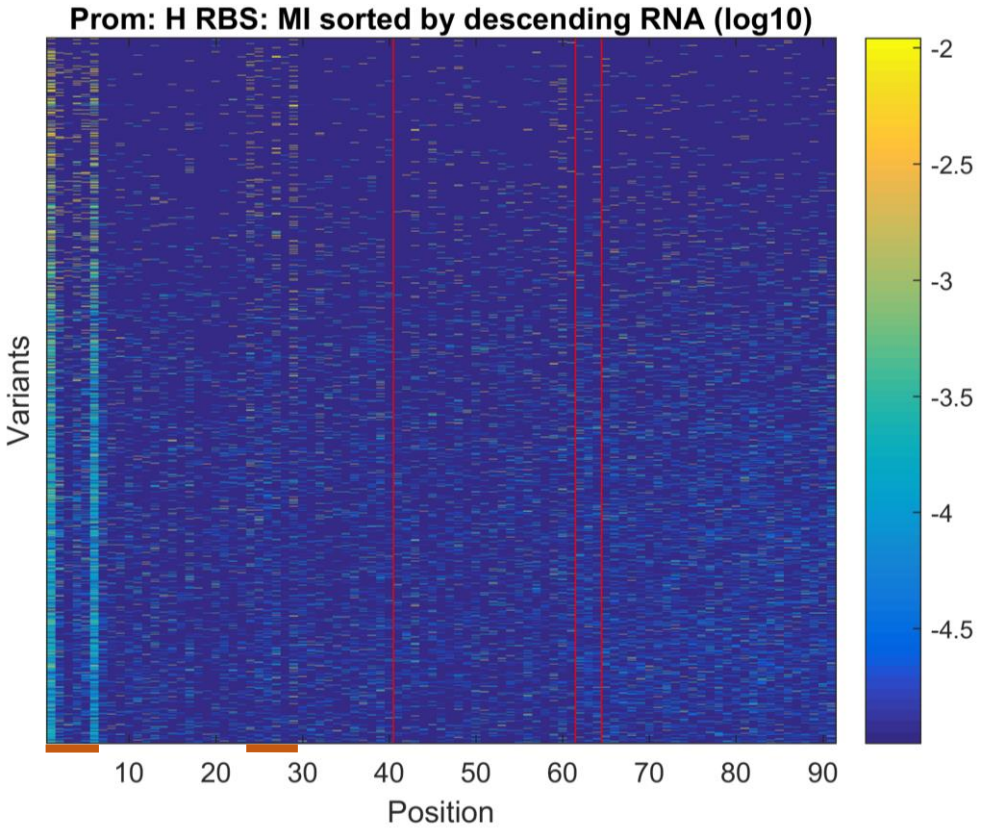


Figure 17. mutations tend to accumulate at distinct sites of the regulatory and coding sequence. Rows indicate variants and are sorted by descending mRNA level. Each position along each row is colored as the log10 of the fraction of mismatches observed. Orange lines indicates -35 and -10 boxes. Red line at (40) indicates end of the promoter; red line at (61) indicates end of RBS and red line at (64) indicates the start codon. Mutations tend to accumulate at functional sites of the promoter, beginning of the RBS and the CDS.

One approach to discriminate between beneficial and neutral mutations is by examining the change in their frequency over time. We conducted the analysis for the promoters and the 3 core RBS components, all of which have a given sequence (and hence each position has a specific contribution to fitness for the different variants), as well as WT RBS. Each position was given a score (annotated S_i for position i) based on the ratio of fraction of variants containing a mutation at that position at generation 84 and 0 (eq. 1).

$$(Equation\ 1) \qquad S_i = \mathbb{E}\left(\log_2\left(\frac{f_{i,84}}{f_{i,0}}\right)\right)$$

Where $f_{i,j}$ is the fraction of variant with a specific mutation at position i and generation j. Our results (fig. 18.) indicate that mutations at the -35 and -10 sites had larger effect on the fitness of the mutated variant, and furthermore, mutations at conserved positions that lead to deviations from the consensus sequence have provided fitness advantage to the variant, whereas for positions at the -35 and -10 sites that did not confer with the consensus sequence, variants that have accumulated mutations at these positions had lower fitness compared to their WT variants. Notably, variants with high promoter have demonstrated higher variance in fitness modification caused by mutations, across the promoter sequence and specifically at the -35 and -10 sites. The results for the RBS (fig. 19, high-promoter variants only) were inconclusive; variants with strong and mid RBS have demonstrated an elevated mutation fixation rate at the beginning of the RBS, however the sequence associated with that region of the RBS does not correspond to the Shine-Dalgarno consensus sequence (AGGAGGU in *E. coli*), or any other motif known to us. Variants with weak and WT RBS does not show this phenomenon.

