מכון ויצמן למדע

WEIZMANN INSTITUTE OF SCIENCE

עבודת גמר (תזה) לתואר

דוקטור לפילוסופיה

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

מאת
עידן פרומקין

מנגנוני אדפטציה של מכונות תאיות המקיימות את הדוגמה המרכזית של הביולוגיה המולקולרית

Thesis for the degree
Doctor of Philosophy

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

By
Idan Frumkin

Evolutionary mechanisms of cellular machineries of the central dogma

מנחה:
פרופ' יצחק פלפל

Advisor:
Prof. Yitzhak Pilpel

אדר תשע"ח

March 2018

Acknowledgments

On the one hand, science is a fundamental endeavor for any democratic, progressive society. It is only with the scientific method that we can improve the quality and well-being of humanity as a whole – and therefore it is important that we continue to practice it and past it through the generations. I am forever grateful to the many people who have taught me along the way how to think and ask biological questions – including my lab peers, collaborators, classmate, and friends I have made along the way. Specifically, I am thankful to Prof. Maya Schuldiner for accompanying me ever since my first steps in science with her always-brilliant advice. Most importantly, I am forever grateful to my mentor, Prof. Yitzhak Pilpel. With his original mind and brilliance, calmness, and ability to always surprise me – Tzachi is the best mentor a young scientist could ever hope for. I also appreciate the Azrieli foundation for a PhD fellowship that has provided me independence and opportunity to mature as a scientist.

On the other hand, science is a privilege that is not accessible to many. Since the best science is practiced with a curiosity-driven approach, one must enjoy a certain degree of freedom and personal happiness to do it well. I am forever thankful to my family and friends for providing me with these needs. First and foremost, I am grateful to my parents, Ronit and Gil Frumkin, who have always, throughout my entire life, supported and helped me to fulfil my dreams and desires. It is only with their constant nurture, care, and love that I could have walked this road. I am also thankful to my sister, Shani Frumkin, who can always truly understand me, and with whom I share a unique relationship of both protection and fun. Finally, I am grateful to my life partner, Yoni Yomtov, who is my best friend in the world, the person that can always make me laugh, and with whom I am always safe.

I devote my PhD thesis to my father, Gil Frumkin.
I love you dad and miss you every day.

Abstract

Cells are complex biological entities that perform all the functions of life and show remarkable ability to change and adapt to their surroundings. While lots of efforts have been devoted to the characterization of cellular pathways and functions, we still did not fully characterize the mechanisms by which cells evolve. The field of Evolutionary Cell Biology merges both evolutionary and cell biological thinking and aims to illuminate how cellular processes and phenotypes change during evolution.

Inspired by the ideas of evolutionary cell biology, I devoted my PhD to reveal evolutionary mechanisms of cellular machineries. This question has been relatively ignored, and I aspired to demonstrate how the combination of genome engineering and genomics tools, lab-evolution methodology, and computational analyses can yield important insights about the reasons behind the evolution of molecular machines that perform fundamental cellular functions. My PhD encompasses three main projects, all of which have been the product of fruitful collaborations with my lab peers.

First, I studied the evolution of the translation machinery by manipulating codon demand or tRNA supply. We revealed a new adaptation mechanism for tRNA genes, which is widely used in nature and is based on strategic mutations that switch one anticodon to another according to cellular needs. Thus, we showed that the tRNA genes are inter-connected in an evolutionary mutational network that allows them to serve as backups for one another, and that this network provides evolutionary plasticity to the translation machinery. We then studied the phenomenon of codon usage bias of highly expressed genes. Most past studies focused on local effects of codon choice in these genes, and how it governs their efficiency and accuracy of translation. By contrast, we revealed that codon usage of highly expressed genes has global implications on the cellular translation machinery, and that manipulating codon choice results in *trans* effects on other genes, which are affected in a codon-dependent manner. Namely, we revealed that introducing a non-optimal codon on highly expressed genes leads to the reduction in translation efficiency of genes that naturally use this codon due to an increased tRNA demand. Thus,

we show that codon usage bias in highly expressed gene was not selected by evolution only to maintain the translation of these genes – but also to maintain the integrity of the entire cellular translation process.

Second, we revealed various molecular mechanisms that cells evolved to reduce the cost of gene expression. Using synthetic DNA libraries, we were able to test in parallel the effects of many different gene architectures on energetic and resource cost per protein molecule. By comparing cost-effective and ineffective architectures, we found that cost per protein molecule could be minimized by lowering transcription levels, regulating translation speeds, and utilizing amino acids that are cheap to synthesize and that are less hydrophobic. We then examined natural bacterial genes and found that highly expressed genes have evolved more forcefully to minimize costs associated with their expression. We thus elucidated gene design elements that improve the economy of protein expression in natural and heterologous systems.

Lastly, I followed the lab-adaptation of the splicing machinery in yeast to reveal molecular means by which the system changes due to a burden of an in-efficiently spliced intron. We identified mutations in *cis* that improved the intron's splicing efficiency and increased the overall expression level of the entire gene. Additionally, we observed adaptations in *trans*, which both increased the cellular availability of the splicing machinery and changed proteins that facilitate splicing. Our work here revealed novel molecular means by which the splicing machinery is changed by natural selection to optimize gene-expression patterns of cells.

Ultimately, my PhD studies offer a new perspective on the evolution of cells by focusing on the interactions between complex systems, cellular demands, and molecular adaptation. With this point of view, more questions about the evolution of cells arise, which I intend to address in the future.

Introduction

Evolutionary Cell Biology is a powerful paradigm to study basic questions on biological systems

Although the cell is commonly referred to as "the most basic unit of life", it is actually so complex that despite over 350 years of research we are still far from fully understanding its structural, functional and evolutionary workings. Even the simplest unicellular organisms routinely carry out complicated tasks that remain incompletely understood, including processing and responding to signals from the environment, maintaining an efficient metabolic network, communicating with other cells, and adapting to new evolutionary challenges[1].

Studies in cell biology have illuminated many molecular pathways and proteins that govern cells, but we still do not yet fully understand how these systems are established and adapted. This is true perhaps because melding understandings of evolutionary processes with the observed variation among cells has been less common. Such dissection of evolutionary mechanisms that produce cellular functions might significantly advance our understanding of the fundamental principles governing cell biological systems by providing the rationale behind alterations and variations in cellular functions[2,3].

Bringing the fields of cell biology and evolution together into an integrated field of "evolutionary cell biology" has provided fundamental ideas on cellular innovation, complexity, and adaptation. There are two common ways by which evolutionary cell biological approaches are implemented: First, and most commonly used, exploratory studies of genomic and cell biological diversity are performed to reveal novel cellular components and pathways. Second, organisms with divergent cellular structures are compared and studied genetically and molecularly, to better understand the evolutionary mechanisms that drive diversity of cellular functions. Here, I review some of the major insights that evolutionary cell biology has provided as an introduction to my work on the evolution of cellular machineries.

Evolution of cellular complexity: the transition from prokaryotes to eukaryotes

Evidence from both molecular phylogeny and fossil suggest that prokaryotes predated eukaryotes[4,5]. Hence, eukaryotes probably have arisen from a prokaryotic state, in a transition from an organism lacking internal membranes to an organism that possesses some membrane-bound organelles[6]. The transition from prokaryote to the many eukaryotic species that exist today probably went through a last eukaryotic common ancestor (LECA), which evolved to form the major lineages of extant eukaryotes. Phylogenetic reconstructions show that LECA was a sophisticated cellular entity, possessing cytoskeleton proteins, an elaborate endomembrane system, a nucleus, mitochondria, and complex machineries that could implement processes such as intron splicing and meiosis[7]. Therefore, a central motivation of evolutionary cell biology is to reveal mechanisms that drive such increase in cellular complexity.

The most comprehensively studied question in evolutionary cell biology is the origin of compartmentalized cells. Two main mechanisms are thought to result in the birth of organelles. First, endosymbiotic acquisition, the incorporation and residence of one organism, the endosymbiont, inside another, the host. This is how the mitochondria and chloroplasts have been suggested to have originated[6]. The second mechanism is termed autogenous origin, in which factors of the pre-eukaryotic ancestor evolved gradually into new membrane-bound compartments without endosymbiotic events[8]. The organelles of the endomembrane system (nuclear membrane, endoplasmic reticulum, Golgi apparatus, lysosomes, and cellular vesicles) are the most established example of organelles derived via an autogenous mechanism[9].

<u>Origin of mitochondria and chloroplasts via endosymbiosis</u>

Cell biological and genomic studies have revealed much about the endosymbiosis process that led to the birth of mitochondria and chloroplasts from alpha-proteobacteria and cyanobacteria, respectively[10]. Interestingly, beyond these two well-characterized examples, an array of organisms with organelles derived from additional endosymbiotic events, termed plastids, also exists[11]. This observed variation allowed for the

characterization of various methods of endosymbiotic acquisition: (i) Primary endosymbiosis, in which a cyanobacteria is engulfed by a heterotrophic eukaryote, resulting in establishment of chloroplasts. (ii) Secondary endosymbiosis, in which a photosynthetic eukaryote is engulfed by a heterotrophic eukaryote, resulting in establishment of chloroplasts that are engulfed by a double-layer membrane. (iii) Tertiary endosymbiosis, in which a photosynthetic organism containing a secondary plastid is itself engulfed by another eukaryote. (iv) Serial endosymbiosis, in which a photosynthetic eukaryote is engulfed by another photosynthetic eukaryote, resulting in the replacement of the chloroplast[12].

Notably, an endosymbiotic event leads to biological changes in both the endosymbiont and the host that must accommodate one another. Some proteins of the host are probably retargeted to the new organelle, while genes from the endosymbiont's genome are in turn adapted to support the biology of the host[6,12]. Our understanding of the molecular processes governing these evolutionary scenarios are incomplete, and it is one of the major, current aims of evolutionary cell biology to fill this gap in our knowledge.

Origin of endomembrane system via autogenous processes

In addition to endosymbiosis, autogenous processes are also a powerful force in the evolutionary origin of the eukaryotic cell, specifically for organelles that are composed of a single lipid bilayer and devoid of genetic material. The membrane-trafficking system is the best example for an autogenous origin, including the organelles: endoplasmic reticulum (ER), Golgi apparatus, endosomes, and plasma membrane[13]. Relying on comparative genomics of membrane-trafficking systems of a wide range of eukaryotes, it has been suggested that the last eukaryotic common ancestor possessed a complex set of trafficking components[8]. According to the organelle-paralogy hypothesis (OPH)[9,14], the membrane-trafficking system became more and more complex with time, encompassing a greater variety of organelles and cellular structures. This process presumably occurred via gene duplication and divergence of specificity-encoding protein families, which define organelle properties such as tethering, docking, fission, or fusion[6].

In addition to comparative genomics, comparative structural analyses were also used to illuminate the origin of organelles. This approach revealed an evolutionary relationship between what were considered two distinct families of membrane-deformation proteins - coated vesicles and nuclear pore complexes[8]. While these two systems do not share a high degree of amino acid sequence similarity, they have a common, structural molecular architecture – which yielded the "protocoatomer hypothesis" that unites the evolutionary origin of vesicle coats with parts of the nuclear envelope[15]. Further investigation into the diversity of the eukaryotic domain may continue to reveal the complex evolutionary origin of the membrane-trafficking system and allow us to better characterize which of its parts are ancient versus those that have evolved more recently and why.

Evolution of cellular processes

While evolutionary cell biology has mostly influenced the study of organelle origin, it has the potential of providing important insights about the evolution of other cellular processes. Using natural variation might reveal different ways to perform same cellular tasks and revealing these dissimilarities could provide insights about fundamental requirements for such processes. If we aim to completely understand a certain cellular function, we need to reveal not only its origin, but also how it keeps evolving.

For example, the evolutionary interactions between hosts and pathogens can leave biological traces in cells[6]. Such functional changes could occur in host proteins or even in the external repertoire of cellular glycan, which is often used by pathogens for cellular recognition[16]. Another example is chromosome segregation, for which an evolutionary cell biology approach has been used to illuminate various molecular ways by which cells accurately separate their replicated DNA between two daughter cells[17]. One interesting observation in nematodes showed the evolutionary intersection between cell and spindle sizes. Mutational accumulation lines along with genetic variations in natural strains revealed that the "normal" range of spindle size is larger than expected and that a very simple scaling relationship with cell size may explain a great deal of variation in subcellular structure among species[18,19]. From an evolutionary cell biology point of view,

this notion suggests that even if a certain phenotype is under a strong selection (here, proper chromosomal segregation) – it does not force cells to converge to exactly same solutions and behaviors.

Lastly, cellular machines that carry out basic processes for cells, which demonstrate increased complexity with time, can also be studied with an evolutionary cell biology perspective. The translation machinery of eukaryotic cells, for example, is diverged from its prokaryotic counterpart, and has continued to evolve and change even within eukaryotes. While the evolution of this system has been studied at the protein level quite extensively[20], how it evolutionarily interacts with cellular structures and which cellular demands lead to its adaptation are largely unknown. *Characterizing molecular evolutionary mechanisms of such cellular machineries could improve our ability to understand the reasons behind the broad cellular diversity in nature, and help us to better reveal the molecular mechanisms and functions of these machineries. This motivation is the focus of this thesis work*.

Evolutionary cell biology of bacterial cells

While most studies in evolutionary cell biology focus on the eukaryotic cell, probably because of its increased cellular complexity, bacterial cells can also be studied with an evolutionary perspective. Even if they are "simpler", bacterial cells are still complex biological entities that evolve, change with time, and demonstrate a large degree of variability that should be explained.

Bacterial morphology and cellular shape have been studied fruitfully with an evolutionary cell biology perspective. Bacteria can be found in multiple shapes and sizes, from simple structures of spheres, rods, and spirals to unconventional chains, coils, stars, and more complex shapes such as branching filaments. Interestingly, some bacterial species can alter their shape, an attribute termed "morphological plasticity", to optimize their ability to survive in different environmental conditions, or as part of their life cycle. Why a particular bacterium has a given shape is unclear, given that same ecological niches

harbor various cellular morphologies. Ultimately, shape will be influenced by a combination of factors, including: nutrient availability, attachment strategies, motility requirements, and more. How this diversity of cellular forms evolved remains one of the most fundamental questions in cell biology. The mechanisms that regulate cellular shape changes are beginning to be understood, but the mechanisms by which new morphologies evolved from ancestral ones mostly remain to be described[21].

Another aspect of bacterial cells that was inspected at the context of evolutionary cell biology is cellular division. Most bacteria divide by binary fission using a mechanism that is based on the interaction between the FtsZ protein and the peptidoglycan (PG) biosynthesis machinery. Assembly of FtsZ into a ring structure at the cell division site is the earliest event in cell division. While FtsZ is abundant across the bacterial domain, it is interestingly missing from a superphylum termed PVC (Planctomycetes, Verrucomicrobia, and Chlamydiae). Therefore, these species show a diverse cell division phenotypes including division by budding[22,23].

How a binary fission mechanism based on FtsZ evolved into an FtsZ-independent mechanism of division is a major question in cell biology that could only be asked due to an evolutionary perspective on the bacterial world. Recent studies are beginning to characterize FtsZ-independent division, suggesting that MreB, a bacterial homolog of actin, interacts with PG synthesis enzymes to define the division plane instead of FtsZ and promote division. However, MreB does not seem to be an early cell division protein like FtsZ, because it is recruited late into the process. Thus, FtsZ-independent division mechanisms share similarities but also have important differences with model organisms that use FtsZ – still left to be characterized[22,23].

Neutrality as an evolutionary force that shapes cells and diversity

While natural selection undoubtedly plays a major role in the evolution of cells, other neutral forces are at play as well, which create a drift-barrier that selection must be strong enough to cross in order to drive molecular refinement[24,25]. Interestingly,

stochasticity and neutrality do not only prevent cellular complexity, but can also directly contribute to such intricacy.

The theory of "constructive neutral evolution" (CNE)[26] argues that biological functions can be changed by neutral forces that influence random interactions between cellular factors in the following manner: (i) Consider factor 'A' that performs a certain function. (ii) A stochastic interaction with a factor 'B' occurs that has little or no effect on the activity of factor A. (iii) A mutation occurs in factor A that reduces its activity, but due to its interaction with factor B, the mutation is suppressed, and the activity (and cellular fitness) is maintained. (iv) Subsequent mutations in factor A, and compensatory mutations in factor B, further integrate factor B in the cellular pathway of factor A via a ratchet-like mechanism that may also lead to the recruitment of additional factors[27].

Interestingly, one of the most complex machineries in cells, the spliceosome, is suggested to have emerged in such a neutral process. It has long been suggested that the spliceosome underlying process of RNA editing could have evolved from a simple self-splicing intron. Rather than being the result of selective forces, evolution of the spliceosome can be explained as a product of neutrality processes in which mutations in the self-splicing RNA molecule were masked by interactions with other RNA/protein factors. These RNA-protein interactions accumulated over time, led to the addition of increasing number of factors that eventually originated the spliceosome while maintaining its basic function of RNA splicing. *Complementary to this idea, the last chapter of this thesis deals with other forms of splicing evolution that are based on natural selection.*

Notably, rewiring of regulatory and metabolic networks may also be driven by neutral processes[28,29]. These processes happen when proteins acquire multiple functions due to the prevalence of duplication of entire genes, their regulatory regions, and the promiscuity of many proteins. Subsequent duplications and sub-functionalization events may lead to alternative evolutionary paths that may be driven not only by selection but also by drift[2].

A recent, elegant study confirmed experimentally that neutral processes can increase complexity of a cellular machinery[30]. The V-ATPase proton pump of many yeast strains is composed of three paralogous proteins, yet, ancestral gene resurrection revealed that this structure evolved from a two-paralogue complex. This process occurred via a gene duplication that was followed by loss of specific interfaces by which each daughter copy of the duplication event interacts with other ring proteins – making them obligate components of the complex. The authors confirmed this hypothesis by expressing the common ancestral gene and observed no alterations in the cell's ability to produce the phenotype. This work shows that increased complexity in an essential molecular machine evolved because of simple, high-probability evolutionary processes, without the apparent evolution of novel functions.

Thus, the existence of complex cellular features does not necessitate that all evolutionary changes in such structures were adaptive. Therefore, the task of identifying and characterizing the evolutionary mechanisms that shape cells is more challenging than one might naively perceive - highlighting the benefits biologists can gain from an evolutionary cell biology perspective.

New model organisms provide deeper understanding of cellular diversity

Often, when biologists recognize that a protein is conserved "from yeast to humans", they think that it is universally conserved across eukaryotes. However, eukaryotes are divided into six supergroups, and the popular model organisms (human, fungi, worms, flies, etc.) all belong to a single supergroup, Opisthokonta, and are hence closely related from an evolutionary perspective. This notion raises the possibility that cellular functions that appear to be essential from their presence in conventional model organisms may actually be divergent or lineage specific. Hence, the lack of evolutionary perspective may lead to unjustified extrapolation of cell biological principles[17]. It is thus essential to examine eukaryotes with a wider evolutionary distance if we want to reveal the extent of conservation in the eukaryotic kingdom. Such broader perspective, and the develop of

new model organisms, might reveal more of the diversity among cells, and new cellular functions and evolutionary trajectories of species[31].

In the past, the usage of a nontraditional organism has revealed the molecular mechanism of chromosome maintenance. This was achieved by exploiting the ciliated protozoan *Tetrahymena*, which has a large number of tiny, linear chromosomes. Thus, each cell is more enriched with telomere sequences than is a typical eukaryotic cell, which allowed the identification of CCCCAA repeats at the ends of chromosomes[32].

Today, biologists are developing new model organisms that will hopefully allow us to address cell biology questions that are un-accessible in full with the current model organisms. For example, how animal cells can survive extreme conditions could be thoroughly studied with tardigrades, which were shown to survive freezing to near absolute zero and exposure to the vacuum of outer space[33]. Another example is cellular and whole-body regeneration that are perfected in planaria, which can regenerate any part of their body[34]. The study of this emerging model organism could reveal how cellular and tissue organizations are renewed upon damage, which of course could have immense applications for human health.

Lastly, one of the biggest questions in evolutionary cell biology is the transition from a unicellular organisms to multicellularity[35]. How cells developed the ability to communicate well enough to be perfectly synchronized and build a biological entity which sum is bigger than its individual parts – is not well understood. Some answers to this fundamental question are beginning to emerge with the study of choanoflagellates, the closest living relatives of animals, that can alternate between unicellular and simple multicellular forms[35]. Hence, such further development of new model organisms could be the hallmark of future evolutionary cell biology studies.

Main Thesis Results

Inspired by the ideas of evolutionary cell biology, I devoted my PhD to reveal evolutionary mechanisms of cellular machineries. I combined genome engineering and genomics tools, lab-evolution methodology, and computational analyses to ask various questions about the evolution of cellular machineries of the central dogma.

1| How does the translation machinery evolve?

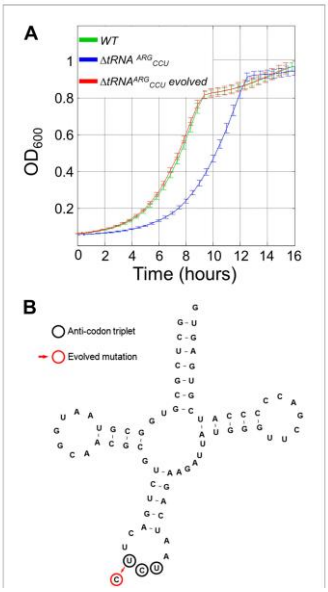
Translation is one of the three cellular processes that form the central dogma of molecular biology, and as such it is highly conserved and regulated across the tree of life[36–38]. The translation process can be viewed as an economic-like model, in which the demand is governed by the usage frequencies of the sense codon, and the supply is governed by the cellular tRNA pool[39]. Specifically, "Translational demand" of a codon is defined by the sum of all the occurrences in which the ribosome translates that codon. Because each codon is represented differently across genes[40], translational demands vary significantly among the 61 sense codons at a given physiological state[41]. Additionally, the demand for a given codon can vary across various physiological states[42]. Further, "Translational supply" refers to the tRNAs that are charged with an amino-acid and are available to translate their corresponding codon(s). It has been widely documented that codon demand and tRNA supply are correlated with one another, leading to a translational balance[39,41,43]. While various regulatory mechanisms of translational elongation have been previously reported[44–46], the implication of the translational balance on synthesis of new proteins and its potential effects on the cellular proteome have not been investigated yet. Hence, we asked how translational imbalance between supply and demand affects the adequacy of the translation machinery and why translational balance exists as a wide biological phenomenon.

At the early stages of my PhD we studied how the translational machinery adapts by perturbed the tRNA pool of the yeast *Saccharomyces cerevisiae* with a tRNA gene deletion. While ribosomal genes do not exhibit appreciable changes in response to environmental alterations[36–38], tRNA genes may provide an important source of

evolutionary plasticity for fine-tuning translation. tRNAs constitute a fundamental component in the process of translation, linking codons to their corresponding amino acids[47]. tRNA genes are classified into gene families according to their anticodon, with each gene family containing between one to several copies scattered throughout the genome. Evolutionary changes to the tRNA pool were appreciated mainly via bioinformatics studies[48–52] and only a handful of experimental findings have been reported, which rely on genetic manipulations[53–55] or direct mutagenesis[56]. However, the fitness effects of an unmet translational demand and its potential role in shaping the tRNA pool were not fully characterized.

Deletion of singleton tRNA gene breaks the translational balance

To demonstrate the importance of the balance between codon usage and the cellular tRNA pool we created a yeast strain in which the single copy of the arginine tRNA gene, *tR(CCU)J*, was deleted (designated $\Delta tRNA^{Arg}_{CCU}$). Consequently, in this deletion strain the arginine codon AGG cannot be translated with its fully-matched tRNA and it is presumably translated by another arginine tRNA, $tRNA^{Arg}_{UCU}$, owing to a wobble interaction[57]. Indeed, the $\Delta tRNA^{Arg}_{CCU}$ strain showed a severe growth defect compared to the wild-type strain (Figure 1A). This growth difference demonstrates the effect of translational imbalance on cellular growth. Although the deletion mutant of this single copy tRNA is viable[58,59], its severe growth defect also reveals the inability of the wobble interactions to fully compensate for the tRNA gene deletion.



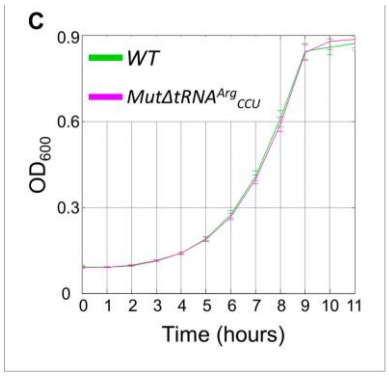


Figure 1. The growth defect associated with deletion of a singleton tRNA gene was rapidly rescued during the lab-evolution experiment. (**A**) Growth curve measurements of wild-type (WT) (green), ΔtRNA$^{Arg}_{CCU}$ (blue) and the evolved deletion (red) are shown in optical density (OD) values over time during continuous growth on rich medium at 30°C. (**B**) The mutation that was found to recover the deletion phenotype in the evolved strains is shown on the secondary structure of tRNA$^{Arg}_{UCU}$. The UCU anticodon nucleotides are marked with black circles, and the red circle indicates the mutation that occurred in the anticodon, that is T→C transition. (**C**) *MutΔtRNA*$^{Arg}_{CCU}$ in which the same mutation that was found in the evolved strain was deliberately engineered, exhibits similar growth as the WT. Growth curve measurements of WT (green) and of *MutΔtRNA*$^{Arg}_{CCU}$ (magenta) are shown in OD_{600} values over time during continuous growth on rich medium at 30°C.

The tRNA pool can rapidly evolve to meet translational demands

To learn how genomes adapt to translational imbalances, we performed lab evolution experiments on the ΔtRNA$^{Arg}_{CCU}$ strain, employing the procedure of daily growth and dilution to fresh a medium[60]. Strikingly, after ~200 generations we observed a full recovery of the growth defect of the ancestor strain ΔtRNA$^{Arg}_{CCU}$, as the growth curve of the evolved population was indistinguishable from that of the wild-type strain (Figure 1A).

In search of the potential genetic adaptations underlying this rapid recovery, we first looked for genetic alterations in other arginine tRNA genes. We found a single point mutation in another arginine tRNA gene which codes for tRNA$^{Arg}_{UCU}$. This mutation changed the anticodon triplet of tRNA$^{Arg}_{UCU}$ from UCU to CCU (i.e. T→C transition). Consequently, the evolved tRNA$^{Arg}_{UCU}$ perfectly matches the AGG codon (Figure 1B). Unlike the singleton tRNA$^{Arg}_{CCU}$, there are 11 copies of tRNA$^{Arg}_{UCU}$ in the yeast genome. Although each of the four-independent lab-evolution experiments showed the exact same solution, i.e. a mutation in the anticodon of a tRNA$^{Arg}_{UCU}$ gene, three different copies of this gene were changed among the four lines. To confirm that a single point mutation in the anticodon of tRNA$^{Arg}_{UCU}$ is sufficient to fully compensate for the growth

defect of $\Delta tRNA^{Arg}_{CCU}$, we artificially inserted the same T→C mutation into the deletion $\Delta tRNA^{Arg}_{CCU}$ mutant. We inserted the mutation into one of the 11 copies of the $tRNA^{Arg}_{UCU}$ genes, a copy that resides on chromosome XI, and was spontaneously mutated in one of the evolution lines. Indeed, the artificially mutated strain, termed here *Mut$\Delta tRNA^{Arg}_{CCU}$*, showed a full recovery of the deletion adverse phenotype (Figure 1C). This indicates that the T→C mutation in the anticodon is sufficient for the full recovery of the $tRNA^{Arg}_{CCU}$ deletion phenotype.

Anticodon switching is a widespread phenomenon in nature

Although the anticodon of a tRNA gene was rapidly mutated in our laboratory conditions, it is unclear to what extent this mechanism naturally occurs in species across the tree of life. To address this question, we performed a systematic bioinformatics screen for tRNA switching events in nature. We defined an anticodon switching event as a case of a tRNA whose nucleotide sequence is closer to a tRNA gene with a different anticodon than to a tRNA gene with the same anticodon. To this end, we downloaded all the known tRNA sequences from the Genomic tRNA Database[61], a collection that stores the tRNA pools of 524 species. We masked the anticodon triplet as "NNN" in all tRNA genes, aligned all tRNA sequences from each species individually and inferred a maximum likelihood phylogenetic tree of each alignment. For each tRNA sequence, we calculated the shortest phylogenetic distance to another tRNA with the same anticodon (designated d_{same}) and the shortest distance to another tRNA with a different anticodon (designated d_{diff}). For each species, we defined its set of tRNA switching events as those in which $d_{diff} < d_{same}$.

Our analysis included 416 eubacterial, 68 eukaryotic and 40 archaeal species. We found that tRNA switching events are present in all domains of life, as we detected at least one tRNA switching events per species in 8 bacteria, 58 eukarya and 1 archaeal species (Figure 2A). A retrospective counting revealed that most switching events occurred due to a mutation in the first position in the anti-codon triplet that corresponds to the 3rd codon position. For comparison we masked as 'NNN' additional triplets of nucleotides within the

tRNA molecule and found higher percentage of discrepancies compared to the anticodon triplet.

Figure 3 demonstrates two examples of tRNA switching events, the first in *Mus musculus* and the second in *Homo sapiens*. In the first example, the phylogeny of tRNA sequences with glutamic acid anticodons is presented (Figure 2B). Notably, 6 out of the 8 tRNAs with a UUC anticodon in *M. musculus* were clustered together in our analysis, while 2 other copies of the same anticodon identity were clustered closer to tRNA genes with a CUC anticodon (Figure 2C). The second example demonstrates a switching event for tRNA genes encoding for valine anticodons. Here, a tRNA with a UAC anticodon was clustered with CAC and AAC tRNA genes and not with the other 4 UAC tRNAs (Figure 2D+E). Interestingly, the CAC and AAC tRNA genes are intermixed in the tree, suggesting that anticodon switching was prevalent in the evolution of CAC and AAC tRNA genes in *H. sapiens* (Figure 2D+E). Notably, the switching events shown in mouse were not found in human and *vice versa*. Thus, in each of these two mammals the switching examples shown here probably occurred after they split from their common ancestor. In general, inspecting the relationship across species between the size of the tRNA pool and the number of detected switching events revealed a modest correlation, and in particular species with same size of tRNA repertoire manifested tRNA switching to different extents.

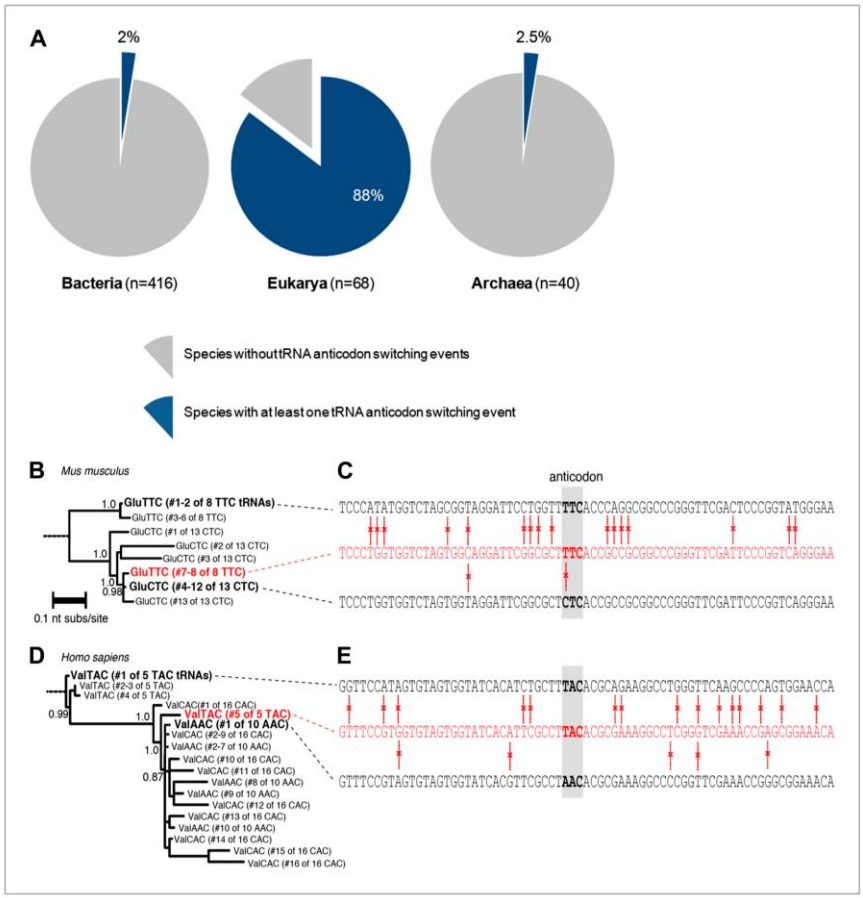


Figure 2 Anticodon switching is a widespread phenomenon in nature. (**A**) Number of species with at least one tRNA switching event in each domain of life. (**B**) The anticodon UUC convergently evolved in *Mus musculus*. A maximum likelihood phylogeny of tRNA sequences in *M. musculus* that decode glutamic acid (Glu) codons. Branch lengths express average nucleotide substitutions per site. Decimals on internal branches express branch support. (**C**) A comparison of nucleotide sequences for glutamic acid tRNA genes in *M. musculus* with anticodon UUC (top, tRNA1547 and tRNA359), 'switched' UUC tRNAs (middle, tRNA286 and tRNA754), and CUC tRNAs (bottom, tRNA1002, tRNA745, tRNA303, tRNA999, tRNA996 tRNA709, tRNA1001, tRNA1912 and tRNA81). The anticodon triplet is boxed in gray. Red vertical bars indicate differences between sequences. (**D**) The anticodon UAC convergently evolved in *Homo sapiens*. A maximum likelihood phylogeny of tRNA sequences in *H. sapiens* encoding for valine (Val) is shown. (**E**) A comparison of nucleotide sequences for *H. sapiens* tRNAs with anticodons UAC (top, tRNA6), a 'switched' UAC tRNA (middle, tRNA40), and an AAC tRNA (bottom, tRNA136). The number of genes is according to the tRNA database.

In conclusion, Genomic duplications, deletions and anticodon mutations shape tRNA gene families, yet the evolutionary scenarios that trigger changes in the tRNA pool have not been thoroughly explored. In our evolution experiments, a translational imbalance was imposed by a tRNA gene deletion, which compromised growth and drove the tRNA pool to adapt to a novel translational demand. Importantly, organisms may experience equivalent imbalances when their gene expression changes due to altered environmental conditions or upon migrating to a new ecological niche[42]. This scenario is particularly feasible given that the genes needed in various environments do show differences in codon usage, e.g. respiration as opposed to fermentation in yeast[62].

When facing the need to adapt, the tRNA pool, i.e. the supply, provides evolutionary plasticity to the translation machinery. The ability of the tRNA pool to change rapidly can be mainly attributed to its unique architecture in the form of multi-member gene-families. Only on a much longer evolutionary time-scale, will the genome-wide codon usage of genes change so as to further fine-tune the translational balance. Notably, the plasticity of the tRNA genes is constrained by the need to maintain proper protein folding[63]. Thus, the need to accommodate changes in codon usage demands acts together with protein folding constrains to shape the tRNA pool in living cells.

Codon Usage of Highly Expressed Genes Affects Proteome-Wide Translation Efficiency

Later, we decided to ask what drives the evolution of codon usage bias in cells. Though differential codon usage can result from neutral processes of mutational biases and drift[64–66], certain codon choices could be specifically favored as they increase efficiency[41,44,46,67,68] or accuracy[63,69–72] of protein synthesis. These forces would typically lead to codon biases in a gene because they locally exert their effect on the gene on which the codons reside. In addition to such *cis* effects, it is possible that codon usage also acts in *trans*, namely, that codon choice of some genes would affect translation of others due to a "shared economy" of the entire translation apparatus[73–75]. Previous theoretical works have suggested that increase in elongation rate may reduce the number of ribosomes on mRNAs and therefore may indirectly increase the rate of initiation of other transcripts due to an increase of the pool of free ribosomes[65,76]. In addition, a recent computational study in yeast has also examined the indirect effects of synonymous codon changes on the translation of the entire transcriptome[77]. Yet, experimental evidences of such changes are absent. Here we ask how manipulating the frequency of a single codon on a small subset of genes influences the synthesis of other proteins.

To tackle this question, we replaced common codons with a synonymous, rare counterpart in several highly expressed genes. We then asked how this massive change in the codon representation in the transcriptome would affect the manipulated genes, other genes, and the physiology and well-being of the cell (Figure 1).

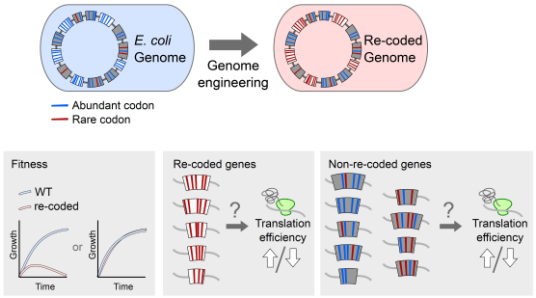


Figure 1 – Does the codon usage of a sub-set of genes affect translation efficiencies of other genes?
Upper panel: Hypothetical genomes of wild-type and re-coded strains are shown. Using genome engineering, we replaced abundant codons ("origin codon", blue lines) with rare codons ("destination codon", red lines) in highly expressed genes (white background).
Bottom left: two potential effects of re-coding on fitness: either reduce, or not affect the fitness.
Bottom middle: The translation efficiency of re-coded genes could be increased, decreased or not changed at all.
Bottom right: The translation efficiency of non-recoded genes that have the origin (blue) or destination (red) codon could be increased, decreased or not changed at all.

Codon usage manipulation leads to proteome-wide changes in translation efficiencies in a codon-dependent manner

We manipulated the frequency of the arginine codon CGG since it is the only codon in *E. coli* that is translated by a single-copy tRNA gene, and whose tRNA does not translate other codons (see Figure 2 for codon-anticodon interactions for CGN codons in E. coli)[43]. Using genome editing, we were able to introduce 60 synonymous mutations into a single genome of an E. coli strain that converted CGU and CGC ("origin codons") to CGG ("destination codon"). To maximize the effects of our manipulations, we introduced synonymous mutations in the eight genes with the highest ribosome-profiling occupancy score that are not essential and that do not relate directly to translation functions. Following our manipulation, the translation demand for the CCG anticodon is elevated by ~3.5-fold and our re-coded genes constitute ~70% of the new total demand for this codon in the cell.

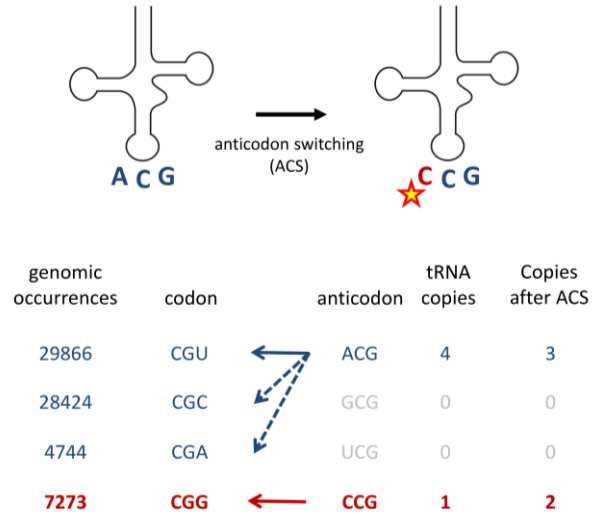


Figure 2 – The arginine CGN box
We re-coded CGU and CGC ("origin codons") to CGG ("destination codon"). In *E. coli*, both origin codons are translated by $tRNA^{ACG}$ with the anticodon ACG due to an A-to-I modification that is mediated by the enzyme tRNA-specific adenosine deaminase (tadA). The destination codon is solely translated by $tRNA^{CCG}$, which translate no other codons. $tRNA^{ACG}$ and $tRNA^{CCG}$ appear in the genome with four and one copies, respectively. A direct arrow symbolizes fully-match interactions between codon and anticodon, while dashed arrows represent wobble interactions, which are enabled by modifying the ACG anticodon to ICG.

We then asked how does our manipulation on the CGG representation in the transcriptome influence translation efficiency in the cell. To this end, we analyzed the transcriptome (by RNA-sequencing) and proteome (by mass-spectrometry) of the original wild-type and the re-coded strains (each strain was analyzed with three independent repetitions for both the transcriptome and proteome). Then, we calculated the translation efficiency of each gene by normalizing the protein level to its corresponding mRNA level.

Notably, only one of the eight recoded genes showed reduced translation efficiency (Figure 3A), suggesting that the effects of our codon-usage manipulation on the genes that harbor the manipulation are weak. A possible reason for this weak effect is that in the current experiment only a single codon type has been manipulated in each re-coded gene, in contrast to prior studies in which entire ORFs have been manipulated[78,79]. It is also possible that our manipulations did affect translation efficiency in *cis*, though some compensatory effect, e.g. acting on the initiation level, may have acted to counter-act the reduction in elongation.

We postulated that the increased usage of CGG at the expense of the CGU and CGC codons might reduce the translation efficiency of other genes in the genome, which were not mutated, in particular genes that naturally have high usage of CGG. Indeed, we

observed 455 genes with increased and 566 genes with decreased translation efficiency at a fold change of above or below 1.5 in the re-coded strain compared to the wild-type (Figure 3A). Strikingly, genes with high occurrences of the CGG codon (>5 occurrences) that were not engineered by us demonstrated lower translation efficiencies in the re-coded strain compared to the WT strain, compared to genes that do not use this codon (Figure 3A inset). This observation suggests that our CGG codon manipulation affected in *trans* the translation of other, non-recoded genes in the re-coded strain. In support of this result, the hundreds of genes that showed reduced translation efficiency demonstrated higher occurrences of the CGG codon compared to the genes with increased translation efficiency (Figure 3B). On the other hand, we observed that genes with increased translation efficiency were enriched with the CGU, CGC, and CGA codons (Figure 3C). We thus conclude that the increased demand on the CGG codon due to our recoding reduced the translation efficiency of genes that were enriched with this codon, while the relief of demand from the CGU, CGC, and CGA codons increased the translation efficiency of genes that utilize these codons. While most studies measure the resulted change in expression level of a gene whose different codons were synonymously manipulated[78,79], our results demonstrate for the first time how a frequency manipulation of a codon can affect global translation patterns by changing the translation efficiency of other genes according to their codon usage.

Theory predicts that changes of elongation rate should have the largest expression effects on genes with high rates of translation initiation because these genes are more likely to suffer from traffic jams and ribosomal collisions[44,77]. Thus, we hypothesized that genes with reduced translation efficiency in the re-coded strain should have higher translation initiation rates compared to genes whose translation efficiency did not decrease. Indeed, reduce translation efficiency genes demonstrate higher initiation rates as calculated with the Ribosome Binding Site Calculator[80] compared to un-effected or increased translation efficiency genes (Figure 3G). The observations that genes with reduced translation efficiency are more enriched with the CGG codon, on one hand, and have higher initiation rates on the other, strengthens our conclusion that the re-coded strain suffers from

ribosomal elongation changes compared to WT cells. In line with theoretical predictions[44,77], increasing the dwell time of ribosome during elongation reduces translation efficiency provided that initiation rate is sufficiently high.

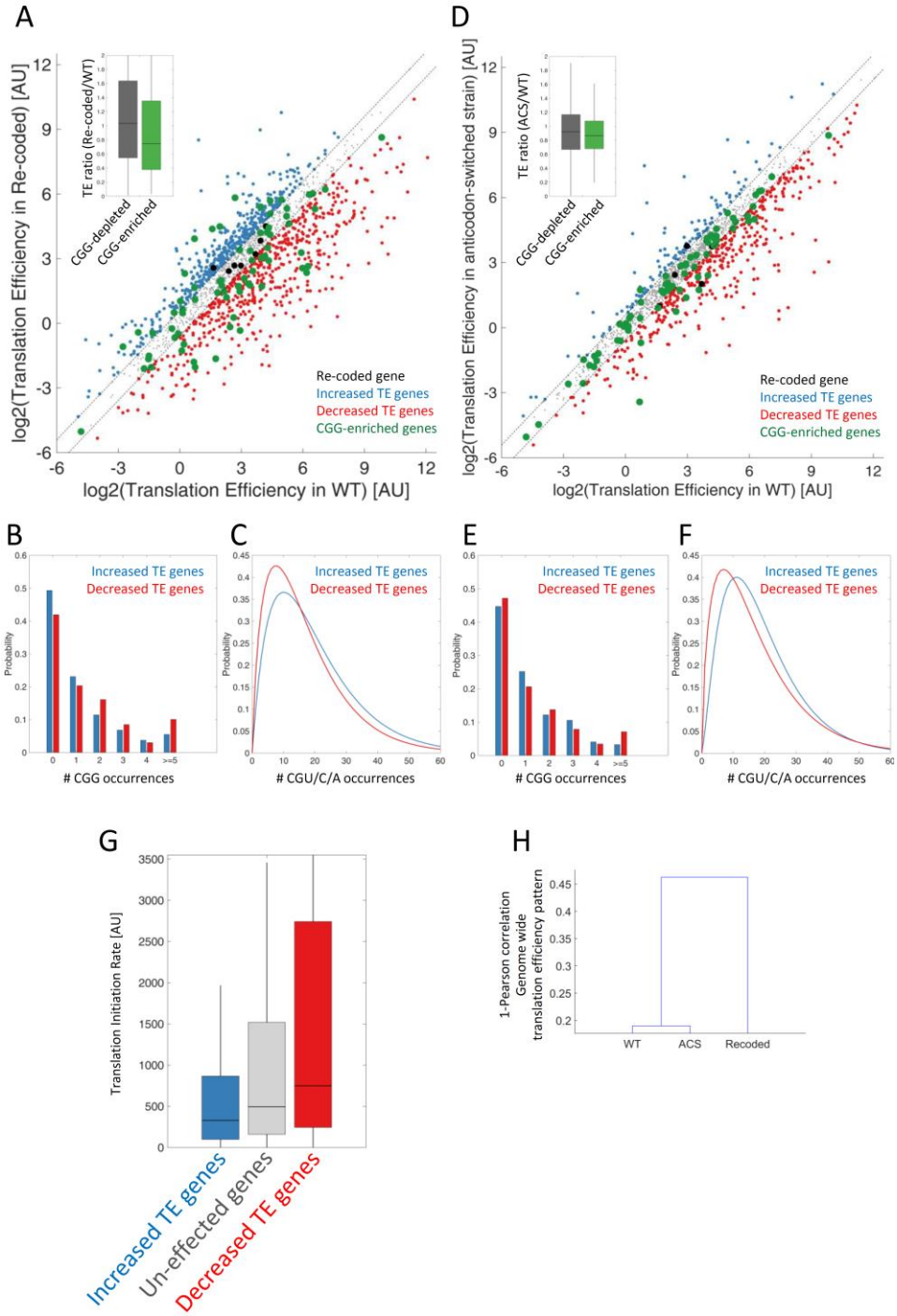


Figure 3 – legend on next page.

Figure 3 – Manipulating codon frequency of CGG results in global translation efficiency changes

A| We carried RNA-sequencing analysis of the transcriptome and mass-spectrometry analysis of the proteome for both the wild-type and re-coded strains. This allowed us to calculate translation efficiency (protein/mRNA) for each gene and classify two gene groups of increased or decreased translation efficiency with a fold change threshold of 1.5. The eight re-coded genes are colored in black, increased translation efficiency group is colored in blue, decreased translation efficiency group is colored in red and CGG-enriched genes are colored in green.

Inset| Ratios of translation efficiency between re-coded and wild-type cells for CGG-enriched genes (>5 occurrences) and CGG-depleted genes (no occurrences). CGG-enriched genes show lower translation efficiency ratios, p-Value=0.01.

B| Distribution of CGG occurrences, translated by $tRNA^{CCG}$, for increased (blue) or decreased (red) translation efficiency genes in re-coded strain compared to the wild-type strain. The group of decreased translation efficiency genes demonstrates higher CGG occurrences (p-Value=0.0018).

C| Distribution of CGU+CGC+CGA occurrences, all translated by $tRNA^{ACG}$, for increased (blue) or decreased (red) translation efficiency genes in re-coded strain compared to the wild-type strain. The group of increased translation efficiency genes demonstrates more codon CGU+CGC+CGA occurrences (p-Value=$6.79*10^{-5}$).

D| To increase $tRNA^{CCG}$ supply, we mutated the anticodon of $tRNA^{ACG}$ from ACG to CCG on the background of the re-coded strain, and termed this new strain as anticodon switched strain. We then analyzed its transcriptome and proteome. Note that much less genes are now deviating from the diagonal, particularly the CGG-enriched genes in green, suggesting that the anticodon switching mutation alleviated the translational difficulty of the re-coded strain. Color code is the same as in A.

Inset| CGG-enriched genes now show similar translation efficiency ratios as CGG-depleted genes, p-Value>0.05.

E| Same as B, but for the increased and decreased translation efficiency genes in anticodon-switched strain compared to the wild-type strain. In contrast to the previous comparison in B, these two groups utilize the CGG codon to the same extend (p-Value>0.05).

F| Same as C, but for the increased and decreased translation efficiency genes between the wild-type and anticodon-switched strain. In contrast to the previous comparison in C, these two groups utilize the CGU+CGC+CGA codon to the same extend (p-Value>0.05).

G| Translation initiation rates for increased, decreased and un-affected genes between re-coded and wild-type strains, as defined in A. Note that decreased translation efficiency genes, which are also enriched with CGG, also show higher initiation rates (p-Value=0.01) – in agreement with theory's prediction.

H| The translation efficacy pattern of the anticodon-switched strain clustered closer to the wild-type strain and away from the re-coded strain.

Proteome-wide changes in translation efficiencies are alleviated by increased tRNA supply

To confirm our hypothesis that the changes in translation efficiencies resulted from the increased cellular demand for $tRNA^{CCG}$, the tRNA which translates CGG, we decided to elevate the availability of this tRNA and examine the effect on the translation phenotype.

We, and others, have recently shown that a mechanism to increase tRNA availability is a mutation in the anticodon that changes the codon specificity of the tRNA[81,82]. We have shown that such anticodon switching mutations can maintain the functionality of tRNA genes, and are utilized by many species as an adaptive mechanism of the cellular tRNA pool.

Thus, we mutated the anticodon of one of the four copies of $tRNA^{ACG}$ gene from ACG to CCG on the background of the re-coded strain (Figure 2). We then analyzed the transcriptome and proteome of this anticodon-switched strain (based on three independent repetitions) and compared it to both the re-coded and wild-type strains. Strikingly, although the genome of the anticodon-switched strain is more similar to the re-coded strain, its global translation efficiency pattern clustered together with the wild-type strain and away from the re-coded strain (Figure 3H). This observation suggests that manipulating the tRNA pool of the re-coded strain restored translation efficiency of genes back to their normal states.

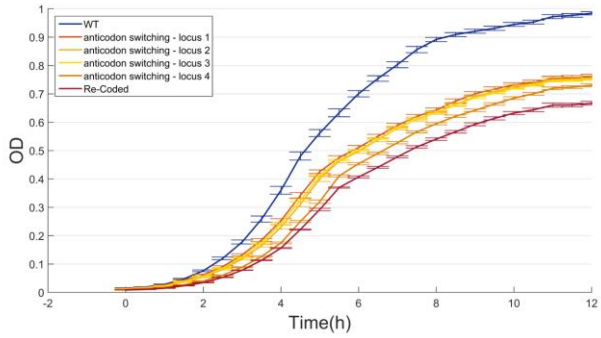
Indeed, only 124 and 408 genes with increased or decreased translation efficiency were respectively identified between the wild-type and the anticodon-switched strains (Figure 3D), further demonstrating that the translation efficiency defect in the re-coded strain was alleviated upon anticodon switching. Strikingly, while CGG-enriched genes particularly tended to have reduced translation efficiencies in the re-coded strain, they demonstrated similar efficiencies to the wild-type in the anticodon-switched strain, and the difference in translation efficiency ratios between these genes and CGG-depleted genes was not observed (Figure 3D inset). Consistently, the genes with increased or decreased translation efficiency between the wild-type and anticodon-switched strain demonstrated the same distribution of codon occurrences for CGG or CGU+CGC+CGA (Figures 3E+F). These observations suggest that the additional supply of $tRNA^{CCG}$, at the expense of $tRNA^{ACG}$ in the anticodon-switched strain, resulted in a more efficient translation of CGG-enriched genes.

Increased codon usage of a rare codon reduces cellular fitness due to excessive use of tRNA molecules

The physiological effects between the wild-type and re-coded strains encouraged us to ask whether these global translation efficiency changes disturb cellular growth and reduce fitness. We thus tested whether introducing the rare codon CGG on highly expressed genes is deleterious to the cell. We compared the growth of the wild-type and re-coded strains and observed that the re-coded strain suffers from a growth defect (Figure 4A). We used a recent logistic growth model[83] that calculates relative fitness from growth curves and observed that the relative fitness of the re-coded strain is 0.87 compared to the wild-type strain.

We next hypothesized that the growth reduction of the re-coded strain is the result of a lack in sufficient tRNA supply that leads to changes in translation efficiency of many genes. However, cellular fitness could also be affected by the off-target mutations that the re-coded strain accumulated following our genome engineering efforts. To test our hypothesis, we compared the growth of all four anticodon-switched strains, in which $tRNA^{CCG}$ levels are increased, and observed that they all demonstrated increased relative fitness in comparison to the re-coded strain (Figure 4A). Importantly, when the same anticodon mutation was inserted on the background of the wild-type strain, a reduction in relative fitness was observed (Figure 4B). These results suggest that introducing a rare codon on highly expressed genes reduces cellular fitness not because of its effects on the manipulated genes themselves, but as it hampers translation of other genes due to an excessive use of tRNA molecules and result in global physiological perturbations.

A



B

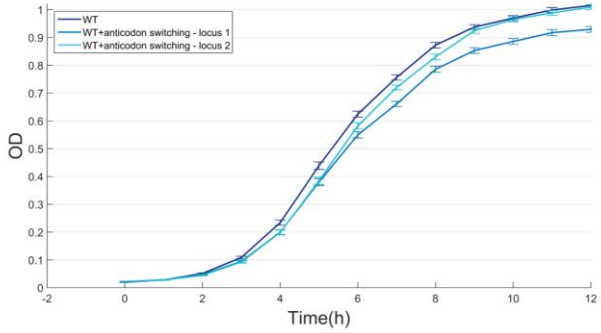


Figure 4 – Change in global translation efficiency patters is deleterious

A| Growth experiment (OD vs. time) of the wild-type strain (blue), the re-coded strain (red) and the four anticodon-switched strains ($tRNA^{ACG}$ argQ in dark orange, $tRNA^{ACG}$ argZ in dark yellow, $tRNA^{ACG}$ argY in bright yellow, $tRNA^{ACG}$ argV in bright orange). The re-coded strain demonstrates reduction in relative fitness to 0.87 compared to the wild-type strain (p-Value<10^{-10}). The four strains with anticodon switching (increased $tRNA^{CCG}$ supply) on the background of the re-coded strain demonstrate a higher fitness compared to the re-coded strain itself, demonstrating that restored translation efficiencies patters also alleviated the growth defect (relative fitness compared to re-coded strain of switched argQ = 1.06, argZ=1.08, argY=1.02 and argV=1.04).

B| Switching the anticodon of $tRNA^{ACG}$ from ACG to CCG on the background of the wild-type strain reduces fitness (relative fitness compared to wild-type strain of switched argQ = 0.95 and argZ=0.96).

In conclusion, this work raises the question of whether changes in global translation efficiencies could pose a challenge to the translation machinery both physiologically and evolutionarily. Previous works have demonstrated how codon-to-tRNA balance reacts to changes in the environment[84–86], to the formation of cancerous tumor[87], or to an evolutionary challenge[81,88]. In agreement with these works, we observed that the re-coded strain suffers from a growth defect, providing a need for selection to optimize the translation economy in the cell. Interestingly, we could alleviate these translation and growth phenotypes by providing more tRNA supply that could meet the new CGG demand. Thus, our work demonstrates that codons and tRNA genes may co-evolve not only to tune the expression level of individual (highly expressed) genes, but also to maintain the efficiency of global protein translation in the cell.

2| How do cells minimize the cost of gene expression?

Cells express different genes in a regulated manner at levels that maximize the benefit from the gene's product on one hand, and minimize the production costs of transcription and translation on the other hand[89,90]. Costs of mRNA and protein expression originate from spending cellular resources such as building blocks of polymers (amino acids and nucleotides), from allocation of cellular machineries (RNA polymerase and ribosome), and from energy and reducing power consumption[91–94]. Understanding what molecular processes determine expression cost, its relation to both cellular growth and gene regulation, and how costs evolutionarily shape the genome - is a key aspect of cell biology that remains largely elusive. While numerous studies investigated molecular mechanisms and gene sequence architectures that regulate expression level[39,73,78,84,95], very little is known about design elements that govern expression costs.

Different works have studied expression costs in unicellular organisms by imposing the expression of an unneeded protein, such as GFP[89,91,96–99]. The production of such unneeded proteins diverts resources from synthesis of other, functional proteins and decreases cellular fitness[100–102]. Central to this body of work is the characterization of the correlation between the imposed expression level of the unneeded proteins to the cost. Yet, ultimately natural selection dictates the expression level of natural genes according to the required concentration of a protein in its cellular localization. Thus, a fundamental question, which has not been addressed before, is how cells can achieve a specific expression level of a gene while minimizing its expression cost.

This question has not been addressed before because changes in sequence could affect concomitantly both expression level and expression costs. To disentangle expression level and expression costs, and expose mechanisms that affect cost per protein molecule, we utilized a synthetic reporter library of ~14,000 different sequence variants, each fused up-stream to a GFP gene[79]. We then combined competition assays and deep-sequencing to measure the fitness of all variants in parallel, a procedure that enabled us to elucidate gene architectures that minimize expression cost at a given protein expression level.

5’ gene-architecture affects cost of gene expression

We ask whether different sequence elements that compose the gene architecture can minimize cost of expression per protein molecule. We focus on sequence features at the 5' region of a gene by utilizing a previously published, synthetic gene library[79] composed from ~14,000 different variants expressing a GFP gene. Each variant in the library holds a unique variable 5’ gene architecture that includes a promoter, RBS and an 11-amino acid long N-terminus fusion (Figure 1A).

To reveal the expression cost of each variant we measured relative fitness of all variants in parallel in a competition assay in six independent biological repeats. We then deep-sequenced the variable region of the pool of variants at several time points, and calculated relative fitness of each variant (Figure 1B).

We regressed fitness values against GFP expression levels and observed a negative, linear correlation (Figure 1C). The linear decline in fitness with expression is in agreement with previous studies[97,99], though others observe more-than-linear decline, especially at very high expression[89]. The regression line, which outlines the relations between fitness and expression, allowed us to estimate the expected fitness for each library variant according to its GFP expression level. Variants whose fitness does not deviate consistently across repeats from this regression line are deduced not to utilize mechanisms that enhance or reduce the production cost per protein molecule.

Yet, many variants did deviate from the linear-regression line, demonstrating fitness that is higher or lower than expected given their GFP expression levels. We hypothesized that variants that repeatedly deviated from the expected fitness might utilize gene architectures that either reduce or increase the cost of GFP production per protein molecule. Hence, we calculated for each variant its “fitness residual”, which we defined as the difference between the fitness measured in our experiment for the variant and the fitness expected for that variant according to its GFP expression level and the linear regression between fitness and GFP expression level (Figure 1C). A positive fitness residual means that a given variant showed higher fitness than expected by its GFP

expression level, suggesting that it can produce this GFP level with lower costs. A negative fitness residual means that the variant showed lower fitness than expected given its GFP expression level.

We then classified each variant as either positive or negative according to its fitness residual sign (Figure 1C). Since the observed fitness residual is sensitive to biological noise (*i.e.* drift during competition) and experimental errors (*i.e.* sampling errors), we only classified variants as positive or negative if their fitness residual sign was identical in at least five out of the six repeats of the experiments in each of the two final sampling points of the populations. This approach resulted in 975 positive and 815 negative variants (significantly higher than expected by chance even at very high levels of measurement errors). Indeed, classifying variants to either positive or negative fitness-residual groups allowed us to eliminate the effect of GFP expression level on fitness as these two groups demonstrate the same expression distribution (Figure 1C, inset).

While inspecting fitness residuals, we noticed a set of 80 library variants, which we termed 'underachievers', and whose fitness residual scores were repeatedly at the bottom 5% of the entire library (Figure 1C). There appeared to be no 'overachievers' in these data. We hypothesized that these underachiever variants show extremely low fitness residuals because they produce GFP even more wastefully, and we expected them to show stronger usage of low-efficiency gene architectures compared to the negative fitness residual group.

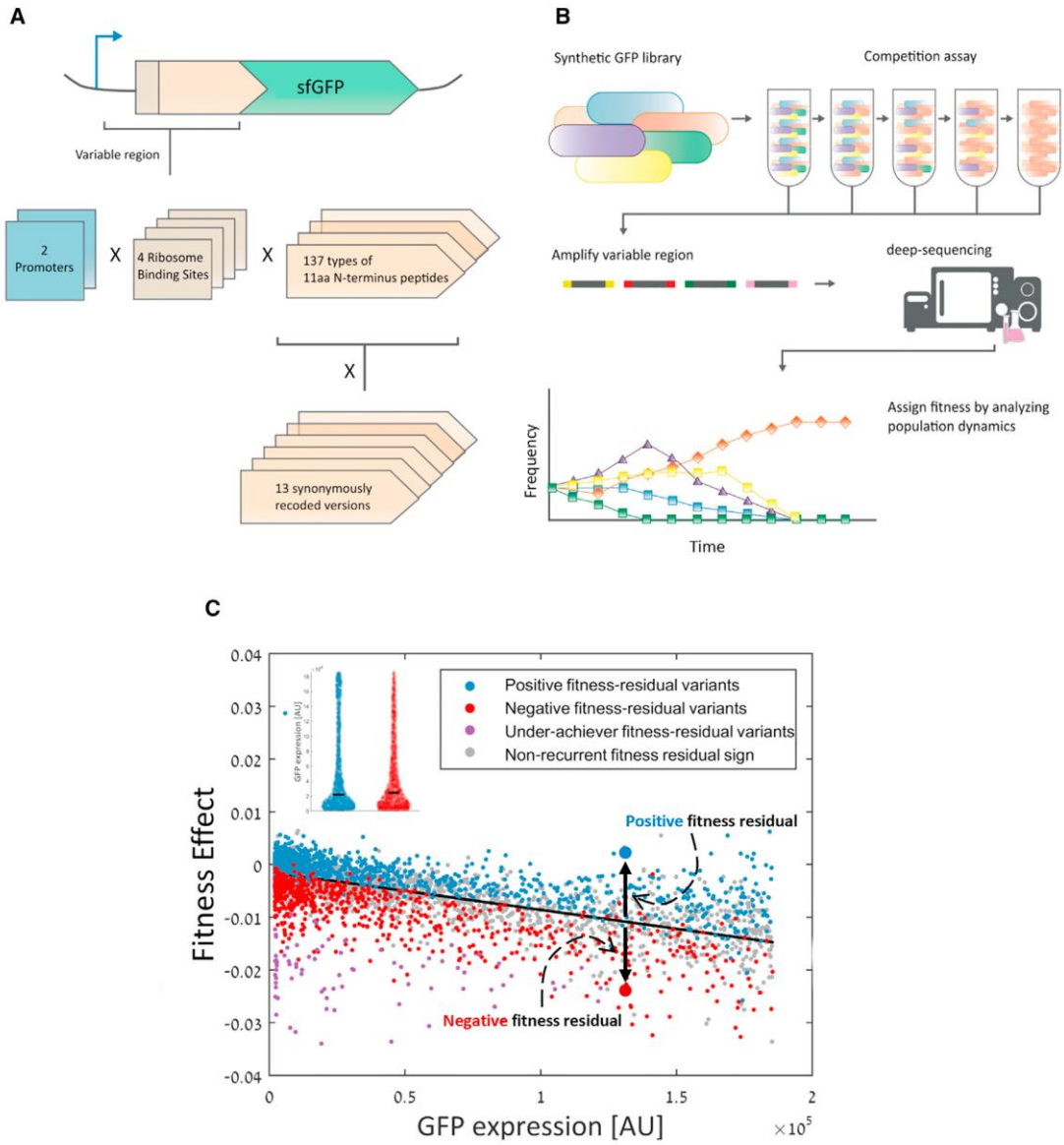


Figure 1 – 5' gene architectures affect cost of gene expression at a given expression level

A| We utilized a synthetic library of ~14K *E. coli* strains, each expressing a GFP construct with a unique 5' architecture that includes a promoter, ribosome binding site (RBS) and an 11-amino acid fused peptide. There were two different promoter types, four RBS and 137 amino acid fusions that were each synonymously re-coded to 13 different versions (see *Goodman et al*. for full details).

B| FitSeq methodology to measure relative fitness of strains in a pooled synthetic library: First, the library was grown six independent times for ~84 generations and samples were taken at generations 0, ~28, ~56 and ~84. Then, unique 5' gene architectures were simultaneously amplified and sent for deep-sequencing, which allowed following the frequency of each variant in the population over the course of the experiment. Finally, a relative fitness score was assigned for each variant based on its frequency dynamics.

C| GFP expression level (as measured by *Goodman et al.*, x-axis) vs. fitness effect (based on results of repetition C, y-axis) of each variant in the library (Pearson correlation r=-0.79, p-Value<10^{-200}). Fitness effect comes from the burden of expressing unneeded proteins on cellular growth, and is calculated by analyzing the frequency dynamics of each variant (see Methods). We defined fitness residual as the difference between a variant's observed and expected fitness. The expected fitness is calculated from the regression line between GFP expression and fitness. Some variants consistently demonstrated positive (blue dots, n=975) or negative (red dots, n=815) fitness residual sign. Other variants showed extremely low fitness residual, and we termed those variants as "underachievers" (purple dots, n=80). The group size of positive, negative and underachiever variants are significantly much higher than expected by chance (Supplementary File 1). These results suggest that certain 5' gene architectures can increase or reduce the cost of gene expression.
Inset: positive (blue violin-plot) and negative (red violin-plot) fitness residual variants come from the same distribution of GFP expression level (Wilcoxon rank-sum p-Value=0.46). Black line represents the median value. Thus, the effect of GFP levels on fitness was successfully factored out, thus allowing us to elucidate other molecular mechanisms that tune expression cost at given expression levels.

Production of more proteins per mRNA molecule is an economic regime that minimize expression costs

We first hypothesized that reaching the same GFP level with lower levels of mRNA of the GFP gene could be beneficial. We compared GFP mRNA levels between positive and negative fitness residual variants and observed that positive variants demonstrated lower levels (Figure 2A). The observation that positive variants have equal GFP protein levels but lower GFP mRNA levels indicates that they are able to produce more GFP proteins per mRNA molecule. Since initiation rate is usually rate limiting in translation[80], we postulated that high translation initiation rate could be a mechanism for maintaining same GFP levels while keeping low mRNA levels in positive variants. We calculated initiation rates for all library variants using a common prediction model, the "Ribosome Binding Site (RBS) Calculator"[80], and observed that indeed positive variants had higher initiation rates (Figure 2B). Indeed, when examining translation efficiency per variant (using measured protein levels divided by mRNA levels), positive variants demonstrated higher translation efficiencies than negative fitness ones (Figure 2C). Moreover, we found that underachiever variants demonstrated even higher mRNA levels and lower translation efficiencies compared to the negative variants (Figures 2A and 2C). These observations suggest that by increasing translation efficiency, cells can reduce transcription costs and ultimately reduce costs per protein molecule.

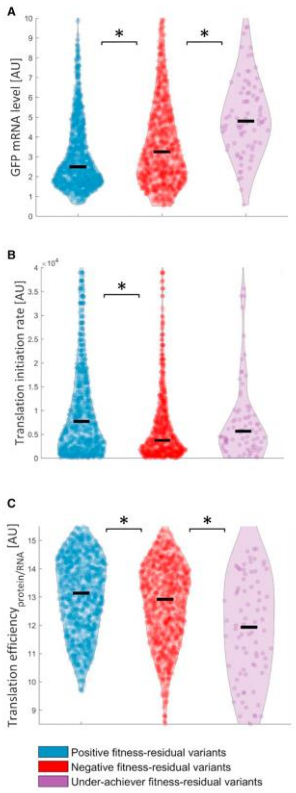


Figure 2 – Higher ratio of GFP protein/mRNA minimizes cost of gene expression

A| Although coming from the same distribution of GFP levels, positive variants (blue violin-plot) demonstrate lower mRNA levels of the GFP gene compared to negative variants (red violin-plot) (Effect size=58.26%, Wilcoxon rank-sum p-Value=$1.6 \cdot 10^{-9}$). Consistently, underachiever variants (purple violin-plot) show higher mRNA levels compared to negative variants (Effect size=68.04%, Wilcoxon rank-sum p-Value=$9.6 \cdot 10^{-8}$). Black line represents the median value.

B| Positive variants show higher translation initiation rates compared to negative variants (Effect size=61.9%, Wilcoxon rank-sum p-Value=$3.7 \cdot 10^{-18}$).

C| Positive variants demonstrate higher translation efficiencies (protein/mRNA) compared to negative variants (Effect size=55.67%, Wilcoxon rank-sum p-Value=$3.4 \cdot 10^{-5}$). Consistently, underachiever variants (purple violin-plot) further show lower translation efficiencies compared to negative variants (Effect size=63.06%, Wilcoxon rank-sum p-Value=$1.1 \cdot 10^{-4}$).

Statistically significant differences (p-Value<0.05) are marked with an asterisk.

Slower translation speed at early elongation of coding region, amino acid synthesis cost, and hydrophobicity affect cost of gene expression

We next aimed to elucidate other cellular mechanisms that affect cost per protein molecule. We first examined translation codon decoding speeds by the ribosome. The prevalence of slowly translated codons at the 5' of ORFs has been suggested to support the efficiency of gene translation[44]. This "ramp model" proposes that delaying ribosomes at the beginning of the elongation phase decreases downstream ribosomal pauses and collisions, which can therefore reduce ribosome jamming, and perhaps also pre-mature termination events.

Although contradicting evidence were reported for the existence and relevance of this mechanism to expression level[77,103–107], the main prediction of the model – that 5' ramping reduces cost of expression at a given expression level – has not been tested so-

far. Here, we had the first opportunity to test this hypothesis in a controlled manner as only the 5' variable region of the GFP varied in the library, while all other parameters remained constant. Thus, we asked whether slow 5' translation speed is associated with positive fitness residual. We were able to show that three independent mechanisms that slow down ribosome progression are more prevalent in positive variants compared to negative variants. First, positive variants show lower values of "Mean of the Typical Decoding Rates" (MTDR)[107], a measure of codon decoding time derived empirically from ribosome profiling data in *E. coli* (Figure 3A). Second, positive variants demonstrated tighter secondary structures compared to negative variants, which can slow down ribosomes[46,79,108,109] (Figure 3B+C). Third, low ribosome speed due to affinity to the anti-Shine Dalgarno (aSD) motif of the ribosome[45] coincided with positive fitness residual variants (Figure 3D).

We thus provide the first experimental evidence for a set of three gene architecture factors - codon decoding time, mRNA structure and affinity to the anti-Shine Dalgarno motif - that could each implement 5' ramping by slowing down ribosomes and by that allow cells to reduce the cost of gene expression at a given expression level.

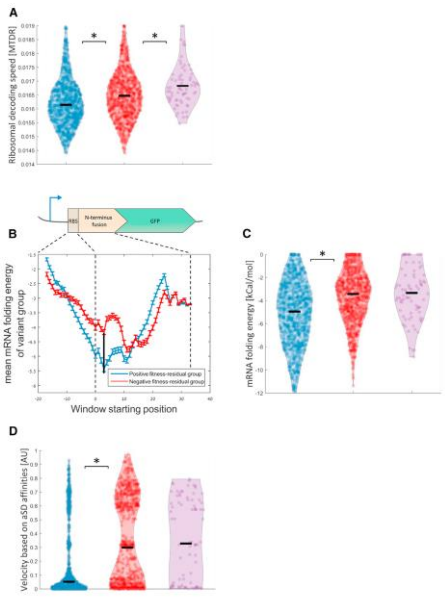


Figure 3 – Slow translation speed at early elongation, achieved by diverse molecular means, reduces expression cost

A+C+D| Positive variants show lower values of codon-decoding speed, stronger mRNA structures and lower speeds due to higher anti- Shine Dalgarno affinities compared to negative variants (Effect size=59.55%, 65.03% and 63.82%, Wilcoxon rank-sum p-Value=$3 \cdot 10^{-12}$, $5.4 \cdot 10^{-28}$ and $6.3 \cdot 10^{-24}$, respectively). Statistically significant differences (p-Value<0.05) are marked with an asterisk.

B| Mean folding energy of mRNA secondary structure according to window's start position for positive (blue curve) and negative (red curve) variants, error bars represent standard error of mean. Dashed lines mark different positions along the variable region up-stream to the GFP. Black vertical line marks the beginning of window with the largest observed difference, which is found at nucleotide positions +4 of the ORF, just after the first AUG codon. The distributions at this window position are seen in C.

Next, we explored the possibility that the amino acid composition of the N-terminus fusion to the GFP influences cellular fitness. Amino acids differ by the metabolic costs associated with their biosynthesis - predominantly energy and reducing power determinants invested in their metabolic production, as was computed in details for *E. coli*[110]. We thus hypothesized that usage of energetically-expensive amino acids may cause a heavier burden at a given expression level. Indeed, lower-cost of the N-terminus fusions were found to associate with positive fitness residual variants (Figure 4A).

We further examined the relation between fitness residual and amino acid energetic cost, by calculating the frequency ratio of each individual amino acid between the positive and negative fitness residual groups. Remarkably, this frequency ratio was found to negatively correlate with the metabolic cost of each amino acid (Figure 4B), demonstrating that expensive amino acids are less frequent in variants with a positive fitness residual. Taken together, these observations suggest that expensive-to-synthesize amino acids not only burden cells during their costly production but also when they are incorporated into proteins by the ribosome, presumably due to a feedback that increases their synthesis in response to consumption.

We next reasoned that an additional factor by which a protein could affect fitness is its toxicity, and in particular the tendency of proteins to form aggregates. As aggregation is driven by hydrophobic interactions, we turned to a conventional measure of amino acid hydrophobicity[111] to examine whether it is predictive of positive or negative fitness residual designations. We computed the hydrophobicity of each N-terminus fusion in positive and negative variants and found that positive variants tended to have significantly less hydrophobic amino acids fused to the GFP (Figure 4C). Since the heterologously expressed GFP protein resides in the cytoplasm of *E. coli*, where hydrophobic amino acids might form toxic aggregates[112], this result suggests that incorporation of hydrophobic residues in cytosolic proteins can increase the cost production per protein molecule. Notably, unlike the three architectural factors described above, this observed expression cost is not caused directly by the protein production process, but rather as a subsequent effect of the protein's toxic outcome.

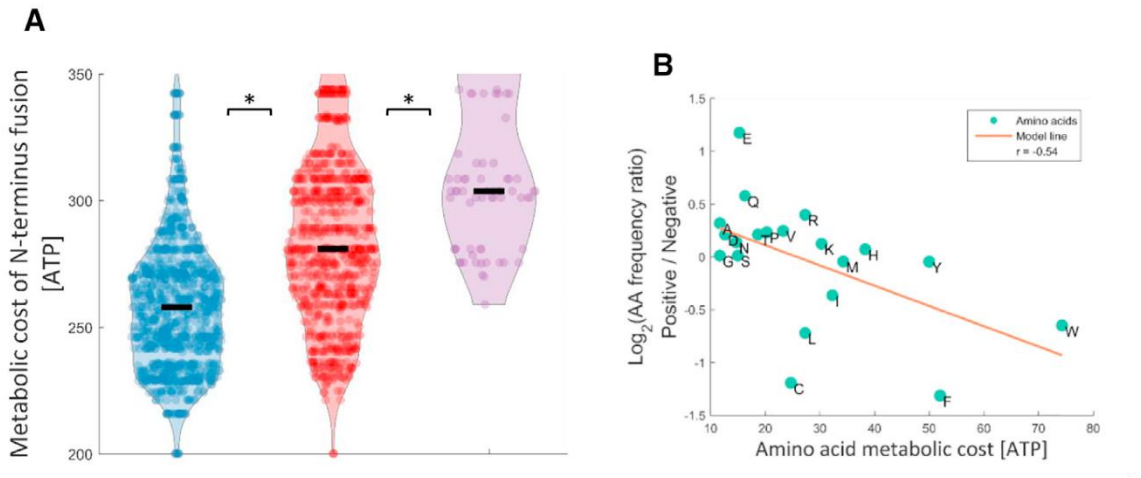


Figure 4 – Usage of expensive-to-synthetize and hydrophobic amino acids decreases fitness residual

A| N-terminus amino acid fusions of negative variants are more expensive to synthesize compared to positive variants (Effect size=72.74%, Wilcoxon rank-sum p-Value=$7.4 \cdot 10^{-62}$). Underachievers utilize even more expensive amino acids (Effect size=72.75%, Wilcoxon rank-sum p-Value=$1.7 \cdot 10^{-11}$).

B| The frequency ratio of amino acids between positive and negative variants is negatively correlated with the energetic cost of amino acids (Pearson correlation r=-0.54, p-Value=0.01).

C| N-terminus amino acid fusions of negative variants are more hydrophobic than positive variants (Effect size=69.11%, Wilcoxon rank-sum p-Value=$3.2 \cdot 10^{-44}$). N-terminus fusion of underachievers are even more hydrophobic (Effect size=81.67%, Wilcoxon rank-sum p-Value=$7.7 \cdot 10^{-21}$).

A regression model calculates relative contribution of each feature and predicts fitness residual scores

We next aimed to predict actual fitness residual values of the library variants from their gene architecture features, using a multiple linear regression model. We trained the model on a randomly chosen subset of 70% of the library variants, cross validated it on all other variants by comparing their predicted and observed fitness residual and found a good correlation (Figure 5A).

When the regression was performed on a scrambled library, which randomly links feature values and variants, the correlation between observed and predicted fitness residual was

practically eliminated. We concluded that a gene architecture that utilizes more of the features we discovered and to a greater extent typically gives rise to higher fitness residuals as expression costs are further minimized.

Additionally, this regression model allowed us to calculate the relative contribution of each feature by comparing the coefficients assigned by the regression model (Figure 5B). This analysis revealed that the features contributing to fitness residual the most are hydrophobicity and metabolic cost of the N-terminus fusion, while codon decoding speed contributing the least.

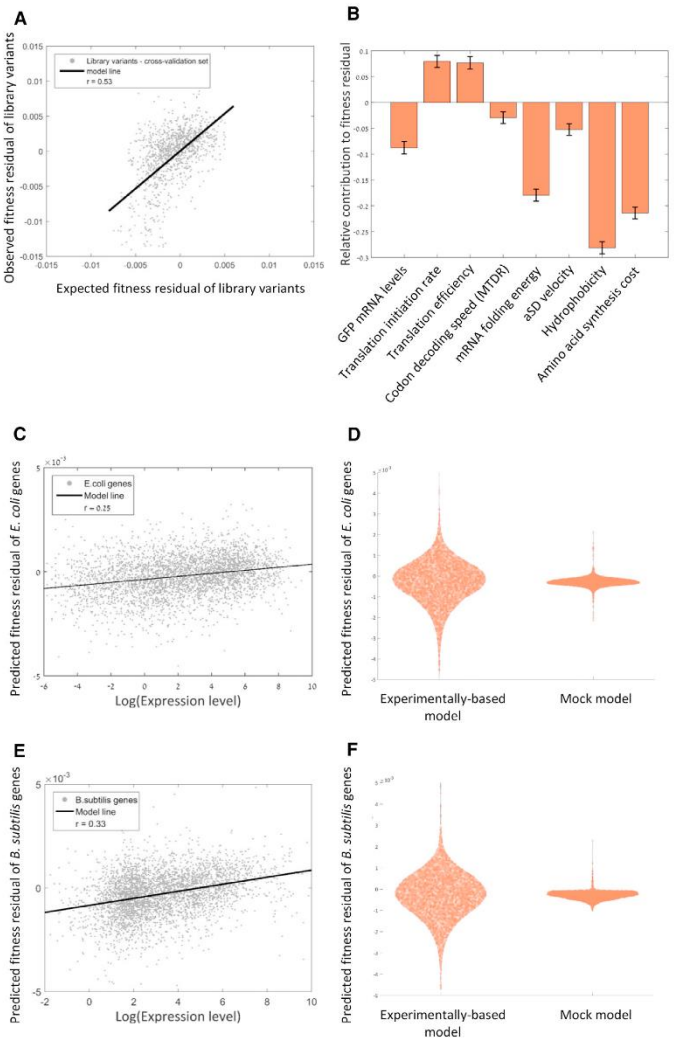


Figure 5 – A model that predicts fitness residual accurately reveals that fitness residual of natural *E. coli* genes is correlated with their expression level

A| A linear regression model based on all eight features predicts fitness residual accurately in a cross-validation test (Pearson correlation r=0.53, p-Value<10^{-200}).

B| The weighted coefficients of each feature in the regression model, demonstrating the relative contribution of each feature to fitness residual (p-Value for regression coefficient of mRNA level=$3.5 \cdot 10^{-11}$, initiation rate=$2.5 \cdot 10^{-12}$, $TE_{GFP\ protein/mRNA}$=$2.7 \cdot 10^{-9}$, codon decoding speed=$8.7 \cdot 10^{-3}$, mRNA folding energy=$1.5 \cdot 10^{-50}$, aSD velocity=$8.7 \cdot 10^{-3}$, hydrophobicity<10^{-200}, amino acid synthesis cost=$5.4 \cdot 10^{-80}$). The sign of the contribution of each coefficient shows whether a feature is correlated positively or negatively with fitness residuals. Error bars represent standard error of the coefficient estimation.

C| Predicted fitness residuals of *E. coli* genes according to the regression model are correlated with their expression levels (Pearson correlation r=0.25, p-Value=$2 \cdot 10^{-53}$), suggesting that natural selection shapes 5' gene architectures in order to minimize costs of gene expression.

D| Distribution of fitness residual scores for *E. coli* genes, as predicted by regression model that was trained on either experimental or mock data. The experimentally-based model predicts a significant, higher range of fitness residual (p-Value<10^{-5}), suggesting that the mechanisms we elucidate with the synthetic library also apply on natural genes.

E| Predicted fitness residuals of *B. subtilis* genes according to the regression model are correlated with their expression levels (Pearson correlation r=0.33, p-Value=10^{-93}), suggesting that our model also applies for other bacteria species.

F| Same as D, only for *B. subtilis* genes.

Highly expressed natural *E. coli* genes have evolved gene architectures that minimize their production costs

With these findings from the synthetic library, we next asked whether the mechanisms we revealed as cost-reducing were also utilized by natural selection to optimize *E. coli*'s native genes. We thus calculated for each *E. coli* gene its scores with respect to the relevant features and used the regression model to predict its fitness residual score. Since higher expression level results in higher expression cost, we hypothesized that *E. coli* genes with higher expression levels are more likely to be endowed with cost reducing architectures. Indeed, we found a significant correlation between predicted fitness residual of *E. coli* genes and their protein expression levels (Figure 5C), demonstrating a stronger selection for optimizing the 5' gene architecture for highly expressed genes.

Finally, to generalize our predictive model for fitness residual, we decided to test it on a different bacterium than the gram-negative *E. coli* and chose the gram-positive *Bacillus subtilis*. We calculated for each of the genes in this species its scores with respect to the relevant features, predicted a fitness residual score and computed its correlation with the genes' protein expression levels. Remarkably, in *B. subtilis* too we observed a significant correlation between expression level and optimization of the gene architecture (Figure 5E). Taken together, these observations suggest that the same molecular mechanisms we found to minimize expression costs in *E. coli* are also utilized by natural selection to reduce production cost per protein molecule in *B. subtilis*.

In conclusion, here we focused on revealing molecular mechanisms that minimize expression cost at a given expression level. Indeed, we found architectures and motifs that govern such costs, and reveal their function even beyond a direct effect on the process of expression.

First, we show that regulating initiation and mRNA levels also affects expression cost, as increasing the number of proteins that are produced per mRNA is associated with a positive fitness residual. This architecture could be beneficial because it reduces energy and resource consumption that are devoted to mRNA production.

Second, we show that three nucleotide-based features that reduce elongation speed at the 5' of the coding region are likely beneficial: low ribosomal codon-decoding speed, occurrence of Shine-Dalgarno like sequences and strong secondary structures.

Next, we revealed that the amino acid composition of a gene can also affect expression cost at a given expression level by showing that hydrophobic amino acids reduce fitness residual, perhaps due to their increased tendency to form toxic aggregates in the cytoplasm.

Finally, our observations are relevant to biotechnology and synthetic biology. Many times in such non-natural systems there is a need to express a foreign gene, whose expression could deprive resources from the hosting cell. Our results allow the design of an optimized nucleotide sequence version for heterologous expression that minimizes the cost of production, and by that reduces the burden on the cell while not compromising expression level.

3| How does the splicing machinery evolve?

Surprisingly, although the process of splicing is central to the maturation and regulation of mRNAs in eukaryotes[113–117], its role in adapting to novel demands on gene expression has not been thoroughly investigated. During mRNA splicing, precursor mRNAs are processed to remove introns while fusing exons together to create the mature transcript. This process provides an evolutionary means to diversify the proteome towards phenotypic novelty, as the choice of intron to be excluded, as well as the exons which are found in the mature transcript, can both be regulated based on the cell's needs[115,118,119]. One aspect of splicing evolution that has been extensively studied is gain and loss of intronic DNA, for which several molecular models have been proposed, mainly Reverse-Transcription and recombination-mediated intron loss, intron transposition and also exonization and intronization via mutations[120–123]. While intron loss and gain have been demonstrated experimentally[124,125], other forms of splicing evolution, such as alterations in splicing efficiency under changing conditions, have not.

Here, we set out to reveal whether introns or the splicing apparatus can evolve so as to alter the expression levels of genes in a timely and adaptive manner, and ask whether and how splicing evolves in *cis* and in *trans* to regulate gene expression. To this end, we generated a reporter construct in yeast cells that could simultaneously be read out and be selected for splicing efficiency. Namely, we introduced an inefficiently-spliced intron to a reporter gene that was fused to an antibiotic resistance gene. Using this approach, we could carry out a lab-evolution experimental setup to study the adaptation of splicing in the presence of the corresponding antibiotics.

Low splicing efficiency leads to stressed cells under restrictive conditions

We hypothesized that tuning splicing of genes could serve as a means to optimize their expression levels. To test this hypothesis, we used the yeast *Saccharomyces cerevisiae* in which ~30% of the transcriptome must be spliced, at a range of splicing efficiencies[117,126], to form mature mRNAs[127]. We built a synthetic gene construct that consists of two fused

domains: A fluorescent reporter (YFP), which includes two alternative natural introns - with either high or low splicing efficiency - near the YFP's fluorescence site[126], fused to an antibiotics resistance gene (Kanamycin resistance gene). Specifically, we created three strains: (i) WT YFP strain without an intron; (ii) "Splicing^{High}" in which the YFP harbors the natural intron of *OSH7* and was previously reported to have high splicing efficiency within this YFP context[126]; and (iii) "Splicing^{Low}" in which the natural intron of *RPS26B*, with a low splicing efficiency[126], was inserted in the same location (Figure 1A).

We first hypothesized that cellular growth of each strain in the presence of the antibiotics, geneticin (G418), will associate with the YFP-Kan expression levels. We followed the growth of the three strains in the presence of the antibiotics and found that the WT strain had the highest fitness, Splicing^{High} grew slower, and Splicing^{Low} demonstrated a severe growth defect compared to the two other strains (Figure 1B+C). We then measured florescence intensity of the YFP-Kan reporter in the presence of the drug. In line with the growth measurements, we observed that WT cells demonstrated the highest fluorescence levels, followed by Splicing^{High}, and with Splicing^{Low} cells showing the lowest YFP-Kan levels (Figure 1D). These results demonstrate that the inefficiently-spliced intron in Splicing^{Low} reduces cellular levels of YFP-Kan and hence lead to a reduced fitness.

Since YFP-Kan expression level in Splicing^{Low} were significantly lower compared to the other strains, we hypothesized that Splicing^{Low} cells did not reach the needed concentration to sufficiently neutralize the antibiotics, and hence resulted in stressed cells. To test this hypothesis, we performed mRNA sequencing of exponentially growing WT and Splicing^{Low} cells in an antibiotics containing medium, and analyzed their transcriptome profiles. Indeed, we observed that ribosomal genes were down-regulated in Splicing^{Low} compared to the control strain – a clear signature of stressed cells[128] (Figure 1E). Notably, the reduction in ribosomal expression levels (~8%) we observed here due to growth rate differences between WT and Splicing^{Low} cells is accurately predicted by a recent study, which calculated the linear correlation between growth rate and ribosomal

expression levels in yeast cells[129]. In parallel, stress-related genes[130] were up-regulated in the Splicing^{Low} compared to the control strain (Figure 1E). We thus concluded that the general stress response was activated in Splicing^{Low} cells.

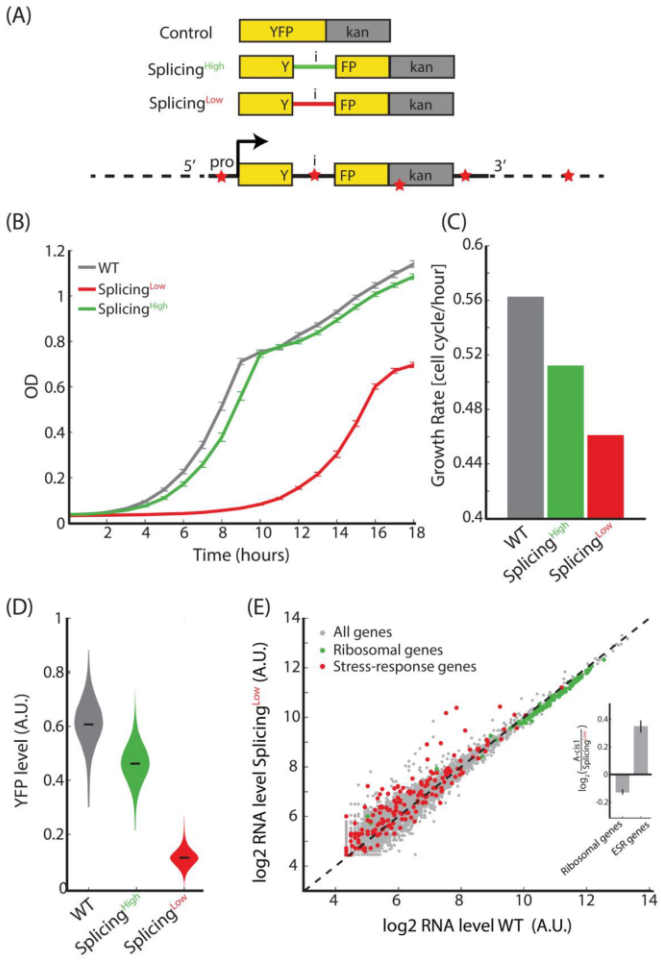


Figure 1 – Inefficient intron splicing leads to lower gene expression levels and compromised antibiotics resistance.

A| We introduced two alternative introns into a YFP domain that was fused to a kanamycin resistance domain - to generate three strains: (i) WT without an intron; (2) Splicing^{High} with an efficiently spliced intron; and (iii) Splicing^{Low} with an inefficiently spliced intron. Evolving cells at the presence of the antibiotics could adapt by mutating different parts of the YFP-Kan construct (evolution in cis) or other loci, evolution in *trans* (red stars represent potential locations of such putative mutation sites).

B+C| Splicing^{Low} suffers from a severe growth defect compared to WT or Splicing^{High} cells when the antibiotic is supplemented to the medium. The growth defect is manifested as a longer lag phase and a lower maximal growth rate.

D| Florescence intensity of the YFP-Kan reporter for all three strains shows that Splicing^{Low} cells have lower expression levels of YFP-Kan. This observation links between YFP-Kan expression levels and cellular fitness.

E| Transcriptome profiling shows that ribosomal genes were down-regulated (green dots, p-Value=4.62x10^{-26}, paired t-test) and stress-related genes were up-regulated (red dots, p-Value=3.40x10^{-5}, paired t-test) in Splicing^{Low} compared to WT cells. This observation suggests that Splicing^{Low} cells are stressed because of compromised resistance to the antibiotics and that the general stress response was activated in them. **Inset|** Mean $\log_2$ ratio of ribosomal and ESR gene groups.

Rapid evolutionary adaptation increases expression level of the resistance gene

Our experimental system mimics an evolutionary scenario in which there is an immediate and continuous selection pressure to up-regulate the expression level of a gene of interest. How would the system evolve to better resist the antibiotics? Possible means to adapt include mutations in the gene's promoter to increase transcription, mutations that increase translation initiation, or mutations inside the gene itself that increase the functional efficiency of the protein. Additionally, the splicing machinery may also take

part in adaptation of gene expression levels. To find which evolutionary track would be used by cells, we evolved the three strains by daily serial dilution on a medium supplemented with G418 for ~560 generations, in four independent cultures for each strain (Figure 2A). Interestingly, only the cultures of Splicing^{Low} cells demonstrated a significant improvement in fitness at the end of the experiment (Figure 2B+C). This observation implies that only Splicing^{Low} experienced a sufficiently strong selective pressure to adapt to the presence of the antibiotics in the medium, in contrast to the WT and Splicing^{High} strains which originally had much higher levels of the resistance gene.

Consistent with the fitness measurements, YFP measurements of the evolved cultures showed that expression levels of the resistance-YFP fusion gene increased in all four evolved cultures of Splicing^{Low} compared to the ancestral strain (Figure 2D). Conversely, the increase in YFP-Kan expression levels in the evolved WT populations was smaller, and only one culture of the evolved Splicing^{High} cells demonstrated strong elevation of the YFP-Kan levels (Figure 2D). These results further support that Splicing^{Low} cells experienced the strongest selective pressure to adapt rapidly to the presence of the antibiotics in our experimental setup, and that they achieved this goal by increasing the levels of the resistance gene. We next moved to reveal the molecular mechanisms underlying this evolutionary process.