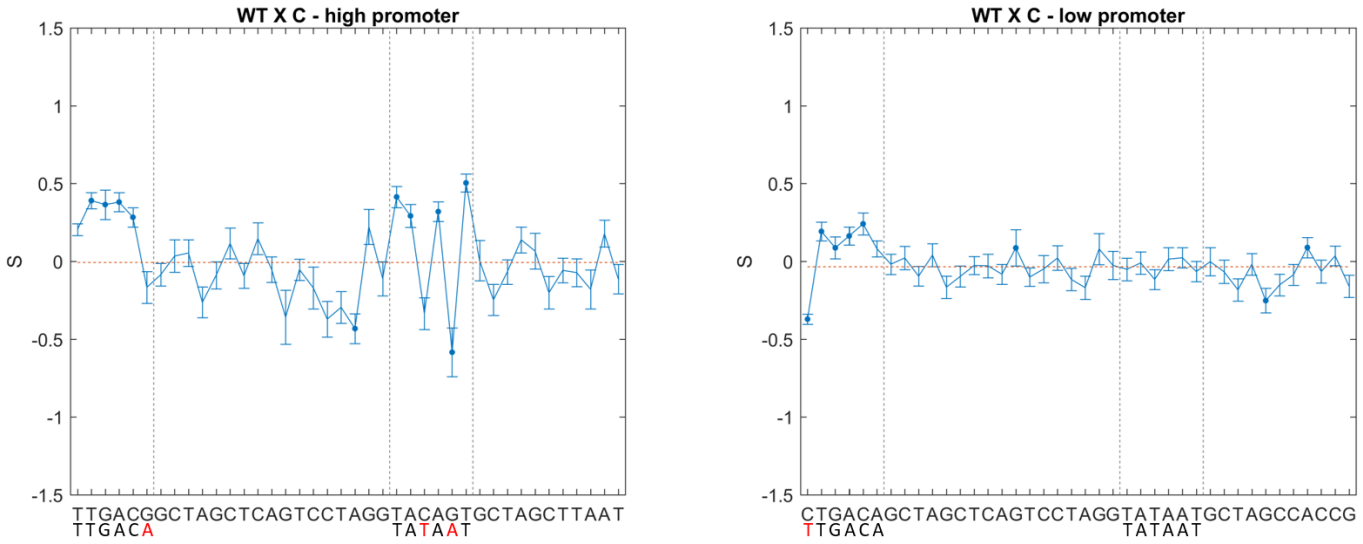


Figure 18. average change in mismatch to WT-variant over time at each position of the promoter. Plotted are s values (± SE), which represent the average ratio of change in mismatch frequency per position (see text). -35 and -10 boxes are marked with gray lines, and the consensus sequence appears beneath the actual promoter sequence. Positions that deviate from the consensus are colored in red. Significant s values (p <.1) are indicated by filled markers.