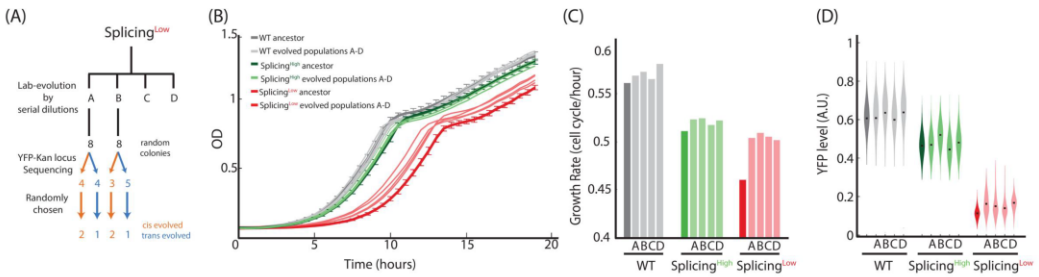


Figure 2 – Rapid adaptation to the presence of the antibiotics is observed only for Splicing^{Low} cells.

A| We evolved WT, Splicing^{High}, and Splicing^{Low} cells for ~560 generations with the presence of the antibiotics in four independent cultures for each strain. We measured fitness and YFP-Kan expression levels for all evolved lines (see below), and also randomly chose 16 colonies from two evolved lines of Splicing^{Low}. We sequenced the YFP-Kan locus of those colonies and observed that around half showed mutations in the YFP-Kan construct (indication of evolution in *cis*) and the other half did not (indication of evolution in *trans*). Of those colonies, we randomly chose two *cis*-evolved and one *trans*-evolved colonies from each evolved population for further examinations (see figure 3 onwards).

B+C| Growth of evolved populations compared to the three ancestors. Only evolved Splicing^{Low} cells demonstrate significant improvement in growth for all four independent evolution lines. This observation suggests that the inefficiently spliced intron led to a rapid adaptation of Splicing^{Low} cells.

D| Florescence intensity of the YFP-Kan reporter for all evolved cultures show that expression levels were much increased in all four evolved cultures of Splicing^{Low} compared to the ancestral strain (effect sizes = 78.67, 79.54, 75.17, 83.19). Conversely, the increase in expression levels in the evolved WT and of Splicing^{High} populations were smaller (WT effect sizes = 64.66, 68.44, 63.51, 67.74; Splicing^{High} effect sizes = 54.33, 70.66, 52.43 and 58.27). This observation suggests that adaptation of Splicing^{Low} cells was based on their ability to increase expression levels of the resistance proteins.

Adaptation in *cis* and *trans* leads to increased splicing efficiency

We hypothesized that improving the low splicing efficiency of the intron in Splicing^{Low} could be exploited by natural selection as an adaptation mechanism towards increasing the resistance gene levels. We therefore sequenced the YFP-Kan locus in 16 randomly chosen colonies from two evolved populations (termed here population A and population B) of Splicing^{Low}. Interestingly, we found that the colonies were split into two types – either with or without a mutation in the YFP-Kan locus. In population A, we found that the same mutation occurred in four out of eight colonies, changing adenine to cytosine inside the intron, 97 nucleotides up-stream to its 3' end (Figure 3A). In population B, we identified an exonic non-synonymous mutation that changed a valine at position 61 of the YFP protein into alanine (a thymine to cytosine 14 nucleotides up-stream of the intron) in three out of eight colonies. In the five other colonies from this population there were no mutations in the YFP-Kan locus.

Notably, none of the colonies demonstrated a mutation in the construct's promoter, terminator or in the sequence of the Kan resistance gene itself. These results propose that different mutations in the intron, or its vicinity, were adaptive and might affect splicing efficiency of the intron. Surprisingly, the observed mutations did not occur in the 5' donor, 3' acceptor, nor in the intron branch point – suggesting that other position of the intron can also be selected in evolution increase fitness by affecting splicing. While the intron- and exon-mutated colonies represent an evolutionary adaptation in *cis*, the colonies that showed no mutation in the entire gene construct potentially found adaptive solutions in *trans* that may have occurred elsewhere in the genome.

We randomly chose six colonies: four colonies with a *cis* mutation and two colonies that showed no mutations in *cis*, for which we reasoned that such colonies may have adapted in *trans*. We termed these colonies according to the evolution lines from which they were derived: A-cis1, A-cis2, B-cis1, B-cis2, A-trans and B-trans. We followed the growth of these evolved colonies in the presence of G418 and found, as expected, that all grew faster than the Splicing^{Low} ancestor (Figure 3B). We then performed RNA-seq and

transcriptome analysis of all colonies, which revealed relaxation of the stress response that was featured in the ancestor. Namely, the general stress response genes were reduced and ribosomal proteins were up- regulated in five evolved colonies (Figure 3C). These observations suggest that the cells indeed adapted to the presence of the antibiotics in the environment and that the stress experienced by them was partially alleviated.

We next hypothesized that cellular fitness might correlate with mRNA levels of the YFP-Kan construct because increased transcript levels should result in higher concentrations of the Kan protein. Indeed, maximal growth rates of the control and Splicing^{Low} ancestors and for the six evolved colonies correlate with mRNA levels of the YFP-Kan construct, as deduced from the RNA-seq (Figure 3D) – supporting our conclusion that adaptation was based on increasing expression levels of the YFP-Kan gene. Since the observed *cis* mutations occurred at the vicinity of the intron, we hypothesized that they increased splicing efficiency of the YFP-Kan transcript. To test this possibility, we performed, for both *cis*- and *trans*-evolved colonies, a splicing efficiency assay with qPCR - targeting the un-spliced and spliced transcript versions. Interestingly, the ratio of spliced to un-spliced transcripts was higher in all evolved colonies compared to the Splicing^{Low} ancestor, suggesting that at least some of the mRNA level increase we observed in the evolved colonies results from increased splicing efficiency (Figure 3E).

To prove that adaptation of the colonies actually led to higher protein levels of the resistance gene, we measured fluorescence intensity using flow cytometry. We found that the two *cis*-colonies from population A (A-cis1 & A-cis2) and the two *trans*-colonies (A-trans & B-trans) showed higher YFP-Kan levels compared to the ancestor. However, the two *cis*-colonies from population B (B-cis1 & B-cis2) demonstrated decreased fluorescence intensity values (Figure 3F). These observations suggest that the non-synonymous, exon mutation reduced the fluorescence-per-protein value of the YFP-Kan construct in these colonies. Indeed, this position corresponds to a position that was recently reported to reduce florescence when mutated in the highly similar GFP[131].

Because YFP functionality was not selected for or against in our setup, it was free to mutate as long as it helps achieve a higher expression level of the entire construct by increasing the intron's splicing efficiency. It thus seems that modular domain-architecture of a protein may increase its evolvability under relevant conditions as it allows the optimization of each domain in isolation from the other.

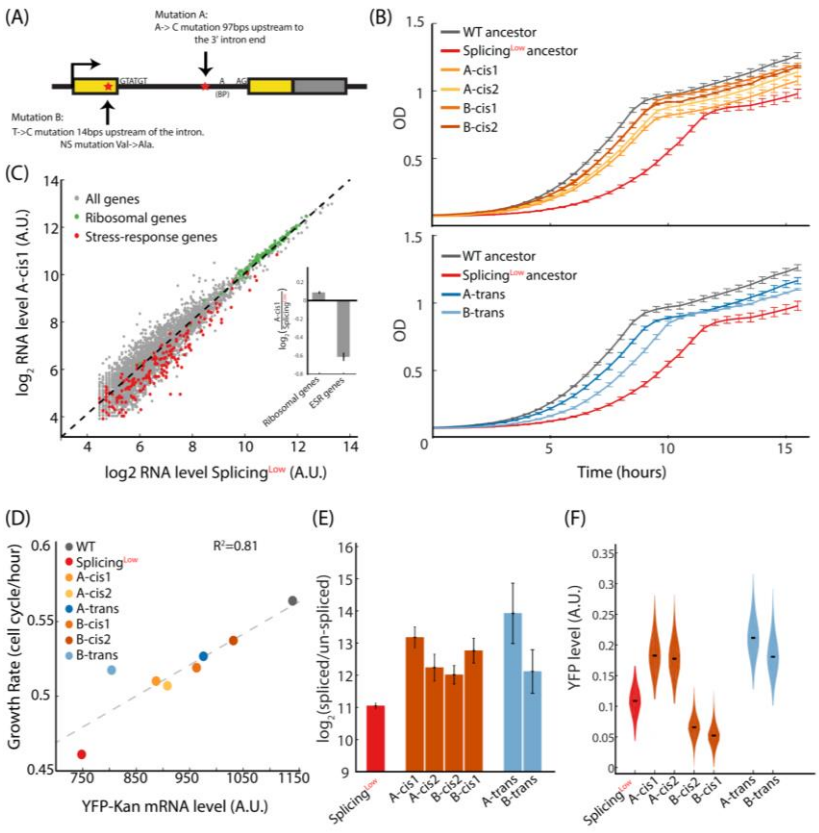


Figure 3 – Evolved colonies demonstrate increased splicing efficiency that results in higher transcript levels and relieved stress.

A| Sequencing of the YFP-Kan construct in the evolved colonies revealed two mutation types: (i) in the intron itself and (ii) in the up-stream exon – see text for full description. These mutations did not occur in the intron 5' donor, 3' acceptor, or the branching point – suggesting that other positions of the intron and its vicinity are phenotypically functional and may affect splicing efficiency.

B| All *cis*-evolved colonies (upper graph) and *trans*-evolved colonies (lower graph) show increased fitness compared to the Splicing^{Low} ancestor, yet still lower than the WT ancestor.

C| Transcriptome profiling reveals that ribosomal genes were up-regulated (green dots, p-Value=4.94 x10^{-18}, paired t-test) and stress-related genes were down-regulated (red dots, p-Value=3.64 x10^{-15}, paired t-test) in the evolved colony A-cis1 compared to the Splicing^{Low} ancestor. This trend was observed in 5 out of 6 evolved colonies (Supplementary Figure 1). **Inset|** Mean $\log_2$ ratio of ribosomal and ESR gene groups.

D| mRNA levels of YFP-Kan transcripts correlate with growth rate – suggesting that cellular fitness in our set-up is indeed determined by the availability of Kanamycin-resistance proteins to overcome the antibiotics.

E| All *cis*- and *trans*-evolved colonies demonstrate increased splicing efficiency of the YFP-Kan mRNA compared to the Splicing^{Low} ancestor. This result suggests that all adaptation trajectories led to the adaptation of the splicing process to better mature the un-spliced YFP-Kan transcript.

F| Florescence intensity of the YFP-Kan reporter show increased levels for the two *cis*-evolved colonies with the mutation in the intron and for the two *trans*-evolved colonies. In contrast, the two *cis*-evolved colonies with the non-synonymous mutation in the exon demonstrate decreased YFP-Kan levels. This observation suggests that the non-synonymous mutation hampered the ability of the YFP domain to florescent and reduced the Florescence intensity per protein molecule (see text for full explanation).

It is possible that additional beneficial mutations exist in the genome of the *cis*-evolved colonies, which account for the phenotypes we observed. To directly assess the effects of the *cis* mutations, we generated two rescue strains, termed rescue-A and rescue-B, in which these *cis*-acting mutations were introduced individually to the ancestral Splicing^{Low} background. Notably, the two rescue strains grew better than Splicing^{Low} cells in the presence of the antibiotics (Figure 4A), though not as good as the wild-type, and the stress experienced by the Splicing^{Low} cells was relieved upon insertion of each individual

cis mutation (Figure 4B). Finally, we measured splicing efficiencies and fluorescence intensity levels for both rescue strains, and found that they resembled the results of the evolved single colonies (Figure 4C-D, in comparison to Figure 3E-F). These observations strengthen our conclusions that the *cis*-acting mutations are sufficient to elevate YFP-Kan levels through an increased splicing efficiency, yet the non-synonymous mutation of population D also hampers the function of the YFP domain and reduces its florescence-per-protein ratio.

Our results thus far provide direct evidence that intron splicing takes part in the adaptation and optimization of gene expression patterns to environmental needs. Although intron sequences are much less conserved compared to exons, and are believed to be less functional, we demonstrate that their sequence can be used by natural selection as a molecular mechanism to regulate splicing efficiency and adjust gene expression patterns.

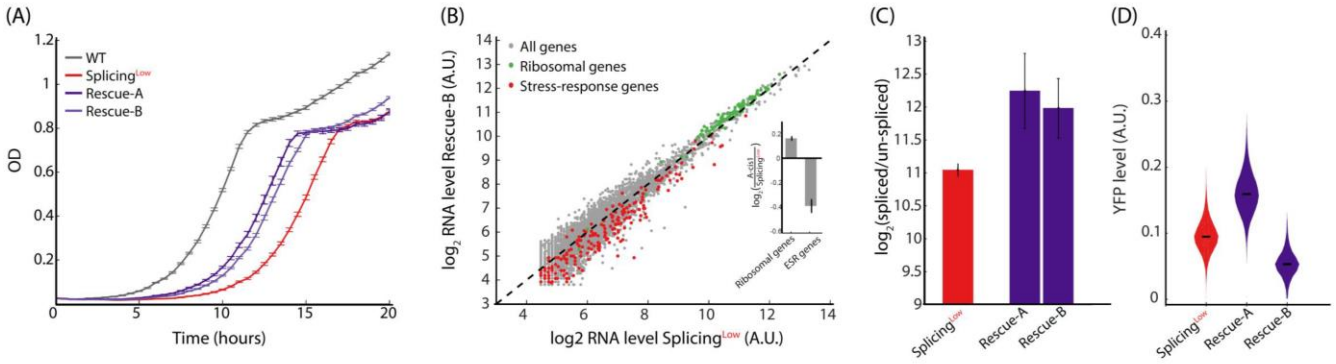


Figure 4 – cis-acting mutations are sufficient to increase fitness by elevating splicing efficiency.

A| We created two rescue strains, each harboring one of the mutations that appeared spontaneously in the evolved populations. Growth of the two rescue strains show that a single mutation in the YFP-Kan construct is sufficient to increase fitness compared to Splicing^{Low}.

B| The exonic mutation is also sufficient to alleviate stress, as ribosomal genes were up-regulated (green dots, p-Value=1.02x10^{-18}, paired t-test) and stress-related genes were down-regulated (red dots, p-Value=9.02x10^{-12}, paired t-test) in Rescue-B compared to Splicing^{Low}. The same trend was also observed for the intronic mutation for Rescue-A cells. **Inset|** Mean $\log_2$ ratio of ribosomal and ESR gene groups.

C| The two rescue strains demonstrate higher splicing efficiency of the YFP-Kan mRNA compared to the Splicing^{Low} ancestor. This result suggests that a single mutation is sufficient to improve splicing efficiency.

D| Florescence intensity of the YFP-Kan reporter for the Rescue-A and Rescue-B strains show similar trends as the colonies in Figure 3D - supporting earlier conclusions.

Increasing cellular availability of the splicing machinery can be adaptive

We finally aimed to decipher the mechanism behind the increased YFP-Kan levels in the *trans*-evolved colonies that showed no mutations in *cis*, i.e. within the reporter gene or in its vicinity. We reasoned that elevating availability of the splicing machinery as a global resource could be a means to increase splicing efficiency of the YFP-Kan transcript, and thus could be used as an adaptive mechanism to the antibiotics challenge. Increased splicing-availability could be achieved by increasing the expression of the splicing machinery genes. In addition, as with other cellular machineries whose functioning depends on supply-to-demand economy[39,73,81,132] reducing expression levels of the intron-containing genes, namely the "demand", could increase the availability of the machinery towards the intron under selection here.

To test if any of these evolutionary routes were indeed taken by the evolved cells, we calculated the expression level ratio of genes between the evolved colonies and their ancestor. In colony A-trans, we observed that while the average expression-ratio of the splicing machinery genes (the "supply") increased, that of the non-ribosomal intron-containing genes (the "demand") decreased (Figure 5A). This observation suggests that indeed the cellular availability of the splicing machinery was elevated in this evolved colony, which might have allowed for the observed increased splicing efficiency of the YFP-Kan gene. Next, we hypothesized that the *cis*-evolving colonies may have also adapted in *trans* and used this adaptation mechanism as well. Indeed, in all other evolved colonies we observed a similar trend, in which the overall supply-to-demand measurement of the splicing machinery was increased (Figure 5B). Importantly, the two rescue strains, which did not evolve and only harbor our artificially introduced *cis*-acting mutation, did not show any change in splicing availability (Figure 5B), strongly supporting our conclusion that this phenotype was achieved by further adaptation of the cells during our lab-evolution experiment. Thus, we concluded that both *cis* and *trans* adaptation routes can co-occur in the same genome towards optimization of its gene expression patterns.

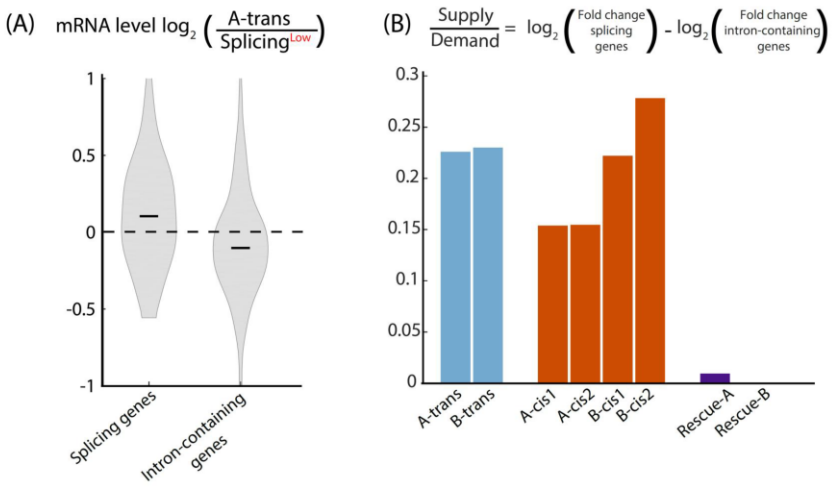


Figure 5 – Increasing cellular availability of the splicing machinery is an adaptive mechanism of splicing.
A| The groups of splicing genes and intron-containing genes were increased (p-Value=1.36x10^{-3}, paired t-test) and decreased (p-Value=1.67x10^{-2}, paired t-test), respectively, in the trans-evolved colony A-trans compared to Splicing^{Low} ancestor. This observation suggests that the supply-to-demand ratio of the splicing machinery was increased in A-trans colony, which allowed its increased splicing efficiency of the YFP-Kan transcript.
B| Supply-to-demand ratios for the splicing machinery were calculated to all evolved colonies and to the rescue strains as the difference between the mean fold-change of splicing genes to the mean fold-change of intron-containing genes. While supply-to-demand ratios were increased in all evolved colonies, they remained the same for the two rescue strains. These results suggest that indeed the cellular availability of the splicing machinery was elevated in the evolved colonies – a *trans*-adaptation mechanism to optimize gene expression using the splicing process.

Mutations in SR-like proteins drive *trans* adaptation of the splicing machinery

To reveal mutations that happened in *trans*, namely in other positions in the genome rather than the YFP-Kan construct, we sequenced the entire genomes of the *trans*-evolved colonies and compared them to the genome of their ancestral strain. This approach revealed two non-synonymous mutations in the SR-like proteins Npl3 and Gbp2, which occurred in one of these proteins' RNA recognition domains. Gbp2 has been shown to work as quality control factors for spliced mRNA, it interacts with Mex67, a key adaptor in the mRNA export pathway. This interaction only occurs upon efficient splicing; else Gbp2 remains associated to the RNA degradation machinery TRAMP and the transcript is degraded in the nucleus[133]. Interestingly, Npl3 is recruited during the early stages of transcription as part of mRNP biogenesis, and has been shown to support efficiency of splicing by stabilizing the U1 snRNP[134,135].

In conclusion, here we study the role of the splicing machinery in optimization of gene expression programs by placing selective pressure on cells to improve the splicing efficiency of a specific gene. Our results provide molecular evidence for the relevance of splicing as another instrument in the cellular toolbox towards adjusting its gene expression patterns. To the best of our knowledge, we demonstrate the first experimental evidence of splicing efficiency adaptation, confirming that this adaptation can occur in *cis* and *trans* similarly to adaptations of other means of gene regulation.

Interestingly, we found that different adaptive means co-occurred in the evolved populations – independently in different cells or even simultaneously in the same genome. In particular, we saw that evolutionary lines that adapted in *cis* appear to also have had adaptations that are not encoded in the evolving gene, hence pointing to changes that must have occurred in *trans*. Further investigations will reveal which of these solutions, *cis* or *trans*, proves to be more evolutionarily stable - to fully reveal the dynamics of splicing adaptation when cells optimize their gene expression.

Discussion

The field of Evolutionary cell biology aims to reveal how cells evolve. While cell biology focuses on components, interactions, and processes at the cellular, rather than the molecular or organismal level, evolutionary cell biology is the study of how such components and complexity emerged and are changed with time.

Historically, this goal has been implemented by focusing on studying the origin of eukaryotic organelles and multicellularity. The common approach by which evolutionary cell biology studies have been conducted is using natural species diversity to elucidate different mechanisms that evolved to allow complex cellular functions. This path has been fruitful and illuminated various insights about the mechanisms of cellular evolution (see introduction for discussion).

However, our understanding of the evolution of cells is still incomplete and we are far from being able to predict accurately which environmental and genetic circumstances drive the evolution of cellular structure and function. Indeed, how cellular pathways, structure, and functions respond to a new challenge at an evolutionary time scale has not been fully revealed. During my PhD, I was inspired by concepts of evolutionary cell biology and aimed to reveal new layers of cellular evolution. Specifically, I decided to combine this conceptual framework with the emerging power of high-throughput technologies and lab-evolution methodology to study the evolution of cellular machineries.

Our ability to follow the changes after many generations of growth in the lab of both prokaryotes and eukaryotes has been improved immensely in the past decade[136]. This developed “lab-evolution” methodology has been used to look at the dynamic of early adaptation[137], the types of genomic alterations during a long-term adaptation to a constant environment[138], the advantage of sex[139], and more. However, changes at the cellular level have not been the major focus of such studies. I have thus decided that an interesting line of investigation could be to challenge fundamental cellular functions either genetically or environmentally, grow them for several hundreds of generations,

and characterize the mechanisms by which cells adapted the perturbed cellular machineries. Such experimental designs, I reasoned, could illuminate new insights on the functions of these machineries and reveal how they were shaped during evolution. Some specific questions I asked were: By which molecular mechanisms do cellular machineries adapt? What are their flexible parts? What are harsh\easy challenges for the systems to face?

While these kinds of questions could be asked to diverse molecular machines in cells, in my PhD I focused on cellular machineries of the central dogma that perform gene expression – and studied mainly the molecular adaptation of the translation and splicing machineries. First, I used genome editing in both bacteria and yeast to reveal the economic interplay in translation between codon usage (demand) tRNA genes (supply). Our experiments in yeast reveled the tRNA genes provide evolutionary plasticity to the translation apparatus, and allows it to respond rapidly to accommodate new translational demand. The strategic mutation that we identified in the anticodon of a tRNA gene inspired us to look for such mutations in wild species. Indeed, we could find such mutations, which suggests that this anticodon switching mechanisms is widely used in nature. This example demonstrates how lab-evolution has the potential to accurately mimic wild adaptation to illuminate mechanisms by which cells evolve. It also shows why indeed it is fruitful to use lab-evolution methodology when studying evolutionary cell biology.

Next, we used a genome engineering technology in bacteria to massively manipulate the codon usage of highly-expressed genes. This approach allowed us to study an evolutionary question that was hard to test experimentally because of the vast changes ones had to introduce into the genome. Why highly-expressed genes show an intensified codon bias compared to the rest of the genomes has been thoroughly studied – focusing on *cis* effects of the codons on the genes on which they reside. However, our perturbation of the system illuminates that in addition to such changes, other *trans* effects are also observed and might have drove the evolution of these genes. We showed

how non-optimal codons in highly-expressed genes result in a cellular-wide translation perturbation that affects translation efficiency of genes in a codon-dependent manner. These two works demonstrate how a cellular perspective of demand and supply allowed us to reveal new aspects of an otherwise intensely studied machinery.

In addition to lab-evolution, I also used synthetic DNA libraries as a means to ask evolutionary cell biology questions. Such libraries allow to simultaneously test many hypotheses as they generate a great genetic diversity within a population that yield many strains, each with a unique phenotype. We combined this technique with fitness measurements of various strains, which express a protein that burdens cellular growth, to find molecular mechanisms that improve efficiency of protein expression and minimize its associated costs. This approached revealed several molecular mechanisms that indeed reduce cost per protein molecule – and importantly from an evolutionary perspective, these mechanisms were selected by evolution to be used by natural bacterial genes. This work shows how new technologies can revolutionize the ways by which we ask evolutionary cell biology question and study the variation of cells.

Finally, I turned to apply a combined approach of cellular evolution and lab-evolution to the splicing apparatus. Splicing evolution has been studied in the past, yet mostly by characterizing how its generates phenotypic diversity with alternative splicing and by characterizing mechanisms of intron gain and loss. Here, I tested whether splicing evolution could occur when a need for increased expression level of a specific gene is presented to cells. Evolving these cells revealed molecular adaptation of the slicing machinery in both *cis* and *trans*. Interestingly, the *cis* adaptation was based on mutations not only in the intron itself of the gene under selection – but also in adjacent exons. Complementary, the *trans* adaptation was based on non-synonymous mutations in associated proteins of the splicing machineries that facilitate the transport of mature, spliced transcripts from the nucleus to the cytosol. This story demonstrates that complex cellular machine that have been evolved for millions of years can still change and help cells adapt to new challenges.

Two general concepts emerge from my work. First, supply-to-demand balance seems to be at play in different cellular pathways. In translation, I showed how the balance between tRNA levels and codon demand in the transcriptome can influence the evolution of the translation machinery. A shortage in a specific tRNA type can lead to accumulation of mutations in other tRNA genes that increase the levels of the needed tRNA by anticodon manipulation. Additionally, manipulating codon demand hampers translation efficiency globally in the cell and may hence might lead to alterations in the translation system in the form of tRNA mutations or manipulation of mRNA levels that will restore balance. Interestingly, I also observed how supply-to-demand balance is used as an adaptation mechanism of the splicing process. One route of adaptation was based on inducing the genes of the splicing apparatus (supply) and reducing the levels of other, intron-containing genes (demand). These changes increased the total supply-to-demand ratio of splicing in cells because there were more spliceosome complexes available to perform splicing and less pre-mature mRNA molecules of them to splice. Therefore, this increased supply-to-demand ratio might have improved the splicing of the gene under selection in our system and hence improved cellular fitness. Changes in supply-to-demand ratio might also be relevant in the evolution of other biological systems, like metabolism. It might be possible that enzymatic efficiencies and expression levels are determined by optimal balance supply-to-demand levels of the various products. It might be interesting to examine whether evolutionary challenges to the metabolic network could result in adaptations that are based on supply-to-demand ratios.

The second conceptual insight concerns the cost of gene expression. As mentioned above, we used a synthetic library and fitness measurements, to characterize several mechanisms that reduce cost per protein molecule. To achieve this goal, we defined “fitness residual” as the difference between expected and measured fitness of any given gene architecture. The concept of fitness residual was also observed when we manipulated the codon usage of highly-expressed genes. This manipulation revealed how changing the codon usage of a gene may lead to residual effects on other genes and hence indirectly increase the cost of expression. It is therefore tempting to postulate that

other pathways also evolve not only to locally achieve a desired phenotype but to accommodate other needs of the cells. An example for that could be the interplay between the different transcription factors. A gene might evolve an interaction with a specific factor, yet this may hold a residual effect on other genes that rely on this factor as well for their transcription.

In conclusion of this thesis, it is my belief that further combination of emerging technologies with evolutionary perspective on cells will yield fascinating questions about the molecular mechanisms of adaptation of cellular structure, pathways, and functions. Such questions I particularly find interesting to address are: How protein-protein interactions change with time? Can cellular localization of proteins and mRNAs be challenged to change during adaptation? Can we follow the evolution of organelles' contact site in the lab? I intend to study such questions in my future scientific path.

Bibliography

1. Harvey Lodish *et al. Molecular Cell Biology*. (Freeman & Company, W. H., 2016).
2. Lynch, M. *et al.* Evolutionary cell biology: two origins, one objective. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 16990–4 (2014).
3. Brodsky, F. M., Thattai, M. & Mayor, S. Evolutionary cell biology: Lessons from diversity. *Nat. Cell Biol.* **14,** 651–651 (2012).
4. Javaux, E. J. The early eukaryotic fossil record. *Adv. Exp. Med. Biol.* **607,** 1–19 (2007).
5. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504,** 231–6 (2013).
6. Mast, F. D., Barlow, L. D., Rachubinski, R. A. & Dacks, J. B. Evolutionary mechanisms for establishing eukaryotic cellular complexity. *Trends Cell Biol.* **24,** 435–442 (2014).
7. Koumandou, V. L. *et al.* Molecular paleontology and complexity in the last eukaryotic common ancestor. *Crit. Rev. Biochem. Mol. Biol.* **48,** 373–96
8. Klute, M. J., Melançon, P. & Dacks, J. B. Evolution and diversity of the Golgi. *Cold Spring Harb. Perspect. Biol.* **3,** a007849 (2011).
9. Dacks, J. B. & Field, M. C. Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode. *J. Cell Sci.* **120,** 2977–2985 (2007).
10. Zimorski, V., Ku, C., Martin, W. F. & Gould, S. B. Endosymbiotic theory for organelle origins. *Curr. Opin. Microbiol.* **22,** 38–48 (2014).
11. Keeling, P. J. The endosymbiotic origin, diversification and fate of plastids. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **365,** 729–48 (2010).
12. Richardson, E., Zerr, K., Tsaousis, A., Dorrell, R. G. & Dacks, J. B. Evolutionary cell biology: functional insight from 'endless forms most beautiful'. *Mol. Biol. Cell* **26,** 4532–4538 (2015).
13. de Duve, C. The origin of eukaryotes: a reappraisal. *Nat. Rev. Genet.* **8,** 395–403 (2007).
14. Dacks, J. B., Poon, P. P. & Field, M. C. Phylogeny of endocytic components yields insight into the process of nonendosymbiotic organelle evolution. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 588–93 (2008).
15. Devos, D. *et al.* Components of coated vesicles and nuclear pore complexes share a common molecular architecture. *PLoS Biol.* **2,** e380 (2004).
16. Varki, A. Nothing in Glycobiology Makes Sense, except in the Light of Evolution. *Cell* **126,** 841–845 (2006).
17. Akiyoshi, B. & Gull, K. Evolutionary cell biology of chromosome segregation: insights from trypanosomes. *Open Biol.* **3,** 130023–130023 (2013).
18. Farhadifar, R. *et al.* Scaling, selection, and evolutionary dynamics of the mitotic

spindle. *Curr. Biol.* **25,** 732–40 (2015).

19. Phillips, P. C. & Bowerman, B. Cell biology: Scaling and the emergence of evolutionary cell biology. *Curr. Biol.* **25,** R223–R225 (2015).
20. Petrov, A. S. *et al.* History of the ribosome and the origin of translation. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 15396–401 (2015).
21. Jiang, C., Caccamo, P. D. & Brun, Y. V. Mechanisms of bacterial morphogenesis: evolutionary cell biology approaches provide new insights. *Bioessays* **37,** 413–25 (2015).
22. Rivas-Marín, E., Canosa, I. & Devos, D. P. Evolutionary Cell Biology of Division Mode in the BacterialPlanctomycetes-Verrucomicrobia-ChlamydiaeSuperphylum. *Front. Microbiol.* **7,** 1964 (2016).
23. van Niftrik, L. & Devos, D. P. Editorial: Planctomycetes-verrucomicrobia-chlamydiae bacterial superphylum: New model organisms for evolutionary cell biology. *Front. Microbiol.* **8,** 1–3 (2017).
24. Lynch, M. The lower bound to the evolution of mutation rates. *Genome Biol. Evol.* **3,** 1107–18 (2011).
25. Lynch, M. Evolutionary layering and the limits to cellular perfection. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 18851–6 (2012).
26. Stoltzfus, A. On the possibility of constructive neutral evolution. *J. Mol. Evol.* **49,** 169–81 (1999).
27. Lukeš, J., Archibald, J. M., Keeling, P. J., Doolittle, W. F. & Gray, M. W. How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life* **63,** 528–537 (2011).
28. Tuch, B. B., Li, H. & Johnson, A. D. Evolution of eukaryotic transcription circuits. *Science* **319,** 1797–9 (2008).
29. Lynch, M. The evolution of genetic networks by non-adaptive processes. *Nat. Rev. Genet.* **8,** 803–13 (2007).
30. Finnigan, G. C., Hanson-Smith, V., Stevens, T. H. & Thornton, J. W. Evolution of increased complexity in a molecular machine. *Nature* **481,** 360–364 (2012).
31. Goldstein, B. & King, N. The Future of Cell Biology: Emerging Model Organisms. *Trends Cell Biol.* **26,** 818–824 (2016).
32. Blackburn, E. H. & Gall, J. G. A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. *J. Mol. Biol.* **120,** 33–53 (1978).
33. Jönsson, K. I., Rabbow, E., Schill, R. O., Harms-Ringdahl, M. & Rettberg, P. Tardigrades survive exposure to space in low Earth orbit. *Curr. Biol.* **18,** R729–R731 (2008).
34. Tanaka, E. M. & Reddien, P. W. The cellular basis for animal regeneration. *Dev. Cell*

21, 172–85 (2011).

35. Brunet, T. & King, N. The Origin of Animal Multicellularity and Cell Differentiation. *Dev. Cell* **43,** 124–140 (2017).

36. Müller, E. C. & Wittmann-Liebold, B. Phylogenetic relationship of organisms obtained by ribosomal protein comparison. *Cell. Mol. Life Sci.* **53,** 34–50 (1997).

37. Itoh, T., Takemoto, K., Mori, H. & Gojobori, T. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.* **16,** 332–46 (1999).

38. Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. & Koonin, E. V. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.* **11,** 356–72 (2001).

39. Gingold, H. & Pilpel, Y. Determinants of translation efficiency and accuracy. *Mol. Syst. Biol.* **7,** 481 (2011).

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

41. Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* **20,** 237–43 (2013).

42. Gingold, H., Dahan, O. & Pilpel, Y. Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome. *Nucleic Acids Res.* **40,** 10053–63 (2012).

43. 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).

44. Tuller, T. *et al.* An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141,** 344–54 (2010).

45. Li, G.-W., Oh, E. & Weissman, J. S. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484,** 538–41 (2012).

46. 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).

47. Widmann, J., Harris, J. & Lozupone, C. Stable tRNA-based phylogenies using only 76 nucleotides. *RNA* 1469–1477 (2010). doi:10.1261/rna.726010.tity

48. Withers, M., Wernisch, L. & dos Reis, M. Archaeology and evolution of transfer RNA genes in the Escherichia coli genome. *RNA* **12,** 933–42 (2006).

49. Rogers, H. H., Bergman, C. M. & Griffiths-Jones, S. The evolution of tRNA genes in Drosophila. *Genome Biol. Evol.* **2,** 467–77 (2010).

50. Bermudez-Santana, C. *et al.* Genomic organization of eukaryotic tRNAs. *BMC*

Genomics **11,** 270 (2010).

51. Rawlings, T. A., Collins, T. M. & Bieler, R. Changing identities: tRNA duplication and remolding within animal mitochondrial genomes. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 15700–5 (2003).

52. Higgs, P. G. & Ran, W. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Mol. Biol. Evol.* **25,** 2279–91 (2008).

53. Byström, A. & Fink, G. A functional analysis of the repeated methionine initiator tRNA genes (IMT) in yeast. *Mol. Gen. Genet. MGG* **216,** 276–86 (1989).

54. von Pawel-Rammingen, U., Aström, S. & Byström, A. S. Mutational analysis of conserved positions potentially important for initiator tRNA function in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **12,** 1432–42 (1992).

55. Aström, S. U., von Pawel-Rammingen, U. & Byström, A. S. The yeast initiator tRNAMet can act as an elongator tRNA(Met) in vivo. *J. Mol. Biol.* **233,** 43–58 (1993).

56. Saks, M. E., Sampson, J. R. & Abelson, J. Evolution of a transfer RNA gene through a point mutation in the anticodon. *Science* **279,** 1665–70 (1998).

57. Begley, U. *et al.* Trm9-catalyzed tRNA modifications link translation to the DNA damage response. *Mol. Cell* **28,** 860–70 (2007).

58. Kawakami, K. *et al.* A rare tRNA-Arg(CCU) that regulates Ty1 element ribosomal frameshifting is essential for Ty1 retrotransposition in Saccharomyces cerevisiae. *Genetics* **135,** 309–20 (1993).

59. Clare, J. J., Belcourt, M. & Farabaugh, P. J. Efficient translational frameshifting occurs within a conserved sequence of the overlap between the two genes of a yeast Ty1 transposon. *Proc. Natl. Acad. Sci. U. S. A.* **85,** 6816–20 (1988).

60. Lenski, R. E., Rose, M. R., Simpson, S. C. & Tadler, S. C. Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. *Am. Nat.* **138,** 1315 (1991).

61. Chan, P. P. & Lowe, T. M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* **37,** D93-7 (2009).

62. Man, O. & Pilpel, Y. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat. Genet.* **39,** 415–21 (2007).

63. Drummond, D. A. & Wilke, C. O. C. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134,** 341–52 (2008).

64. Shah, P. & Gilchrist, M. A. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 10231–6 (2011).

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

66. Duret, L. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet.*

Dev. **12,** 640–9 (2002).

67. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).

68. Subramaniam, A. R., Zid, B. M. & O'Shea, E. K. An integrated approach reveals regulatory controls on bacterial translation elongation. *Cell* **159,** 1200–11 (2014).

69. Zhou, T., Weems, M. & Wilke, C. O. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol. Biol. Evol.* **26,** 1571–80 (2009).

70. Wilke, C. O. & Drummond, D. A. Signatures of protein biophysics in coding sequence evolution. *Curr. Opin. Struct. Biol.* **20,** 385–9 (2010).

71. Akashi, H. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* **136,** 927–35 (1994).

72. Stoletzki, N. & Eyre-Walker, A. Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Mol. Biol. Evol.* **24,** 374–81 (2007).

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

74. Akashi, H. & Eyre-Walker, A. Translational selection and molecular evolution. *Curr. Opin. Genet. Dev.* **8,** 688–93 (1998).

75. Akashi, H. Translational selection and yeast proteome evolution. *Genetics* **164,** 1291–303 (2003).

76. Andersson, S. G. & Kurland, C. G. Codon preferences in free-living microorganisms. *Microbiol. Rev.* **54,** 198–210 (1990).

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

78. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324,** 255–8 (2009).

79. Goodman, D. B., Church, G. M. & Kosuri, S. Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342,** 475–9 (2013).

80. Salis, H. M. The ribosome binding site calculator. *Methods Enzymol.* **498,** 19–42 (2011).

81. Yona, A. H. *et al.* tRNA genes rapidly change in evolution to meet novel translational demands. *Elife* **2,** e01339 (2013).

82. Rogers, H. H. & Griffiths-Jones, S. tRNA anticodon shifts in eukaryotic genomes. *RNA* **20,** 269–81 (2014).

83. Ram, Y. *et al. Predicting microbial relative growth in a mixed culture from growth curve data*. *bioRxiv* (Cold Spring Harbor Labs Journals, 2015). doi:10.1101/022640

84. 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).

85. Dittmar, K. a, Sørensen, M. a, Elf, J., Ehrenberg, M. & Pan, T. Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO Rep.* **6,** 151–7 (2005).

86. Wiltrout, E., Goodenbour, J. M., Fréchin, M. & Pan, T. Misacylation of tRNA with methionine in Saccharomyces cerevisiae. *Nucleic Acids Res.* 1–13 (2012). doi:10.1093/nar/gks805

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

88. Yona, A. H., Frumkin, I. & Pilpel, Y. A Relay Race on the Evolutionary Adaptation Spectrum. *Cell* **163,** 549–559 (2015).

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

90. Wagner, A. Energy constraints on the evolution of gene expression. *Mol. Biol. Evol.* **22,** 1365–74 (2005).

91. Rang, C., Galen, J. E., Kaper, J. B. & Chao, L. Fitness cost of the green fluorescent protein in gastrointestinal bacteria. *Can. J. Microbiol.* **49,** 531–7 (2003).

92. Bienick, M. S. *et al.* The interrelationship between promoter strength, gene expression, and growth rate. *PLoS One* **9,** e109105 (2014).

93. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420,** 186–9 (2002).

94. Glick, B. R. Metabolic load and heterologous gene expression. *Biotechnol. Adv.* **13,** 247–61 (1995).

95. Sharp, P. M., Tuohy, T. M. & Mosurski, K. R. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* **14,** 5125–43 (1986).

96. Dong, H., Nilsson, L. & Kurland, C. G. Gratuitous overexpression of genes in Escherichia coli leads to growth inhibition and ribosome destruction. *J. Bacteriol.* **177,** 1497–504 (1995).

97. 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).

98. Bentley, W. E., Mirjalili, N., Andersen, D. C., Davis, R. H. & Kompala, D. S. Plasmid-encoded protein: the principal factor in the 'metabolic burden' associated with recombinant bacteria. *Biotechnol. Bioeng.* **35,** 668–81 (1990).

99. Kafri, M., Metzl-Raz, E., Jona, G. & Barkai, N. The Cost of Protein Production. *Cell Rep.* **14,** 22–31 (2016).

100. Vind, J., Sørensen, M. A., Rasmussen, M. D. & Pedersen, S. Synthesis of proteins in Escherichia coli is limited by the concentration of free ribosomes. Expression from reporter genes does not always reflect functional mRNA levels. *J. Mol. Biol.* **231,**

678–88 (1993).

101. Emilsson, V. & Kurland, C. G. Growth rate dependence of transfer RNA abundance in Escherichia coli. *EMBO J.* **9,** 4359–66 (1990).

102. Marr, A. G. Growth rate of Escherichia coli. *Microbiol. Rev.* **55,** 316–33 (1991).

103. 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,** 218–23 (2009).

104. Heyer, E. E. & Moore, M. J. Redefining the Translational Status of 80S Monosomes. *Cell* **164,** 757–69 (2016).

105. Tuller, T. & Zur, H. Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Res.* **43,** 13–28 (2014).

106. Charneski, C. A. & Hurst, L. D. Positive charge loading at protein termini is due to membrane protein topology, not a translational ramp. *Mol. Biol. Evol.* **31,** 70–84 (2014).

107. Dana, A. & Tuller, T. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res.* 1–11 (2014). doi:10.1093/nar/gku646

108. Wen, J.-D. *et al.* Following translation by single ribosomes one codon at a time. *Nature* **452,** 598–603 (2008).

109. Tholstrup, J., Oddershede, L. B. & Sørensen, M. A. mRNA pseudoknot structures can act as ribosomal roadblocks. *Nucleic Acids Res.* **40,** 303–13 (2012).

110. 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,** 3695–3700 (2002).

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

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

113. Matera, a. G. & Wang, Z. A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.* **15,** 108–121 (2014).

114. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338,** 1593–9 (2012).

115. Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338,** 1587–93 (2012).

116. Petibon, C., Parenteau, J., Catala, M. & Elela, S. A. Introns regulate the production of ribosomal proteins by modulating splicing of duplicated ribosomal protein genes. *Nucleic Acids Res.* **44,** gkw140 (2016).

117. Parenteau, J. *et al.* Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. *Cell* **147,** 320–31 (2011).

118. Reyes, a. *et al.* Drift and conservation of differential exon usage across tissues in primate species. *Proc. Natl. Acad. Sci.* 2–7 (2013). doi:10.1073/pnas.1307202110

119. Bush, S. J., Chen, L., Tovar-Corona, J. M. & Urrutia, A. O. Alternative splicing and the evolution of phenotypic novelty. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **372,** (2017).

120. Irimia, M. & Roy, S. W. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb. Perspect. Biol.* **6,** (2014).

121. Roy, S. W. & Irimia, M. Mystery of intron gain: new data and new models. *Trends Genet.* **25,** 67–73 (2009).

122. Hooks, K. B., Delneri, D. & Griffiths-Jones, S. Intron evolution in Saccharomycetaceae. *Genome Biol. Evol.* (2014). doi:10.1093/gbe/evu196

123. Shabalina, S. a *et al.* Distinct patterns of expression and evolution of intronless and intron-containing mammalian genes. *Mol. Biol. Evol.* **27,** 1745–9 (2010).

124. Lee, S. & Stevens, S. W. Spliceosomal intronogenesis. *Proc. Natl. Acad. Sci. U. S. A.* 201605113 (2016). doi:10.1073/pnas.1605113113

125. Derr, L. K., Strathern, J. N. & Garfinkel, D. J. RNA-mediated recombination in S. cerevisiae. *Cell* **67,** 355–64 (1991).

126. Yofe, I. *et al.* Accurate, Model-Based Tuning of Synthetic Gene Expression Using Introns in S. cerevisiae. *PLoS Genet.* **10,** e1004407 (2014).

127. Ares, M., Grate, L. & Pauling, M. H. A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA* **5,** 1138–9 (1999).

128. de Nadal, E., Ammerer, G. & Posas, F. Controlling gene expression in response to stress. *Nat. Rev. Genet.* **12,** 833–45 (2011).

129. Metzl-Raz, E. *et al.* Principles of cellular resource allocation revealed by condition-dependent proteome profiling. *Elife* **6,** (2017).

130. Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11,** 4241–57 (2000).

131. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533,** 397–401 (2016).

132. Brackley, C. A., Romano, M. C. & Thiel, M. The dynamics of supply and demand in mRNA translation. *PLoS Comput. Biol.* **7,** e1002203 (2011).

133. Martínez-Lumbreras, S., Taverniti, V., Zorrilla, S., Séraphin, B. & Pérez-Cañadillas, J. M. Gbp2 interacts with THO/TREX through a novel type of RRM domain. *Nucleic Acids Res.* **44,** 437–448 (2016).

134. Shepard, P. J. & Hertel, K. J. The SR protein family. *Genome Biol.* **10,** 242 (2009).

135. Kress, T. L., Krogan, N. J. & Guthrie, C. A Single SR-like Protein, Npl3, Promotes Pre-mRNA Splicing in Budding Yeast. *Mol. Cell* **32,** 727–734 (2008).

136. Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nat. Rev. Genet.* **14,** 827–39 (2013).

137. Levy, S. F. *et al.* Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **advance on,** (2015).

138. Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. & Desai, M. M. The dynamics of molecular evolution over 60,000 generations. *Nature* (2017). doi:10.1038/nature24287

139. McDonald, M. J., Rice, D. P. & Desai, M. M. Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature* **531,** 233–6 (2016).

eLIFE
elife.elifesciences.org

RESEARCH ARTICLE

tRNA genes rapidly change in evolution to meet novel translational demands

Avihu H Yona[1†], Zohar Bloom-Ackermann[1†], Idan Frumkin[1†], Victor Hanson-Smith[2,3], Yoav Charpak-Amikam[1], Qinghua Feng[4], Jef D Boeke[5‡], Orna Dahan[1], Yitzhak Pilpel[1]*

[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel; [2]Department of Microbiology, University of California, San Francisco, San Francisco, United States; [3]Department of Immunology, University of California, San Francisco, San Francisco, United States; [4]Department of Pathology, University of Washington, Seattle, United States; [5]Department of Molecular Biology and Genetics, The Johns Hopkins University School of Medicine, Baltimore, United States

Abstract Changes in expression patterns may occur when organisms are presented with new environmental challenges, for example following migration or genetic changes. To elucidate the mechanisms by which the translational machinery adapts to such changes, we perturbed the tRNA pool of *Saccharomyces cerevisiae* by tRNA gene deletion. We then evolved the deletion strain and observed that the genetic adaptation was recurrently based on a strategic mutation that changed the anticodon of other tRNA genes to match that of the deleted one. Strikingly, a systematic search in hundreds of genomes revealed that anticodon mutations occur throughout the tree of life. We further show that the evolution of the tRNA pool also depends on the need to properly couple translation to protein folding. Together, our observations shed light on the evolution of the tRNA pool, demonstrating that mutation in the anticodons of tRNA genes is a common adaptive mechanism when meeting new translational demands.
DOI: 10.7554/eLife.01339.001

***For correspondence:** Pilpel@weizmann.ac.il

[†]These authors contributed equally to this work

[‡]**Present address:** New York University Langone Medical Center, New York, United States

Competing interests: The authors declare that no competing interests exist.

Funding: See page 13

Received: 05 August 2013
Accepted: 30 October 2013
Published: 20 December 2013

Reviewing editor: Michael Laub, Massachusetts Institute of Technology, United States

Introduction

The process of gene translation is fundamental to the function of living cells, and as such its apparatus is highly conserved across the tree of life (***Müller and Wittmann-Liebold, 1997***; ***Itoh et al., 1999***; ***Wolf et al., 2001***). Yet, the capacity of the translation machinery to adaptively evolve is crucial in order to support life in changing environments. Therefore, a key open question is to identify the mechanisms by which the translation machinery adapts to changing conditions.

A thoroughly studied aspect of translation that demonstrates its adaptation capacities is the different proportions by which synonymous codons are used, a phenomenon known as 'codon usage bias'. Although differential use of codons can be the result of neutral processes such as mutational biases and the genomic GC content (***Urrutia and Hurst, 2001***; ***Rao et al., 2011***), natural selection also influences codon usage bias. Indeed, it has been demonstrated that codon choice affects expression level, protein folding, translational accuracy, and other translational features (***Akashi, 1994***; ***Parmley and Hurst, 2007***; ***Zhou et al., 2009***; ***Hudson et al., 2011***). Since both neutral and selective processes govern codon usage bias, the balance between selection, mutational bias and drift is crucial in shaping the codon usage of each species (***Bulmer, 1991***). Importantly, although the selective advantage offered by alternative synonymous codons is considered to be moderate, it was recently demonstrated that selection can still shape codon usage patterns in vertebrates even with their small effective population sizes (***Doherty and McInerney, 2013***).

Notably, the differential usage of codons represents the evolution of the 'demand' aspect of translation, namely the codon usage of all expressed genes. Yet, the adaptation mechanisms of the 'supply',

eLife digest Genes contain the blueprints for the proteins that are essential for countless biological functions and processes, and the path that leads from a particular gene to the corresponding protein is long and complex. The genetic information stored in the DNA must first be transcribed to produce a messenger RNA molecule, which then has to be translated to produce a string of amino acids that fold to form a protein. The translation step is performed by a molecular machine called the ribosome, with transfer RNA molecules bringing the amino acids that are needed to make the protein.

The information in messenger RNA is stored as a series of letters, with groups of three letters called codons representing the different amino acids. Since there are four letters—A, C, G and U—it is possible to form 64 different codons. And since there are only 20 amino acids, two or more different codons can specify the same amino acid (for example, AGU and AGC both specify serine), and two or more different transfer RNA molecules can take this amino acid to the ribosome. Moreover, some codons are found more often than others in the messenger RNA molecules, so the genes that encode the related transfer RNA molecules are more common than the genes for other transfer RNA molecules.

Environmental pressures mean that organisms must adapt to survive, with some genes and proteins increasing in importance, and others becoming less important. Clearly the relative numbers of the different transfer RNA molecules will also need to change to reflect these evolutionary changes, but the details of how this happens were not understood.

Now Yona et al. have explored this issue by studying yeast cells that lack a gene for one of the less common transfer RNA molecules (corresponding to the codon AGG, which specifies the amino acid arginine). At first this mutation resulted in slower growth of the yeast cells, but after being allowed to evolve over 200 generations, the rate of growth matched that of a normal strain with all transfer RNA genes. Yona et al. found that the gene for a more common transfer RNA molecule, corresponding to the codon AGA, which also specifies arginine, had mutated to AGG. As a result, the mutated yeast was eventually able to produce proteins as quickly as wild type yeast. Moreover, further experiments showed that the levels of some transfer RNAs are kept deliberately low in order to slow down the production of proteins so as to ensure that the proteins assume their correct structure.

But does the way these cells evolved in the lab resemble what happened in nature? To address this question Yona et al. examined a database of transfer RNA sequences from more than 500 species, and found evidence for the same codon-based switching mechanism in many species across the tree of life.

DOI: 10.7554/eLife.01339.002

namely the expression level of each tRNA type that is loaded with an amino acid, are not fully understood. While ribosomal genes do not exhibit appreciable changes in response to environmental alterations (***Müller and Wittmann-Liebold, 1997***; ***Itoh et al., 1999***; ***Wolf et al., 2001***), tRNA genes may provide an important source of evolutionary plasticity for fine tuning translation.

tRNAs constitute a fundamental component in the process of translation, linking codons to their corresponding amino acids (***Widmann et al., 2010***). tRNA genes are classified into gene families according to their anticodon, with each gene family containing between one and several copies scattered throughout the genome. Importantly, it has been experimentally observed for *Saccharomyces cerevisiae* (***Tuller et al., 2010***) and *Escherichia coli* (***Dong et al., 1996***) that the cellular concentrations of each tRNA family in the cell (i.e., the tRNA pool) correlate with its genomic tRNA copy number (***Percudani et al., 1997***; ***Kanaya et al., 1999***). Notably, the rate-limiting step of polypeptide elongation is the recruitment of a tRNA that matches the translated codon (***Varenne et al., 1984***). Thus, the translation efficiency is defined as the extent to which the tRNA pool can accommodate the transcriptome (***Sharp and Li, 1987***; ***Dos Reis et al., 2004***; ***Stoletzki and Eyre-Walker, 2007***), thereby affecting protein production and accuracy.

In general, highly expressed genes exhibit a marked codon usage bias toward 'optimal' codons, whose corresponding tRNA gene copy number is high (***Sharp and Li, 1986a***, ***1986b***). The evolutionary force that acts to maintain optimal translation efficiency of such genes was coined 'translational

selection' (**Dos Reis et al., 2004**). It was previously suggested that translational selection acts to maintain a balance between codon usage and tRNA availability. On the one hand, there is a selective pressure to increase the frequency of preferred codons in highly expressed genes. On the other hand, changes in the tRNA pool may also occur, for example duplication of tRNA genes for which high codon demand exists. Thus, codon frequencies and tRNA copy numbers coevolve toward a supply versus demand balance that facilitates optimal protein production (**Higgs and Ran, 2008**; **Gingold et al., 2012**).

The fitness effects of an unmet translational demand and its potential role in shaping the tRNA pool are not fully characterized. Evolutionary changes to the tRNA pool were appreciated mainly via bioinformatics studies (**Rawlings et al., 2003**; **Withers et al., 2006**; **Higgs and Ran, 2008**; **Bermudez-Santana et al., 2010**; **Rogers et al., 2010**) and only a handful of experimental findings have been reported, which rely on genetic manipulations (**Byström and Fink, 1989**; **Von Pawel-Rammingen et al., 1992**; **Aström et al., 1993**) or direct mutagenesis (**Saks et al., 1998**). Sequence analyses of divergent genomes have demonstrated that both the sequence and copy number of tRNA genes may change among various species or strains. However, it is unclear whether the observed variations in the tRNA pool are a consequence of an adaptive process due to unbalanced translational demand or the result of random genomic processes, as tRNA genes are a known source of genomic instability (**McFarlane and Whitehall, 2009**).

Further, the forces that direct and maintain low copy tRNA families remain unclear. Specifically, it is not clear whether translational selection acts only to favor optimal codons or also acts in particular cases to keep other codons deliberately as 'non-optimal' by maintaining their tRNA supply at low level. Encoding genes with optimal codons might not always lead to higher protein expression levels (**Kudla et al., 2009**). Similarly, the use of 'slow codons' may not always result in lower levels of protein expression as they could have functional roles in improving expression, for example when enriched at the beginning of a transcript in order to reduce the energy of the RNA structure (**Goodman et al., 2013**) or to efficiently allocate ribosomes along the mRNA (**Tuller et al., 2010**). Additionally, it has been proposed that non-optimal codons may play a role in governing the process of cotranslational folding by slowing down translation, which supports proper folding between domain boundaries (**Thanaraj, 1996**; **Kramer et al., 2009**; **Cabrita et al., 2010**; **Wilke and Drummond, 2010**; **Pechmann and Frydman, 2012**). Yet, the contribution of non-optimal codons to proper protein folding was observed only for specific genes (**Crombie et al., 1992**; **Komar et al., 1999**; **Cortazzo et al., 2002**; **Tsai et al., 2008**; **Zhang et al., 2009**; **Zhou et al., 2013**). Thus, the extent and relevance of this phenomenon to the global folding state of the proteome needs to be substantiated.

To elucidate the importance of restoring translational equilibrium, we used an experimental evolution approach. To this end, we genetically perturbed the tRNA pool of the budding yeast *S. cerevisiae*. In this yeast, the genetic code is decoded by 42 different tRNA families that make up a total of 274 tRNA genes (**Chan and Lowe, 2009**). Each tRNA gene family ranges from 1–16 copies, with 6 tRNA families consisting of only a single copy. In a recent study (**Bloom-Ackermann et al., In press**), we have systematically manipulated the tRNA pool in *S. cerevisiae* by individually deleting most tRNA genes from its genome. Here, we focus on one particular deletion strain that showed the most extreme fitness reduction among the viable deletion mutants in this tRNA deletion library. This tRNA exists in only one copy in the genome, thus after its deletion the cell is left without a tRNA with the corresponding anticodon. Lab-evolution experiments performed on this strain demonstrated that the translational balance was rapidly restored by mutations in other tRNA genes that compensated for the tRNA deletion. An extensive bioinformatic analysis revealed that a similar evolutionary trend is widespread in nature too, suggesting that the anticodon mutations we observed in the lab recapitulate an existing mechanism that shapes the tRNA pool. To shed light on the constraints that shape the size of tRNA gene families, we artificially overexpressed singleton tRNAs, rather than deleting them. We found that when low copy tRNAs were overexpressed, the protein quality control machinery was challenged due to increased proteotoxic stress. This observation suggests that low tRNA availability for particular codon can serve an essential means to ensure proper cotranslation folding of proteins.

Results

Deletion of singleton tRNA gene breaks the translational balance

To demonstrate the importance of the balance between codon usage and the cellular tRNA pool we created a yeast strain in which the single copy of the arginine tRNA gene, *tR(CCU)J*, was deleted

(designated ΔtRNA$^{Arg}_{CCU}$). Consequently, in this deletion strain, the arginine codon AGG cannot be translated with its fully matched tRNA and it is presumably translated by another arginine tRNA, tRNA$^{Arg}_{UCU}$, owing to a wobble interaction (***Begley et al., 2007***). This shortage in tRNA supply is particularly evident given the demand: AGG is the second most highly used codon for arginine in the yeast genome (***Supplementary file 1A***). Indeed, the ΔtRNA$^{Arg}_{CCU}$ strain showed a severe growth defect compared to the wild-type strain (***Figure 1A***, blue and green curves, respectively). This growth difference demonstrates the effect of translational imbalance on cellular growth. Although the deletion mutant of this single copy tRNA is viable (***Clare et al., 1988***; ***Kawakami et al., 1993***), its severe growth defect also reveals the inability of the wobble interactions to fully compensate for the tRNA gene deletion.

The tRNA pool can rapidly evolve to meet translational demands

To learn how genomes adapt to translational imbalances, we performed lab-evolution experiments on the ΔtRNA$^{Arg}_{CCU}$ strain, employing the procedure of daily growth and dilution to a fresh medium (***Lenski et al., 1991***). The deletion strain was grown under optimal laboratory conditions (rich medium at 30°C) and was diluted daily into a fresh medium by a factor of 120, corresponding to approximately 7 generations per cycle. Every 50 generations, the growth of the evolving population was compared to both the wild-type and the ancestor ΔtRNA$^{Arg}_{CCU}$ strains. Strikingly, after 200 generations we observed a full recovery of the growth defect of the ancestor strain ΔtRNA$^{Arg}_{CCU}$, as the growth curve of the evolved population was indistinguishable from that of the wild-type strain (***Figure 1A***, red curve). Similar dynamics were observed in all four independent evolutionary lines.

In search of the potential genetic adaptations underlying this rapid recovery, we first looked for genetic alterations in other arginine tRNA genes. We found a single point mutation in another arginine tRNA gene that codes for tRNA$^{Arg}_{UCU}$. This mutation changed the anticodon triplet of tRNA$^{Arg}_{UCU}$ from UCU to CCU (i.e., T→C transition). Consequently, the evolved tRNA$^{Arg}_{UCU}$ perfectly matched the AGG codon (***Figure 1B***). Unlike the singleton tRNA$^{Arg}_{CCU}$, there are 11 copies of tRNA$^{Arg}_{UCU}$ in the yeast genome. Although each of the 4 independent lab-evolution experiments showed the exact same solution (that is, a mutation in the anticodon of a tRNA$^{Arg}_{UCU}$ gene), 3 different copies of this gene were changed among the 4 lines (i.e., 1 of the 11 copies of tRNA$^{Arg}_{UCU}$ was mutated in 2 repetitions; see 'Materials and methods'). To confirm that a single point mutation in the anticodon of tRNA$^{Arg}_{UCU}$ is sufficient to fully compensate for the growth defect of ΔtRNA$^{Arg}_{CCU}$, we artificially inserted the same T→C mutation into the deletion ΔtRNA$^{Arg}_{CCU}$ mutant. We inserted the mutation into 1 of the 11 copies of the tRNA$^{Arg}_{UCU}$ genes, a copy that resides on chromosome XI, and thus spontaneously mutated in 1 of the evolution lines. Indeed, the artificially mutated strain, termed as *MutΔtRNA$^{Arg}_{CCU}$*, showed a full recovery of the deletion adverse phenotype (***Figure 1C***). This indicates that the T→C mutation in the anticodon is sufficient for the full recovery of the tRNA$^{Arg}_{CCU}$ deletion phenotype.

Mutated tRNA$^{Arg}_{UCU}$ is functional despite sequence dissimilarities with respect to the deleted tRNA$^{Arg}_{CCU}$

In general, all copies of each tRNA gene family tend to be highly similar in sequence in *S. cerevisiae* (***Chan and Lowe, 2009***). In particular, the sequences of the 11 copies of tRNA$^{Arg}_{UCU}$ are 100% identical to each other. Yet, the 2 arginine tRNA, tRNA$^{Arg}_{UCU}$, and tRNA$^{Arg}_{CCU}$, differ in 21 of their 72 nucleotides (including the third anticodon position, ***Figure 2A***). Thus, the evolutionary solution that occurred in our experiments created a 'chimeric' tRNA with a CCU anticodon, whereas the rest of the tRNA sequence (termed as the 'tRNA scaffold') remained as tRNA$^{Arg}_{UCU}$. The sequence identity among all members of the tRNA$^{Arg}_{UCU}$ family suggests a functional role for the tRNA scaffold in addition to that of the anticodon (***Schultz and Yarus, 1994***; ***Konevega et al., 2004***; ***Cochella and Green, 2005***; ***Olejniczak et al., 2005***; ***Saks and Conery, 2007***; ***Schmeing et al., 2011***). Therefore, it is surprising that the chimeric tRNA performed just as well as the deleted tRNA$^{Arg}_{CCU}$ in terms of cell growth, despite the major sequence differences between the two tRNA scaffolds. Thus, we raised the hypothesis that more challenging growth conditions may expose possible inadequacies in the chimeric tRNA. To test this notion, we compared the rescued strain, *MutΔ*tRNA$^{Arg}_{CCU}$, which carries the chimeric tRNA, to the wild-type strain under an array of challenging conditions. Surprisingly, under all the examined conditions, we observed no significant growth difference between the two strains (***Figure 2B***). Hence, the chimeric tRNA provides a direct in vivo indication that the scaffolds of tRNAs, which encode for the same amino acid, may be interchangeable in terms of their effect on cellular growth under the conditions we tested.

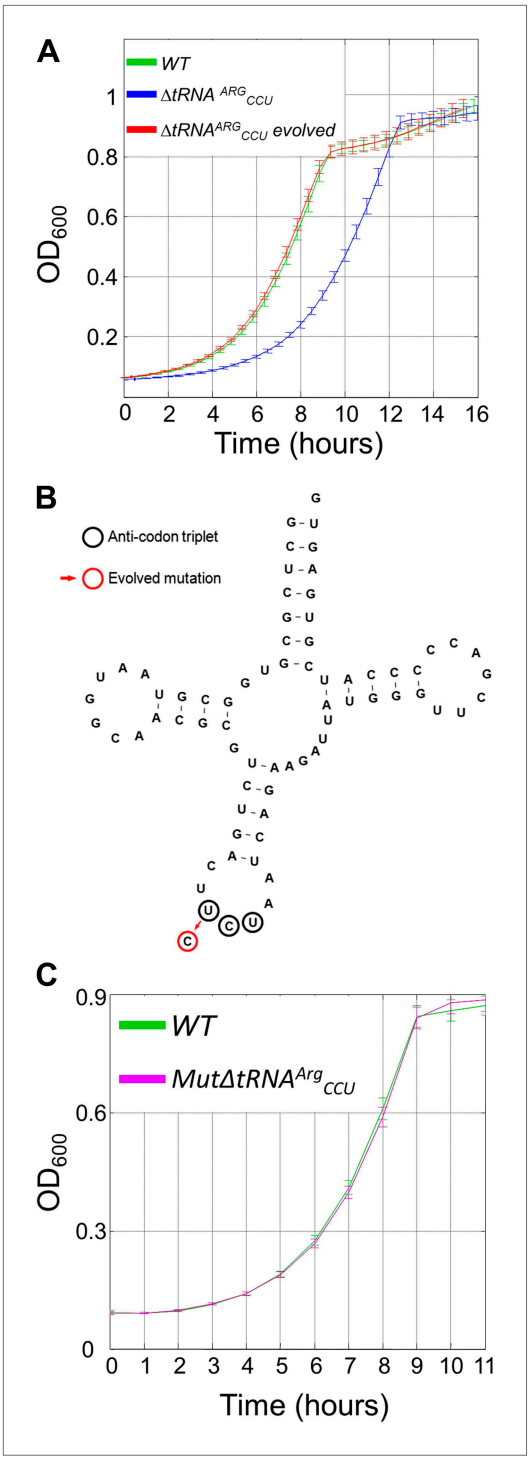


Figure 1. The growth defect associated with deletion of a singleton tRNA gene was rapidly rescued during the lab-evolution experiment. (**A**) Growth curve measurements of wild-type (WT) (green), ΔtRNA$^{Arg}_{CCU}$ (blue) and the evolved deletion (red) are shown in optical density (OD) values over time during continuous growth on rich medium at 30°C. (**B**) The mutation that was found to recover the deletion phenotype in the evolved strains is shown on the secondary structure of tRNA$^{Arg}_{UCU}$.

Figure 1. Continued on next page

To examine the generality of our observation, we once again perturbed the tRNA pool in a wild-type (WT) strain by deletion of an entire serine tRNA family, tRNA$^{Ser}_{GCU}$ that has four copies in the genome. A complete deletion of this gene family is lethal, indicating that the tRNA$^{Ser}_{GCU}$ is essential in *S. cerevisiae*. Although we could not evolve that strain, we did find that this quadruple deletion strain was viable when supplemented with a centromeric plasmid carrying the tRNA$^{Ser}_{GCU}$ gene. Thus, the lethality is conferred directly from the tRNA loss and is not due to other indirect effects (***Figure 2—figure supplement 1***). We hypothesized that, as with tRNA$^{Arg}_{CCU}$, other chimeric serine tRNAs that carry a GCU anticodon, yet with the scaffold of another tRNA for serine, would also prevent the observed lethality. Indeed, a strain carrying a chimeric tRNA with a scaffold of tRNA$^{Ser}_{CGA}$ and the GCU anticodon is viable on the background of the tRNA$^{Ser}_{GCU}$-family deletion (***Figure 2—figure supplement 2***). Therefore, we concluded that the identity of the anticodon is essential for the function of the tRNA$^{Ser}_{GCU}$ gene family. Thus, it appears that for the examined tRNAs the anticodon is a dominant feature in terms of cellular fitness, overshadowing other sequence elements.

Anticodon switching is a widespread phenomenon in nature

Although the anticodon of a tRNA gene was rapidly mutated under our laboratory conditions, thus regaining proper translational equilibrium, it is unclear to what extent this mechanism naturally occurs in species across the tree of life. To address this question, we performed a systematic bioinformatics screen for tRNA switching events in nature. We defined an anticodon-switching event as a case of a tRNA whose nucleotide sequence is closer to a tRNA gene with a different anticodon than to a tRNA gene with the same anticodon. To this end, we downloaded all the known tRNA sequences from the Genomic tRNA Database (***Chan and Lowe, 2009***), a collection that stores the tRNA pools of 524 species. We masked the anticodon triplet as 'NNN' in all tRNA genes, aligned all tRNA sequences from each species individually and inferred a maximum likelihood phylogenetic tree for each alignment. For each tRNA sequence, we calculated the shortest phylogenetic distance to another tRNA with the same anticodon (designated d_{same}) and the shortest distance to another tRNA with a different anticodon (designated d_{diff}). For each species, we defined its set of tRNA switching events as those in which $d_{diff}<d_{same}$ (see 'Materials and methods', ***Figure 3—source data 1***).

Figure 1. Continued

The UCU anticodon nucleotides are marked with black circles, and the red circle indicates the mutation that occurred in the anticodon, that is T→C transition. (**C**) *MutΔtRNA*$^{Arg}_{CCU}$ in which the same mutation that was found in the evolved strain was deliberately engineered, exhibits similar growth as the WT. Growth curve measurements of WT (green) and of *MutΔtRNA*$^{Arg}_{CCU}$ (magenta) are shown in OD_{600} values over time during continuous growth on rich medium at 30°C.

DOI: 10.7554/eLife.01339.003

Our analysis included 416 eubacterial, 68 eukaryotic, and 40 archaeal species. We found that tRNA switching events are present in all domains of life, as we detected at least 1 tRNA switching event per species in 8 bacteria, 58 eukarya, and 1 archaeal species (***Figure 3A***). A retrospective counting revealed that most switching events occurred due to a mutation in the first position in the anticodon triplet that corresponds to the third codon position (see details in 'Materials and methods'). For comparison, we masked as 'NNN' additional triplets of nucleotides within the tRNA molecule, and found a higher percentage of discrepancies compared to the anticodon triplet (***Figure 3—figure supplement 1***; ***Supplementary file 1B***).

Figure 3 demonstrates two examples of tRNA switching events, the first in *Mus musculus* and the second in *Homo sapiens*. In the first example, the phylogeny of tRNA sequences with glutamic acid anticodons is presented (***Figure 3B***). Notably, six out of the eight tRNAs with a UUC anticodon in *M. musculus* were clustered together in our analysis, while two other copies of the same anticodon identity were clustered closer to tRNA genes with a CUC anticodon (***Figure 3C***). The second example demonstrates a switching event for tRNA genes encoding for valine anticodons. In this study, a tRNA with a UAC anticodon was clustered with CAC and AAC tRNA genes and not with the other four UAC tRNAs (***Figure 3D,E***). Interestingly, the CAC and AAC tRNA genes are intermixed in the tree, suggesting that anticodon switching was prevalent in the evolution of CAC and AAC tRNA genes in *H. sapiens* (***Figure 3D,E***). Also of interest, the switching events shown in mouse were not found in human and vice versa. Thus, in each of these two mammals the switching examples shown here probably occurred after they split from their common ancestor. In general, inspecting the relationship across species between the size of the tRNA pool and the number of detected switching events revealed a modest correlation, and in particular species with same size of tRNA repertoire manifested tRNA switching to different extents (not shown). This analysis suggests that future examination of the tRNA switching phenomenon in individual species could be of interest.

Multiple copies of rare tRNAs are deleterious to the cell

After demonstrating the prevalence of anticodon switching, we refocused on our lab-evolution results. The switching events that we observed (from tRNA$^{Arg}_{UCU}$ to tRNA$^{Arg}_{CCU}$) suggest that the effective gene copy number of each tRNA anticodon set can change during evolution, presumably due to demand-to-supply changes. Given that a single point mutation can functionally convert one tRNA into another, an interesting question emerges: why does the genome maintain a single copy of tRNA$^{Arg}_{CCU}$? T to C mutations must have occurred in evolution but they appear to have been selected against so as to preserve only a single copy of the CCU anticodon tRNA. Consistent with this hypothesis is the observation that other yeast species maintain tRNA$^{Arg}_{CCU}$ at a single copy (***Supplementary file 1C***). We thus reasoned that an artificial increase in the copy number of a rare tRNA, but not of an abundant one, might result in a deleterious effect on the cells.

Indeed, transformation of a multi-copy plasmid containing a tRNA$^{Arg}_{CCU}$ gene to a wild-type strain (termed as *WTmultiCCU*) resulted in a substantial growth reduction when compared to wild-type cells carrying an empty multi-copy plasmid (termed *WTmultiControl*). In contrast, when we created a strain with a similar multi-copy plasmid that contains the abundant tRNA$^{Arg}_{UCU}$ gene, designated as *WTmultiUCU*, a growth profile much closer to that of *WTmultiControl* was observed (***Figure 4A***). These findings are consistent with the evolutionary tendency for yeast to keep a low copy number of tRNA$^{Arg}_{CCU}$ and suggest that a high copy number of such rare tRNAs is deleterious to cells.

To demonstrate the generality of our findings, we employed the same assays in two additional cases. First, we examined 2 serine tRNAs, the singleton tRNA$^{Ser}_{CGA}$ and tRNA$^{Ser}_{AGA}$ that is found in the genome in 11 copies. In the second case, we focused on two glutamine tRNAs, the singleton tRNA$^{Gln}_{CUG}$ and tRNA$^{Gln}_{UUG}$ that is found in the genome in nine copies. In both the cases, we observed that the wild-type strain supplemented with multiple copies of a singleton tRNA exhibit impaired growth compared to the same strain supplemented with the abundant tRNA for the same amino acid (***Figure 4—figure supplements 1 and 2***). Since the changes in tRNA family sizes during evolution likely occur gradually,

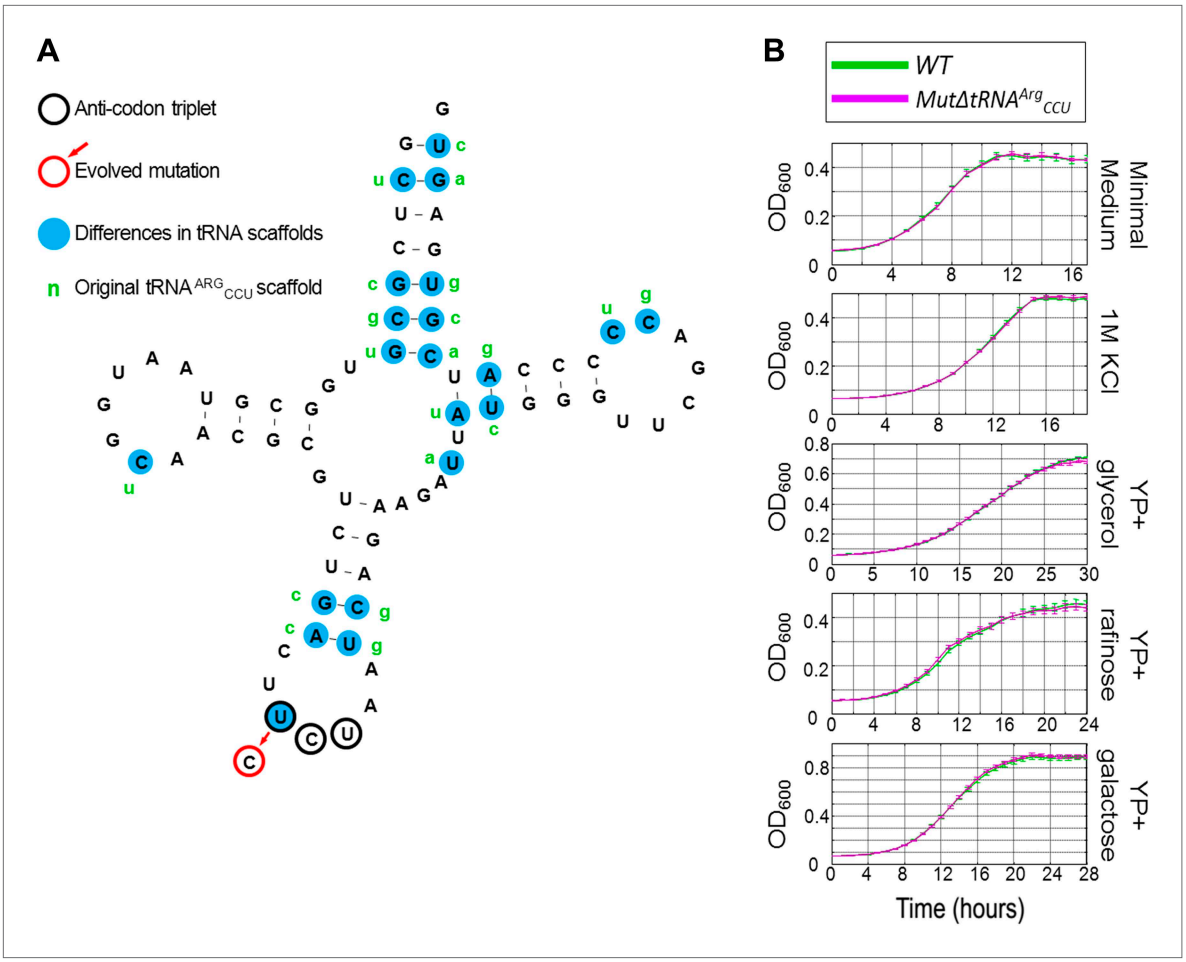


Figure 2. The growth of MutΔtRNA$^{Arg}_{CCU}$ carrying the chimeric tRNA compared to wild-type (WT) under different conditions. (**A**) The sequence of the chimeric tRNA is drawn showing the scaffold of tRNA$^{Arg}_{UCU}$ with the mutated CCU anticodon. The anticodon triplet is marked with black circles. The evolved mutation is marked with a red circle. All 20 nucleotide differences between tRNA$^{Arg}_{UCU}$ and tRNA$^{Arg}_{CCU}$ are marked with blue background, next to which, in green letters, the original nucleotide of tRNA$^{Arg}_{CCU}$ are written. (**B**) Growth curve measurements of WT (green) and of *MutΔtRNA$^{Arg}_{CCU}$* (magenta) are shown in OD_{600} values over time during continuous growth.

DOI: 10.7554/eLife.01339.004

The following figure supplements are available for figure 2:

Figure supplement 1. Quadruple deletion of tRNA$^{ser}_{GCU}$ is lethal.

DOI: 10.7554/eLife.01339.005

Figure supplement 2. A chimeric serine tRNA can rescue the lethality of the quadruple deletion.

DOI: 10.7554/eLife.01339.006

perhaps one copy at a time, we also examined the effect of adding low copy number plasmids carrying either tRNA$^{Arg}_{UCU}$ or tRNA$^{Arg}_{CCU}$. The cells with the tRNA$^{Arg}_{CCU}$ plasmid showed a modest growth defect compared to the cells with tRNA$^{Arg}_{UCU}$ plasmid, yet only when grown at 39°C (***Figure 4—figure supplement 3***).

Multiple copies of the rare tRNA$^{Arg}_{CCU}$ induce proteotoxic stress

Why is it essential to keep certain tRNAs at a low level? One interesting possibility is that rare tRNAs are essential for the process of cotranslation folding, presumably because low abundance tRNAs provide a pause in translation that might be needed for proper folding (***Thanaraj, 1996***; ***Drummond and Wilke, 2008***; ***Cabrita et al., 2010***; ***Pechmann and Frydman, 2012***). Other deleterious effects that may stem from a high copy number of tRNA$^{Arg}_{CCU}$ could be misincorporation of arginine into non-arginine codons, or the misloading of arginine tRNA molecules with other amino acids. These potential sources of errors are not mutually exclusive and can each contribute to the observed growth defect by exerting a protein folding stress.

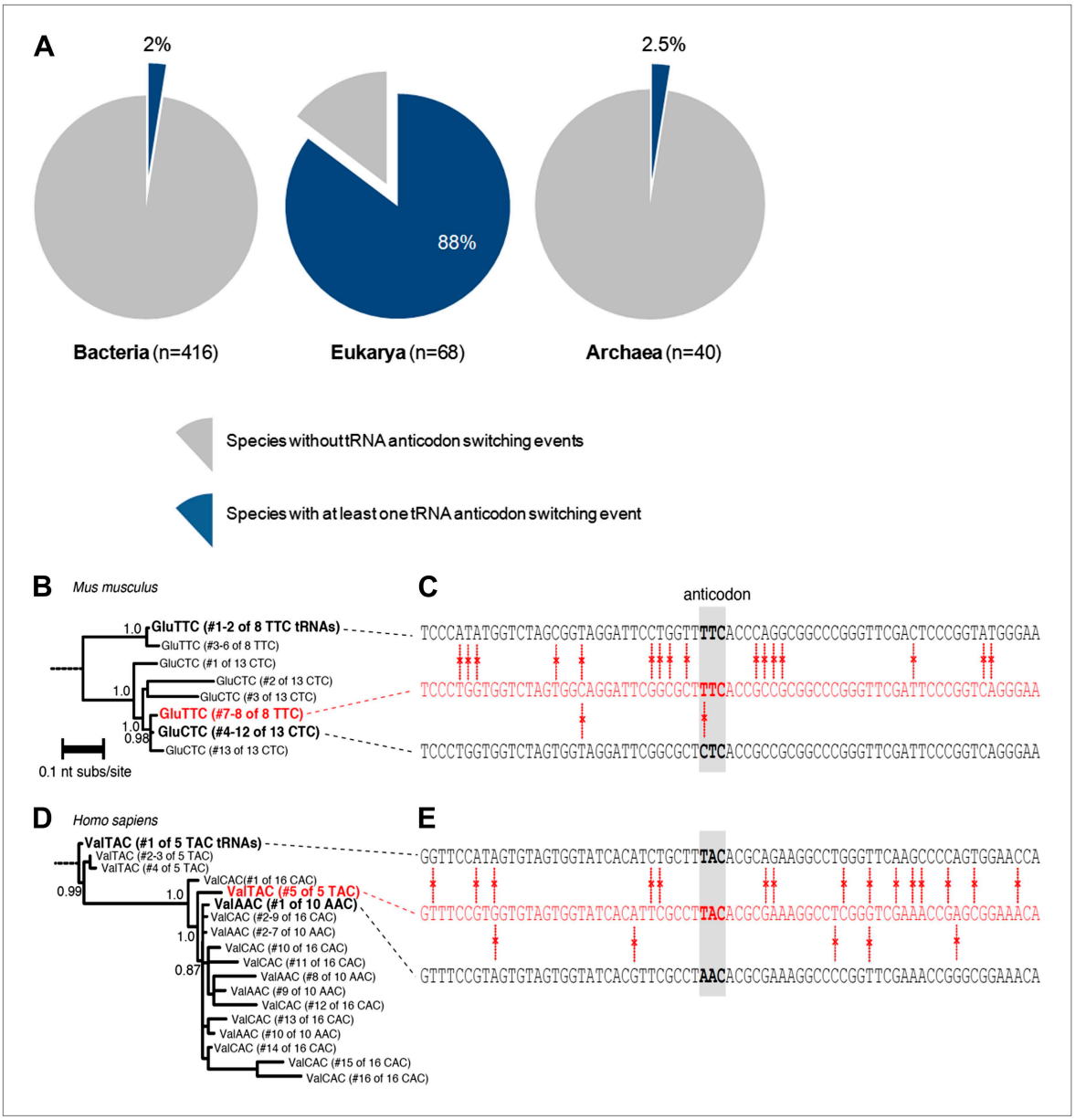


Figure 3. Anticodon switching is a widespread phenomenon in nature. (**A**) Number of species with at least one tRNA switching event in each domain of life. (**B**) The anticodon UUC convergently evolved in *Mus musculus*. A maximum likelihood phylogeny of tRNA sequences in *M. musculus* that decode glutamic acid (Glu) codons. Branch lengths express average nucleotide substitutions per site. Decimals on internal branches express branch support. (**C**) A comparison of nucleotide sequences for glutamic acid tRNA genes in *M. musculus* with anticodon UUC (top, tRNA1547 and tRNA359), 'switched' UUC tRNAs (middle, tRNA286 and tRNA754), and CUC tRNAs (bottom, tRNA1002, tRNA745, tRNA303, tRNA999, tRNA996 tRNA709, tRNA1001, tRNA1912 and tRNA81). The anticodon triplet is boxed in gray. Red vertical bars indicate differences between sequences. (**D**) The anticodon UAC convergently evolved in *Homo sapiens*. A maximum likelihood phylogeny of tRNA sequences in *H. sapiens* encoding for valine (Val) is shown. (**E**) A comparison of nucleotide sequences for *H. sapiens* tRNAs with anticodons UAC (top, tRNA6), a 'switched' UAC tRNA (middle, tRNA40), and an AAC tRNA (bottom, tRNA136). The number of genes is according to the tRNA database.

DOI: 10.7554/eLife.01339.007

The following source data and figure supplements are available for figure 3:

Source data 1. Table of anticodon switchings in different species across the tree of life.

DOI: 10.7554/eLife.01339.008

Figure supplement 1. A comparison of discrepancy proportions at the anticodon triplet vs control triplets.

DOI: 10.7554/eLife.01339.009

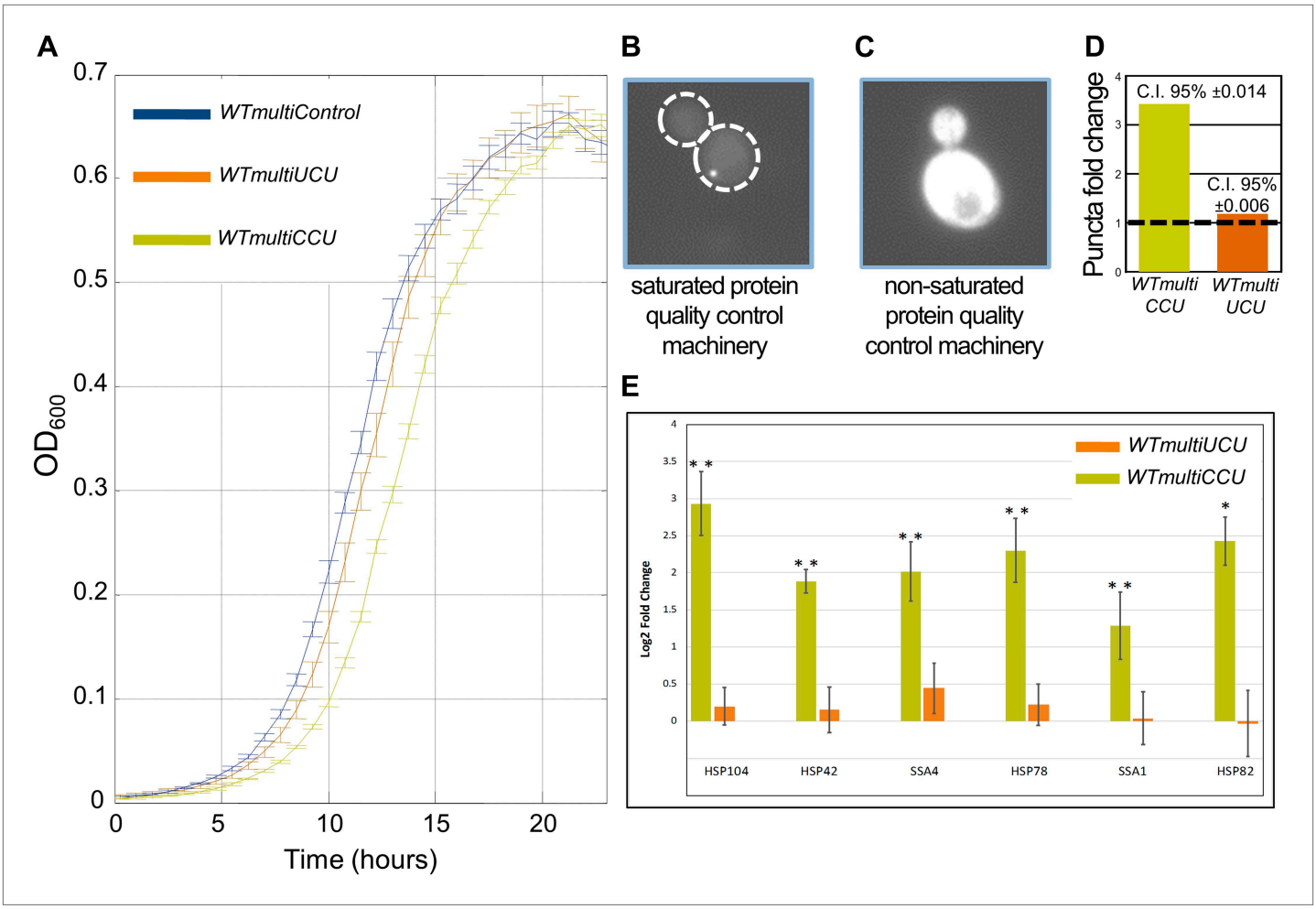


Figure 4. *WTmultiCCU* experiences a growth defect compared to *WTmultiUCU* and demonstrates higher levels of misfolded proteins. (**A**) Growth curve measurements of *WTmultiControl* (blue), *WTmultiUCU* (brown) and *WTmultiCCU* (khaki) are shown in optical density (OD) values over time during continuous growth. The *WTmultiCCU* strain carrying a high copy number plasmid harboring $tRNA^{Arg}_{CCU}$ demonstrates slower growth compared to cells with an empty plasmid or with a $tRNA^{Arg}_{UCU}$ plasmid that is mainly characterized by a longer growth delay (lag phase). (**B**) A demonstration of a *WTmultiCCU* cell in which the mCherry-Von Hippel–Lindau (VHL) proteins appear with a punctum phenotype when the protein quality control machinery is saturated with misfolded proteins. (**C**) A demonstration *WTmultiUCU* cell in which the quality control machinery is not occupied with other proteins; mCherry-VHL is localized to the cytosol. (**D**) *WTmultiCCU*, *WTmultiUCU* and *WTmultiControl* were transformed with a VHL-mCherry containing plasmid and visualized under the microscope; 1000 cells per strain were counted for either cytosolic or punctum localization of the VHL protein. The fold change in the number of cells containing puncta was then deduced by normalization to the *WTmultiControl* population. The 95% confidence interval is indicated. (**E**) The mRNA fold change of six representative heat-shock genes measured by real-time quantitative PCR (RT-qPCR). Presented values are the mean of two biological repetitions ± SEM. The significance of the fold change differences was examined using a *t* test, with *p<0.001 or **p<0.0001.

DOI: 10.7554/eLife.01339.010

The following figure supplements are available for figure 4:

Figure supplement 1. Multiple copies of rare $tRNA^{Ser}_{CGA}$ gene are deleterious compared to abundant $tRNA^{Ser}_{AGA}$.

DOI: 10.7554/eLife.01339.011

Figure supplement 2. Multiple copies of the rare $tRNA^{Gln}_{CUG}$ gene are deleterious compared to abundant $tRNA^{Gln}_{UUG}$.

DOI: 10.7554/eLife.01339.012

Figure supplement 3. Addition of low copy number $tRNA^{Arg}_{CCU}$ is deleterious compared to low copy number $tRNA^{Arg}_{UCU}$ when grown in heat.

DOI: 10.7554/eLife.01339.013

To examine the possibility that the growth defect associated with multiple copies of $tRNA^{Arg}_{CCU}$ is indeed associated with such a proteotoxic stress, we used an established method that examines the load on the protein quality control machinery of the cell (see 'Materials and methods') (***Kaganovich et al., 2008***). In this assay, we transformed cells with a plasmid harboring the human gene, von Hippel–Lindau (VHL), fused to a fluorescent tag (mCherry). Fluorescently tagged VHL that is present as

aggregated puncta (***Figure 4B***), and not as a disperse cytosolic localization (***Figure 4C***), indicates that the protein quality control machinery is saturated due to high levels of misfolding in the cell's endogenous proteins. We transformed the VHL-mCherry plasmid to each of the multi-copy tRNA strains, *WTmultiCCU*, *WTmultiUCU* and *WTmultiControl,* and monitored the level of proteotoxic stress by quantifying the number of cells with puncta phenotype in each population. The fold change in those cells was then deduced by normalization to the *WTmultiControl* population. We found that while *WTmultiUCU* exhibited similar amount of cells with puncta as the *WTmultiControl*, the *WTmultiCCU* exhibited a threefold increase (***Figure 4D***).

The proteotoxic stress experienced by the two strains overexpressing tRNAs was further assessed by measuring the induction level of an array of heat-shock proteins (HSPs) using real-time quantitative PCR (RT-qPCR). Since the HSPs have been shown to undergo induction under proteotoxic stress, they are an excellent indicator for this stress (***McClellan et al., 2005***; ***Kaganovich et al., 2008***). Indeed, we found a significant upregulation in mRNA levels for all the examined HSP genes in the *WTmultiCCU* strain, but not in the *WTmultiUCU* strain (***Figure 4E***). These findings further demonstrate that increasing the copy number of a rare tRNA gene, but not of an already abundant one, results in proteotoxic stress in the cell.

Discussion

Genomic duplications, deletions, and anticodon mutations shape tRNA gene families, yet the evolutionary scenarios that trigger changes in the tRNA pool have not been thoroughly explored. In our evolution experiments, a translational imbalance was imposed by a tRNA gene deletion that compromised growth and drove the tRNA pool to adapt to a novel translational demand. Importantly, organisms may experience equivalent imbalances when their gene expression changes due to altered environmental conditions or upon migrating to a new ecological niche (***Gingold et al., 2012***). This scenario is particularly feasible given that the genes needed in various environments do show differences in codon usage, for example respiration as opposed to fermentation in yeast (***Man and Pilpel, 2007***).

Indeed, when faced with different environmental challenges, transcriptional changes affect the codon usage of the transcriptome (***Gingold et al., 2012***), and hence the demand for the various tRNAs, and may thus cause translational imbalances. To maintain optimal protein production, the tRNA pool is under pressure to restore the translational balance by accommodating the new translational demands. On a short timescale, the tRNA pool might respond non-genetically by changing expression profiles of the tRNAs (***Tuller et al., 2010***; ***Saikia et al., 2012***; ***Pavon-Eternod et al., 2013a***, ***2013b***). Yet, if changes in demand-to-supply persist, a genetic change in the tRNA pool might become beneficial evolutionarily. In this work, we demonstrate how anticodon mutations provide a rapid mechanism to alter the tRNA pool. We propose that during evolution, novel translational requirements can be addressed by anticodon shifting of tRNA copies more readily than by duplications and deletions of tRNA genes. The tRNA pool can evolve to meet new translation demands by adjusting the ratios of tRNA families that code for the same amino acid. Within a single mutational event, anticodon switching holds the potential to rapidly change the ratios of tRNAs within the pool, by increasing the copy number of one tRNA family at the expense of a counterpart. A similar solution could be obtained by a sequence of genomic duplications and deletions of tRNA genes. These alternatives are likely to fixate less frequently than anticodon switching, as they may carry negative effects due to duplications or deletions of adjacent unrelated genetic features. Furthermore, our systematic search for tRNA switching events throughout the tree of life revealed the prevalence of tRNA anticodon mutations in nature. This observation is consistent with the results of our lab-evolution experiment and may be the evolutionary outcome to novel translational demands in the wild.