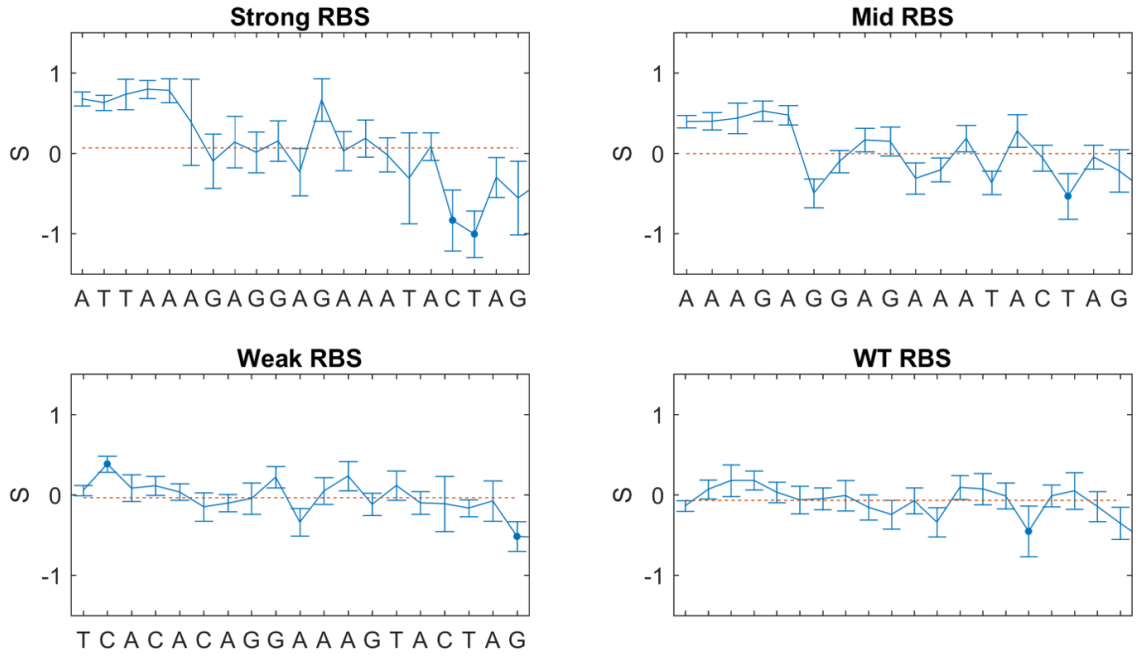


Figure 19. average change in mismatch to WT-variant over time at each position of the RBS. Plotted are s values (± SE), which represent the average ratio of change in mismatch frequency per position (see text). Significant s values (p <.1) are indicated by filled markers.

In a second strategy, we predicted the fitness effect of mutations from their inferred protein sequence change. Clearly most beneficial mutations introduced to the coding sequence should be those that eliminate production of the foreign GFP. In particular nonsense mutations that introduce a STOP codon early on, in the variable region, should be beneficial. Synonymous mutations were expected to affect fitness the least, and hence regarded as neutral, though effects of such mutations are known to happen[11]. Non-synonymous mutations could either increase or decrease fitness, for example if they convert between amino acids that are costly or cheap to synthesize, or if they change the hydrophobicity of the peptide[4]. We thus classified all mutations observed in the CDS into these four categories and examined for each the fitness of the original variant to the fitness of the mutate that emerges from that variant (fig 20). Yet, since the observed dispersion is similar regardless of mismatch type, we cannot conclude that one type of mismatches is actually beneficial than others. The lack of difference between the effect of the various type of mismatches on fitness can be either due to low effect of a single substitution of an amino acid on the fitness of the variant or because of insufficient sensitivity of our method – a sequencing coverage too low to determine the fitness of a single mutation – because of a relatively high sampling and sequencing biases occurring at very low frequencies.

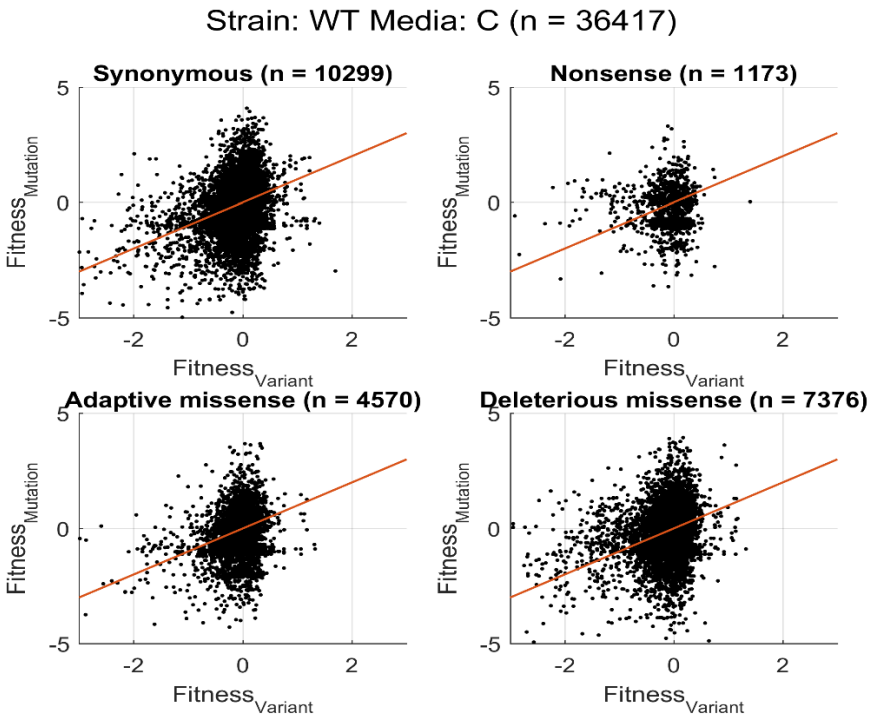


Figure 20. Mutations have similar contribution to fitness, regardless of their nature. Mutations' fitness is plotted as a function of the original variants' fitness. Despite having different theoretical contribution to fitness, none of the mutations' types introduces a distinguishable change to the observed fitness of the original variants.

Our dataset contains information on both mutation distribution and RNA levels. We thus have considered to examine the controversial notion of transcription-coupled mutagenesis; briefly, since the DNA is unwound during transcription, it may be more vulnerable to mutagenic agents, and this tendency to acquire mutations is expected to be correlated with transcription level. Several analyses (not shown) have failed to demonstrate such correlation. We would carefully argue that the transcription-coupled mutagenesis hypothesis is not supported by our data. However, our experimental system was not designed to answer such questions.