Studies on methionine tRNAs have previously shown that the scaffold sequences determine their function as either initiator or elongator Met tRNA (***Von Pawel-Rammingen et al., 1992***; ***Aström et al., 1993***; ***Kolitz and Lorsch, 2010***). Yet, the initiator and the elongator represent extreme cases of tRNAs that are used at different stages in translation. The present chimeric tRNAs that emerged in our lab-evolution experiments successfully replaced the deleted tRNA, despite differences in 20 nucleotide positions between the 2 tRNA scaffolds. If tRNA scaffolds are interchangeable in terms of the effect of their function on the fitness, what can explain the high sequence similarity observed among tRNA gene copies of the same family in yeast? It is possible that the sequence of the tRNA scaffold is indeed important under specific conditions that were not examined in this work, or that our measurements

were not sensitive enough to detect small selective disadvantages that can act against chimeric tRNAs in nature. Under these scenarios, the high sequence similarity can be explained by purifying selection that maintains sequence identity within tRNA families. Yet, there is also a possibility that the sequence similarity is not due to purifying selection but the result of 'concerted evolution', an evolutionary process that maintains sequence identity by frequent recombination events among copies of the same gene family (**Munz et al., 1982**; **Amstutz et al., 1985**; **Teshima and Innan, 2004**). This possibility implies that the high conservation observed within tRNA gene families is not due to functionality, but is rather the result of neutral evolution. At present, it is not possible to determine which of the two possibilities explains best the observed sequence identity.

If a single point mutation in one of the $tRNA^{Arg}_{UCU}$ copies enables it to function like a tRNA from a different family, what were the evolutionary constraints that have left some families with more members while others with fewer? Is purifying selection acting to deliberately maintain low levels of certain tRNAs? Such selection would render their corresponding codons 'non-optimal'. To examine potential adaptive functions of tRNA family sizes, we tested the consequences of increasing the sizes of several tRNA families. We found that keeping low copy tRNA families is adaptive, as increasing their copy number can result in a proteotoxic stress due to problems in protein folding.

Most of the published work on the functionality of codons that correspond to rare tRNAs have so far tested how modified codon usage of specific proteins influences their proper folding (**Crombie et al., 1992**; **Komar et al., 1999**; **Cortazzo et al., 2002**; **Tsai et al., 2008**; **Zhang et al., 2009**; **Zhou et al., 2013**). In contrast, we took a different approach, in which no protein coding gene sequence is modified, but rather the tRNA supply is manipulated. Thus, the effect we generated could be exerted on all genes, and we could indeed detect it as a global proteotoxic stress in the cell. Our observations are consistent with the theory that programmed pauses in the translation process could promote proper folding during translation (**Thanaraj, 1996**; **Kramer et al., 2009**; **Cabrita et al., 2010**; **Wilke and Drummond, 2010**). The overexpression of the rare tRNA could have thus impaired with cotranslation folding. Yet, there could be additional reasons for the observed proteotoxic stress, which are not necessarily mutually exclusive. First, overexpression of $tRNA^{Arg}_{CCU}$ may result in misincorporation of arginine into non-arginine codons. Second, other aminoacyl tRNA synthetases may aminoacylate an incorrect amino acid to the highly expressed tRNA. Misloading will result in the incorporation of a different amino acid instead of arginine. A potential part of the observed proteotoxic stress due to misincorporation still remains to be studied. Yet, such an effect should be relevant not only for the overexpression of the rare $tRNA^{Arg}_{CCU}$ but also for the overexpression of the abundant $tRNA^{Arg}_{UCU}$. Our results show a sever proteotoxic stress only upon expression of the rare tRNA, thus landing more support to the intriguing hypothesis, that the proteotoxic phenotypes observed are due to converting a slow translating codon, scattered in many genes in the genome, into a fast one. This notion is consistent with and complementary to the picture that emerges from single gene-based analyses.

When facing the need to adapt, the tRNA pool (that is the supply) provides evolutionary plasticity to the translation machinery. The ability of the tRNA pool to change rapidly can be mainly attributed to its unique architecture in the form of multimember gene families. Only on a much longer evolutionary timescale, will the genome-wide codon usage of genes change so as to further fine tune the translational balance. Notably, the plasticity of the tRNA genes is constrained by the need to maintain proper protein folding (**Drummond and Wilke, 2008**). Thus, the need to accommodate changes in codon usage demands acts together with protein folding constrains to shape the tRNA pool in the living cells.

Materials and methods

Yeast strains and plasmids

The following *S. cerevisiae* strains were used in this study: $\Delta tRNA^{Arg}_{CCU}$ (based on Y5565, genetic background: *ΔtR(CCU)J::Hyg, MATα, can1Δ::MFA1pr-HIS3 mfα1Δ::MFα1pr-LEU2 lyp1Δ ura3Δ0 leu2Δ0)* (**Bloom-Ackermann et al., In press**) was used for lab-evolution experiments. $Mut\Delta tRNA^{Arg}_{CCU}$ is based on $\Delta tRNA^{Arg}_{CCU}$ and carries a mutation (T→C transition) in *tR(UCU)K* gene plus a *URA3* selection marker. BY384 (*MATa leu2Δ1 lys2Δ202 trp1Δ63 ura3-52 his3Δ200*) was used to generate a complete deletion of the $tRNA^{Ser}_{GCU}$ gene family. BY4741 (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) and BY4742 (*MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*) were used for examining the effect of increasing tRNA gene copy number.

Plasmids used in this study to express tRNA genes were pRS316 (CEN, *URA3*), pRS425 (2μ, *LEU2*), and pRS426 (2μ, *URA3*). For the rescue assays of the quadruple serine deletion, the pQF50 (2μ, *URA3*) and pQF150 (2μ, *LEU2*) were used. For the protein quality control assays the pGAL-VHL-mCherry (2μ, *LEU2)* plasmid was used (***Kaganovich et al., 2008***). For additional information on plasmids and primers see ***Supplementary file 2***.

Media

Cultures were grown at 30°C in either rich medium (1% bacto-yeast extract, 2% bacto-peptone and 2% dextrose [YPD]) or synthetic medium (0.67% yeast nitrogen base with ammonium sulfate and without amino acids and 2% dextrose, containing the appropriate supplements for plasmid selection). Protein quality control assays were performed on synthetic medium supplemented with 2% galactose as a carbon source. All chemicals used to create the media were manufactured by BD. All sugars, nucleic acids and amino acids were manufactured by Sigma-Aldrich.

Evolution experiments

Lab-evolution experiments were carried out by serial dilution. Cells were grown on 1.2 ml of YPD at 30°C until reaching stationary phase and then diluted by a factor of 1:120 into fresh media (6.9 generations per dilution). This procedure was repeated daily until population growth under the applied condition matched the wild type. In all measurements of evolved populations, we used a population sample and not selected clones.

Liquid growth measurements

The cultures were grown at the relevant condition, and optical density $(OD)_{600}$ measurements were taken during the growth at 30–45 min intervals until reaching early stationary phase. Qualitative growth comparisons were performed using 96-well plates (Thermo Scientific) in which 2 strains were divided on the plate in a checkerboard manner on the plate to cancel out positional geographical effects. For each strain, a growth curve was obtained by averaging over 48 wells.

Growth on 5-fluoro-orotic acid (5-FOA) plates

Strains were grown for 2 days in a non-selective liquid medium, which contains uracil (YPD), to allow growth of cells that lost the plasmid containing the *URA3* counterselectable marker (***Boeke et al., 1984***). Then, 100 μl were plated on a YPD plate and replicated on the following day on either YPD or standard 5-fluoro-orotic acid (5-FOA) (US Biological) plates to identify potential colonies that lost the plasmid. Following 2 days of incubation at 30°C, growth of the colonies was scored.

Measurements for saturation of the protein quality control machinery

We used a previously published method that allows examination of the protein quality control of the cell (***Kaganovich et al., 2008***). This assay provides an indication for the protein unfolding stress in cells by assessing the load on the protein quality control machinery. In this assay, the cells were introduced with a high copy number plasmid that contains the human gene VHL fused to a fluorescent tag (mCherry). VHL is a naturally unstructured protein and is dependent on two additional proteins (Elongin B and C) for proper folding in human cells. Expressing VHL in yeast cells, which lack VHL's complex partners, leads to misfolding of the translated proteins. Under normal conditions, the misfolded VHL proteins are handled by the cell's quality control machinery. When the quality control machinery is not saturated, the fluorescently tagged VHL appears in the cytosol. However, under stress, in which the quality control machinery is fully occupied, misfolded proteins in the cytosol are processed into dedicated inclusions (JUNQ and IPOD) and form punctum structures. Hence, a punctum phenotype of the VHL-mCherry construct is an indication for cells that experience proteotoxic stress and saturation of the protein quality control machinery.

Wild-type yeast cells harboring the pGAL-VHL-mCherry (CHFP) fusion plasmid (***Kaganovich et al., 2008***) and either an additional empty plasmid or the tRNA overexpression plasmid, were grown overnight on SC+2% raffinose, diluted into SC+2% galactose and grown at 30°C for 6 hr. The cells were visualized using an Olympus IX71 microscope (Olympus) controlled by Delta Vision SoftWoRx 3.5.1 software, with X60 oil lens. Images were captured by a Photometrics Coolsnap HQ camera with excitation at 555/28 nm and emission at 617/73 nm (mCherry). The images were scored using ImageJ image processing and analysis software. The percentage of cells harboring VHL-CHFP foci (Puncta) in the overexpression strains were normalized to the level in a control strain carrying an empty plasmid.

Computational identification of anticodon switching events

To characterize the extent of anticodon switching across the tree of life, we first downloaded all tRNA sequences from the Genomic tRNA Database (***Chan and Lowe, 2009***). The sequences in this database were discovered with the tRNAscan algorithm (***Lowe and Eddy, 1997***), which finds tRNA sequences by scanning genomic DNA. We removed psuedogene tRNAs, which are defined as those tRNAs with a COVE score less than 40.0 (***Lowe and Eddy, 1997***) . For each remaining tRNA sequence, we masked its anticodon triplet as 'NNN'. We next grouped all tRNA sequences by their species, and then aligned the sequences for each species using Muscle with default settings (***Edgar, 2004***). For each species, we inferred a maximum likelihood phylogeny of its tRNA sequences using RAxML with the GTRCAT model (***Stamatakis, 2006***). We calculated statistical support for tree branches using SH-like approximate likelihood ratio test (***Anisimova and Gascuel, 2006***). We next interrogated each species' tRNA phylogeny using DendroPy (***Sukumaran and Holder, 2010***). Specifically, we identified those tRNA sequences harboring an anticodon that appears in the genome more than once; for each of these tRNA sequences, we found the shortest distance to another tRNA with the same anticodon (d_{same}) and the shortest distance to another tRNA with a different anticodon (d_{diff}). We labeled tRNAs as putatively 'switched' if $d_{diff}<d_{same}$.

RT-qPCR measurements of HSP genes in strains overexpressing tRNAs

Cultures were grown in rich medium at 30°C to a cell concentration of 1×10^7 cells/ml. Then, RNA was extracted using MasterPure kit (Epicentre-illumina) (EPICENTER Biotechnologies), and used as a template for quantitative RT-PCR using light cycler 480 SYBR I master kit (Roche Applied Science) and the LightCycler 480 system (Roche Applied Science), according to the manufacturer's instructions.

Genomic copies of $tRNA^{Arg}_{UCU}$ mutated during lab-evolution experiments

In all, 4 independent lab-evolution experiments that started with $\Delta tRNA^{Arg}_{CCU}$ as the ancestral strain showed full recovery of the deletion phenotype after 200 generations. In each of the evolved populations a mutation in one of the copies of $tRNA^{Arg}_{UCU}$ was found to change the anticodon from UCU to CCU. The genomic copies of $tRNA^{Arg}_{UCU}$ that were found to carry the mutation were: *tR(UCU)K*, *tR(UCU)G1* and *tR(UCU)D* that was changed in two of the independent cultures.

The contribution of different anticodon positions to tRNA switching events

Out of 4245 anticodon switching events that we detected, the first position in the anticodon was changed in 2540 cases while the second and third were only involved in 1448 and 1330 cases, respectively.

Acknowledgements

We thank M Schuldiner, D Kaganovich, and S Leidel for kindly providing plasmids. We also thank the Pilpel lab, and especially H Gingold, for fruitful discussions. We acknowledge T Ast, Y Cohen and R Ackermann for critical reading of the manuscript. We thank the European Research Council (ERC) (YP), the Ben-May Charitable Trust (YP) and the NIH (grant 1P50GM107632) (JDB) for grant support. YP is an incumbent of the Ben May Professorial Chair.

Additional information

Funding

Funder	Grant reference number	Author
European Research Council	ERC-2007-StG 205199-ERNBPTC	Yitzhak Pilpel
Ben-May Charitable Trust		Yitzhak Pilpel

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

AHY, IF, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; ZB-A, Design, Acquisition of data, Analysis and interpretation of data, Drafting or

revising the article; VH-S, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; YC-A, Acquisition of data; QF, Acquisition of data, Drafting or revising the article, Contributed unpublished essential data or reagents; JDB, Analysis and interpretation of data, Drafting or revising the article, Contributed unpublished essential data or reagents; OD, Design, Analysis and interpretation of data, Drafting or revising the article; YP, Conception and design, Analysis and interpretation of data, Drafting or revising the article

Additional files

Supplementary files

• Supplementary file 1. Usage of arginine codons in *Saccharomyces cerevisiae*. (**A**) The AGG codon constitutes approximately 21% of the arginine codons in the yeast genome and approximately 16% of the arginine codons in the yeast transcriptome under standard lab conditions. In comparison, the AGA codon constitutes approximately 47.5% of the arginine codons in the genome and approximately 56% of the arginine codons in the transcriptome (***Gingold et al., 2012***). (**B**) Statistical significance of differences between proportions of discrepancies when masking the anticodon triplet versus control triplets. (a) Six control triplets were masked (***Figure 3—figure supplement 1***), and the analysis to identify switched tRNAs was repeated (see 'Materials and methods'). For each masked triplet, we computed the percentage of tRNA sequences that are not clustered according to their triplet content in each species. (b) The proportion of switched tRNAs was averaged across all species, and (c) the standard error of this average was computed. (d) The distribution of tRNA switching proportions from the original anticodon analysis was compared to the distribution of discrepancies from the control triplet, using a paired *t* test. The p value from this comparison is reported here. (**C**) Various yeast species tend to keep tRNAArgCCU in a single copy. All examined yeast species maintain a single copy of $tRNA^{Arg}_{CCU}$ compared to $tRNA^{Arg}_{UCU}$, which is mostly found in multiple copies. The copy number of $tRNA^{Arg}_{UCU}$ and $tRNA^{Arg}_{CCU}$ genes is shown together with the codon usage of AGG and AGA codons in the different yeast genomes.
DOI: 10.7554/eLife.01339.014

• Supplementary file 2. Plasmids and primers.
DOI: 10.7554/eLife.01339.015

Major dataset

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset ID and/or URL	Database, license, and accessibility information
Chan PP, Lowe TM	2009	Genomic tRNA database	http://gtrnadb.ucsc.edu/	We downloaded all the known tRNA sequences that are publicly available from the Genomic tRNA Database.

References

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**:927–935.

Amstutz H, Munz P, Heyer WD, Leupoid U, Kohli J. 1985. Concerted evolution of tRNA genes: intergenic conversion among three unlinked serine tRNA genes in *S. pombe*. *Cell* **40**:879–886. doi: 10.1016/0092-8674(85)90347-2.

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Systematic Biology* **55**:539–552. doi: 10.1080/10635150600755453.

Aström SU, von Pawel-Rammingen U, Byström AS. 1993. The yeast initiator tRNAMet can act as an elongator tRNA(Met) in vivo. *Journal of Molecular Biology* **233**:43–58. doi: 10.1006/jmbi.1993.1483.

Begley U, Dyavaiah M, Patil A, Rooney JP, DiRenzo D, Young CM, Conklin DS, Zitomer RS, Begley TJ. 2007. Trm9-catalyzed tRNA modifications link translation to the DNA damage response. *Molecular Cell* **28**:860–870. doi: 10.1016/j.molcel.2007.09.021.

Bermudez-Santana C, Attolini CS, Kirsten T, Engelhardt J, Prohaska SJ, Steigele S, Stadler PF. 2010. Genomic organization of eukaryotic tRNAs. *BMC Genomics* **11**:270. doi: 10.1186/1471-2164-11-270.

Bloom-Ackermann Z, Navon S, Gingold H, Towers R, Pilpel Y, Dahan O. 2014. A Comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool. *Public Library of Science Genetics* **10**:e1004084. doi: 10.1371/journal.pgen.1004084.

Boeke JD, LaCroute F, Fink GR. 1984. A positive selection for mutants lacking orotidine-5′-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance. *Molecular and General Genetics* **197**:345–346. doi: 10.1007/BF00330984.
Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**:897–907.
Byström AS, Fink GR. 1989. A functional analysis of the repeated methionine initiator tRNA genes (IMT) in yeast. *Molecular and General Genetics* **216**:276–286. doi: 10.1007/BF00334366.
Cabrita LD, Dobson CM, Christodoulou J. 2010. Protein folding on the ribosome. *Current Opinion In Structural Biology* **20**:33–45. doi: 10.1016/j.sbi.2010.01.005.
Chan PP, Lowe TM. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research* **37**:D93–D97. doi: 10.1093/nar/gkn787.
Clare JJ, Belcourt M, Farabaugh PJ. 1988. Efficient translational frameshifting occurs within a conserved sequence of the overlap between the two genes of a yeast Ty1 transposon. *Proceedings of the National Academy of Sciences of the United States of America* **85**:6816–6820. doi: 10.1073/pnas.85.18.6816.
Cochella L, Green R. 2005. An active role for tRNA in decoding beyond codon:anticodon pairing. *Science* **308**:1178–1180. doi: 10.1126/science.1111408.
Cortazzo P, Cerveñansky C, Marín M, Reiss C, Ehrlich R, Deana A. 2002. Silent mutations affect in vivo protein folding in *Escherichia coli*. *Biochemical and Biophysical Research* **293**:537–541. doi: 10.1016/S0006-291X(02)00226-7.
Crombie T, Swaffield JC, Brown AJ. 1992. Protein folding within the cell is influenced by controlled rates of polypeptide elongation. *Journal of Molecular Biology* **228**:7–12. doi: 10.1016/0022-2836(92)90486-4.
Doherty A, Mclnerney JO. 2013. Translational selection frequently Overcomes genetic drift in shaping synonymous codon usage patterns in vertebrates. *Molecular Biology and Evolution* **30**:2263–2267. doi: 10.1093/molbev/mst128.
Dong H, Nilsson L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *Journal of Molecular Biology* **260**:649–663. doi: 10.1006/jmbi.1996.0428.
Dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Research* **32**:5036–5044. doi: 10.1093/nar/gkh834.
Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**:341–352. doi: 10.1016/j.cell.2008.05.042.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**:1792–1797. doi: 10.1093/nar/gkh340.
Gingold H, Dahan O, Pilpel Y. 2012. Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome. *Nucleic Acids Research* **40**:1–11. doi: 10.1093/nar/gks772.
Goodman DB, Church GM, Kosuri S. 2013. Causes and effects of N-Terminal codon bias in bacterial genes. *Science* **342**:475–479. doi: 10.1126/science.1241934.
Higgs PG, Ran W. 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Molecular Biology and Evolution* **25**:2279–2291. doi: 10.1093/molbev/msn173.
Hudson NJ, Gu Q, Nagaraj SH, Ding YS, Dalrymple BP, Reverter A. 2011. Eukaryotic evolutionary transitions are associated with extreme codon bias in functionally-related proteins. *Public Library of Science One* **6**:e25457. doi: 10.1371/journal.pone.0025457.
Itoh T, Takemoto K, Mori H, Gojobori T. 1999. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Molecular Biology and Evolution* **16**:332–346. doi: 10.1093/oxfordjournals.molbev.a026114.
Kaganovich D, Kopito R, Frydman J. 2008. Misfolded proteins partition between two distinct quality control compartments. *Nature* **454**:1088–1095. doi: 10.1038/nature07195.
Kanaya S, Yamada Y, Kudo Y, Ikemura T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238**:143–155. doi: 10.1016/S0378-1119(99)00225-5.
Kawakami K, Pande S, Faiola B, Moore DP, Boeke JD, Farabaugh PJ, Strathern JN, Nakamura Y, Garfinkel DJ. 1993. A rare tRNA-Arg(CCU) that regulates Ty1 element ribosomal frameshifting is essential for Ty1 retrotransposition in *Saccharomyces cerevisiae*. *Genetics* **135**:309–320.
Kolitz SE, Lorsch JR. 2010. Eukaryotic initiator tRNA: finely tuned and ready for action. *FEBS Letters* **584**:396–404. doi: 10.1016/j.febslet.2009.11.047.
Komar AA, Lesnik T, Reiss C. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Letters* **462**:387–391. doi: 10.1016/S0014-5793(99)01566-5.
Konevega AL, Soboleva NG, Makhno VI, Semenkov YP, Wintermeyer W, Rodnina MV, Katunin VI. 2004. Purine bases at position 37 of tRNA stabilize codon-anticodon interaction in the ribosomal A site by stacking and Mg2+-dependent interactions. *RNA* **10**:90–101. doi: 10.1261/rna.5142404.
Kramer G, Boehringer D, Ban N, Bukau B. 2009. The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. *Nature Structural and Molecular Biology* **16**:589–597. doi: 10.1038/nsmb.1614.
Kudla G, Murray A, Tollervey D, Plotkin J. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**:255–258. doi: 10.1126/science.1170160.
Lenski RE, Rose MR, Simpson SC, Tadler SC. 1991. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *The American Naturalist* **138**:1315. doi: 10.1086/285289.
Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**:955–964.

Man O, Pilpel Y. 2007. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nature Genetics* **39**:415–421. doi: 10.1038/ng1967.
McClellan AJ, Scott MD, Frydman J. 2005. Folding and quality control of the VHL tumor suppressor proceed through distinct chaperone pathways. *Cell* **121**:739–748. doi: 10.1016/j.cell.2005.03.024.
McFarlane RJ, Whitehall SK. 2009. tRNA genes in eukaryotic genome organization and reorganization. *Cell Cycle* **8**:3102–3106. doi: 10.4161/cc.8.19.9625.
Müller EC, Wittmann-Liebold B. 1997. Phylogenetic relationship of organisms obtained by ribosomal protein comparison. *Cellular and Molecular Life Sciences* **53**:34–50. doi: 10.1007/PL00000578.
Munz P, Amstutz H, Kohli J, Leupold U. 1982. Recombination between dispersed serine tRNA genes in *Schizosaccharomyces pombe*. *Nature* **300**:225–231. doi: 10.1038/300225a0.
Olejniczak M, Dale T, Fahlman RP, Uhlenbeck OC. 2005. Idiosyncratic tuning of tRNAs to achieve uniform ribosome binding. *Nature Structural & Molecular Biology* **12**:788–793. doi: 10.1038/nsmb978.
Parmley JL, Hurst LD. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Molecular Biology and Evolution* **24**:1600–1603. doi: 10.1093/molbev/msm104.
Pavon-Eternod M, David A, Dittmar K, Berglund P, Pan T, Bennink JR, Yewdell JW. 2013a. Vaccinia and influenza A viruses select rather than adjust tRNAs to optimize translation. *Nucleic Acids Research* **41**:1914–1921. doi: 10.1093/nar/gks986.
Pavon-Eternod M, Gomes S, Rosner MR, Pan T. 2013b. Overexpression of initiator methionine tRNA leads to global reprogramming of tRNA expression and increased proliferation in human epithelial cells. *RNA* **19**:461–466. doi: 10.1261/rna.037507.112.
Pechmann S, Frydman J. 2012. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature Structural & Molecular Biology* **20**:1–8. doi: 10.1038/nsmb.2466.
Percudani R, Pavesi A, Ottonello S. 1997. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *Journal of Molecular Biology* **268**:322–330. doi: 10.1006/jmbi.1997.0942.
Rao Y, Wu G, Wang Z, Chai X, Nie Q, Zhang X. 2011. Mutation bias is the driving force of codon usage in the Gallus gallus genome. *DNA Research* **18**:499–512. doi: 10.1093/dnares/dsr035.
Rawlings TA, Collins TM, Bieler R. 2003. Changing identities: tRNA duplication and remolding within animal mitochondrial genomes. *Proceedings of the National Academy of Sciences of the United States of America* **100**:15700–15705. doi: 10.1073/pnas.2535036100.
Rogers HH, Bergman CM, Griffiths-Jones S. 2010. The evolution of tRNA genes in *Drosophila*. *Genome Biology and Evolution* **2**:467–477. doi: 10.1093/gbe/evq034.
Saikia M, Krokowski D, Guan BJ, Ivanov P, Parisien M, Hu GF, Anderson P, Pan T, Hatzoglou M. 2012. Genome-wide identification and quantitative analysis of cleaved tRNA fragments induced by cellular stress. *Journal of Biological Chemistry* **287**:42708–42725. doi: 10.1074/jbc.M112.371799.
Saks ME, Conery JS. 2007. Anticodon-dependent conservation of bacterial tRNA gene sequences. *RNA* **13**:651–660. doi: 10.1261/rna.345907.
Saks ME, Sampson JR, Abelson J. 1998. Evolution of a transfer RNA gene through a point mutation in the anticodon. *Science* **279**:1665–1670. doi: 10.1126/science.279.5357.1665.
Schmeing TM, Voorhees RM, Kelley AC, Ramakrishnan V. 2011. How mutations in tRNA distant from the anticodon affect the fidelity of decoding. *Nature Structural & Molecular Biology* **18**:432–436. doi: 10.1038/nsmb.2003.
Schultz DW, Yarus M. 1994. tRNA structure and ribosomal function. I. tRNA nucleotide 27-43 mutations enhance first position wobble. *Journal of Molecular Biology* **235**:1381–1394. doi: 10.1006/jmbi.1994.1095.
Sharp PM, Li WH. 1986a. An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution* **24**:28–38. doi: 10.1007/BF02099948.
Sharp PM, Li WH. 1986b. Codon usage in regulatory genes in Escherichia coli does not reflect selection for "rare" codons. *Nucleic Acids Research* **14**:7737–7749. doi: 10.1093/nar/14.19.7737.
Sharp PM, Li WH. 1987. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* **15**:1281–1295. doi: 10.1093/nar/15.3.1281.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**:2688–2690. doi: 10.1093/bioinformatics/btl446.
Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Molecular Biology and Evolution* **24**:374–381. doi: 10.1093/molbev/msl166.
Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**:1569–1571. doi: 10.1093/bioinformatics/btq228.
Teshima KM, Innan H. 2004. The effect of gene conversion on the divergence between duplicated genes. *Genetics* **166**:1553–1560. doi: 10.1534/genetics.166.3.1553.
Thanaraj TA, Argos P. 1996. Ribosome-mediated translational pause and protein domain organization. *Protein Science* **5**:1594–1612. doi: 10.1002/pro.5560050814.
Tsai C-J, Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM, Nussinov R. 2008. Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *Journal of Molecular Biology* **383**:281–291. doi: 10.1016/j.jmb.2008.08.012.
Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**:344–354. doi: 10.1016/j.cell.2010.03.031.
Urrutia AO, Hurst LD. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* **159**:1191–1199.

Varenne S, Buc J, Lloubes R, Lazdunski C. 1984. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *Journal of Molecular Biology* **180**:549–576. doi: 10.1016/0022-2836(84)90027-5.

Von Pawel-Rammingen U, Aström S, Byström AS. 1992. Mutational analysis of conserved positions potentially important for initiator tRNA function in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* **12**:1432–1442.

Widmann J, Harris J, Lozupone C. 2010. Stable tRNA-based phylogenies using only 76 nucleotides. *RNA* 1469–1477. doi: 10.1261/rna.726010.

Wilke CO, Drummond DA. 2010. Signatures of protein biophysics in coding sequence evolution. *Current Opinion In Structural Biology* **20**:385–389. doi: 10.1016/j.sbi.2010.03.004.

Withers M, Wernisch L, dos Reis M. 2006. Archaeology and evolution of transfer RNA genes in the *Escherichia coli* genome. *RNA* **12**:933–942. doi: 10.1261/rna.2272306.

Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV. 2001. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Research* **11**:356–372. doi: 10.1101/gr.161901.

Zhang G, Hubalewska M, Ignatova Z. 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature Structural & Molecular Biology* **16**:274–280. doi: 10.1038/nsmb.1554.

Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Molecular Biology and Evolution* **26**:1571–1580. doi: 10.1093/molbev/msp070.

Zhou M, Guo J, Cha J, Chae M, Chen S, Barral JM, Sachs MS, Liu Y. 2013. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* **495**:111–115. doi: 10.1038/nature11833.

Codon Usage of Highly Expressed Genes Affects Proteome-Wide Translation Efficiency

Idan Frumkin[1], Marc J. Lajoie[2], Christopher J. Gregg[2], Gil Hornung[3], George M. Church[2] & Yitzhak Pilpel[1,#]

1- Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel.
2- Department of Genetics, Harvard Medical School, Boston, MA 02115.
3- The Nancy & Stephen Grand Israel National Center for Personalized Medicine, Rehovot, Israel.
- correspondence: pilpel@weizmann.ac.il

Major Category: Biological Sciences
Minor Category: Systems Biology
Keywords: evolution of translation machinery; tRNA; evolution of codon usage; translation efficiency; codon-to-tRNA balance; genome engineering

Abstract

Although the genetic code is redundant, synonymous codons for the same amino acid are not used with equal frequencies in genomes, a phenomenon termed codon usage bias. Previous studies have demonstrated that synonymous changes in a coding sequence can exert significant *cis* effects on the gene's expression level. Yet, whether the codon composition of a gene can also affect the translation efficiency of other genes has not been thoroughly explored. To study how codon usage bias influences the cellular economy of translation, we massively converted abundant codons to their rare synonymous counterpart in several highly expressed genes in *Escherichia coli*. This perturbation reduces both cellular fitness and the translation efficiency of genes that have high initiation rates and are naturally enriched with the manipulated codon – in agreement with theory perditions. Interestingly, we could alleviate the observed phenotypes by increasing the supply of the tRNA for the highly-demanded codon, thus demonstrating that the codon usage of highly expressed genes was selected in evolution to maintain the efficiency of global protein translation.

Significance Statement:

Highly expressed genes are encoded by codons that correspond to abundant tRNAs, a phenomenon thought to ensure high expression levels. An alternative interpretation is that highly expressed genes are codon-biased so-as to support efficient translation of the rest of the proteome. Until recently, it was impossible to examine these alternatives, since statistical analyses providing correlations but not causal mechanistic explanations. Massive genome engineering now allows recoding genes and examining effects on cellular physiology and protein translation. We engineered the *E. coli* genome by changing codon bias of highly expressed genes. The perturbation affected translation of other genes, depending on their codon demand – suggesting that codon bias of highly expressed genes ensures translation integrity of the rest of the proteome.

Introduction

Since there are 61 sense codons but only 20 amino acids, most amino acids are encoded by more than a single codon. Yet, synonymous codons for the same amino acid are not utilized to the same extent across different genes or genomes. This phenomenon, termed codon usage bias, has been the subject of intense research and was shown to affect gene expression and cellular function through varied processes in bacteria, yeast and mammals(1–4).

Though differential codon usage can result from neutral processes of mutational biases and drift(5–7), certain codon choices could be specifically favored as they increase efficiency(8–12) or accuracy(13–17) of protein synthesis. These forces would typically lead to codon biases in a gene because they locally exert their effect on the gene on which the codons reside. Indeed, there is a positive correlation between a gene's expression level and the degree of its codon bias(1). Various systems have demonstrated how altering the codon usage synonymously can alter expression levels of the manipulated genes(18–21), an effect that could reach more than 1,000-fold(22).

In addition to such *cis* effects, it is possible that codon usage also acts in *trans*, namely, that codon choice of some genes would affect translation of others due to a "shared economy" of the entire translation apparatus(23–25). Previous theoretical works have suggested that increase in elongation rate may reduce the number of ribosomes on mRNAs and therefore may indirectly increase the rate of initiation of other transcripts due to an increase of the pool of free ribosomes(6, 26). In addition, a recent computational study in yeast has also examined the indirect effects of synonymous codon changes on the translation of the entire transcriptome(27). Yet, experimental evidences of such changes are absent. Here we ask how manipulating the frequency of a single codon on a small subset of genes influences the synthesis of other proteins.

To tackle this question, we replaced common codons with a synonymous, rare counterpart in several highly expressed genes. We then asked how this massive change in the codon representation in the transcriptome would affect the manipulated genes, other

genes, and the physiology and well-being of the cell (Figure 1). Interestingly, our genetic manipulation has not consistently affected the translation efficiency of the mutated genes, yet it did show a profound proteome-wide effect on the translation process. Importantly, translation efficiency of genes changed in a way that was dependent on the extent to which they contain the affected codons. These observations demonstrate that *trans* effects of codon usage could have strong implications in the cell. We could alleviate these physiological and molecular defects by increasing tRNA supply for the manipulated codon in a manner that restored codon-to-tRNA balance. Our work demonstrates that codon choice does not only tune the expression level of individual genes, but also maintains the efficiency of global protein translation in the cell.

Results

Codon usage manipulation leads to proteome-wide changes in translation efficiencies in a codon-dependent manner

We ask how the codon usage of a small subset of genes affects the translation of other genes. To this end, we manipulated the frequency of the arginine codon CGG since it is the only codon in *E. coli* that is translated by a single-copy tRNA gene, and whose tRNA does not translate other codons (see Box 1 for codon-anticodon interactions for CGN codons in *E. coli*)(28). Using genome editing, we were able to introduce 60 synonymous mutations into a single genome of an *E. coli* strain that converted CGU and CGC ("origin codons") to CGG ("destination codon"). To maximize the effects of our manipulations on the proteome and on the cell, we re-coded genes that show high mRNA levels and are highly occupied by ribosomes. Notably, we avoided any re-coding of essential, ribosomal or global regulatory genes, as manipulating these genes might influence the cell directly, hence masking potential effects due to changes in codon usage. We introduced synonymous mutations in the eight genes with the highest ribosome-profiling occupancy score that are not essential and that do not relate directly to the above functions (Table 1). Following our manipulation, the translation demand for the ACG anticodons is reduced by ~5%, the demand for the CCG anticodon is elevated by ~3.5-fold and our re-coded genes constitute ~70% of the new total demand for this codon in the cell (Table 1). See Materials and Methods for a full description of the re-coded process.

We then asked how does our manipulation on the CGG representation in the transcriptome influence translation efficiency in the cell. To this end, we analyzed the transcriptome (by RNA-sequencing) and proteome (by mass-spectrometry) of the original wild-type and the re-coded strains (each strain was analyzed with three independent repetitions for both the transcriptome and proteome, see Materials and Methods). Then, we calculated the translation efficiency of each gene by normalizing the protein level to its corresponding mRNA level based on the three independent repetitions.

Notably, only one of the eight recoded genes showed reduced translation efficiency (Figure 2A), suggesting that the effects of our codon-usage manipulation on the genes that harbor the manipulation are weak. A possible reason for this weak effect is that in the current experiment only a single codon type has been manipulated in each re-coded gene, in contrast to prior studies in which entire ORFs have been manipulated(18, 21). It is also possible that our manipulations did affect translation efficiency in *cis*, though some compensatory effect, e.g. acting on the initiation level, may have acted to counter-act the reduction in elongation. Ultimately, this observation reassures that our codon manipulations successfully increased translation demand for the CGG codon and provides a unique opportunity to elucidate any *trans* effects of codon usage in highly expressed genes.

We postulated that the increased usage of CGG at the expense of the CGU and CGC codons might reduce the translation efficiency of other genes in the genome, which were not mutated, in particular genes that naturally have high usage of CGG. Indeed, we observed 455 genes with increased and 566 genes with decreased translation efficiency at a fold change of above or below 1.5 in the re-coded strain compared to the wild-type (Figure 2A). Strikingly, genes with high occurrences of the CGG codon (>5 occurrences) that were not engineered by us demonstrated lower translation efficiencies in the re-coded strain compared to the WT strain, compared to genes that do not use this codon (Figure 2A inset). This observation suggests that our CGG codon manipulation affected in *trans* the translation of other, non-recoded genes in the re-coded strain. In support of this result, the hundreds of genes that showed reduced translation efficiency demonstrated higher occurrences of the CGG codon compared to the genes with increased translation efficiency (Figure 2B). On the other hand, we observed that genes with increased translation efficiency were enriched with the CGU, CGC, and CGA codons (Figure 2C). We thus conclude that the increased demand on the CGG codon due to our recoding reduced the translation efficiency of genes that were enriched with this codon, while the relief of demand from the CGU, CGC, and CGA codons increased the translation efficiency of genes that utilize these codons. While most studies measure the resulted change in expression

level of a gene whose different codons were synonymously manipulated(18, 21), our results demonstrate for the first time how a frequency manipulation of a codon can affect global translation patterns by changing the translation efficiency of other genes according to their codon usage.

Theory predicts that changes of elongation rate should have the largest expression effects on genes with high rates of translation initiation because these genes are more likely to suffer from traffic jams and ribosomal collisions(10, 27). Thus, we hypothesized that genes with reduced translation efficiency in the re-coded strain should have higher translation initiation rates compared to genes whose translation efficiency did not decrease. Indeed, reduce translation efficiency genes demonstrate higher initiation rates as calculated with the Ribosome Binding Site Calculator(29) compared to un-effected or increased translation efficiency genes (Figure 2G). The observations that genes with reduced translation efficiency are more enriched with the CGG codon, on one hand, and have higher initiation rates on the other, strengthens our conclusion that the re-coded strain suffers from ribosomal elongation changes compared to WT cells. In line with theoretical predictions(10, 27), increasing the dwell time of ribosome during elongation reduces translation efficiency provided that initiation rate is sufficiently high.

Proteome-wide changes in translation efficiencies are alleviated by increased tRNA supply

To confirm our hypothesis that the changes in translation efficiencies resulted from the increased cellular demand for $tRNA^{CCG}$, the tRNA which translates CGG, we decided to elevate the availability of this tRNA and examine the effect on the translation phenotype. We, and others, have recently shown that a mechanism to increase tRNA availability is a mutation in the anticodon that changes the codon specificity of the tRNA(30, 31). We have shown that such anticodon switching mutations can maintain the functionality of tRNA genes, and are utilized by many species as an adaptive mechanism of the cellular tRNA pool.

Thus, we mutated the anticodon of one of the four copies of $tRNA^{ACG}$ gene from ACG to CCG on the background of the re-coded strain (Box 1). We then analyzed the transcriptome and proteome of this anticodon-switched strain (based on three independent repetitions) and compared it to both the re-coded and wild-type strains. Strikingly, although the genome of the anticodon-switched strain is more similar to the re-coded strain, its global translation efficiency pattern clustered together with the wild-type strain and away from the re-coded strain (Figure 2H). This observation suggests that manipulating the tRNA pool of the re-coded strain restored translation efficiency of genes back to their normal states.

Indeed, only 124 and 408 genes with increased or decreased translation efficiency were respectively identified between the wild-type and the anticodon-switched strains (Figure 2D), further demonstrating that the translation efficiency defect in the re-coded strain was alleviated upon anticodon switching. Strikingly, while CGG-enriched genes particularly tended to have reduced translation efficiencies in the re-coded strain, they demonstrated similar efficiencies to the wild-type in the anticodon-switched strain, and the difference in translation efficiency ratios between these genes and CGG-depleted genes was not observed (Figure 2D inset). Consistently, the genes with increased or decreased translation efficiency between the wild-type and anticodon-switched strain demonstrated the same distribution of codon occurrences for CGG or CGU+CGC+CGA (Figures 2E+F). These observations suggest that the additional supply of $tRNA^{CCG}$, at the expense of $tRNA^{ACG}$ in the anticodon-switched strain, resulted in a more efficient translation of CGG-enriched genes.

We next wanted to examine whether particular codons, especially those involved in the re-coding process (CGN codons), were enriched or depleted from the proteome of the re-coded strain. We defined a “proteomic codon usage” as the multiplication of codon occurrences in each gene and the measured expression level of its protein product. We then calculated this index for each codon and calculated its re-coded/WT ratio (Figure 3A). Remarkably, the CGG codon has the lowest re-coded/WT ratio from all 61 codons,

further showing how the introduction of this codon on highly expressed genes resulted in a global proteomic effect. Notably, the two origin codons, namely CGU and CGC, behaved similarly to all other sense codons in this measurement. When the same comparison was performed between the re-coded and the anticodon-switched strains, the observed ratio for CGG was significantly increased, while the ratio for the CGU and CGC codons was reduced (Figure 3B). These observations are consistent with the additional supply of $tRNA^{CCG}$ at the expense of $tRNA^{ACG}$ in the anticodon-switched strain. Interestingly, the codon CGA, which is translated by $tRNA^{ACG}$, demonstrated a similar behavior to CGG and not CGU or CGC. This trend is probably the result of the fact the CGA is the rarest CGN codon and usually co-occurs with CGG on the same genes.

The re-coded strain suffers from reduced ability to translate transcripts with the CGG codon

To directly demonstrate that the codon manipulation in the re-coded strain indeed hampered the translation of other genes in a codon-dependent manner, we sought a reporter that would read-out the effects of re-coding on translation of the CGG codon. To this end, we used two previously published versions of a YFP reporter with six occurrences of arginine(32), each version with an alternative codon choice - either CGU or CGG (Figure 4A). Since both YFP variants are not under any specific regulation by the cell and were shown to have similar mRNA levels(32), they can serve as a direct proxy for protein synthesis in each strain.

Low-copy plasmids carrying either YFP-CGU or YFP-CGG were transformed to the wild-type and re-coded strains. Then, YFP production was measured (see Materials and Methods) and a YFP-CGG/YFP-CGU ratio was calculated for each strain (Figure 4A). The wild-type strain demonstrated max YFP production values of 1224 (AU) and 1405 (AU) for YFP-CGU and YFP-CGG, respectively, leading to a YFP-CGG/YFP-CGU ratio of 1.15 (Figure 4B), in agreement with a previous measurement of codon translation speeds in *E. coli*(33).

In comparison, the re-coded strain showed max YFP production values that were consistent with the phenotypes we observed for natural genes, namely, an increased value of 1316 (AU) for YFP-CGU and a reduced YFP-CGG value of 1328 (AU). Thus, the YFP-CGG/YFP-CGU ratio was significantly reduced in the re-coded strain, and was measured to be 0.99 (Figure 4B). This result further supports the view that the increased translational demand for CGG in the re-coded strain hampers the production of proteins that utilize the CGG codon.

Since the anticodon switching mutation alleviated the translation difficulty of the re-coded strain, we next asked whether it would also restore translation efficiency of the YFP-CGG reporter. Hence, we generated three more anticodon-switched strains, each with a different $tRNA^{ACG}$ copy that we mutated, and measured their YFP-CGG/YFP-CGU expression ratio. Indeed, all four strains with anticodon-switching mutation showed YFP-CGG/YFP-CGU ratios closer to the wild-type strain and above the ratio of the re-coded strain (Figure 4B). These observations further support our conclusion that the re-coded strain suffers from low availability of $tRNA^{CCG}$ due to the codon manipulation of CGG, which hampers protein production in a codon-specific manner. This perturbation could be alleviated upon increased tRNA supply in the cell.

Increased codon usage of a rare codon reduces cellular fitness due to excessive use of tRNA molecules

The physiological effects between the wild-type and re-coded strains encouraged us to ask whether these global translation efficiency changes disturb cellular growth and reduce fitness. We thus tested whether introducing the rare codon CGG on highly expressed genes is deleterious to the cell. We compared the growth of the wild-type and re-coded strains (see Materials and Methods) and observed that the re-coded strain suffers from a growth defect (Figure 5A). We used a recent logistic growth model(34) that calculates relative fitness from growth curves and observed that the relative fitness of the re-coded strain is 0.87 compared to the wild-type strain.

We next hypothesized that the growth reduction of the re-coded strain is the result of a lack in sufficient tRNA supply that leads to changes in translation efficiency of many genes. However, cellular fitness could also be affected by the off-target mutations that the re-coded strain accumulated following our genome engineering efforts. To test our hypothesis, we compared the growth of all four anticodon-switched strains, in which $tRNA^{CCG}$ levels are increased, and observed that they all demonstrated increased relative fitness in comparison to the re-coded strain (Figure 5A). Importantly, when the same anticodon mutation was inserted on the background of the wild-type strain, a reduction in relative fitness was observed (Figure 5B). These results suggest that introducing a rare codon on highly expressed genes reduces cellular fitness not because of its effects on the manipulated genes themselves, but as it hampers translation of other genes due to an excessive use of tRNA molecules and result in global physiological perturbations.

Changes in codon usage lead to mis-translation at modified positions

We did not observe changes in translation efficiency for the recoded genes (Figure 2A), suggesting that the manipulation of CGG demand leads to stronger *trans* effects on global cellular patterns of translation efficiency. However, we hypothesized that our re-coding might have other *cis* effects in the form of mis-translation. To test this idea, we used our new developed methodology that uses mass-spectrometry proteomics data to identify peptides that harbor mis-translation events that results in the replacement of the correct amino-acid with a different one(35). Strikingly, we identified such events for two of the recoded genes, ompC and ompA, exactly at the positions in which CGU or CGC were respectively mutated into CGG (Figure 6). The mis-translation event for ompC was found in the re-coded strain, and changed the coded arginine with glutamine (which has a near-cognate codon, CAG, to CGG) at position 238 of the protein. The mis-translation event for ompA was found in the ACS strain, and it changed the coded arginine with lysine at position 329 of the protein – suggesting that the additional tRNA(ACG) supply in this strain did not fully alleviated the mis-translation phenotype, similarly to other phenotypes we observed in this study.

Discussion

Often in biology, a correlation between two factors could be explained either by a physiological causal link or by an evolutionary one(36). This is particularly relevant to the correlation that is broadly observed between codon usage and expression level(1, 2, 37–39). On the one hand, optimal codon usage could lead to higher translation speeds(1, 28, 40), suggesting that some proteins enjoy higher expression levels because of their codon usage. On the other hand, highly expressed genes could be strongly evolutionarily selected for codon optimization compared to lowly expressed genes because the fitness cost of not optimizing them is greater, and hence they force the genome to optimize their codon usage(13, 16, 18). Furthermore, specific codons may be selected for or against for reasons other than their effect on translation itself, for example to maintain mRNA structure(21), splicing signals(4), degradation rate of the transcript(41) or to minimize the cost of gene expression(42).

An additional reason for the strong codon bias of highly expressed genes could be their massive representation in the transcriptome and overall impact on the translation machinery. Thus, a non-optimal codon presents on a gene with a high mRNA level could disturb the translation of other genes that utilize this codon(27).

Here we study this possibility and our results provide the first experimental evidence that introducing a rare codon into highly expressed genes indeed hampers protein production of other genes, especially those that are encoded with that codon. In our experimental system, we introduced 60 new occurrences of a rare arginine codon, CGG, on highly expressed genes and this manipulation led to a reduction in translation efficiency of CGG-containing genes. Importantly, we confirmed the protein synthesis difficulties of such genes by using two versions of a reporter gene that either use the CGG codon or avoid it. Thus, our results demonstrate that translation of a certain gene is not only influenced by its own regulation in the form of codon usage or mRNA level – but also by the translation efficiency of all other genes in the cellular genetic network.

One limitation of the genome engineering approach we took here is the accumulation of off-target mutations, in addition to the planned mutations (see also Materials and Methods). Yet, we argue that our observed phenotypes are mostly due to the on-target mutations for the following reasons. First, the off-target mutations in the re-coded strain did not manipulate CGG specifically (in direct contrast to the on-target mutations), and they are of diverse nature: half are synonymous mutations, 7 are inter-genic, 20% occurred inside un-validated or un-characterized proteins and none occurred in genes that are part of the translation machinery. It is extremely unlikely that the off-target mutations, which are diverse and do not show any pattern, would lead to the CGG-specific phenotype we observed. Second, out of the 1021 genes with increased or decreased translation efficiency only 8 had off-target mutations in them. Importantly, these genes were either enriched or depleted from the CGG codon in agreement with the overall increased translational demand in the re-coded strain. This phenotype, together with the observation for the YFP production as discussed above, are extremely unlikely to occur as a result of the off-targets – especially given the fact that we planned the on-target changes to directly manipulate CGG. Third, and most importantly, we could cure most of the phenotypic and molecular defects of the re-coded strain by tRNA manipulation in the form of anticodon switching. This is a very clear indication that the re-coded strain mainly suffers from a direct effect of recoding, and less so due to off-target mutations. In particular, we observed the following phenotypes for the anticodon switched strain that support the direct effects of the on-target mutations: a) The anticodon switching strain has an additional copy of tRNA(CCG) and therefore the translation efficiency of the perfectly-matching codon CGG-containing genes is increased. b) The anticodon switching stain has one less copy of tRNA(ACG) and therefore the translation efficiency of CGU/CGC-containing genes is decreased. c) The translation efficiency pattern of the anticodon switched strain clusters with that of the WT, although it is genetically closer to the re-coded strain. This result shows that the re-coded strain indeed suffers from imbalance between CGG demand and tRNA(CCG) supply that anticodon switching fixes. If off-targets were dominant in determining the phenotypes we

observe, we would have obtained exactly the opposite result – that the anti-codon switched recoded strain would have clustered with the recoded strain since they share the off-target mutations.

Our observations are also relevant to the context of heterologous gene expression. The codon usage of a gene is most relevant to its successful expression in a foreign system(1, 22). Yet, the effects of artificially expressing a gene that is not native to the genome, usually to high levels, on cell physiology has not been explored thoroughly. Our results allow to measure and appreciate the proteome-wide changes under such conditions. Although in our systems the re-coded genes are natural to the genome, it is likely that they apply to heterologous proteins, which are thus predicted here to affect the translation apparatus in a similar manner as in our case due to changes in codon demands.

Finally, this work raises the question of whether changes in global translation efficiencies could pose a challenge to the translation machinery both physiologically and evolutionarily. Previous works have demonstrated how codon-to-tRNA balance reacts to changes in the environment(32, 43, 44), to the formation of cancerous tumor(45), or to an evolutionary challenge(30, 46). In agreement with these works, we observed that the re-coded strain suffers from a growth defect, providing a need for selection to optimize the translation economy in the cell. Interestingly, we could alleviate these translation and growth phenotypes by providing more tRNA supply that could meet the new CGG demand. Thus, our work demonstrates that codons and tRNA genes may co-evolve not only to tune the expression level of individual (highly expressed) genes, but also to maintain the efficiency of global protein translation in the cell.

Materials and Methods

Genome Engineering

To introduce synonymous mutations that replace origin codons (CGU & CGC) to the destination codon (CGG), we used Co-Selection Multiplex Automated Genome Engineering (CoS-MAGE) as previously described(47–49). The background *E. coli* strain was EcM2.1, an especially designed strain for high MAGE efficiency(50). Each day one CoS-MAGE cycle was perform with ten 90mer oligos and selecting either for or against the tolC marker. Briefly, cells were grown overnight at 34°C. Then, 30 µl of the saturated culture were transferred into fresh 3ml of LBL medium until reaching OD=0.4 and then moved to a shaking water bath (350 RPM) at 42°C for 15 min after which it was moved immediately to ice. Next, 1ml was transferred to an Eppendorf tube and cells were washed twice with sterile water at centrifuge speed of 13,000g for 30 seconds. Next, the bacterial pellet was dissolved in 50µl of ddW containing 4µM of ten SS-DNA oligos + a tolC dedicated oligo and transferred into a cuvette. Electroporation was performed in 1.78kV, 200ohms, 25µF. After electroporation, the bacteria were transferred into 1ml of fresh LBL for recovery and then moved to selection medium. Selection for tolC was performed on liquid LBL + 0.005%SDS and selection against tolC was with LBLCCoV plates that contain 50µg/ml Carbenicillin + 64µg/ml Vancomycin + purified Colicin E1(51). Every four CoS-MAGE cycles, random colonies were screened for on-target mutations via multiplex allele specific colony PCR (mascPCR), and colonies with highest number of mutations were sequenced for further verification. Then, the best colony was picked for successive engineering via additional CoS-MAGE cycles.

To facilitate re-coding efforts, we split the eight targeted genes into two groups according to their genomic loci (Supplementary Figure 1). Strain A was re-coded for genes: ahpC, cspE, pal, ompX, ompF & ompA. Strain B was re-coded for genes: atpE & ompC. After engineering was completed for both strains, we merged their genome by following the conjugative assembly genome engineering (CAGE) protocol(52, 53). Strain A was the donor, and thus was transformed with the pRK29 plasmid while strain B was the recipient. Selection for final strain was done on LBL plates with 0.005% SDS + 100µg/ml

Spectinomycin + 5μg/ml Gentamycin. To maintain a similar genetic background as possible between the re-coded and the wild-type strains, we also transformed the resistance markers for SDS, Spectinomycin and Gentamycin to the same loci as in the re-coded strain.

We confirmed the successful introduction of all 60 planned genomic changes by whole-genome sequencing. We also revealed additional 58 off-target mutations as typically happens with this genome-editing technology(54). Any off-targeted gene, which was mutated unintentionally, was therefore excluded from all our down-stream analyses. Importantly, the off-target mutations in the re-coded strain did not manipulate CGG specifically (in direct contrast to the on-target mutations), and they are of diverse nature: half are synonymous mutations, 7 are inter-genic and did not affect any gene, 20% occurred inside un-validated or un-characterized proteins and none occurred in genes that are part of the translation machinery. Thus, it is unlikely that the off-target mutations, which are diverse and do not show any pattern, would lead to the CGG-specific phenotype we observed. See full discussion on off-target mutations in Discussion section.

See list of off-target mutations in Supplementary File 1 and a list of strains, CoS-MAGE oligos, mascPCR primers and tolC information in Supplementary File 2.

Liquid growth measurements

Cultures were grown for 48hours in LB medium at 30^{o}C, back diluted in a 1:100 ratio and dispensed on 96-well plates in a checkerboard manner. Wells were measured for optical density at OD_{600} and measurements were taken during growth at 30min intervals until reaching stationary phase. For each strain, a growth curve was obtained by averaging all wells. Then, we converted these curves to relative fitness using the Curveball approach(34).

YFP production measurements

Each strain was transformed with the plasmid pZS*11-YFP-Kan harboring a Kan resistance cassette and a YFP gene with six occurrences of either CGG or CGU. Growth was measured as described in Liquid growth measurements section only YFP measurements (excitation=500±25nm, emission=540±25nm) were taken in addition to OD_{600}. YFP production rate was measured as previously described(55) by subtracting the YFP value at time t by the YFP value in time t-1 and dividing the result by OD_{600} value at time t. Maximal production rate was defined as the highest value on this curve. We follow the YFP production along the entire growth curve (from lag to saturation) as it includes all the different physiological states the cells experience under these growth conditions. Then, a YFP-CGG/YFP-CGU ratio was calculated for each strain.

Harvesting cells for transcriptome and proteome analyses

To compare the transcriptome of the wild-type, re-coded and anticodon-switched strains, we grew each strain with three independent repetitions in LB at 30^{o}C over-night. Then, for each repetition 400µl of culture was diluted in 50ml of LB and grew until cells reached OD_{600} of 0.4. Cells were flash frozen in liquid nitrogen and pellets were used for either RNA-sequencing or mass-spectrometry.

Transcriptome analysis

RNA-sequencing was performed as described in Dar et al(56). RNA was extracted with standard protocol. Then, samples were treated with DNase using TURBO DNA-free™ Kit by Ambion and rRNA was depleted by using epicenter's Ribo-Zero rRNA Removal Kit. Next, Strand-specific RNA-seq was performed with the NEBNext Ultra Directional RNA Library Prep Kit. Libraries were sequenced by using the Illumina Nextseq with a read length of 50 nucleotides.

Proteome analysis

Sample preparation for mass-spectrometry

Cell pellets were subjected to in-solution tryptic digestion using a modified Filter Aided Sample Preparation protocol (FASP). All chemicals are from Sigma Aldrich, unless stated otherwise. Sodium dodecyl sulfate buffer (SDT) included: 4%(w/v) SDS, 100mM Tris/HCl pH 7.6, 0.1M DTT. Urea buffer (UB): 8M urea (Sigma, U5128) in 0.1M Tris/HCl pH 8.0 and UC buffer: 2M Urea, pH 7.6-8.0 (dilute UB X 4 with 0.1M Tris-HCl pH 7.6). Cells were dissolved in 100µl SDT buffer and lysed for 3min at 95°C. Then, centrifuged at 16,000RCF for 10min. 30µl were mixed with 200µl UB and loaded onto 30kDa molecular weight cutoff filters and spun down. 200µl of UA were added to the filter unit and centrifuge at 14,000g for 40min. Trypsin was then added and samples incubated at 37°C overnight. Digested proteins were then spun down, acidified with trifluoroacetic acid and stored in -80°C until analysis.

Liquid chromatography

ULC/MS grade solvents were used for all chromatographic steps. Each sample was fractionated using high pH reversed phase followed by low pH reversed phase separation. 200µg digested protein was loaded using high Performance Liquid Chromatography (Agilent 1260 uHPLC). Mobile phase was: A) 20mM ammonium formate pH 10.0, B) acetonitrile. Peptides were separated on an XBridge C18 column (3x100mm, Waters) using the following gradient: 3% B for 2 minutes, linear gradient to 40% B in 50min, 5 min to 95% B, maintained at 95% B for 5 min and then back to initial conditions. Peptides were fractionated into 15 fractions. The fractions were then pooled: 1 with 8, 2 with 9, 3 with 10, 4 with 11, 5 with 12, 6 with 13 and 7 with 14-15. Each fraction was dried in a speedvac, then reconstituted in 25µl in 97:3 acetonitrile:water+0.1% formic acid. Each pooled fraction was then loaded using split-less nano-Ultra Performance Liquid Chromatography (10 kpsi nanoAcquity; Waters, Milford, MA, USA). The mobile phase was: A) H2O + 0.1% formic acid and B) acetonitrile + 0.1% formic acid. Desalting of the

samples was performed online using a reversed-phase C18 trapping column (180µm internal diameter, 20mm length, 5µm particle size; Waters). The peptides were then separated using a T3 HSS nano-column (75µm internal diameter, 250mm length, 1.8µm particle size; Waters) at 0.35µl/min. Peptides were eluted from the column into the mass spectrometer using the following gradient: 4% to 35%B in 150 min, 35% to 90%B in 5min, maintained at 95% for 5min and then back to initial conditions.

Mass Spectrometry

The nanoUPLC was coupled online through a nanoESI emitter (10 µm tip; New Objective; Woburn, MA, USA) to a quadrupole orbitrap mass spectrometer (Q Exactive Plus, Thermo Scientific) using a FlexIon nanospray apparatus (Proxeon).

Data was acquired in DDA mode, using a Top20 method. MS1 resolution was set to 60,000 (at 400m/z) and maximum injection time was set to 20msec. MS2 resolution was set to 17,500 and maximum injection time of 60msec.

Data processing and analysis

Raw data was imported into the Expressionist software (Genedata) and processed as previously described(57). The software was used for retention time alignment and peak detection of precursor peptides. A master peak list was generated from all MS/MS events and sent for database searching using Mascot v2.5 (Matrix Sciences). Data was searched against *Escherichia coli* K12 protein database (http://www.uniprot.org/) appended with 125 common laboratory contaminant proteins. Fixed modification was set to carbamidomethylation of cysteines and variable modification was set to oxidation of methionines. Search results were then filtered using the PeptideProphet algorithm(58) to achieve maximum false discovery rate of 1% at the protein level. Peptide identifications were imported back to Expressions to annotate identified peaks. Quantification of proteins from the peptide data was performed using an in-house script(57).

Data was normalized base on the total ion current. Protein abundance was obtained by summing the three most intense, unique peptides per protein. Principal Component Analysis was used to assess global integrity of the data and search for outlier samples.

Acknowledgments

We thank Arvind R Subramaniam for suppling plasmids and helpful discussions and Nir Fluman for help with ribosome-profiling analysis. We also appreciate Tslil Ast, Raz Bar-Ziv, Hila Gingold, Daniel B. Goodman, Gleb Kuznetsov, Michael Napolitano, Dan Bar-Yaacov, Avihu Yona and Emmanuel Levy for helpful discussions and critical reading of the manuscript. Special thanks to Maya Schuldiner and Ron Milo for many supportive discussions. We also thank Rotem Sorek, Maya Shamir and Shany Doron for help with RNA-seq protocol, and Ernest Mordret and Avia Yehonadav for help with the mis-translation pipeline. Our gratitude goes to Shlomit Gilad, Sima Benjamin, Barak Markus, Alon Savidor and Yishai Levin from the Nancy & Stephen Grand Israel National Center for Personalized Medicine (G-INCPM) for assistance with high-throughput data and to Yoav Ram for help with Curveball. Special thanks to Nitai Steinberg for figure design. IF was supported by an EMBO short-term fellowship and thanks the Azrieli Foundation for the Azrieli PhD Fellowship award. This study was supported by the Minerva Foundation who funded the Minerva Center for Live Emulation of Evolution in the Lab.

References

1. Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* 12(1):32–42.
2. Gingold H, Pilpel Y (2011) Determinants of translation efficiency and accuracy. *Mol Syst Biol* 7(481):481.
3. Chamary J V., Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* 7(2):98–108.
4. Hershberg R, Petrov D a (2008) Selection on codon bias. *Annu Rev Genet* 42(iv):287–99.
5. Shah P, Gilchrist MA (2011) Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proc Natl Acad Sci U S A* 108(25):10231–6.
6. Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129(3):897–907.
7. Duret L (2002) Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev* 12(6):640–9.
8. Pechmann S, Frydman J (2013) Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat Struct Mol Biol* 20(2):237–43.
9. Dixon JR, et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398):376–380.
10. Tuller T, et al. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* 141(2):344–54.
11. Tuller T, Waldman YY, Kupiec M, Ruppin E (2010) Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci U S A* 107(8):3645–50.
12. Subramaniam AR, Zid BM, O'Shea EK (2014) An integrated approach reveals regulatory controls on bacterial translation elongation. *Cell* 159(5):1200–11.
13. Drummond DA, Wilke COC (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134(2):341–52.
14. Zhou T, Weems M, Wilke CO (2009) Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol* 26(7):1571–80.
15. Wilke CO, Drummond DA (2010) Signatures of protein biophysics in coding sequence evolution. *Curr Opin Struct Biol* 20(3):385–9.
16. Akashi H (1994) Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* 136(3):927–35.

17. Stoletzki N, Eyre-Walker A (2007) Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Mol Biol Evol* 24(2):374–81.

18. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324(5924):255–8.

19. Zhou M, et al. (2013) Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* 495(7439):111–5.

20. Navon S, Pilpel Y (2011) The role of codon selection in regulation of translation efficiency deduced from synthetic libraries. *Genome Biol* 12(2):R12.

21. Goodman DB, Church GM, Kosuri S (2013) Causes and effects of N-terminal codon bias in bacterial genes. *Science* 342(6157):475–9.

22. Gustafsson C, Govindarajan S, Minshull J (2004) Codon bias and heterologous protein expression. *Trends Biotechnol* 22(7):346–53.

23. Qian W, Yang JR, Pearson NM, Maclean C, Zhang J (2012) Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet* 8(3). doi:10.1371/journal.pgen.1002603.

24. Akashi H, Eyre-Walker A (1998) Translational selection and molecular evolution. *Curr Opin Genet Dev* 8(6):688–93.

25. Akashi H (2003) Translational selection and yeast proteome evolution. *Genetics* 164(4):1291–303.

26. Andersson SG, Kurland CG (1990) Codon preferences in free-living microorganisms. *Microbiol Rev* 54(2):198–210.

27. Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB (2013) Rate-limiting steps in yeast protein translation. *Cell* 153(7):1589–601.

28. dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* 32(17):5036–44.

29. Salis HM (2011) The ribosome binding site calculator. *Methods Enzymol* 498:19–42.

30. Yona AH, et al. (2013) tRNA genes rapidly change in evolution to meet novel translational demands. *Elife* 2:e01339.

31. Rogers HH, Griffiths-Jones S (2014) tRNA anticodon shifts in eukaryotic genomes. *RNA* 20(3):269–81.

32. Subramaniam AR, Pan T, Cluzel P (2013) Environmental perturbations lift the degeneracy of the genetic code to regulate protein levels in bacteria. *Proc Natl Acad Sci U S A* 110(6):2419–24.

33. Chevance FF V, Le Guyon S, Hughes KT (2014) The effects of codon context on in vivo translation speed. *PLoS Genet* 10(6):e1004392.

34. Ram Y, et al. (2015) *Predicting microbial relative growth in a mixed culture from*

growth curve data (Cold Spring Harbor Labs Journals) doi:10.1101/022640.

35. Mordret E, et al. (2018) Systematic detection of amino acid substitutions in proteome reveals mechanistic basis of ribosome errors. *Submitted*.

36. Karmon A, Pilpel Y (2016) Biological causal links on physiological and evolutionary time scales. *Elife* 5(APRIL2016):1–6.

37. Grosjean H, Fiers W (1982) Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* 18(3):199–209.

38. Sharp PM, Tuohy TM, Mosurski KR (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* 14(13):5125–43.

39. Sharp PM, Li W-HH (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15(3):1281–95.

40. Navon S, Pilpel Y (2011) The role of codon selection in regulation of translation efficiency deduced from synthetic libraries. *Genome Biol* 12(2):R12.

41. Presnyak V, et al. (2015) Codon optimality is a major determinant of mRNA stability. *Cell* 160(6):1111–24.

42. Frumkin I, et al. (2017) Gene Architectures that Minimize Cost of Gene Expression. *Mol Cell* 65(1):142–153.

43. Dittmar K a, Sørensen M a, Elf J, Ehrenberg M, Pan T (2005) Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO Rep* 6(2):151–7.

44. Wiltrout E, Goodenbour JM, Fréchin M, Pan T (2012) Misacylation of tRNA with methionine in Saccharomyces cerevisiae. *Nucleic Acids Res*:1–13.

45. Gingold H, et al. (2014) A dual program for translation regulation in cellular proliferation and differentiation. *Cell* 158(6):1281–92.

46. Yona AH, Frumkin I, Pilpel Y (2015) A Relay Race on the Evolutionary Adaptation Spectrum. *Cell* 163(3):549–559.

47. Wang HH, et al. (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460(7257):894–8.

48. Gallagher RR, Li Z, Lewis AO, Isaacs FJ (2014) Rapid editing and evolution of bacterial genomes using libraries of synthetic DNA. *Nat Protoc* 9(10):2301–2316.

49. Carr P a, et al. (2012) Enhanced multiplex genome engineering through co-operative oligonucleotide co-selection. *Nucleic Acids Res* 40(17):e132.

50. Gregg CJ, et al. (2014) Rational optimization of tolC as a powerful dual selectable marker for genome engineering. *Nucleic Acids Res* 42(7):4779–90.

51. Schwartz SA, Helinski DR (1971) Purification and characterization of colicin E1. *J Biol Chem* 246(20):6318–27.

52. Ma NJ, Moonan DW, Isaacs FJ (2014) Precise manipulation of bacterial chromosomes by conjugative assembly genome engineering. *Nat Protoc* 9(10):2285–2300.

53. Isaacs FJ, et al. (2011) Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. *Science* 333(6040):348–53.

54. Lajoie MJ, et al. (2013) Genomically Recoded Organisms Expand Biological Functions. *Science (80-)* 342(6156):357–360.

55. Zeevi D, et al. (2011) Compensation for differences in gene copy number among yeast ribosomal proteins is encoded within their promoters. *Genome Res* 21(12):2114–28.

56. Dar D, et al. (2016) Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science* 352(6282):aad9822.

57. Shalit T, Elinger D, Savidor A, Gabashvili A, Levin Y (2015) MS1-Based Label-Free Proteomics Using a Quadrupole Orbitrap Mass Spectrometer. *J Proteome Res* 14:1979–1986.

58. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 74(20):5383–5392.

Figure Legends

Figure 1 – Does the codon usage of a sub-set of genes affect translation efficiencies of other genes?
Upper panel: Hypothetical genomes of wild-type and re-coded strains are shown. Using genome engineering, we replaced abundant codons ("origin codon", blue lines) with rare codons ("destination codon", red lines) in highly expressed genes (white background).
Bottom left: two potential effects of re-coding on fitness: either reduce, or not affect the fitness.
Bottom middle: The translation efficiency of re-coded genes could be increased, decreased or not changed at all.
Bottom right: The translation efficiency of non-recoded genes that have the origin (blue) or destination (red) codon could be increased, decreased or not changed at all.

Figure 2 – Manipulating codon frequency of CGG results in global translation efficiency changes
A| We carried RNA-sequencing analysis of the transcriptome and mass-spectrometry analysis of the proteome for both the wild-type and re-coded strains. This allowed us to calculate translation efficiency (protein/mRNA) for each gene and classify two gene groups of increased or decreased translation efficiency with a fold change threshold of 1.5. The eight re-coded genes are colored in black, increased translation efficiency group is colored in blue, decreased translation efficiency group is colored in red and CGG-enriched genes are colored in green.
Inset| Ratios of translation efficiency between re-coded and wild-type cells for CGG-enriched genes (>5 occurrences) and CGG-depleted genes (no occurrences). CGG-enriched genes show lower translation efficiency ratios, p-Value=0.01.
B| Distribution of CGG occurrences, translated by $tRNA^{CCG}$, for increased (blue) or decreased (red) translation efficiency genes in re-coded strain compared to the wild-type strain. The group of decreased translation efficiency genes demonstrates higher CGG occurrences (p-Value=0.0018).
C| Distribution of CGU+CGC+CGA occurrences, all translated by $tRNA^{ACG}$, for increased (blue) or decreased (red) translation efficiency genes in re-coded strain compared to the wild-type strain. The group of increased translation efficiency genes demonstrates more codon CGU+CGC+CGA occurrences (p-Value=$6.79*10^{-5}$).
D| To increase $tRNA^{CCG}$ supply, we mutated the anticodon of $tRNA^{ACG}$ from ACG to CCG on the background of the re-coded strain, and termed this new strain as anticodon switched strain. We then analyzed its transcriptome and proteome. Note that much less genes are now deviating from the diagonal, particularly the CGG-enriched genes in green, suggesting that the anticodon switching mutation alleviated the translational difficulty of the re-coded strain. Color code is the same as in A.
Inset| CGG-enriched genes now show similar translation efficiency ratios as CGG-depleted genes, p-Value>0.05.
E| Same as B, but for the increased and decreased translation efficiency genes in anticodon-switched strain compared to the wild-type strain. In contrast to the previous comparison in B, these two groups utilize the CGG codon to the same extend (p-Value>0.05).
F| Same as C, but for the increased and decreased translation efficiency genes between the wild-type and anticodon-switched strain. In contrast to the previous comparison in C, these two groups utilize the CGU+CGC+CGA codon to the same extend (p-Value>0.05).
G| Translation initiation rates for increased, decreased and un-affected genes between re-coded and wild-type strains, as defined in A. Note that decreased translation efficiency genes, which are also enriched with CGG, also show higher initiation rates (p-Value=0.01) – in agreement with theory's prediction.

H| The translation efficacy pattern of the anticodon-switched strain clustered closer to the wild-type strain and away from the re-coded strain.

Figure 3 – Codon manipulation affects proteomic codon usage

A| We defined the codon proteomic usage as the multiplication of codon occurrences in each gene and the measured expression level of its protein product. We calculated the re-coded/WT ratio of this index for each of the 61 sense codons and observed that the CGG codon has the lowest value.

B| The same as A, but comparing the anticodon-switched strain with the wild-type. Due to the additional supply of $tRNA^{CCG}$, at the expense of $tRNA^{ACG}$, the CGG codon demonstrates similar values to other codons. Here, CGU and CGC show lower values than in A.

Figure 4 – Increased translational demand for CGG hampers protein synthesis of a reporter gene

A| To directly link frequency manipulation of CGG with protein synthesis of other genes, we utilized two versions of a YFP reporter-gene with six occurrences of either CGU or CGG. These YFP reporters were introduced separately to either the wild-type or the re-coded strain. Following the production of YFP vs. time along the growth cycle allowed us to derive the maximal YFP production for each combination of strain & YFP version.

B| For each strain, a YFP-CGG/YFP-CGU ratio is shown for maximal YFP production. The re-coded strain demonstrates lower ratios for both these parameters compared to the wild-type strain (p-Value=$5.6*10^{-5}$), supporting our observation that changing the codon usage of small subset of genes hampers the production of other genes that contain the CGG codon. Upon anticodon switching on the background of the re-coded strain, maximal YFP rate is restored to similar values of the wild-type strain.

Figure 5 – Change in global translation efficiency patters is deleterious

A| Growth experiment (OD vs. time) of the wild-type strain (blue), the re-coded strain (red) and the four anticodon-switched strains ($tRNA^{ACG}$ argQ in dark orange, $tRNA^{ACG}$ argZ in dark yellow, $tRNA^{ACG}$ argY in bright yellow, $tRNA^{ACG}$ argV in bright orange). The re-coded strain demonstrates reduction in relative fitness to 0.87 compared to the wild-type strain (p-Value<10^{-10}). The four strains with anticodon switching (increased $tRNA^{CCG}$ supply) on the background of the re-coded strain demonstrate a higher fitness compared to the re-coded strain itself, demonstrating that restored translation efficiencies patters also alleviated the growth defect (relative fitness compared to re-coded strain of switched argQ = 1.06, argZ=1.08, argY=1.02 and argV=1.04).

B| Switching the anticodon of $tRNA^{ACG}$ from ACG to CCG on the background of the wild-type strain reduces fitness (relative fitness compared to wild-type strain of switched argQ = 0.95 and argZ=0.96).

Figure 6 – introducing CGG on highly-expressed genes results in mis-translation events

Using a new methodology to identify translation errors from mass-spectrometry data, we identified such events for two of the recoded genes, ompC and ompA, exactly at the positions in which CGU or CGC were respectively mutated into CGG. While we did not find errors in the wild-type strain, we did observe them in the re-coded and anticodon-switched strains. The mis-translation event for ompC was found in the re-coded strain, and it changed the coded arginine with glutamine (which has a near-cognate anticodon, CUG) at position 238 of the protein. The mis-translation event for ompA was found in the anticodon-switched strain, and it changed the coded arginine with lysine at position 329 of the protein.

Supplementary Figure 1 – Genome engineering to create re-coded strain

The eight re-coded genes were split into two groups according to genome loci. Group A = ahpC, cspE, pal, ompX, ompF & ompA; Group B = atpE & ompC. To facilitate the genome editing process, each group was re-coded separately as described in methods section. After each group was successfully re-coded and validated via Sanger sequencing, selection markers were introduced to enable the merger between the two genomes by using the CAGE technology. Additionally, the donor strain was transformed with pRK29 that harbors the conjugation machinery. After conjugation was performed, cells were selected to all three markers and were then grown in permissive conditions that resulted in the loss of pRK29.

Supplementary Figure 2 – Manipulating codon frequency of CGG results in global proteomic changes

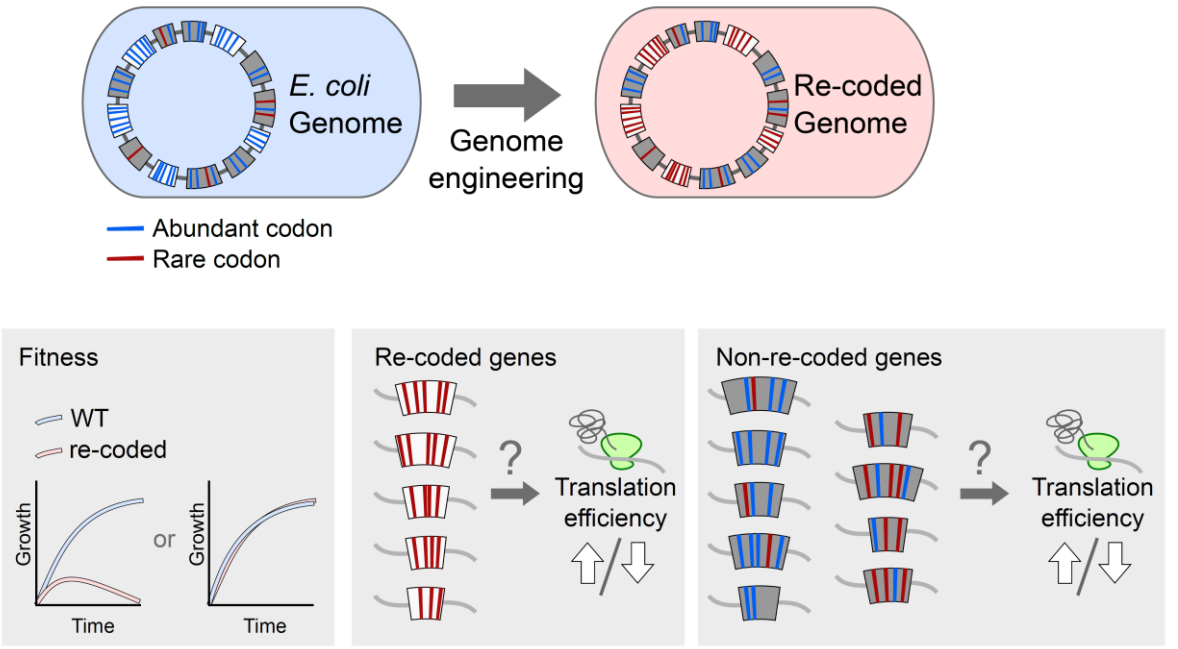
A| The introduction of CGG on highly expressed genes results in massive proteomic changes between the re-coded and wild-type strains.

B| Similarly to translation efficiency, as seen in Figure 2D, increasing tRNA(GGC) supply with an anticodon switching mutation restores the proteome of the cell back to its wild-type form.

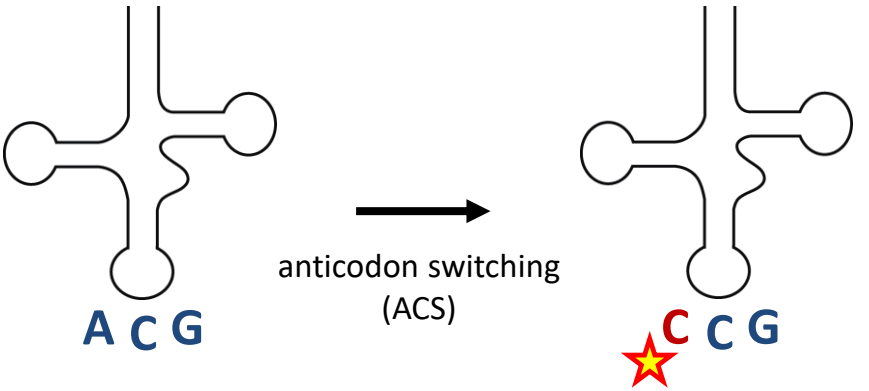
Box 1 – The arginine CGN box

We re-coded CGU and CGC ("origin codons") to CGG ("destination codon"). In *E. coli*, both origin codons are translated by $tRNA^{ACG}$ with the anticodon ACG due to an A-to-I modification that is mediated by the enzyme tRNA-specific adenosine deaminase (tadA). The destination codon is solely translated by $tRNA^{CCG}$, which translate no other codons. $tRNA^{ACG}$ and $tRNA^{CCG}$ appear in the genome with four and one copies, respectively. A direct arrow symbolizes fully-match interactions between codon and anticodon, while dashed arrows represent wobble interactions, which are enabled by modifying the ACG anticodon to ICG.

Figure 1


E. coli
Genome
Genome
engineering
Re-coded
Genome
Abundant codon
Rare codon
Fitness
WT
re-coded
Growth
Time
or
Growth
Time
Re-coded genes
?
Translation
efficiency
Non-re-coded genes
?
Translation
efficiency

Box 1



genomic occurrences	codon		anticodon	tRNA copies	Copies after ACS
29866	CGU	←	ACG	4	3
28424	CGC	⇢	GCG	0	0
4744	CGA	⇢	UCG	0	0
7273	**CGG**	←	**CCG**	**1**	**2**

Table 1

Genes	# CGU Codons	# CGC Codons	% of total CGG translation demand after re-coding
ompA	3	10	25.8
ompC	1	12	18.5
ompF	2	10	9.1
ompX	2	3	5.7
pal	0	8	4.5
ahpC	1	5	3.3
atpE	0	2	2.7
cspE	1	0	1.8
Total:	**10**	**50**	**71.3**

Figure 2

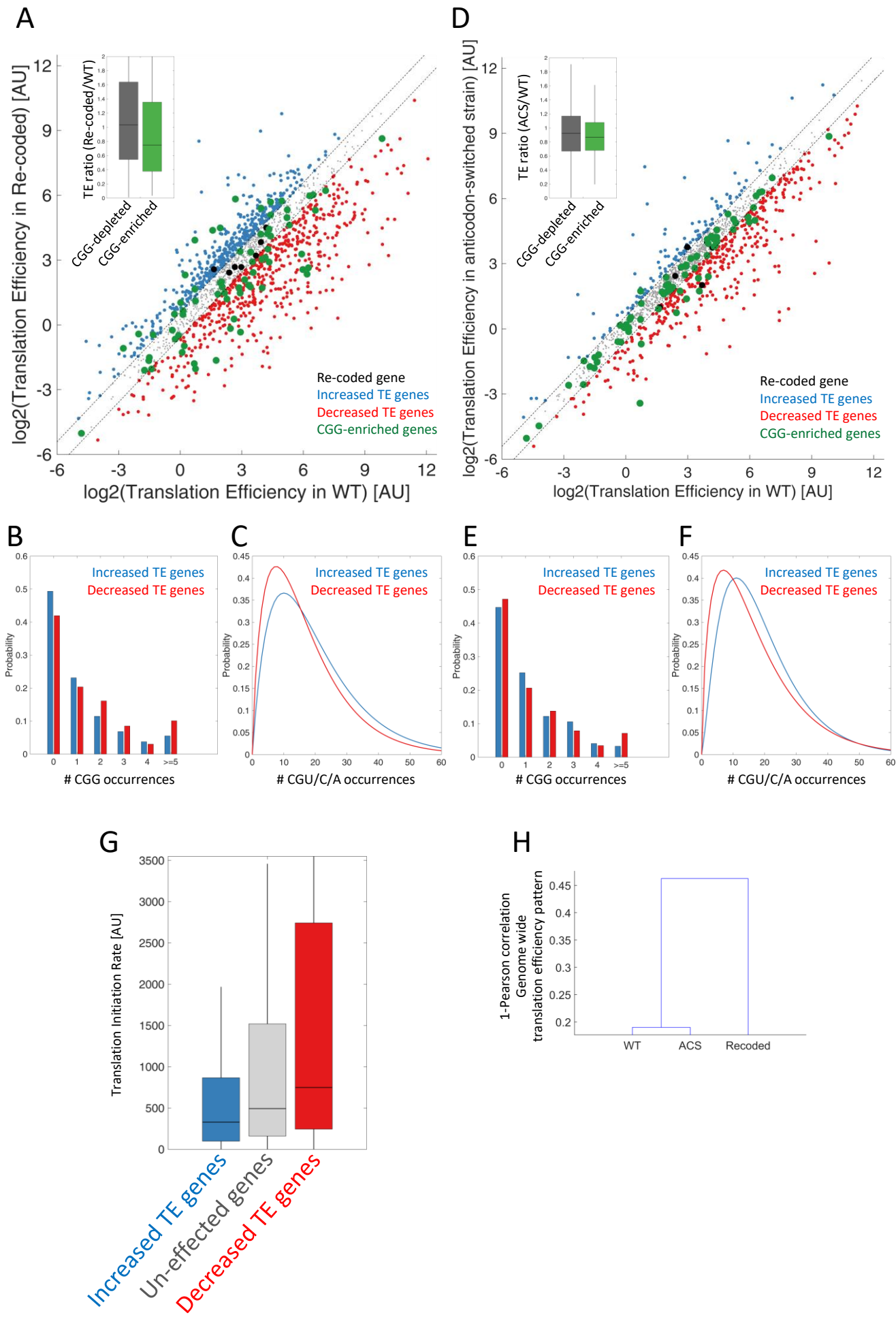
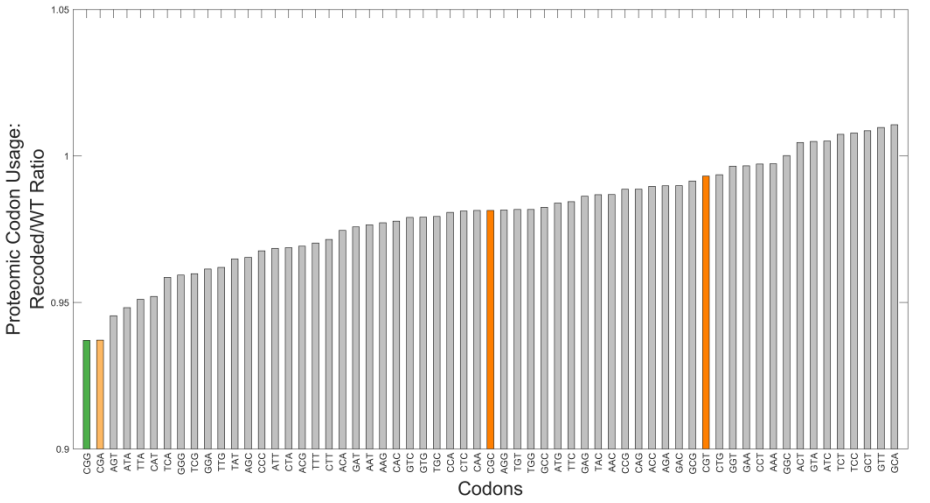

A
log2(Translation Efficiency in Re-coded) [AU]
log2(Translation Efficiency in WT) [AU]
TE ratio (Re-coded/WT)
CGG-depleted
CGG-enriched
Re-coded gene
Increased TE genes
Decreased TE genes
CGG-enriched genes
D
log2(Translation Efficiency in anticodon-switched strain) [AU]
log2(Translation Efficiency in WT) [AU]
TE ratio (ACS/WT)
CGG-depleted
CGG-enriched
Re-coded gene
Increased TE genes
Decreased TE genes
CGG-enriched genes
B
Increased TE genes
Decreased TE genes
Probability
CGG occurrences
C
Increased TE genes
Decreased TE genes
Probability
CGU/C/A occurrences
E
Increased TE genes
Decreased TE genes
Probability
CGG occurrences
F
Increased TE genes
Decreased TE genes
Probability
CGU/C/A occurrences
G
Translation Initiation Rate [AU]
Increased TE genes
Un-effected genes
Decreased TE genes
H
1-Pearson correlation
Genome wide
translation efficiency pattern
WT
ACS
Recoded

Figure 3

A


Proteomic Codon Usage:
Recoded/WT Ratio
1.05
1
0.95
0.9
CGG CGA AGT ATA TTA CAT TCA GGG TCG GGA TTG TAT AGC CCC ATT CTA ACG TTT CTT ACA GAT AAT AAG CAC GTC GTG TGC CCA CTC CAA CGC AGG TGT TGG GCC ATG TTC GAG TAC AAC CCG CAG ACC AGA GAC GCG CGT CTG GGT GAA CCT AAA GGC AGT GTA ATC TCT TCC GCT GTT GCA
Codons

B

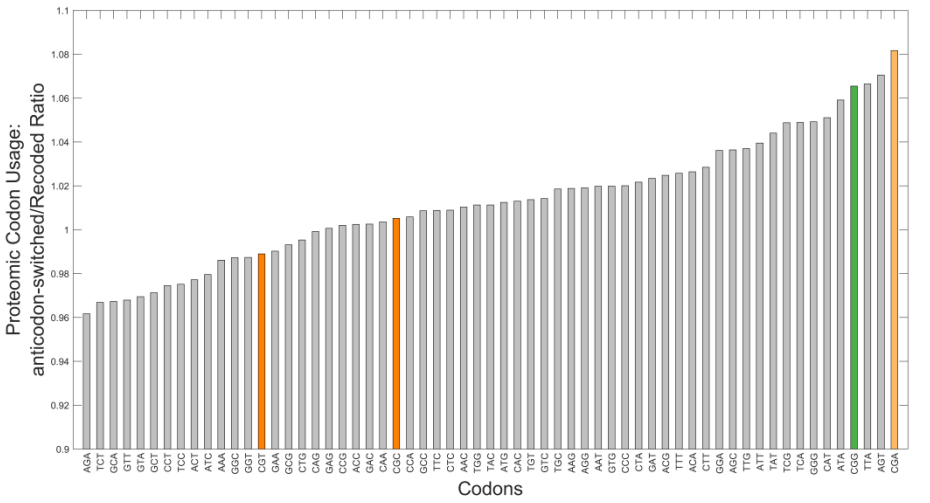
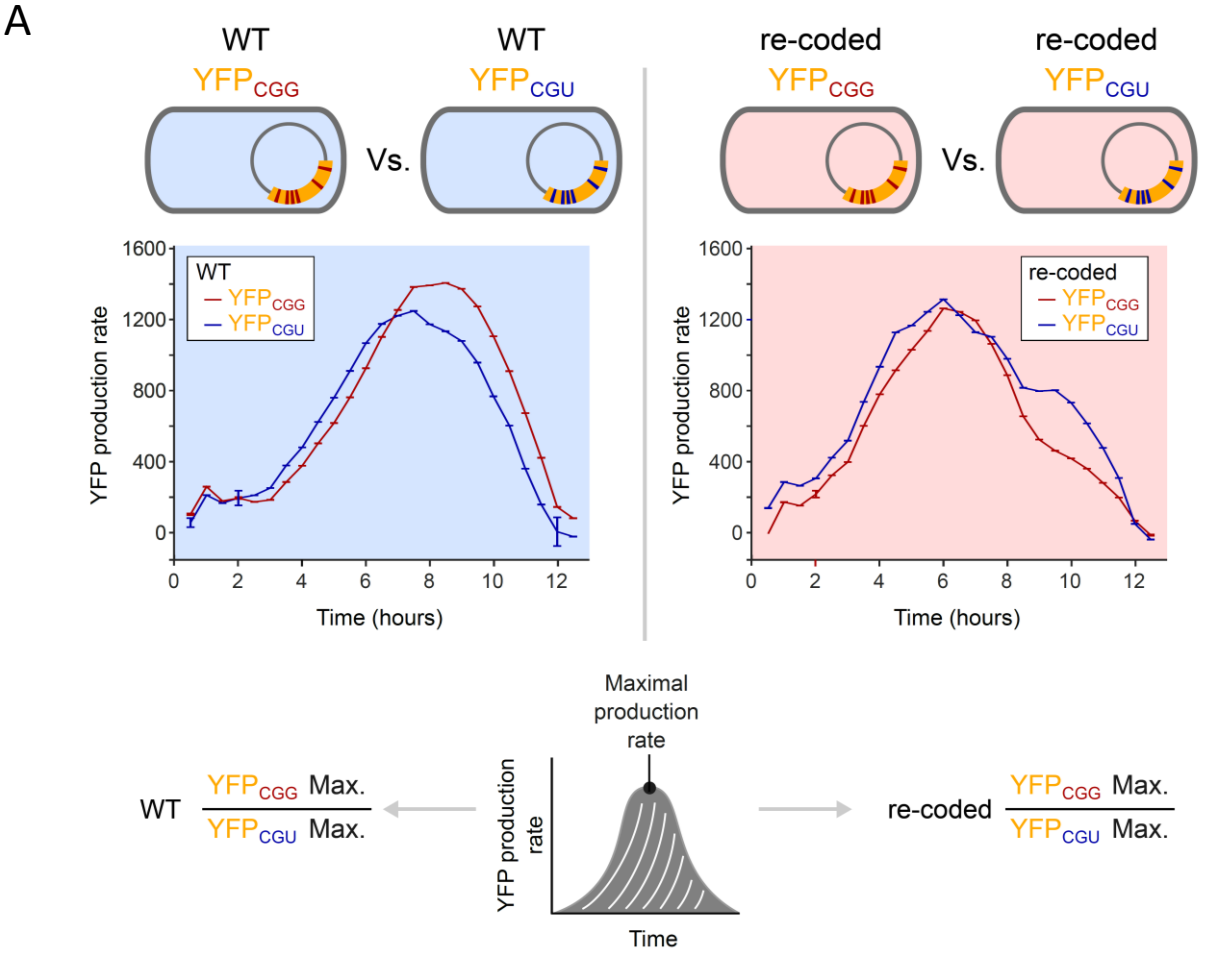

Proteomic Codon Usage:
anticodon-switched/Recoded Ratio
1.1
1.08
1.06
1.04
1.02
1
0.98
0.96
0.94
0.92
0.9
AGA TCA GCA GTT GTA GCT CCT TCC ACT ATC AAA GGC GGT CGT GAA GCG CTG CAG GAG CCG ACC GAC CAA CGC CCA GCC TTC CTC AAC TGG TAC ATG CAC TGT GTC TGC AAG AGG AAT GTG CCC CTA GAT ACG TTT ACA CTT GGA AGC TTG ATT TAT TCG TCA GGG CAT ATA CGG TTA AGT CGA
Codons

Figure 4


A
WT
YFP CGG
WT
YFP CGU
Vs.
re-coded
YFP CGG
re-coded
YFP CGU
Vs.
WT
YFP CGG
YFP CGU
YFP production rate
Time (hours)
re-coded
YFP CGG
YFP CGU
YFP production rate
Time (hours)
Maximal production rate
WT
YFP CGG Max.
YFP CGU Max.
YFP production rate
Time
re-coded
YFP CGG Max.
YFP CGU Max.