Discussion

In the presented study we have established an experimental system to learn the effect of amino acids depletion on the cost of gene expression in strains that either experienced the starvation as a novel challenge or were accustomed to it during evolution.

We chose to use phenylalanine (F) and leucine (L) as the amino acids to be depleted due to considerations of metabolic cost and abundance in the library's variants. We saw that while phenylalanine depletion imposed a little growth defect to the WT strain, starvation to leucine lead to an extreme impairment in growth, as reflected by a longer lag phase. Respectively, the strain evolved in absence of leucine showed a dramatic adaptation to the depleted media, whereas the phenylalanine-starved evo^{F} demonstrated only minor improvements compared to the WT.

To mimic the different genetic architectures that are found in the genome, we used a synthetic GFP library whose variants consisted of different regulatory element, as well as a short peptide sequence. We introduced the library into the WT and evolved strains, and conducted a pooled competition for each strain in presence or absence of the amino acid to which it was starved during the evolution. The WT strain was competed on complemented medium or in absence of both amino acids, separately. Each competition was done in 3 independent replicates. The fitness of each variant in each condition was evaluated as the binary log of its frequencies at the end and beginning of the competition.

We noticed that our variants' fitness across samples tended to cluster according to strain, and then according to media; this serves as an indication that indeed the variants have experienced similar selective forces that were governed by their genetic background and environmental composition. There were a few exceptions: when starved to leucine, the WT strain showed large deviations in variants' fitness compared to the other WT conditions, and the three replicates of that condition has clustered together separately from the other WT

conditions. Given that leucine starvation has caused a great growth deficit to the WT strain, as described earlier, we suspect that the cells harboring the library's variants have followed not only a competitional trajectory, in which the fitness of the variants remains constant and only its frequency is changes, but also an evolutionary process that had modified the cells' basal fitness. The evo^F strain has also shown some fluctuations in fitness, and indeed this is the only strain in which samples do not sub-cluster by media. A possible explanation is that during evolution the evo^F strain had found a solution that, although resulting in better acclimation to F starvation, imposes a fitness cost of its own, that increased the evolvability of the strains towards a better solution. A similar observation has been made in when yeast were allowed to evolve to an steep decline in media temperature; first, chromosomes that are harboring cold-shock genes were duplicated, to increase the amount of cold-shock proteins. However, since that solution includes replication of an entire chromosome and possibly expression of un-needed genes, it has been replaced by specific point mutations later on[12].

We saw that variants with low promoter have benefited if the translation was driven by a stronger RBS. This finding is not trivial, since weak-RBS variants are produce less protein than their strong RBS counterparts; To this point, we have no explanation for that observation.

By examining the fitness of F and L enriched variants over the different conditions, we noticed that while both have a lower-than-average fitness when competed on complemented medium (which was significant for F-enriched variants only), a slight elevation in fitness was observed when the variants were starved for the corresponding amino acid. That finding was later reassured by post-analysis, which revealed that F-containing variants had significantly higher fitness when harbored by WT that was starved for phenylalanine, compared to either WT or evo^F supplied with all amino acids, and slightly better when harbored by evo^F starved compared to evo^F on complete media. Furthermore, tryptophan (W) -enriched variants had significantly higher fitness when harbored by evo^F on complemented media compared to F depleted media, or compared to WT on complemented media. Comparisons done on the leucine-related conditions have failed to demonstrate interesting or inferable interactions, possibly due to the evolution experienced by WT X -L.

Why do phenylalanine-enriched variants perform better when starved to phenylalanine? We had two hypotheses, one regards concerns of supply and demand and the other related to the efficiency of translation.

According to the supply-and-demand hypothesis, when the WT strain encounters phenylalanine depletion, it increases the rate of phenylalanine biosynthesis to overcome the shortage and to create an additional

phenylalanine reservoir. Under these conditions, phenylalanine is already produced in the cells to a high amount such that the synthesis-cost component of phenylalanine becomes irrelevant, as all cells (those that have F-enriched library variants, and those that do not have such variants) produce to equal levels and 'pay' the same in terms of metabolic cost. At such conditions the relative burden in expressing F-enriched variants, compared to expressing other variants, is reduced. Furthermore, the theory argues that during evolution of evo^F, the ancestral clone has adapted by decreasing the internal demand for phenylalanine, hence leaving more available precursors for the synthesis of of tryptophan and tyrosine, two amino acids that share a metabolic precursor with phenylalanine.