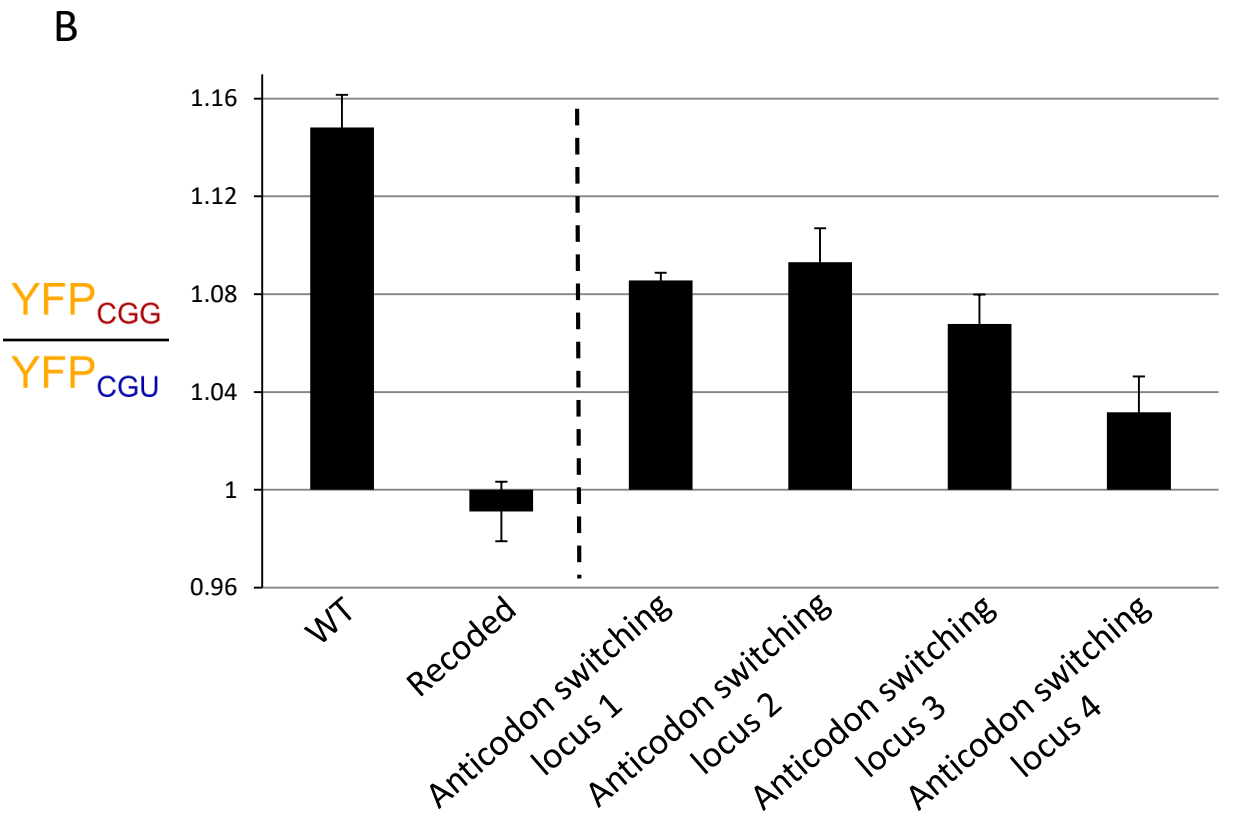
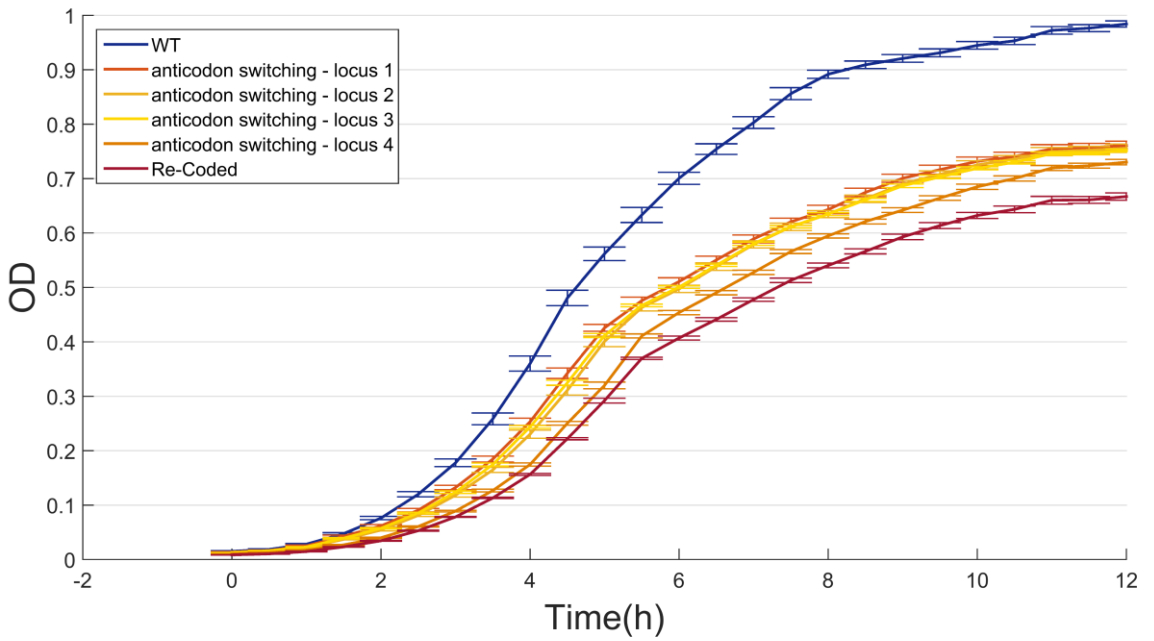

B
YFP CGG
YFP CGU
1.16
1.12
1.08
1.04
1
0.96
WT
Recoded
Anticodon switching locus 1
Anticodon switching locus 2
Anticodon switching locus 3
Anticodon switching locus 4

Figure 5

A


WT
anticodon switching - locus 1
anticodon switching - locus 2
anticodon switching - locus 3
anticodon switching - locus 4
Re-Coded
OD
Time(h)

B

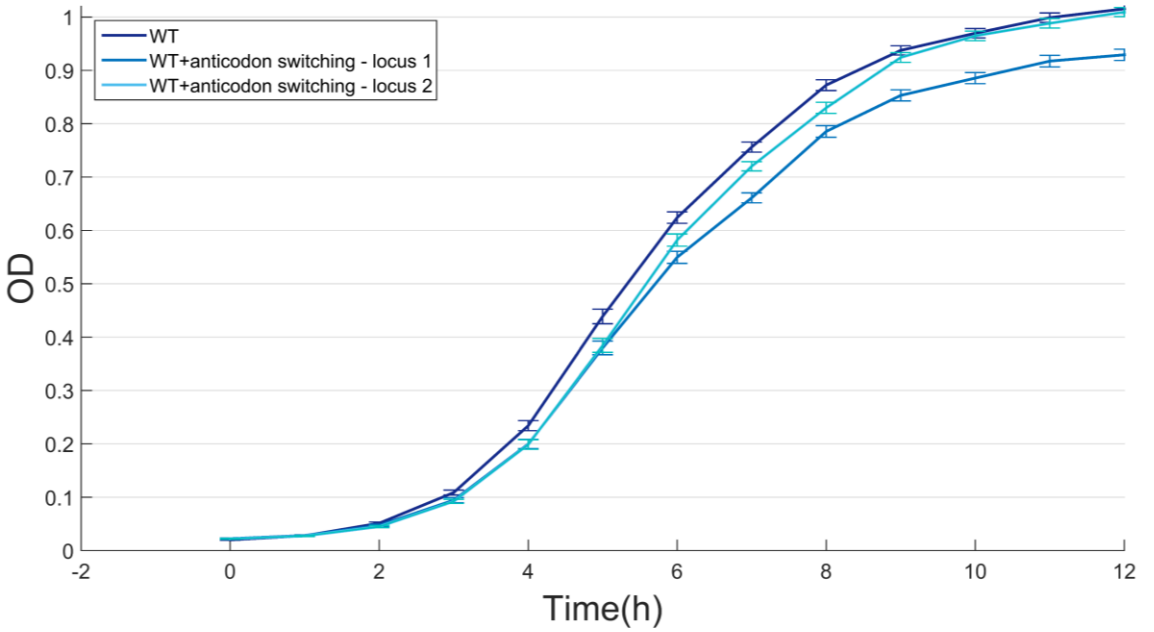
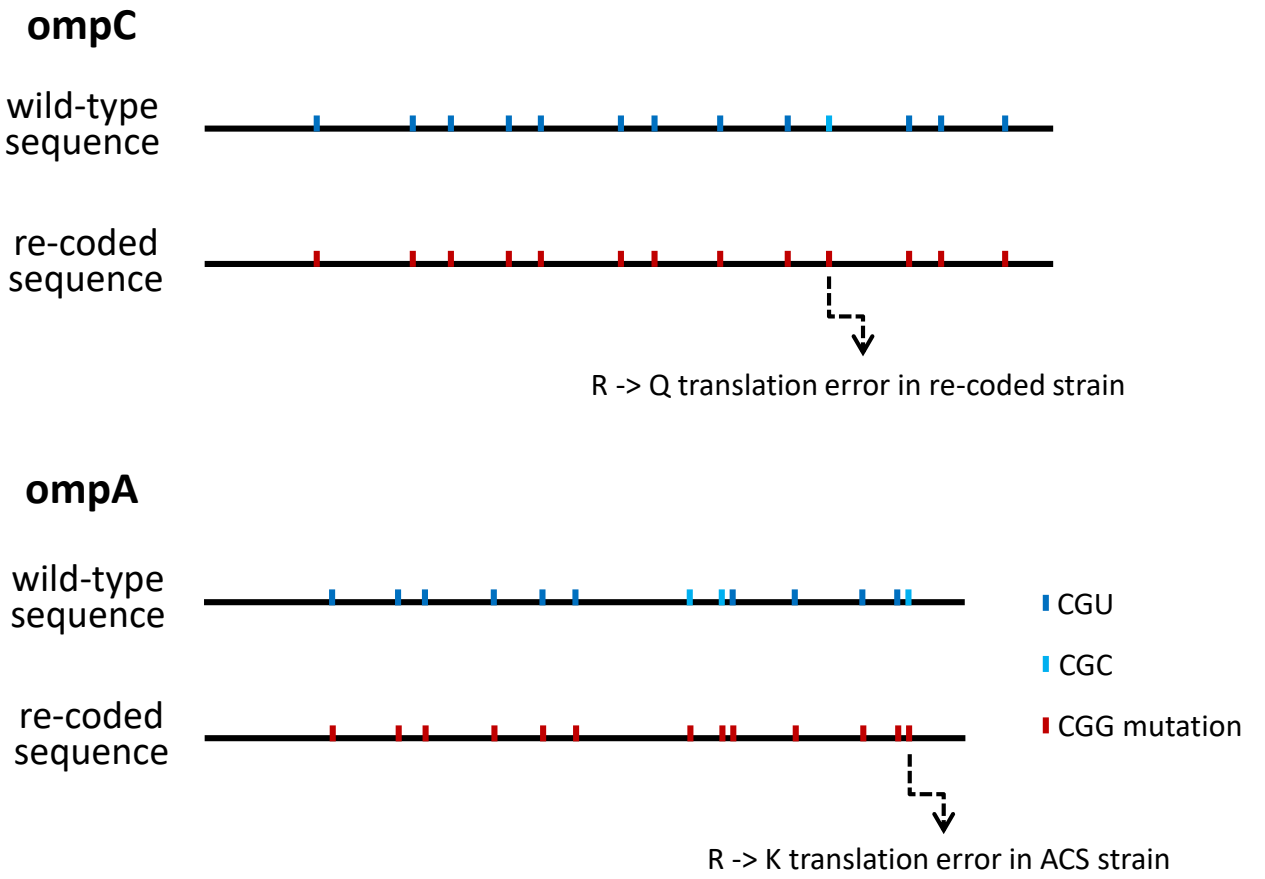
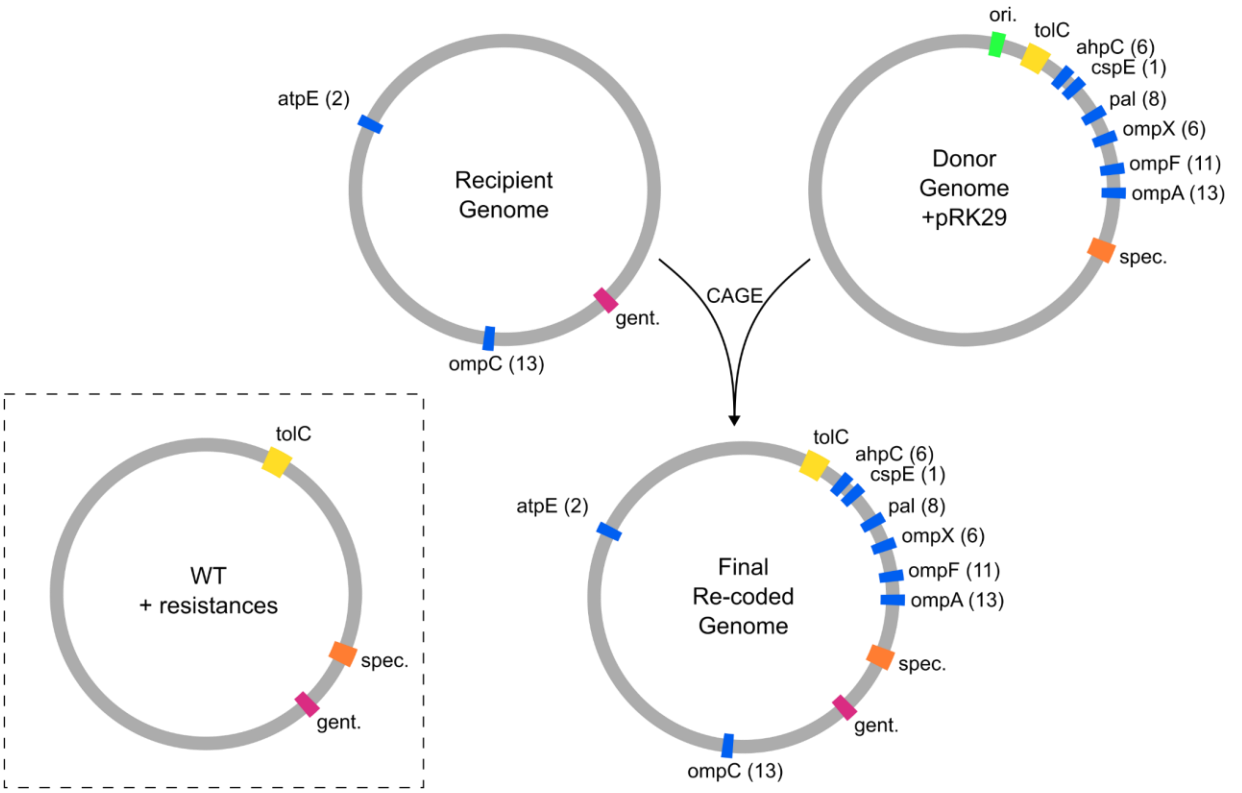

WT
WT+anticodon switching - locus 1
WT+anticodon switching - locus 2
OD
Time(h)

Figure 6

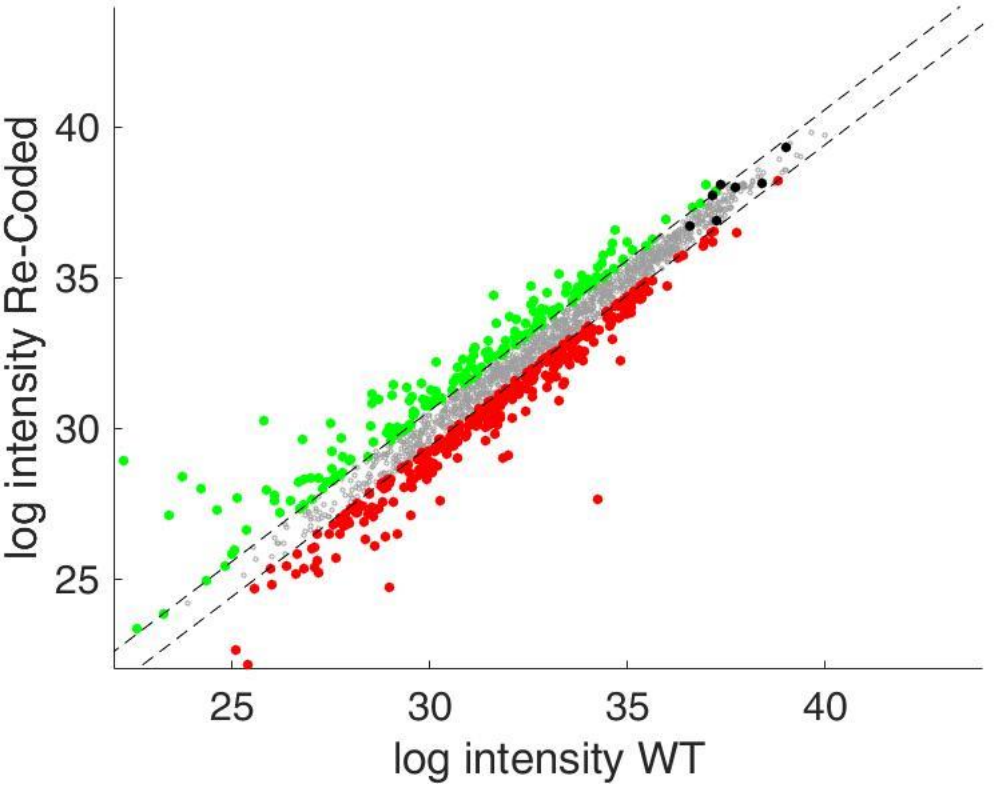

ompC
wild-type sequence
re-coded sequence
R -> Q translation error in re-coded strain
ompA
wild-type sequence
re-coded sequence
CGU
CGC
CGG mutation
R -> K translation error in ACS strain

Supplementary Figure 1

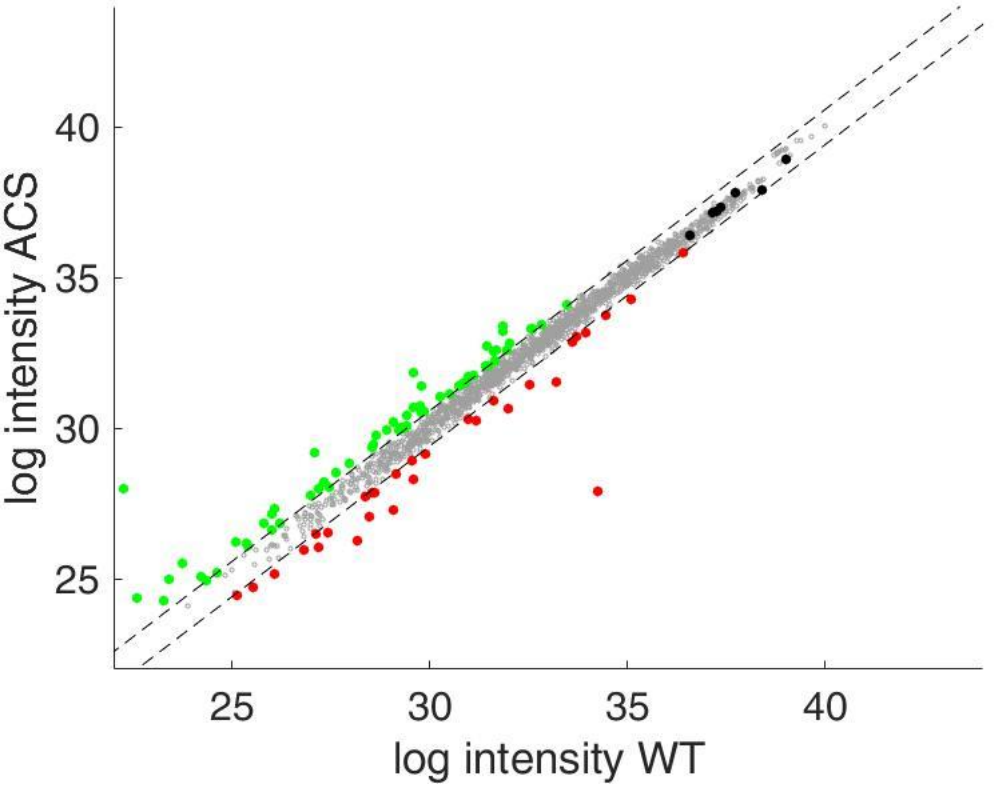


Supplementary Figure 2

A



B



Supplementary File 1

Off-target Mutations in re-coded strain

Position	Ref	Alt	Gene	Impact	Info	Effect
75915	G	A	sgrR	Moderate	P462L	NON_SYNONYMOUS_CODING
291125	G	A	insB1 + insA	Moderate	R10C + G74	NON_SYNONYMOUS_CODING SYNONYMOUS_CODING
344053	G	T		Modifier		INTERGENIC
451664	T	C		Modifier		INTERGENIC
520507	C	T	ybbP	Moderate	A31V	NON_SYNONYMOUS_CODING
553657	G	A	lpxH	Moderate	T95M	NON_SYNONYMOUS_CODING
644379	C	T	rna	Moderate	G209R	NON_SYNONYMOUS_CODING
651008	A	G	citC	Low	T283	SYNONYMOUS_CODING
794306	G	A	modE	Low	S113	SYNONYMOUS_CODING
826233	C	T	ybhS	Low	L340	SYNONYMOUS_CODING
869757	C	T	gsiB	Moderate	A116V	NON_SYNONYMOUS_CODING
871927	A	G	gsiD	Modifier	N13S	NON_SYNONYMOUS_CODING
891137	G	A	ybjC	Low	L75	SYNONYMOUS_CODING
975807	C	T	mukE	Low	N62	SYNONYMOUS_CODING
976037	A	G	mukE	Moderate	N139S	NON_SYNONYMOUS_CODING
987200	G	A		Modifier		INTERGENIC
1019932	T	C	ompA	Moderate	D41G	NON_SYNONYMOUS_CODING
1055317	T	C	torS	Moderate	T288A	NON_SYNONYMOUS_CODING
1198145	T	C	ymfE	Moderate	I31M	NON_SYNONYMOUS_CODING
1358078	A	G	puuP	Low	A370	SYNONYMOUS_CODING

1450180	A	G	tynA	Low	G390	SYNONYMOUS_CODING
1568925	T	C	gadC	Modifier		DOWNSTREAM
2067300	C	T	insH1	Moderate	G8R	NON_SYNONYMOUS_CODING
2099006	T	C	ugd	Moderate	Y203C	NON_SYNONYMOUS_CODING
2262518	G	A	fruB	Low	G326	SYNONYMOUS_CODING
2312925	TAAATGC	T		Modifier		INTERGENIC
2345274	A	G	nrdA	Moderate	H137R	NON_SYNONYMOUS_CODING
2564421	G	GT	yffR	Modifier		DOWNSTREAM
2599853	GT	G		Modifier		INTERGENIC
2602010	G	GC	hyfB	High	-65?	FRAME_SHIFT
2623506	A	G	ppk	Moderate	T155A	NON_SYNONYMOUS_CODING
2650927	C	T ½	yfhM	Moderate	S454N	NON_SYNONYMOUS_CODING
3020579	G	A	ygfM	Moderate	A14T	NON_SYNONYMOUS_CODING
3079014	G	A	cmtB	Moderate	A106V	NON_SYNONYMOUS_CODING
3101032	G	GC	yggN	High	-197?	FRAME_SHIFT
3114683	G	A	yghJ	Moderate	A1477V	NON_SYNONYMOUS_CODING
3123400	C	T	glcB	Moderate	G136S	NON_SYNONYMOUS_CODING
3203932	C	T	bacA	Moderate	R67H	NON_SYNONYMOUS_CODING
3310505	G	A	pnp	Low	L222	SYNONYMOUS_CODING
3356723	G	A	gltB	Moderate	G667S	NON_SYNONYMOUS_CODING
3420670	A	G	yhdX	Low	G179	SYNONYMOUS_CODING
3479581	T	C	kefB	Moderate	T343A	NON_SYNONYMOUS_CODING
3680536	G	A	yhjJ	High	Q469STOP	STOP_GAINED
3962536	A	G	rep	Low	E620	SYNONYMOUS_CODING
3974525	A	G	wecD	Modifier	N20S	NON_SYNONYMOUS_CODING

4173879	A	G	birA	Low	E266	SYNONYMOUS_CODING
4196244	A	G		Modifier		INTERGENIC
4309790	C	T	alsA	Low	A397	SYNONYMOUS_CODING
4347986	A	G	dcuB	Low	G253	SYNONYMOUS_CODING
4356415	T	C	dtpC	Modifier		UPSTREAM
4390374	C	T	psd	Modifier		UPSTREAM
4419725	C	CT		Modifier		INTERGENIC
4472418	C	CA	pyrL	High	-39?	FRAME_SHIFT
4492664	A	G	idnO	Moderate	F230L	NON_SYNONYMOUS_CODING
4561974	C	T	yjiJ	Low	P234	SYNONYMOUS_CODING
4563557	A	G	yjiK	Moderate	I16T	NON_SYNONYMOUS_CODING
4564586	C	T	yjiL	Moderate	R35H	NON_SYNONYMOUS_CODING
4616629	C	T	yjjI	High	W146STOP	STOP_GAINED

Off-target Mutations - Summary

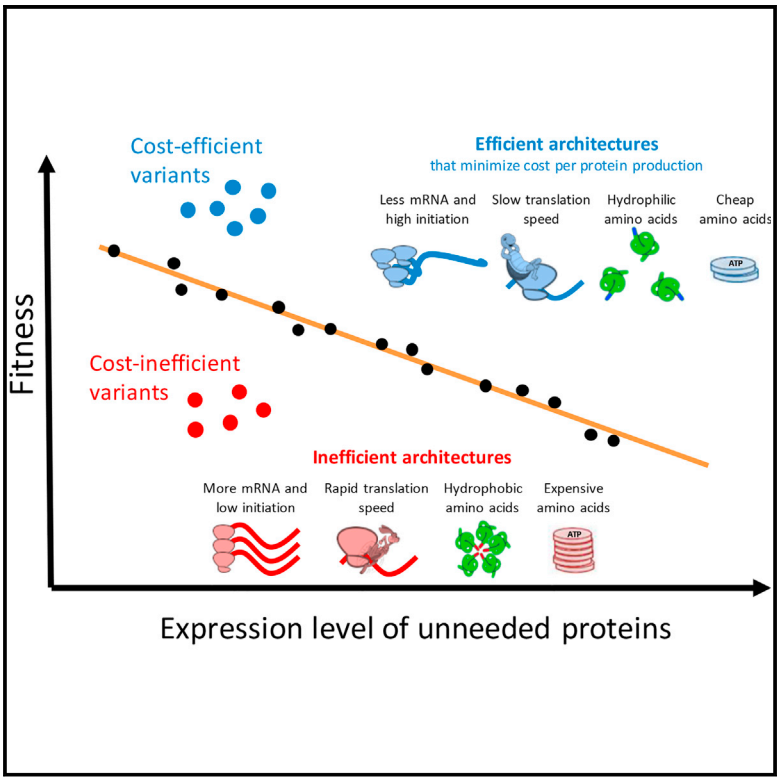
Non-Synonymous	26
Synonymous	15
Intergenic	7
Frame shift	3
Upstream	2
Downstream	2
Stop gained	2
NS & SY	1
Sum	**58**

Article

Molecular Cell

Gene Architectures that Minimize Cost of Gene Expression

Graphical Abstract



Authors

Idan Frumkin, Dvir Schirman, Aviv Rotman, ..., Song Wu, Sasha F. Levy, Yitzhak Pilpel

Correspondence

pilpel@weizmann.ac.il

In Brief

While numerous studies have investigated regulation of expression level, Frumkin et al. study gene design elements that govern expression costs and allow cells to minimize such costs while maintaining a given protein expression level.

Highlights

- Microorganisms can minimize expression cost with diverse molecular means
- Some design elements can produce more unneeded proteins but maintain high fitness
- Such elements optimize use of production machineries and utilize cheap materials
- Natural highly expressed genes evolved more forcefully to lower expression costs

Frumkin et al., 2017, Molecular Cell 65, 1–12
January 5, 2017
http://dx.doi.org/10.1016/j.molcel.2016.11.007

CellPress

Please cite this article in press as: Frumkin et al., Gene Architectures that Minimize Cost of Gene Expression, Molecular Cell (2016), http://dx.doi.org/10.1016/j.molcel.2016.11.007

Molecular Cell
Article

CellPress

Gene Architectures that Minimize Cost of Gene Expression

Idan Frumkin,[1,5] Dvir Schirman,[1,5] Aviv Rotman,[1,5] Fangfei Li,[2,3] Liron Zahavi,[1] Ernest Mordret,[1] Omer Asraf,[1] Song Wu,[3] Sasha F. Levy,[2,4] and Yitzhak Pilpel[1,6,*]
[1]Department of Molecular Genetics, Weizmann Institute of Science, 7610001 Rehovot, Israel
[2]Laufer Center for Physical and Quantitative Biology
[3]Department of Applied Mathematics and Statistics
[4]Department of Biochemistry and Cell Biology
Stony Brook University, Stony Brook, NY 11794, USA
[5]Co-first author
[6]Lead Contact
*Correspondence: pilpel@weizmann.ac.il
http://dx.doi.org/10.1016/j.molcel.2016.11.007

SUMMARY

Gene expression burdens cells by consuming resources and energy. While numerous studies have investigated regulation of expression level, little is known about gene design elements that govern expression costs. Here, we ask how cells minimize production costs while maintaining a given protein expression level and whether there are gene architectures that optimize this process. We measured fitness of ~14,000 *E. coli* strains, each expressing a reporter gene with a unique 5′ architecture. By comparing cost-effective and ineffective architectures, we found that cost per protein molecule could be minimized by lowering transcription levels, regulating translation speeds, and utilizing amino acids that are cheap to synthesize and that are less hydrophobic. We then examined natural *E. coli* genes and found that highly expressed genes have evolved more forcefully to minimize costs associated with their expression. Our study thus elucidates gene design elements that improve the economy of protein expression in natural and heterologous systems.

INTRODUCTION

In nature, cells must express different genes in a regulated manner. On one hand, genes must be expressed at levels that maximize their benefit, and on the other, cells need to minimize the genes' production costs (Dekel and Alon, 2005; Wagner, 2005). Costs of expression originate from spending cellular resources, such as building blocks (amino acids and nucleotides), from allocation of cellular machineries (RNA polymerase and ribosome), and from energy and reducing power consumption (Bienick et al., 2014; Glick, 1995; Ibarra et al., 2002; Rang et al., 2003). Even after their production, proteins might still impose costs when degraded or by exerting toxicity, e.g., due to aggregation (Geiler-Samerotte et al., 2011). Understanding what molecular processes determine expression cost, its relation to cellular growth and gene regulation, and how costs evolutionarily shape the genome are key aspects of cell biology that remain largely elusive. While numerous studies investigated molecular mechanisms and gene sequence architectures that regulate expression level (Gingold and Pilpel, 2011; Kudla et al., 2009; Qian et al., 2012; Sharp et al., 1986; Subramaniam et al., 2013), very little is known about design elements that govern expression costs.

Different works have studied expression costs in unicellular organisms by imposing the expression of an unneeded protein (Bentley et al., 1990; Dekel and Alon, 2005; Dong et al., 1995; Kafri et al., 2016; Rang et al., 2003; Scott et al., 2010). The production of such unneeded proteins diverts resources from synthesis of the cell's own proteins, thus decreasing cellular fitness (Emilsson and Kurland, 1990; Marr, 1991; Vind et al., 1993). Central to these studies is the characterization of the correlation between the imposed expression levels of the unneeded proteins to the cost. Yet, ultimately natural selection dictates the expression level of natural genes according to the required concentration of each protein. Thus, a fundamental question, which has not been addressed before, is how cells can achieve a specific expression level of a gene while minimizing its expression costs.

Addressing this question is challenging because changes in sequence could affect both expression level and expression costs. To disentangle expression level and expression costs and reveal mechanisms that affect cost per protein molecule, we utilized a synthetic reporter library of ~14,000 different sequence variants, each fused upstream to a GFP gene (Goodman et al., 2013). We then combined competition assays and deep sequencing to measure the fitness of all variants in parallel. This procedure allowed us to elucidate gene architectures that minimize expression cost at a given protein expression level. We show that various molecular mechanisms, such as protein/mRNA ratios, ribosome early elongation pauses, amino acid synthesis costs, and peptide hydrophobicity, determine the cost per protein molecule. We then generated a model that predicts the cost effectiveness of gene architectures and applied it to natural *E. coli* genes. We found that highly expressed genes have

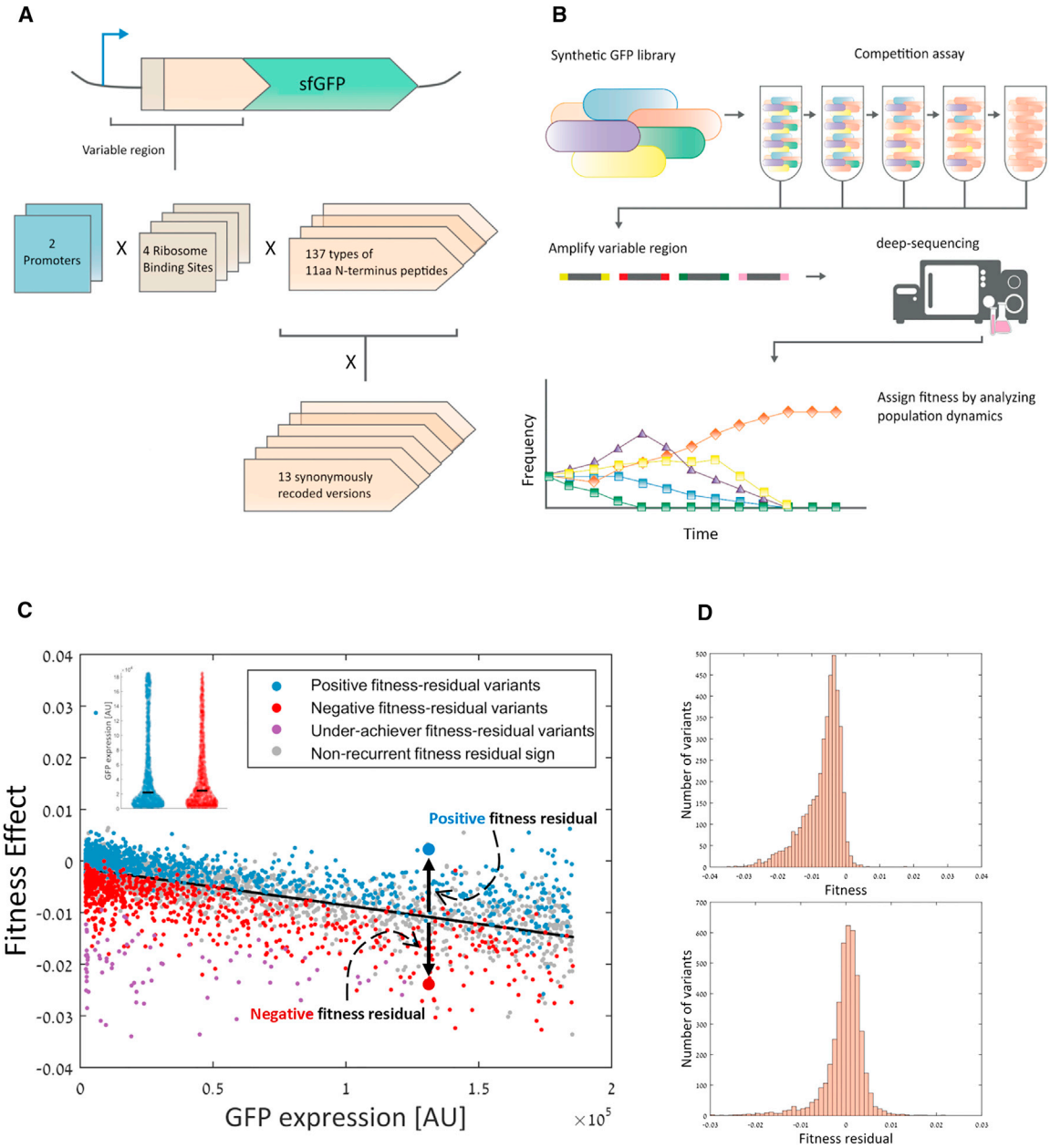


Figure 1. 5′ Gene Architectures Affect Cost of Gene Expression at a Given Expression Level

(A) We utilized a synthetic library of ~14,000 *E. coli* strains, each expressing a GFP construct with a unique 5′ architecture that includes a promoter, ribosome binding site (RBS), and an 11-amino-acid-fused peptide. There were two different promoter types, four RBSs, and 137 amino acid fusions that were each synonymously re-coded to 13 different versions (see Goodman et al., 2013 for full details).

(B) FitSeq methodology to measure relative fitness of strains in a pooled synthetic library. First, the library was grown six independent times for ~84 generations, and samples were taken at generations 0, ~28, ~56, and ~84. Then, unique 5′ gene architectures were simultaneously amplified and sent for deep sequencing, which allowed to follow the frequency of each variant in the population over the course of the experiment. Finally, a relative fitness score was assigned for each variant based on its frequency dynamics.

(C) GFP expression level (as measured by Goodman et al., 2013; x axis) versus fitness effect (based on results of repetition C; y axis) of each variant in the library (Pearson correlation, $r = -0.79$, $p < 10^{-200}$). Fitness effect comes from the burden of expressing unneeded proteins on cellular growth and is calculated by analyzing the frequency dynamics of each variant (see Experimental Procedures). We defined fitness residual as the difference between a variant's observed and expected fitness. The expected fitness is calculated from the regression line between GFP expression and fitness (black line). Some variants consistently demonstrated positive (blue dots, n = 975) or negative (red dots, n = 815) fitness residual sign. Other variants showed extremely low fitness residual, and we termed those variants as "underachievers" (purple dots, n = 80). The group size of positive, negative, and underachiever variants are significantly much higher than expected by chance (Supplemental Information). These results suggest that certain 5′ gene architectures can increase or reduce the cost of gene

(legend continued on next page)

evolved more forcefully to be encoded by cost-minimizing mechanisms. Our observations indicate that natural selection has shaped genes' architectures to reduce cost of gene expression.

RESULTS

5′ Gene Architecture Affects Cost of Gene Expression

Our question is whether different gene sequence elements can minimize cost of expression per protein molecule and hence increase cellular fitness. To focus on sequence features at the 5′ region of a gene, we utilized a previously published synthetic gene library (Goodman et al., 2013) composed from ~14,000 different variants expressing a GFP gene. Each variant holds a unique variable 5′ gene architecture that includes a promoter, a ribosome binding site (RBS), and an 11-amino-acid-long N terminus fusion (Figure 1A; Experimental Procedures).

To reveal the expression cost of each variant, we measured relative fitness of all variants in parallel in a competition assay in six independent repeats. We then deep sequenced the variable region of the pool of variants and calculated relative fitness of each variant (Figure 1B; see Experimental Procedures).

We regressed fitness values against GFP expression levels and observed a negative, linear correlation (Figure 1C, Pearson correlation, $r = -0.79$, $p < 10^{-200}$; Figure S1A). The linear decline in fitness with expression is in agreement with previous studies (Kafri et al., 2016; Scott et al., 2010). The regression line, which outlines the relations between fitness and expression, allowed us to estimate the expected fitness for each library variant according to its GFP expression level. Variants whose fitness does not deviate consistently across repeats from this regression line are deduced not to utilize mechanisms that enhance or reduce the production cost per protein molecule.

Yet, many variants did deviate from the linear regression line, demonstrating fitness that is higher or lower than expected given their GFP expression levels. We hypothesized that variants that repeatedly deviated from the expected fitness might utilize gene architectures that either reduce or increase the cost of GFP production per protein molecule. Hence, we calculated each variant's "fitness residual," which we defined as the difference between the actual fitness that we measured for the variant and the fitness expected for it according to its GFP expression level and the linear regression (Figure 1C). A positive fitness residual means that a given variant showed higher fitness than expected given its GFP expression level, suggesting that it can produce this GFP level with lower costs. A negative fitness residual means that the variant showed lower fitness than expected given its GFP expression level.

We then classified each variant as either positive or negative according to its fitness residual sign (Figure 1C, blue and red dots; see Experimental Procedures). Since the observed fitness residual is sensitive to biological noise (i.e., drift during competition) and experimental errors (i.e., sampling errors), we only classified variants as positive or negative if their fitness residual sign was identical in at least five out of the six repeats of the experiments in each of the two final sampling points of the competition (see Experimental Procedures and Supplemental Experimental Procedures). This approach resulted in 975 positive and 815 negative variants (significantly higher than expected by chance even at very high levels of measurement errors; Supplemental Experimental Procedures). Classification into either positive or negative fitness residual groups allowed us to eliminate the effect of GFP expression level on fitness as these two groups demonstrate the same expression distribution (Figure 1C, inset).

We also noticed a set of 80 library variants, which we termed "underachievers," whose fitness residual scores were repeatedly at the bottom 5% of the entire library (Figure 1C, purple dots; see Experimental Procedures). We hypothesized that these underachiever variants show extremely low fitness residuals because they produce GFP even more wastefully, and we expected them to show stronger usage of low-efficiency gene architectures compared to the negative fitness residual group. There appeared to be no "overachievers" in these data.

Production of More Proteins per mRNA Molecule Is an Economic Means to Minimize Expression Costs

We first hypothesized that reaching the same GFP level with lower levels of mRNA of the GFP gene could be beneficial. While positive and negative fitness residual variants come from the same distribution of GFP expression levels (Figure 1C, inset), we compared their GFP mRNA levels and found positive variants to have lower levels compared to negative variants (Figure 2A; Wilcoxon rank-sum, $p = 1.6 \times 10^{-9}$, effect size = 58.26%; see Experimental Procedures). This difference was independent of GFP level: binning the data according to GFP levels, we observed the reduced levels of mRNA for positive variants in all expression bins (Figure S1B).

The observation that positive variants have equal GFP protein levels but lower GFP mRNA levels indicates that they are able to produce more GFP proteins per mRNA molecule. We postulated that high translation initiation rate could be a mechanism for maintaining the same GFP levels despite low mRNA levels in positive variants. We calculated initiation rates for all library variants using the "Ribosome Binding Site Calculator" (Salis, 2011) and observed that indeed positive variants had higher initiation rates (Figure 2B; effect size = 61.9%, Wilcoxon rank-sum, $p = 3.7 \times 10^{-18}$). This observation holds true when examining mRNA level versus translation initiation rate at the individual variant level (Figure S2A). Indeed, when examining translation efficiency per variant (using measured protein levels divided by mRNA levels), positive variants demonstrated higher translation efficiencies than negative fitness residual variants (Figure 2C; effect size = 55.67%, Wilcoxon rank-sum, $p = 3.4 \times 10^{-5}$). Moreover, we found that underachiever variants demonstrated even

expression. See also Figure S1A. Inset: positive (blue violin plot) and negative (red violin plot) fitness residual variants come from the same distribution of GFP expression level (Wilcoxon rank-sum, p = 0.46). Black line represents the median value. Thus, the effect of GFP levels on fitness was successfully factored out, thus allowing us to elucidate other molecular mechanisms that tune expression cost at given expression levels.

(D) Fitness and fitness residuals demonstrate different distributions. While most variants showed negative fitness values, fitness residual is more similar to a normal distribution, though with a negative tail.

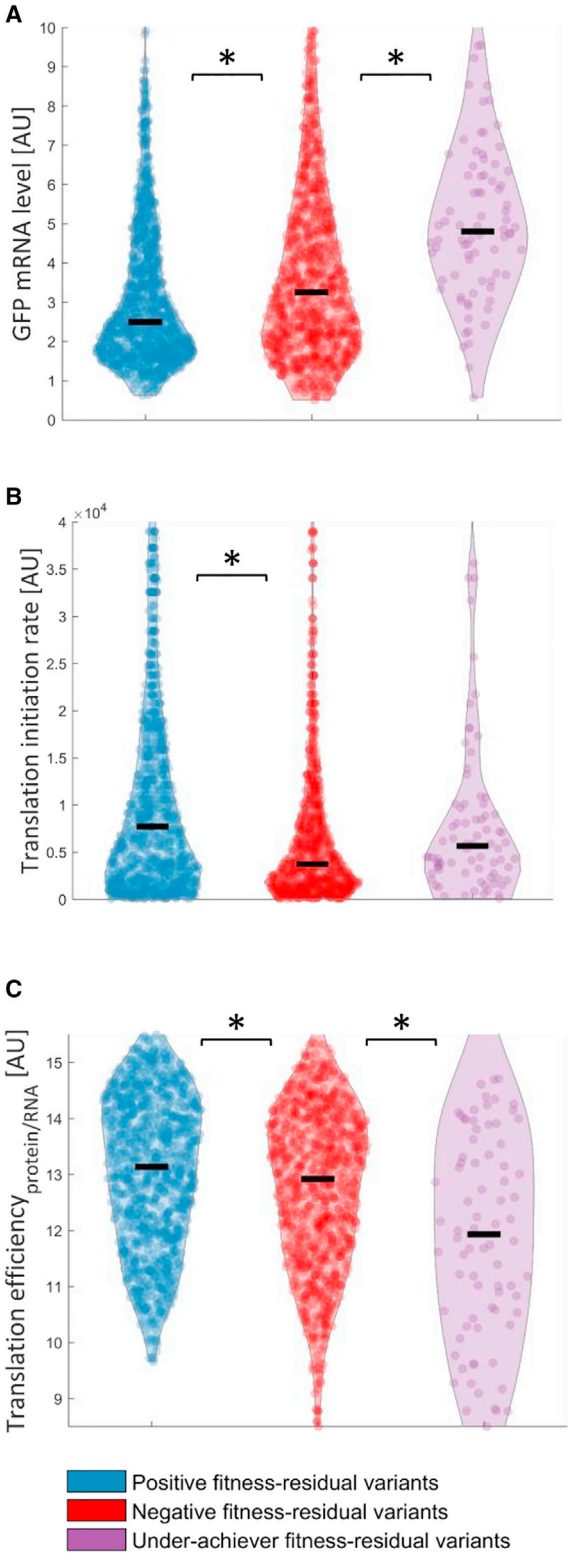


Figure 2. Higher Ratio of GFP Protein/mRNA Minimizes Cost of Gene Expression

(A) Although coming from the same distribution of GFP levels, positive variants (blue violin plot) demonstrate lower mRNA levels of the GFP gene compared to negative variants (red violin plot) (effect size = 58.26%, Wilcoxon rank-sum, $p = 1.6 \times 10^{-9}$). Consistently, underachiever variants (purple violin plot) show higher mRNA levels compared to negative variants (effect size = 68.04%, Wilcoxon rank-sum, $p = 9.6 \times 10^{-8}$). Black line represents the median value. (B) Positive variants show higher translation initiation rates compared to negative variants (effect size = 61.9%, Wilcoxon rank-sum, $p = 3.7 \times 10^{-18}$). (C) Positive variants demonstrate higher translation efficiencies (protein/mRNA) compared to negative variants (effect size = 55.67%, Wilcoxon rank-sum, $p = 3.4 \times 10^{-5}$). Consistently, underachiever variants (purple violin plot) further show lower translation efficiencies compared to negative variants (effect size = 63.06%, Wilcoxon rank-sum, $p = 1.1 \times 10^{-4}$).
Statistically significant differences ($p < 0.05$) are marked with an asterisk. See also Figures S1B and S2A.

higher mRNA levels and lower translation efficiencies compared to the negative variants (Figures 2A and 2C; effect size = 68.04% and 63.06%, Wilcoxon rank-sum, $p = 9.6 \times 10^{-8}$ and 1.1×10^{-4}, respectively). Thus, by increasing translation efficiency, cells reduce transcription costs and hence also cost per protein.

Slower Translation Speed at Early Elongation of Coding Region, Achieved by Diverse Means, Reduces Expression Costs

We next aimed to elucidate other cellular mechanisms that directly regulate the translation machinery and that might reduce expression costs. We first examined codon decoding speeds by the ribosome. Codon adaptation of transcripts to the cellular tRNA pool has been shown to be a regulatory mechanism for translation elongation (Goodarzi et al., 2016; Higgs and Ran, 2008; Kudla et al., 2009; Plotkin and Kudla, 2011; Shah and Gilchrist, 2011; Weinberg et al., 2016; Yona et al., 2013). Specifically, the prevalence of slowly translated codons at the 5′ of open reading frames (ORFs) has been suggested to support the efficiency of gene translation (Tuller et al., 2010a). This "ramp model" proposes that delaying ribosomes at the beginning of the elongation phase decreases downstream ribosomal pauses and collisions, which can therefore reduce ribosome jamming, and perhaps also ribosomal abortion events.