The translation-efficiency hypothesis is based on the notion that when cells are starved for a particular amino acid, ribosomes stall on codons that encode for that amino acid. In that manner, during starvation the cell actually produces less GFP per time unit compared to non-starvation conditions, therefore reducing the overall cost imposed by expression of an un-needed protein. This prediction can be validated by measuring the fluorescence of an F-enriched variant isolated from the library under depleted and complemented media. Importantly, stalled ribosomes are activators of the starvation-induced stringent response, which leads to growth arrest and cell death; the theory assumes that harboring F-enriched variants does not cause an elevation in the activation of the stringent response. Moreover, the theory predicts that by devising methods for efficient production of F-containing proteins during evolution, the benefit of starvation will vanish when comparing fitness across genetic backgrounds.

To summarize, while both theories agree that starvation will benefit with F-enriched variants when harbored by the same strain, the supply and demand theory predicts that they will also perform better when comparing the evolved strain to the WT strain, whereas the stalling theory does not.

Since there's no evidence that being hosted by evo^F strain increases the fitness of F-enriched variants compared to WT when strains were exposed to the same media, and since only the stalling theory can explain why F-enriched variants perform better under starvation when hosted by WT compared to evo^F grown on complete media, we'd argue that the stalling theory is better supported by the data. However, we still consider it very likely that the consumption of phenylalanine is reduced during evolution, and we'd argue that our experimental system not suited to approve or refute such claim.

As for the finding that tryptophan-enriched variants increase in fitness when harbored by evo^F, compared to WT, while both compete on complete media, Tryptophan and phenylalanine share a metabolic synthesis

precursor, and both theories relay on that fact to explain increase in fitness of W-variants. If, during evolution, the demand for phenylalanine has decreased, as proposed by the supply and demand hypothesis, then a larger fraction of the shared precursor can be devoted to tryptophan, hence reducing its synthesis cost. In other words, less precursor is required to produce same amounts of tryptophan. On the other hand, if evolution led to favoring of phenylalnine synthesis over tryptophan, those W-enriched variants may experience 'internal' depletion of tryptophan, which will lead to reduced synthesis of GFP over time according to the stalling theory. While the stalling theory predicts that when harbored by evo^{F}, W-enriched variants will benefit from starvation to phenylalanine due to an increased 'depletion' of W, the supply and demand theory predicts that as the demand for F increases, less precursor would be available for W synthesis, and therefore W-variants will perform worse; the observation supports, thus, the supply and demand theory.

Crucially, the fitness measurement in our study is derived from a costly expression of a gene that is unfunctional and unnecessary to the organism. In practice, if starvation to an amino acid downregulates expression of genes in which it is common, the cell may gain fitness "benefit" due to reduced expression but the overall implications on fitness would be negative due to reduced functionality.

Our data had allowed us to explore the dependence between translation-related components, such as the RBS and 5' of the coding sequence (referred as CDS), and mRNA levels. Two competing hypotheses were proposed to explain the observed relation, the first is that the RBS and CDS act as an unfunctional extension of the promoter, allowing for easier unwinding of the DNA; the second is that high translation initiation rate prevents degradation, which is mediated by reduced access of the degradation machineries to the mRNA due to its occupancy by ribosome. Since correlation between mRNA levels and mRNA free energy is higher than with melting temperature of the DNA, we claim that degradation-protection is the main mechanism by which strong RBS and loosely-structured CDS allow for higher mRNA levels. That argument is supported by recent findings[13]; however, a link between transcription rate and translation initiation rate has been suggested as well[14].

Our data had also enabled us to study the implications of mutations on fitness. We've witnessed that some of our variants indeed gained mutations, and that these mutations increased in frequency when occurring at positions that promote the expression of the GFP. Similarly, mutations at positions that did not obey consensus sequences at -35 and -10 boxes of the promoter were deleterious to the cells and declined in frequency over time. Importantly, this finding indicates of an active evolutionary process that took place

during the competitions, even at conditions that seemed ‘unaffected’ by evolution, and possibly interfered us to estimate fitness correctly for some of the variants.

Our experiment had few drawbacks: first, expression levels of GFP were not re-measured for the different conditions, and therefore we cannot know if certain variants have decreased in GFP production due to depletion or evolution of the harboring strain. We are certain that such measurements would have allowed us to provide better biological insights regarding processes that have occurred in our experiment.

Second, our ancestral sample (as provided to us) demonstrates a clear competitional signal, a result of inevitable competition that took place during preparation of the library. The frequency of the variants at the ancestor spans 2-3 orders of magnitude, and has high anti-correlation with GFP levels.
Since the frequency of the variants at the end of the competition depends not only at its fitness, but also on the frequency at the start of the competition, some of the dynamics of the competition were not captured by our fitness estimation, as a) variants with low initial frequency were more sensitive to genetic drift occurring during evolution, as well as sampling and sequencing biases; for example, a variant with low initial frequency (which have reached the ‘asymptotic’ phase of its decline curve) that was observed at the end of the competition can show a higher fitness compared to a variant that had higher initial frequency and has declined more steeply. Mathematically, such phenomena should not occur; however, due to factors mentioned before (drift, biases and mutations), that I could not model to a reasonable account, I’ve witnessed such variants. A method to remove these variants while maintaining data integrity was not found.

Third, I sequenced our samples at two time points only: generations 0 and 84, as has been demonstrated by a previous experiment to provide a good fitness estimation. In practice, due to an accelerated evolution that occurred at WT X -L condition, I could not derive insightful conclusions from leucine-related conditions.

In light of the above, and in order to provide better answers to the questions raised at the beginning of the experiment, I propose to use an experimental system that is not library based, but relies on evolution: by taking an auxotroph for an amino acid, and evolve it in a gradient of starvation over time. Sequencing the genome of the evolving cells in different time points (until no improvement in growth rate is observed, or until the growth rate is similar to that of the cell when grown in complemented medium, which indicates that the starvation does not impose a severe challenge anymore) would allow the experimenter to investigate how did the genome alter in response to starvation. One could ask if the frequency of codons encoding the depleted amino acid in the genome has decreased over time, and if so, to which codons and amino acids; how

does that rate of substitution depend on the transcription and translation rates of the genes; what were the implications to the tRNA pool – whether certain tRNA genes have increased in copy number, whereas others have decreased; and similarly, are there observed modifications to the tRNA aminoacyl synthetases and synthesis pathways of the depleted amino acids, as can be inferred by their sequence. Of course, one can extend the set of tool he or she uses and sequence the transcriptome or evaluate the proteome of the newly-evolved strain.

Materials and methods

Composition of the synthetic GFP library

The synthetic GFP library was kindly provided by to us by Daniel B. Goodman[15] and is described there. Briefly, each variant contains a variable region fused upstream to super-folder GFP gene. The variable region is composed of a promoter, ribosome binding site (RBS), start codon and a 10 amino acids peptide. The repertoire of the library is either a low- or high- transcription rate promoter; 3 different sequences to the RBS, with different affinities to the ribosome. The 10 amino acids peptide was taken from the 5' of 137 native *E. coli* proteins, and was recoded 13 times using synonymous codons. Each amino acids peptide stretch was fused to its native RBS as well. Overall, the library contains 14,234 sfGFP variants.
Library plasmids contained a kanamycin resistance gene as a selective marker.

Strains and media

E. coli strain MG1655-K12 was used for all experiments. The strain was selected, despite being autotroph for both phenylalanine and leucine, to maintain compatibility with previous experiments.

To control the amino acid content of the media, we used MOPS defined media (MOPS EZ rich defined medium, Teknova). The media ordered contained all amino acids but phenylalanine and leucine, and either phenylalanine (0.4 mM) or leucine (0.8mM) were added to create either leucine- or phenylalanine-depleted media, respectively. Both amino acids were added to create a complemented media.

Lab evolution setup

Both initial evolution on complemented MOPS media and the follow evolutions on leucine\phenylalanine-depleted media were conducted in a serial dilution regime. Cells were grown over night on Mops complete media and then diluted 1:120 into 1.2 ml of appropriate media (see above) and grown at 30˚C for 24h under constant shaking. Each day, cells were diluted 1:120 (~6.9 generations) to fresh media. Each experiment was

done in 3 independent repeats. Every 3-4 days (~21-28 generations) samples of each repeat were mixed with glycerol (final concentration 30%) and frozen at -80˚C.

Growth assays

Growth experiments were conducted in 96-wells plates, 150 µl per well. An overnight culture was diluted to fresh media to OD_{600} of ~0.05. Different strains were then arranged in checkerboard or other equally-representing pattern. Plates were continuously shaken at 30˚C. OD_{600} was read every 15 minutes for 10 hours using a plate reader (Infinity).

Preparation of elecctrocompetent cells

An overnight culture was diluted 1:100 in LB media. Cells were shaken at 30˚C until reaching OD_{600} ~0.7. Cells were then harvested by centrifugation at 2000_g for 10 minutes at 4˚C. Cells were washed twice by resuspension in ice-cold 10% glycerol while on ice and harvesting by centrifugation (2000_g, 10 minutes). Cells were resuspended in 1ml 10% ice cold glycerol per 50 ml of initial culture and divided to aliquots of 50µl. Tubes were stored at -80˚C.

Library Transformation and pooling

For each ancestral sample, 10 electroporation reactions were conducted (see results). Ancestral library culture was grown overnight in 10 ml of LB + kanamycin ($50_{\mu g/ul}$). plasmids were purified using Promega miniprep kit. Library plasmids ($1_{\mu l}$, $\sim 40_{ng/\mu l}$) were mixed with electrocompetent cells from the adaptation ancestral, MOPS complete strain or depletion-evolved strains, and the mixture was transferred to cooled 0.1 cm electroporation cuvettes. Cells were electroporated using a BioRad Xpulser device (voltage: 1800_V, capacitance: 25 $_{\mu F}$, resistance: 200_Ω) and immediately resuspended in 500µl of LB. The culture was inoculated in additional 500µl of LB and was allowed to recover for 1 hour in a shaking incubator at 30˚C. To allow for spreading of all cells in the tube, cells were concentrated by centrifugation (1 min, maximal speed) and resuspended in 400µl LB. We spread 200µl of the recovered culture on kanamycin-containing LB plates, 2 plates per electrotoration reaction to a total of 20 plates. Plates grew overnight at 30˚c and colonies were then pooled by suspension in 1 ml of LB. The pooled culture was mixed with glycerol and kept at -80˚C. A large portion of the frozen cultures was grown overnight in the media it has adapted to and that overnight culture served as an ancestor for the competition assays.

Lab competition assay

Competition assays were performed by serial dilution, similarly to the evolution procedure, for 84 generations (12 days). Kanamycin (50µl/ml) was added to the media to maintain retention of the plasmids.

Library construction

Plasmids were extracted from the 84-generations old cultures, as well as from the ancestral culture using Promega miniprep kit. Each sample was grown overnight in 10 ml of the media that used for the competition assays (ancestral samples were inoculated in complemented MOPS medium). The variable region (91-94 bp) was amplified by a PCR reaction using primers that contained a gene-homology region flanked by a part of Illumina's sequencing adapters. To achieve the sequence complexity required for the sequencing reaction, both forward and reverse primers varied in length of the gene-homology region, from 21 to 25 nucleotides, resulting 5 forward and 5 reverse primers in total. PCR was done as describe in the table (named PCR1). PCR products (~170 bp) were purified using a PCR and gel purification kit (Promega) and a second PCR reaction was carried, to synthesize the remaining fragment of the adapters with primers containing a unique barcode to each sample. Yielded PCR products size was ~270 bp. To remove unspecific bands, PCR products were separated using gel electrophoresis (1.5% TAE gel) and the desired band was cut and purified using a gel purification kit (Promega). Samples were diluted to concentration of 4nM. Sequencing was done using a 125-bp paired end kit on a Hiseq 2500 machine (Illumina). For primer sequences and PCR programs, see attached appendix.

Data processing

Reads were received demultiplexed from INCPM. Merging of paired-end reads was done using Pear (http://sco.h-its.org/exelixis/web/software/pear/, default settings) and primers were trimmed using cutadapt (http://cutadapt.readthedocs.io, default settings). Assignment of trimmed reads to reference sequences was done by a custom code. Analyses were done using MATLAB (MathWorks, Inc.).

Acknowledgements

I'd like to thank:

My mentor, Prof. Yitzhak Pilpel, for the advice, support and patience.

Dr. Orna Dahan, for the assistance accompanied with a smile.

The Pilpel lab members, for fruitful (and ongoing) discussions.
Special thanks go to Ms. Kaminski, a brilliant colleague and a friend.

My friends,

My family.

References

1. Rosenow, C. Prokaryotic RNA preparation methods useful for high density array analysis: comparison of two approaches. *Nucleic Acids Res.* **29,** 112e–112 (2001).

2. Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z. & Hwa, T. Interdependence of Cell Growth and Gene Expression: Origins and Consequences. *Science (80-.).* **330,** 1099–1102 (2010).

3. Gingold, H., Dahan, O. & Pilpel, Y. Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome. *Nucleic Acids Res.* **40,** 10053–10063 (2012).

4. Frumkin, I. *et al.* Gene Architectures that Minimize Cost of Gene Expression. *Mol. Cell* **65,** 142–153 (2017).

5. Chen, W.-H., Lu, G., Bork, P., Hu, S. & Lercher, M. J. Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nat. Commun.* **7,** 11334 (2016).

6. Evans, M. E., Clark, W. C., Zheng, G. & Pan, T. Determination of tRNA aminoacylation levels by high-throughput sequencing. *Nucleic Acids Res.* 1–8 (2017). doi:10.1093/nar/gkx514

7. Elf, J., Nilsson, D., Tenson, T. & Ehrenberg, M. Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science* **300,** 1718–1722 (2003).

8. Subramaniam, A. R., Pan, T. & Cluzel, P. Environmental perturbations lift the degeneracy of the genetic code to regulate protein levels in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 2419–24 (2013).

9. Durfee, T., Hansen, A.-M., Zhi, H., Blattner, F. R. & Jin, D. J. Transcription profiling of the stringent response in Escherichia coli. *J. Bacteriol.* **190,** 1084–96 (2008).

10. Brown, A., Fernández, I. S., Gordiyenko, Y. & Ramakrishnan, V. Ribosome-dependent activation of stringent control. *Nature* **534,** 277–80 (2016).

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

12. Yona, A. H. *et al.* Chromosomal duplication is a transient evolutionary solution to stress. *Proc. Natl. Acad. Sci.* **109,** 21010–21015 (2012).

13. Boël, G. *et al.* Codon influence on protein expression in E . coli correlates with mRNA levels. *Nature*

529, 358–363 (2016).

14. Slobodin, B. *et al.* Transcription Impacts the Efficiency of mRNA Translation via Co-transcriptional N6-adenosine Methylation. *Cell* **169,** 326–337.e12 (2017).

15. Goodman, D. B., Church, G. M. & Kosuri, S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science (80-.).* **342,** 475–479 (2013).

Appendix

PCR programs:

Kapa HiFi hotstart readymix (Kapa Biosystems): 12.5 ul.

ddW: 10 ul.

F primer: .75 ul.

R primer: .75 ul.

Plasmid: 1 ul.

	PCR 1		PCR 2	
	Temperature (˚C)	Time (s)	Temperature (˚C)	Time (s)
Initial denaturaion	95	180	95	180
Denaturation	98	20	98	20
Annealing	58	15	58	15
Elongation	72	15	72	20
	Go to step 2, X25		Go to step 2, X10	
Final elongation	72	60	72	60

Primers:

Primer ID	Sequence
F21*	ACGACGCTCTTCCGATCTATGAAAAGCTTAGTCATGGCG
F22*	ACGACGCTCTTCCGATCTAATGAAAAGCTTAGTCATGGCG
F23*	ACGACGCTCTTCCGATCTCAATGAAAAGCTTAGTCATGGCG
F24*	ACGACGCTCTTCCGATCTACAATGAAAAGCTTAGTCATGGCG
F25*	ACGACGCTCTTCCGATCTGACAATGAAAAGCTTAGTCATGGCG
R21	AGACGTGTGCTCTTCCGATCTCTCTTCGCCTTTACGCATATG
R22	AGACGTGTGCTCTTCCGATCTGCTCTTCGCCTTTACGCATATG
R23	AGACGTGTGCTCTTCCGATCTAGCTCTTCGCCTTTACGCATATG
R24	AGACGTGTGCTCTTCCGATCTCAGCTCTTCGCCTTTACGCATATG
R25	AGACGTGTGCTCTTCCGATCTACAGCTCTTCGCCTTTACGCATATG
* During data inspection and primer trimming, it has been found that the sequence CGCC was added (during synthesis or PCR reactions) to the 3' of all forward primers. Therefore, in the sequencing data the primers read as …CATGCGGCGCC.	

Primers and Illumina barcodes assignment

Sample ID	Illumina barcode	F primer	R primer
WT_C_12_1	CTATGCGT	'F21'	'R23'
WT_C_12_2	GTCCACAG	'F21'	'R24'
WT_C_12_3	TTGTCTAT	'F21'	'R25'
WT_F_12_1	AACAATGG	'F22'	'R23'
WT_F_12_2	ATTCCTCT	'F22'	'R23'
WT_F_12_3	GAAGGAAG	'F22'	'R24'
WT_L_12_1	TCGCTAGA	'F22'	'R25'
WT_L_12_2	ACAGTTGA	'F23'	'R21'
WT_L_12_3	CAATAGTC	'F23'	'R23'
evoF_C_12_1	GACCGTTG	'F23'	'R23'
evoF_C_12_2	TCTGCAAG	'F23'	'R24'
evoF_C_12_3	ACTGTATC	'F23'	'R25'
evoF_F_12_1	CCAGCACC	'F24'	'R21'
evoF_F_12_2	GACCTAAC	'F24'	'R21'
evoF_F_12_3	TGCAAGTA	'F24'	'R23'
evoL_C_12_1	AGCATGGA	'F24'	'R24'
evoL_C_12_2	CGCTATGT	'F25'	'R21'
evoL_C_12_3	GATATCCA	'F25'	'R22'
evoL_L_12_1	TGTAACTC	'F25'	'R23'
evoL_L_12_2	AGGTCGCA	'F21'	'R25'
evoL_L_12_3	CTGCGGAT	'F22'	'R24'
WT_anc_1	GCAGCCGT	'F23'	'R22'
WT_anc_2	TTAATCAG	'F24'	'R23'
F_anc_1	AGGTGCGA	'F25'	'R21'
F_anc_2	CTGTGGCG	'F22'	'R25'
L_anc_1	GCCGCAAC	'F23'	'R24'
L_anc_2	TTATATCT	'F24'	'R22'