Although contradicting evidence were reported for the existence and relevance of this mechanism to expression level (Charneski and Hurst, 2014; Dana and Tuller, 2014; Heyer and Moore, 2016; Ingolia et al., 2009; Shah et al., 2013; Tuller and Zur, 2015), the main prediction of the model—that 5′ ramping reduces cost of expression at a given expression level—has not been tested so far. Here, we had the first opportunity to test this hypothesis as only the 5′ variable region of the GFP varied in the library, while all other parameters remained constant. Thus, we asked whether slow 5′ translation speed is associated with positive fitness residual. We used "mean of the typical decoding rates" (MTDR) (Dana and Tuller, 2014), a measure of codon decoding time derived empirically from ribosome profiling data in *E. coli* (see Experimental Procedures), to calculate translation speed for each library variant. We reasoned that if translational ramp is beneficial, then low MTDR scores, i.e., low ribosome speeds, should be more prevalent among the positive fitness residual variants. Indeed, our results showed that positive variants demonstrate significantly lower translation speeds at the N-terminal fusion (Figure 3A; effect size = 59.55%, Wilcoxon rank-sum, $p = 3 \times 10^{-12}$) and further for

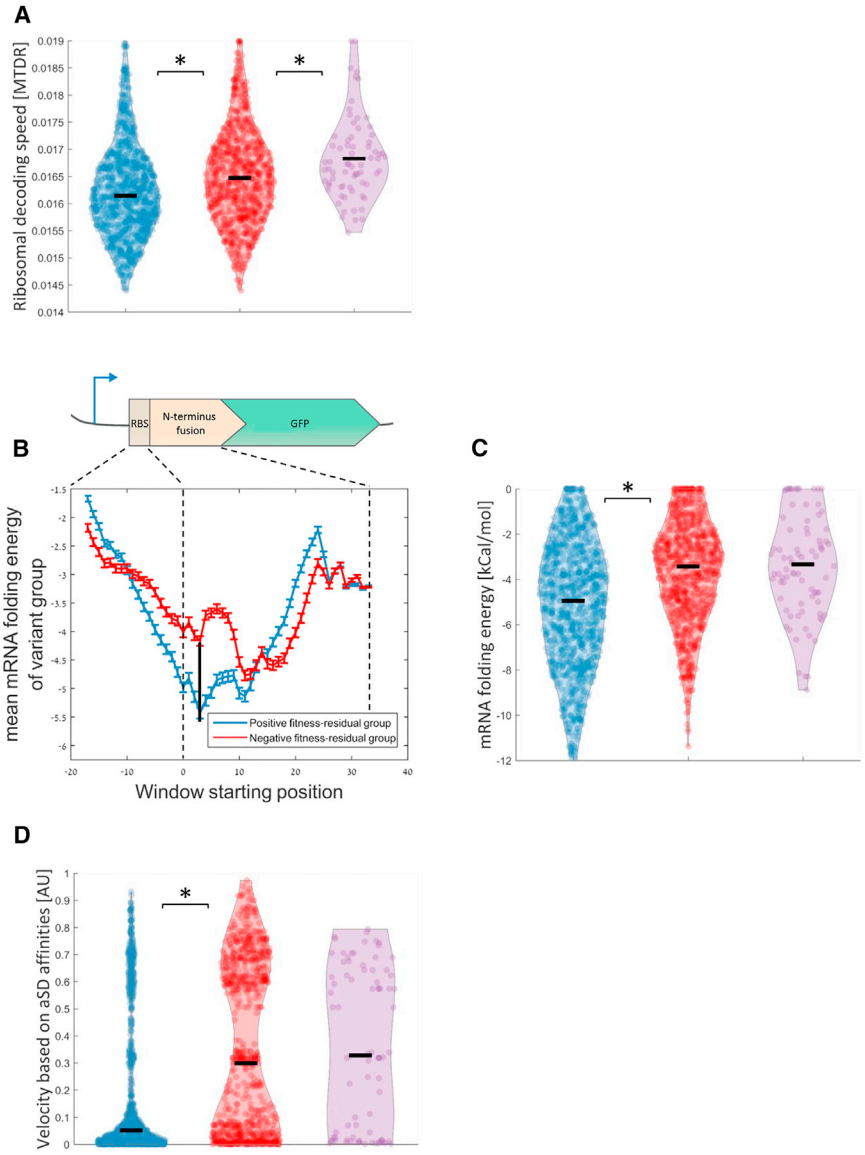


Figure 3. Slow Translation Speed at Early Elongation, Achieved by Diverse Molecular Means, Reduces Expression Cost

(A, C, and D) Positive variants show lower values of codon decoding speed (A), stronger mRNA structures (C), and lower speeds due to higher anti-Shine Dalgarno affinities (D) compared to negative variants (effect size = 59.55%, 65.03%, and 63.82%, Wilcoxon rank-sum, $p = 3 \times 10^{-12}$, 5.4×10^{-28}, and 6.3×10^{-24}, respectively). Statistically significant differences ($p < 0.05$) are marked with an asterisk. See also Figure S1B.

(B) Mean folding energy of mRNA secondary structure according to window's start position for positive (blue curve) and negative (red curve) variants; error bars represent SEM. Dashed lines mark different positions along the variable region upstream to the GFP. Black vertical line marks the beginning of window with the largest observed difference, which is found at nucleotide positions +4 of the ORF, just after the first AUG codon. The distributions at this window position are seen in (C). See also Figure S2B.

the underachievers (effect size = 64.79%, Wilcoxon rank-sum, $p = 1.2 \times 10^{-5}$).

Though in the original ramp model ribosome attenuation was proposed to be obtained by codons that correspond to rare tRNAs, additional mechanisms that can slow down the ribosome at early elongation regions could serve in ramping. These mechanisms include, in particular, tight mRNA secondary structure (Goodman et al., 2013; Tholstrup et al., 2012; Tuller et al., 2010b; Wen et al., 2008) and high affinity to the anti-Shine Dalgarno (aSD) motif of the ribosome (Li et al., 2012). We thus examined each of these factors separately and asked whether they are associated with positive or negative fitness residual.

When we computed folding energies for segments of mRNA nucleotides on a sliding window along the variable region of each variant, we found that positive fitness residual variants demonstrated tighter secondary structures compared to negative variants along many different window positions (Figure 3B; Figure S2B for different window sizes). Strikingly, the maximum difference in folding energy is observed when the window's start position is at the beginning of the translated region of the ORF, excluding the upstream 5′ UTR (Figure 3C; effect size = 65.03%, Wilcoxon rank-sum, $p = 5.4 \times 10^{-28}$). Hence, these results, together with previous ones, reveal the dual role of mRNA folding: on one hand, loose mRNA structure at the RBS is associated with high expression level (Goodman et al., 2013), and on the other hand, utilization of a strong secondary structure at the 5′ end of the ORF can reduce per-protein costs.

It was previously suggested that elongating ribosomes in *E. coli* dwell longer on sequences that have high affinity to the aSD motif in the ribosome (Li et al., 2012). However, this observation has been recently questioned (Mohammad et al., 2016). We next examined the effects of Shine Dalgarno-mediated ribosomal pauses on fitness residuals. We calculated affinities to the aSD along the sequence of each variant, derived a ribosome speed estimation based on these affinities (see Experimental Procedures) and found that positive fitness residual variants are characterized by low ribosome speed early in the ORF (Figure 3D; effect size = 63.82%, Wilcoxon rank-sum test, $p = 6.3 \times 10^{-24}$).

We thus provide the first experimental evidence for a set of three gene architecture factors—codon decoding time, mRNA structure, and affinity to the anti-Shine Dalgarno motif—that could each implement 5′ ramping by slowing down ribosomes and, by that, allow cells to reduce the cost of gene expression at a given expression level.

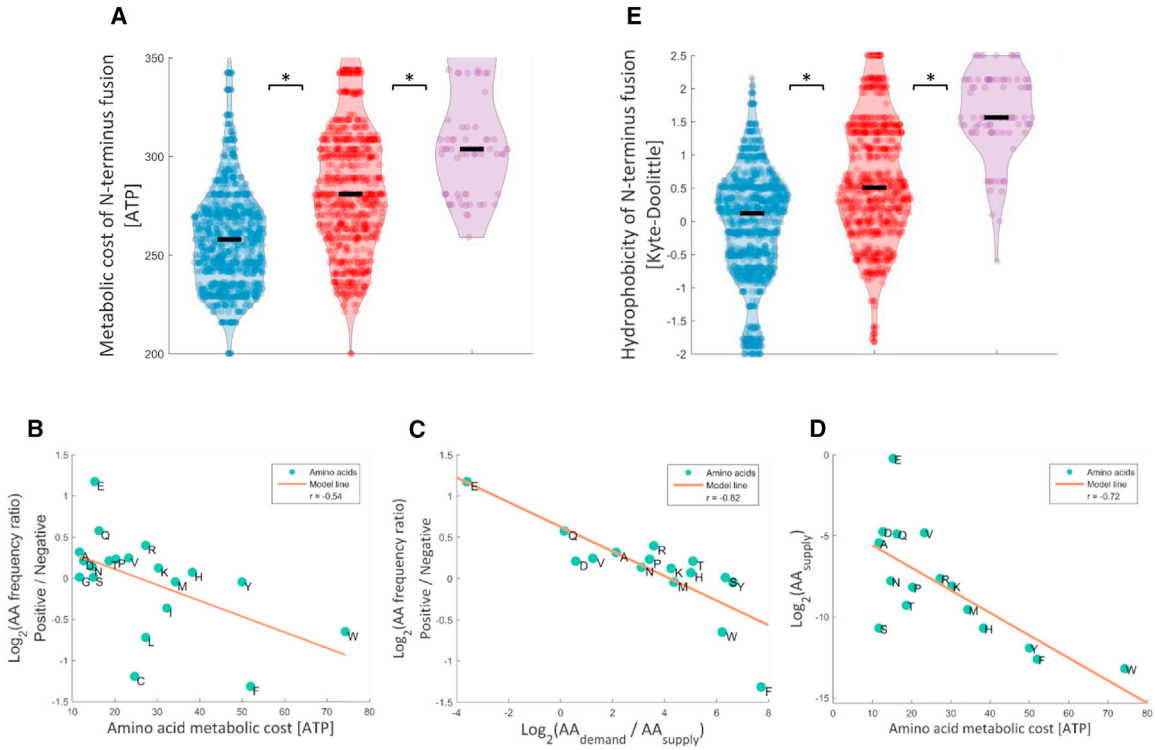


Figure 4. Usage of Expensive-to-Synthetize, Lowly Available, and Hydrophobic Amino Acids Decreases Fitness Residual

(A) N terminus amino acid fusions of negative variants are more expensive to synthesize compared to positive variants (effect size = 72.74%, Wilcoxon rank-sum, $p = 7.4 \times 10^{-62}$). Underachievers utilize even more expensive amino acids (effect size = 72.75%, Wilcoxon rank-sum, $p = 1.7 \times 10^{-11}$). See also Figures S1B and S2C.

(B) The frequency ratio of amino acids between positive and negative variants is negatively correlated with the energetic cost of amino acids (Pearson correlation, $r = -0.54$, $p = 0.01$). Each amino acid is marked according to its one-letter code.

(C) The frequency ratio of amino acids between positive and negative variants is negatively correlated with the demand/supply ratio of amino acids (Pearson correlation, $r = -0.82$, $p = 10^{-4}$). Demand comes from occupancy of ribosomes on each transcript (see Experimental Procedures), and supply is the cellular concentration of each amino acid (Bennett et al., 2009).

(D) Amino acid availability and energetic cost are correlated (Pearson correlation, $r = -0.72$, $p = 1.8 \times 10^{-3}$).

(E) N terminus amino acid fusions of negative variants are more hydrophobic than positive variants (effect size = 69.11%, Wilcoxon rank-sum, $p = 3.2 \times 10^{-44}$). N terminus fusion of underachievers are even more hydrophobic (effect size = 81.67%, Wilcoxon rank-sum, $p = 7.7 \times 10^{-21}$). See also Figures S1B and S2C.

Another means of reducing translation speed that was recently demonstrated (so far in yeast) is the incorporation of positively charged amino acids (Charneski and Hurst, 2013) or proline residues (Artieri and Fraser, 2014) in newly synthesized peptides. Yet, we did not detect any difference in frequency of such amino acids between the positive and negative fitness residual groups.

Amino Acid Synthesis Cost and Hydrophobicity Affect Cost of Gene Expression

So far we have examined features that are based on the nucleotide sequence and how it associates with fitness residual. Next, we aimed to explore the possibility that the amino acid composition of the N terminus fusion to the GFP associates with cellular fitness.

Amino acids differ by the metabolic costs associated with their biosynthesis—predominantly energy and reducing power determinants invested in their metabolic production (Akashi and Gojobori, 2002). We thus hypothesized that usage of energetically expensive amino acids may cause a heavier burden at a given expression level. Indeed, lower cost of the N terminus fusions were found to associate with positive fitness residual variants (Figure 4A; effect size = 72.74%, Wilcoxon rank-sum, $p = 7.4 \times 10^{-62}$). Here, as well, underachiever variants show more expensive amino acid usage compared to the negative group (Figure 4A; effect size = 72.75%, Wilcoxon rank-sum, $p = 1.7 \times 10^{-11}$).

We further examined the relation between fitness residual and amino acid energetic cost by calculating the frequency ratio of each individual amino acid between the positive and negative fitness residual groups (see Experimental Procedures). Remarkably, this frequency ratio was found to negatively correlate with the metabolic cost of each amino acid (Figure 4B; Pearson correlation, $r = -0.54$, $p = 0.01$). These observations suggest that expensive-to-synthesize amino acids burden cells during their costly production due to a potential feedback that increases their synthesis in response to consumption.

In addition to direct metabolic cost, the incorporation of amino acids that appear in low cellular concentrations could reduce

fitness indirectly as it might disturb the synthesis of other native proteins. We used ribosome profiling data (Li et al., 2012) to calculate amino acid demands and utilized previously measured cellular concentrations as amino acid supplies (Bennett et al., 2009) (see Experimental Procedures). Indeed, we found that amino acids with low demand-to-supply ratios are more prevalent in positive variants (Figure 4C; Pearson correlation, $r = -0.82$, $p = 10^{-4}$). This observation implies that utilization of amino acids that are less available to the cell (either due to high demand or low supply) increase expression cost and are associated with negative fitness residual variants. Since metabolic cost of amino acids and their cellular supplies are correlated (Figure 4D; Pearson correlation, $r = -0.72$, $p = 1.8 \times 10^{-3}$), we could not evaluate which mechanism—cost or availability—contributes more to fitness residual.

We next reasoned that an additional factor by which a protein could affect fitness is its toxicity, e.g., due to aggregation. As aggregation is driven by hydrophobic interactions, we turned to a conventional measure of amino acid hydrophobicity (Kyte and Doolittle, 1982) to examine whether it is predictive of fitness residuals. We found that positive fitness residual variants tended to have significantly less hydrophobic amino acids fused to the GFP (Figure 4E; effect size = 69.11%, Wilcoxon rank-sum, $p = 3.2 \times 10^{-44}$). Underachievers showed an even more pronounced effect (Figure 4E; effect size = 81.67%, Wilcoxon rank-sum, $p = 7.7 \times 10^{-21}$). This negative effect of hydrophobic residues in cytosolic proteins could indeed be derived from post-synthesis costs, but it could also reflect an equally interesting possibility: that aggregation-prone peptides reduce the functional level of the GFP (and similarly the fraction of the active form of native proteins). According to this possibility, aggregation is wasteful and must be compensated by further costly production to reach the required expression level of the protein.

We further found that the higher the GFP expression, the more beneficial it should be to utilize cheap or hydrophilic amino acids (Figure S2C).

All Sequence Parameters Contribute Independently to Fitness

We have revealed, so far, a set of mechanisms that affect expression costs and therefore cellular fitness. Although these mechanisms are different in their nature, it is possible that variants that score highly on one of these parameters tend to score highly on others. For example, anti-Shine Dalgarno affinity could correlate with the energy of the secondary structure of the mRNA, as both parameters are influenced by Guanine content. To check this possibility, we computed the correlation among the variants in the library between each pair of sequence parameters: codon decoding speed, mRNA secondary structure, anti-Shine Dalgarno affinity, hydrophobicity, and amino acid energy cost. Reassuringly, no strong correlation was found between any two parameters (Figure 5). Nonetheless, for feature pairs that did demonstrate non-negligible correlations (Pearson correlation, $r > 0.1$), we asked whether the signal of one feature is still observed while controlling for variation in the other. We found that each factor contributed directly to the signal, even upon controlling for other factors as potential confounders (see Figure S3).

Expression Costs Can be Minimized Even at Specified Amino Acid Sequences

Since maintaining a protein's function usually requires keeping its specific amino acid sequence, we next asked whether the mechanisms that we found here can reduce expression costs for a specified peptide sequence by using alternative nucleotide sequences. We defined "Δfitness-residual" as the difference between a variant's fitness residual and the average fitness residual of all library variants who share with that variant the same amino acid sequence. Then, we compared the various architectural features between variants with above-average Δfitness-residual to variants with below-average Δfitness-residual (see Experimental Procedures).

Figures 6A–6E depict, for each of the analyzed features, the difference in feature value between variants with above- or below-average Δfitness-residual. Interestingly, for each feature, the above- and below-average sub-groups had significantly different feature scores, reflecting the same trends as observed in all earlier analyses. For example, mRNA levels tend to be higher in the below-average sub-group in most of the 137 N terminus fusions (t test, p values for GFP mRNA levels = 6.2×10^{-3}, initiation rates = 7×10^{-9}, codon decoding speeds = 4.3×10^{-2}, mRNA folding = 3.5×10^{-16}, and aSD velocity = 7.6×10^{-7}). The conclusion from this analysis is that although amino acid features affect fitness residuals, the other features provide sufficient degrees of freedom to minimize costs even at a specified amino acid sequence.

A Regression Model Calculates Relative Contribution of Each Feature and Predicts Fitness Residual Scores

So far, we have examined fitness residual as a binary classification, namely categorizing variants with either positive or negative fitness residual. Complementing this binary analysis, in Figure S4A, we show that each feature correlates significantly with actual fitness residual values. We next aimed to predict actual fitness residual values of the library variants from their gene architecture features using a multiple linear regression model. We trained the model on a randomly chosen subset of 70% of the library variants, cross validated it on all other variants by comparing their predicted and observed fitness residual, and found a good correlation (see Experimental Procedures; Figure 7A; $r = 0.53$, $p < 10^{-200}$).

When the regression was performed on a scrambled library, which randomly links feature values and variants, the correlation between observed and predicted fitness residual was practically eliminated (Figure S4B; $r = 0.02$). We performed 10^5 such randomizations, and all of them demonstrated such extremely weak correlations. This negative control demonstrates that we obtained a genuine means to predict fitness residual values based on computable gene architecture parameters. We concluded that a gene architecture that utilizes more of the features that we discovered and that, to a greater extent, typically gives rise to higher fitness residuals as expression costs are further minimized.

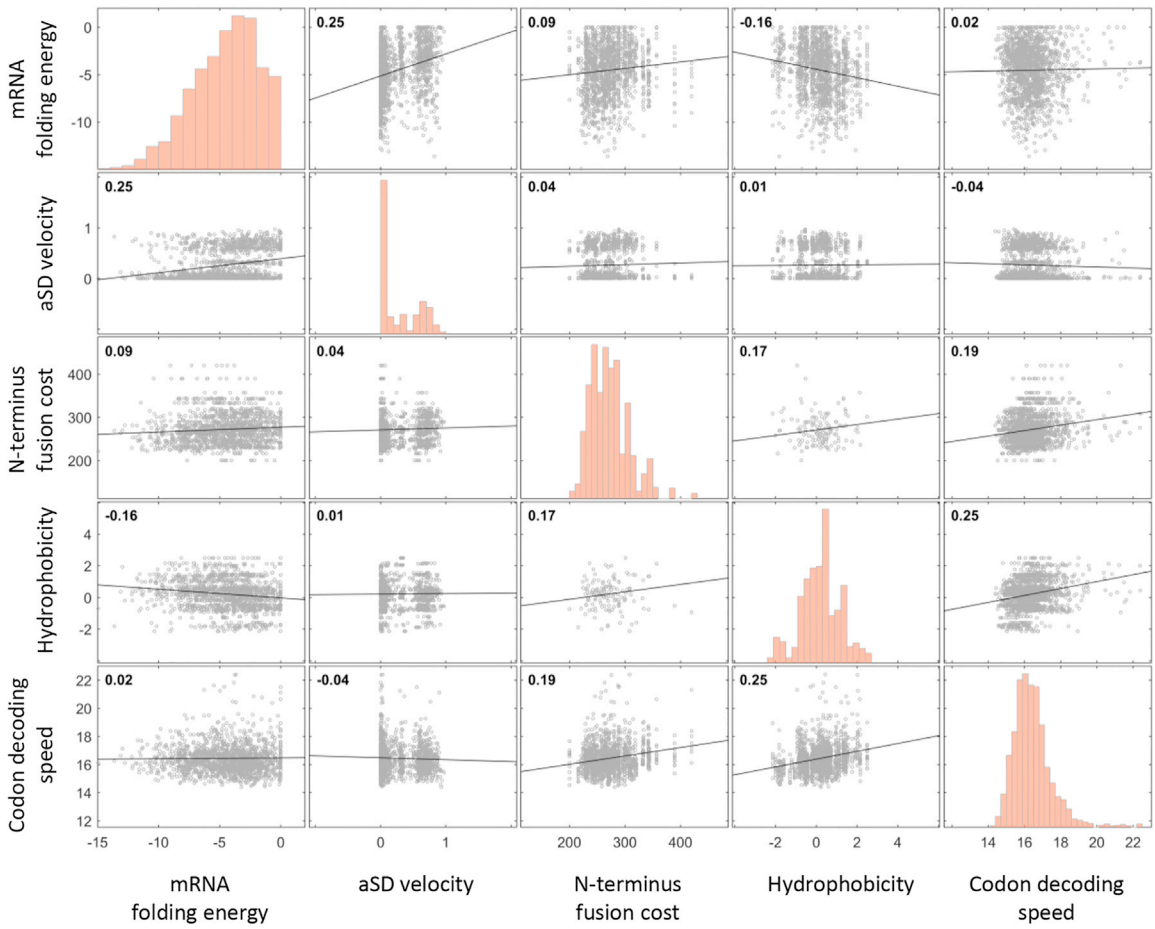


Figure 5. Each Feature Affects Fitness Residual Independently

Correlation plots of each feature pair show lack of correlation in most cases and only weak correlations in other cases. For feature pairs with Pearson correlation of $r > 0.1$, we compared the difference in one feature while controlling for the second and vice versa. See also Figure S3. Black lines are the regression curves between each feature pair. Number at upper-left corner is the Pearson correlation.

Additionally, this regression model allowed us to calculate the relative contribution of each feature by comparing the coefficients assigned by the regression model (Figure 7B). This analysis revealed that the features contributing to fitness residual the most are hydrophobicity and metabolic cost of the N terminus fusion, while codon decoding speed contributes the least. To avoid over-fitting of our model on the library data, we performed feature selection using the Lasso algorithm (see Experimental Procedures). This validation resulted in the exclusion of only codon decoding speed from the model, suggesting that its contribution to fitness residual is indeed lower compared to other features.

Highly Expressed Natural Bacterial Genes Have Evolved Gene Architectures that Minimize Their Production Costs

With these findings from the synthetic library, we next asked whether the mechanisms that we revealed as cost reducing were also utilized by natural selection to optimize *E. coli*'s native genes. We thus calculated each *E. coli* gene's score with respect to the relevant features and used the regression model to predict its fitness residual score (see Experimental Procedures and Table S4, related to Figure 7). Since a higher expression level results in higher expression cost, we next hypothesized that *E. coli* genes with higher expression levels are more likely to be endowed with cost-reducing architectures. Indeed, we found a significant correlation between predicted fitness residual of *E. coli* genes and their protein expression levels (Figure 7C; $r = 0.25$, $p = 2 \times 10^{-53}$), demonstrating a stronger selection for optimizing the 5′ gene architecture for highly expressed genes. We obtained similar results when predicting fitness residuals for all genes in the Gram positive *B. subtilis*, pointing to the generality of the model (Figure 7E; $r = 0.33$, $p = 10^{-93}$; see Experimental Procedures and Table S4, related to Figure 7).

Interestingly, the range of fitness residuals predicted by our model for the *E. coli* and *B. subtilis* genes was significantly larger than the range predicted by a mock regression model that was trained on randomly scrambled data of the synthetic library (see Experimental Procedures; Figures 7D and 7F; $p < 10^{-5}$).

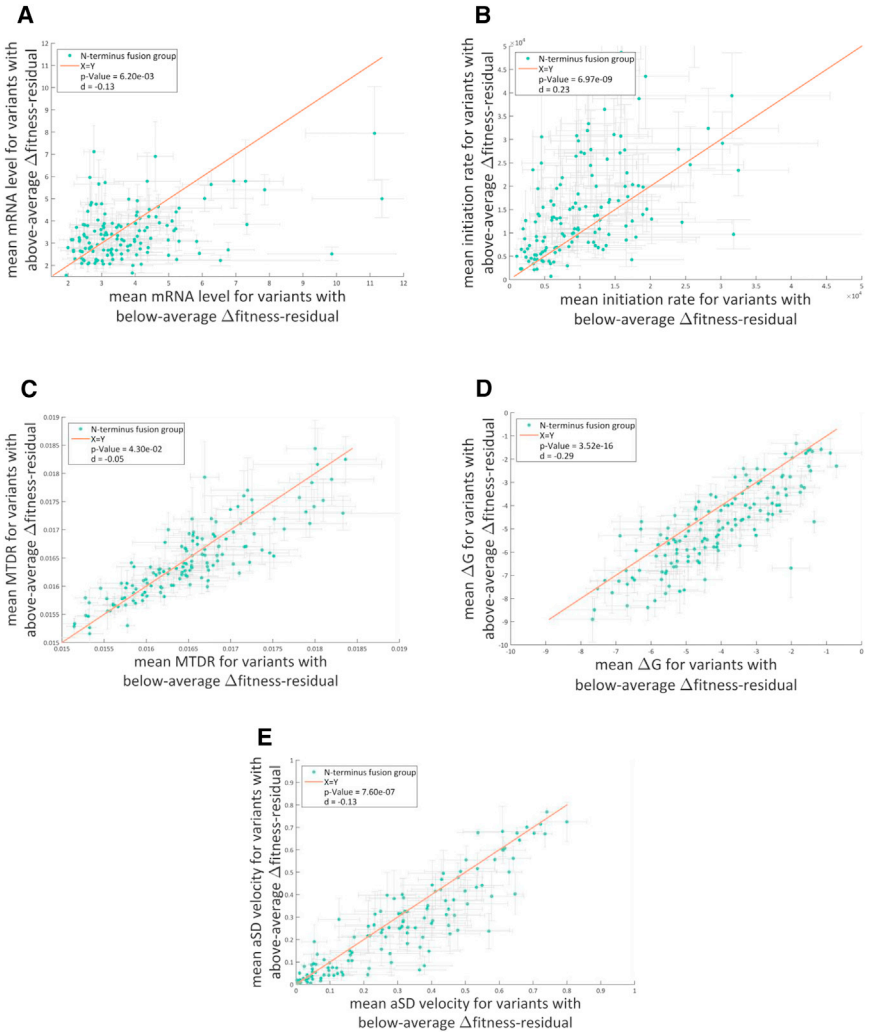


Figure 6. Variant with Same N Terminus Amino Acid Fusion Demonstrate a Range of Fitness Residuals

(A–E) Each dot represents one of the 137 N terminus fusions in the library. The x axis and the y axis represent the mean value of a feature for the variants with either below-average or above-average Δfitness-residual, respectively. The vertical and horizontal error bars represent standard errors for each of the axes. A statistical difference for deviance from the X = Y line was observed for all features, suggesting that even at a given amino acid sequence, these mechanisms affect fitness residual and can minimize expression costs (t test, p values: A, mRNA levels, 6.2×10^{-3}; B, initiation rates, 7×10^{-9}; C, codon decoding speeds, 4.3×10^{-2}; D, mRNA folding, 3.5×10^{-16}; and E, aSD velocity, 7.6×10^{-7}). d is Cohen's d that calculates the effect size.

This observation suggests that the model that we trained on the library data is able to expose the expression-cost optimality of natural 5′ gene architectures.

DISCUSSION

In this study, we found architectures and motifs that govern expression costs and reveal their function even beyond a direct effect on the process of expression. We show that regulating initiation and mRNA levels affects expression cost, as increasing the number of proteins that are produced per mRNA is associated with a positive fitness residual. This architecture could be beneficial because it reduces energy and resource consumption that are devoted to mRNA production. If cost reducing, why do genomes not further utilize the strategy of low transcription and mRNA abundance, combined with high translation initiation? One potential reason is that too low of mRNA levels might lead to increased expression noise (Taniguchi et al., 2010) or increased response time to an environmental signal (Gasch and Werner-Washburne, 2002). It is thus expected that natural genes would show a tradeoff between cost-reducing architectures and designs that satisfy other requirements, such as controlled noise and short response times.

The "translational ramp" theory predicted an effect of ribosome speed at early elongation on expression cost at a given expression level (Tuller et al., 2010a). The theory was never tested as such, since fitness reduction upon expression of an unneeded protein was not systematically measured for different gene sequences at various expression levels. We demonstrate here that slow translation speed at the 5′ end is beneficial in terms of reduced expression cost and increased cellular growth rate. We show that in addition to codon decoding times, there are at least two additional ramping means that are likely beneficial: occurrence of Shine-Dalgarno-like sequences and strong secondary structures.

Recent works showed that 5′ mRNA secondary structure governs expression level of transcripts in bacteria (Goodman et al., 2013; Kudla et al., 2009; Shah et al., 2013). Here, we observed that tight mRNA structures are enriched in positive variants. Consequently, it seems that mRNA structure plays a more complex role than previously thought. On one hand, 5′ mRNA structure, specifically upstream of the AUG start codon, regulates expression levels as it governs initiation rates (Goodman et al., 2013; Salis, 2011). On the other hand, tight structures at the beginning of the ORF, which were previously observed in *E. coli* genes (Tuller et al., 2011), are shown here to be beneficial in minimizing expression cost.

We revealed that the amino acid composition of a gene can also affect expression cost at a given expression level by showing that hydrophobic amino acids reduce fitness residual, perhaps due to their increased tendency to form toxic aggregates in the cytoplasm. In agreement with this, it was shown that mis-folded proteins impose growth reduction to yeast

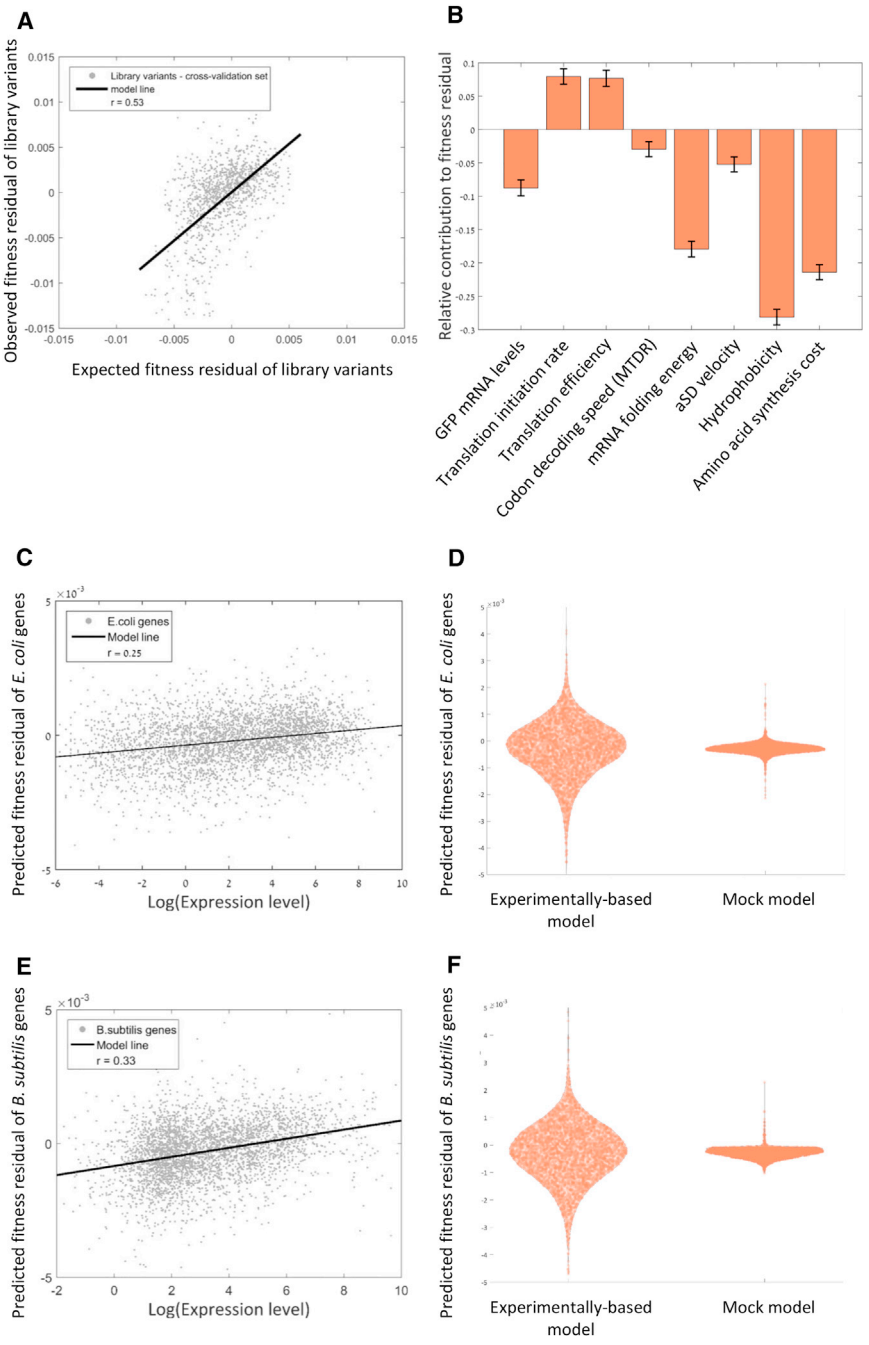


Figure 7. A Model that Predicts Fitness Residual Accurately Reveals that Fitness Residual of Natural Bacterial Genes Is Correlated with Their Expression Level

(A) A linear regression model based on all eight features predicts fitness residual accurately in a cross-validation test (Pearson correlation, r = 0.53, $p < 10^{-200}$). See also Figure S4.

(B) The weighted coefficients of each feature in the regression model demonstrating the relative contribution of each feature to fitness residual (p value for regression coefficient of mRNA level = 3.5×10^{-11}, initiation rate = 2.5×10^{-12}, $TE_{GFP\ protein/mRNA} = 2.7 \times 10^{-9}$, codon decoding speed = 8.7×10^{-3}, mRNA folding energy = 1.5×10^{-50}, aSD velocity = 8.7×10^{-3}, hydrophobicity < 10^{-200}, and amino acid synthesis cost = 5.4×10^{-80}). The sign of the contribution of each coefficient shows whether a feature is associated positively or negatively with fitness residuals. Error bars represent standard error of the coefficient estimation.

(C) Predicted fitness residuals of *E. coli* genes according to the regression model are correlated with their expression levels (Pearson correlation, r = 0.25, $p = 2 \times 10^{-53}$), suggesting that natural selection shapes 5′ gene architectures in order to minimize costs of gene expression.

(D) Distribution of fitness residual scores for *E. coli* genes as predicted by regression model that was trained on either experimental or mock data. The experimentally based model predicts a significant, higher range of fitness residuals ($p < 10^{-5}$), suggesting that the mechanisms that we elucidate with the synthetic library also apply on natural genes.

(E) Predicted fitness residuals of *B. subtilis* genes according to the regression model are correlated with their expression levels (Pearson correlation, r = 0.33, $p = 10^{-93}$), suggesting that our model also applies for other bacteria species.

(F) Same as (D), only for *B. subtilis* genes.

cells in a dosage-dependent manner (Geiler-Samerotte et al., 2011). It is interesting to postulate that hydrophobic residues that promote aggregation can reduce the portion of properly folded, functional protein. Such futile protein synthesis might need to be compensated for by further costly production in order to reach the needed functional level of a certain protein.

We further demonstrate that there are sufficient degrees of freedom for a gene to evolve a cost-reducing architecture, even when its amino acid sequence is constant. Hence, our study suggests design elements that could be utilized both for better heterologous gene expression and by natural selection for the optimization of natural genes.

As such, our observations are also relevant to biotechnology and synthetic biology. Many times in such non-natural systems, there is a need to express a foreign gene, whose expression could deprive resources from the hosting cell. Our results allow the design of an optimized nucleotide sequence version for heterologous expression that minimizes the cost of production and, by that, reduces the burden on the cell while not compromising expression level.

EXPERIMENTAL PROCEDURES

See Supplemental Experimental Procedures for full description.

Library Architecture

The synthetic library was provided to us by Goodman et al. (2013) and is fully described there. In short, each variant in the library harbors a unique 5′ gene architecture that is composed of a promoter, a ribosome binding site, and an N′ terminus amino acid fusion of 11 amino acids followed by a super-folder GFP (sfGFP) gene. The library as a whole includes two promoters with either high or low transcription rates; three synthetic RBSs with strong, medium, or low translation initiation rates, as well as 137 different genomic RBSs that were defined as the 20 bp upstream to the ORF of 137 *E. coli* genes; and, finally, 137 coding sequences (CDSs) consisting of the first 11 amino acids from the same genes. Each CDS appears in the library in 13 different nucleotide sequences representing alternative synonymous forms. All combinations amounted in 14,234 distinct library variants.

Competition Assay

Competition experiment was carried out by serial dilution. The library was grown on 1.2 mL of Lysogeny broth (LB) and 50 μg/mL kanamycin at 30°C, the exact same conditions that were used in Goodman et al. (2013) to measure GFP expression level. We grew six parallel, independent lineages, and each was diluted daily by a factor of 1:120 into fresh media (resulting in ~6.9 generations per dilution). This procedure was repeated for 12 days, and samples were taken from each lineage every 4 days (~27 generations), mixed with glycerol, and kept at −80°C.

Fitness and Fitness Residual Estimations

Fitness effect is derived from the following equation:

$$f(t) = f(anc) \cdot (1+s)^t \approx f(anc) \cdot e^{st}$$

where f is the variant frequency, t is the generation number, and s is the fitness effect.

To extract fitness effect, we took two independent approaches. First, we took the logarithm of the ratio between the frequency of a variant at a certain time point and its frequency at time zero. We then divided this value by the number of generations. This calculation was performed for both generation ~84 and generation ~56. See Supplemental Experimental Procedures for description of fitness calculation based on maximum likelihood. The two fitness-estimation methods were highly correlated (Figures S5A and S5B; $r = 0.99$, $p < 10^{-200}$) and resulted in the same conclusions throughout our analyses.

We then defined "fitness residual" of a variant as the difference between the observed fitness by FitSeq and the fitness predicted by a linear model given the variant's GFP expression level (see Supplemental Experimental Procedures for further details).

Model for Estimating Translation Velocity Based on Anti-Shine Dalgarno Affinity

The Shine-Dalgarno affinity was calculated identically to Li et al. (2012). In short, for each position, we calculated the affinity of 8–11 bp upstream of that position (the distance between the ribosome A site and the aSD site) to the anti-Shine Dalgarno motif. The free energy of interaction between the aSD motif and the mRNA sequence (ΔG) was calculated for all possible 10-mer sequences for that position using the RNA annealing function from the ViennaRNA package algorithm (Lorenz et al., 2011), and the highest affinity (lowest energy) score was used. We calculated the affinity for all positions for which the annealing with the aSD motif resides in the 11 amino acid fusion (positions 19–33) and then transformed all affinities of a given variable sequence to estimated ribosomal velocity, as follows.

We converted the ΔG estimates into the equilibrium constant of the interaction, K, which represents the equilibrium between association (k_f) and dissociation (k_b). The elongation velocity (v) as the ribosome moves from current site n to the n + 1 site is given by the harmonic mean of the dissociation reaction of site n and the association reaction of site $n + 1$:

$$\frac{1}{v_{n\rightarrow n+1}} = \frac{1}{k_{b_n}} + \frac{1}{k_{f_{n+1}}} \qquad \text{Equation 1}$$

$$v_{n\rightarrow n+1} = \frac{k_{b_n} k_{f_{n+1}}}{k_{b_n} + k_{f_{n+1}}} \qquad \text{Equation 2}$$

We further assume that the association reaction rate is not dependent on the sequence, therefore, for every n, $k_{f_n} = k_f$, and that differences in affinity thus only reflect differences in dissociation constant displayed by various sequences. We then get a term for the ribosomal velocity at a specific position by the anti-Shine Dalgarno affinity:

$$v_{n\rightarrow n+1} = \frac{k_f \cdot k_f K^{-1}}{k_f(1+K^{-1})} = k_f \frac{e^{\frac{\Delta G}{RT}}}{1+e^{\frac{\Delta G}{RT}}} \qquad \text{Equation 3}$$

To calculate the average ribosomal velocity across the entire N terminus fusion sequence of each library variant, we calculated the harmonic mean of the velocity values for all positions. See Supplemental Experimental Procedures for full description.

ACCESSION NUMBERS

The accession number for all sequencing data reported in this paper is SRA: SRP092267.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2016.11.007.

AUTHOR CONTRIBUTIONS

I.F., D.S., A.R., and Y.P. conceived and designed the study. I.F., D.S., A.R., F.L., L.Z., E.M., O.A., S.W., and S.F.L. acquired the data. I.F., D.S., A.R., and Y.P. analyzed and interpreted the data. I.F., D.S., and Y.P. wrote the manuscript.

ACKNOWLEDGMENTS

We are grateful to Daniel B. Goodman and George M. Church for providing us with the *E. coli* library and to Daniel B. Goodman for helpful discussions along the way. We are thankful to Tamir Tuller and Gilad Shaham for fruitful discussions about the MTDR concept. We thank Naama Barkai, Maya Schuldiner, Moshe Oren, Tal Galili, Tslil Ast, Avihu Yona, and Hila Gingold for helpful discussions. Our gratitude goes to Shlomit Gilad and Sima Benjamin from the Nancy & Stephen Grand Israel National Center for Personalized Medicine (G-INCPM) for assistance with high-throughput data. I.F. thanks the Azrieli Foundation for the Azrieli Ph.D. Fellowship award. S.F.L. is supported by The Louis and Beatrice Laufer Center and NIH grants R01 HG008354 and U01 HL127522. This study was supported by the Minerva Foundation, which funded the "Minerva Center for Live Emulation of Evolution in the Lab" and a Minerva grant to Y.P.

Received: July 29, 2016
Revised: October 10, 2016
Accepted: November 1, 2016
Published: December 15, 2016

REFERENCES

Akashi, H., and Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc. Natl. Acad. Sci. USA *99*, 3695–3700.

Artieri, C.G., and Fraser, H.B. (2014). Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. Genome Res. *24*, 2011–2021.

Bennett, B.D., Kimball, E.H., Gao, M., Osterhout, R., Van Dien, S.J., and Rabinowitz, J.D. (2009). Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. Nat. Chem. Biol. *5*, 593–599.

Bentley, W.E., Mirjalili, N., Andersen, D.C., Davis, R.H., and Kompala, D.S. (1990). Plasmid-encoded protein: the principal factor in the "metabolic burden" associated with recombinant bacteria. Biotechnol. Bioeng. *35*, 668–681.

Bienick, M.S., Young, K.W., Klesmith, J.R., Detwiler, E.E., Tomek, K.J., and Whitehead, T.A. (2014). The interrelationship between promoter strength, gene expression, and growth rate. PLoS ONE *9*, e109105.

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

Charneski, C.A., and Hurst, L.D. (2014). Positive charge loading at protein termini is due to membrane protein topology, not a translational ramp. Mol. Biol. Evol. *31*, 70–84.

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

Dekel, E., and Alon, U. (2005). Optimality and evolutionary tuning of the expression level of a protein. Nature *436*, 588–592.

Dong, H., Nilsson, L., and Kurland, C.G. (1995). Gratuitous overexpression of genes in Escherichia coli leads to growth inhibition and ribosome destruction. J. Bacteriol. *177*, 1497–1504.

Emilsson, V., and Kurland, C.G. (1990). Growth rate dependence of transfer RNA abundance in Escherichia coli. EMBO J. *9*, 4359–4366.

Gasch, A.P., and Werner-Washburne, M. (2002). The genomics of yeast responses to environmental stress and starvation. Funct. Integr. Genomics *2*, 181–192.

Geiler-Samerotte, K.A., Dion, M.F., Budnik, B.A., Wang, S.M., Hartl, D.L., and Drummond, D.A. (2011). Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. Proc. Natl. Acad. Sci. USA *108*, 680–685.

Gingold, H., and Pilpel, Y. (2011). Determinants of translation efficiency and accuracy. Mol. Syst. Biol. *7*, 481.

Glick, B.R. (1995). Metabolic load and heterologous gene expression. Biotechnol. Adv. *13*, 247–261.

Goodarzi, H., Nguyen, H.C.B., Zhang, S., Dill, B.D., Molina, H., and Tavazoie, S.F. (2016). Modulated expression of specific tRNAs drives gene expression and cancer progression. Cell *165*, 1416–1427.

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

Heyer, E.E., and Moore, M.J. (2016). Redefining the translational status of 80S monosomes. Cell *164*, 757–769.

Higgs, P.G., and Ran, W. (2008). Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol. Biol. Evol. *25*, 2279–2291.

Ibarra, R.U., Edwards, J.S., and Palsson, B.O. (2002). Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature *420*, 186–189.

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

Kafri, M., Metzl-Raz, E., Jona, G., and Barkai, N. (2016). The cost of protein production. Cell Rep. *14*, 22–31.

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

Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.

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

Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.

Marr, A.G. (1991). Growth rate of Escherichia coli. Microbiol. Rev. *55*, 316–333.

Mohammad, F., Woolstenhulme, C.J., Green, R., and Buskirk, A.R. (2016). Clarifying the translational pausing landscape in bacteria by ribosome profiling. Cell Rep. *14*, 686–694.

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

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

Rang, C., Galen, J.E., Kaper, J.B., and Chao, L. (2003). Fitness cost of the green fluorescent protein in gastrointestinal bacteria. Can. J. Microbiol. *49*, 531–537.

Salis, H.M. (2011). The ribosome binding site calculator. Methods Enzymol. *498*, 19–42.

Scott, M., Gunderson, C.W., Mateescu, E.M., Zhang, Z., and Hwa, T. (2010). Interdependence of cell growth and gene expression: origins and consequences. Science *330*, 1099–1102.

Shah, P., and Gilchrist, M.A. (2011). Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. Proc. Natl. Acad. Sci. USA *108*, 10231–10236.

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

Sharp, P.M., Tuohy, T.M., and Mosurski, K.R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. *14*, 5125–5143.

Subramaniam, A.R., Pan, T., and Cluzel, P. (2013). Environmental perturbations lift the degeneracy of the genetic code to regulate protein levels in bacteria. Proc. Natl. Acad. Sci. USA *110*, 2419–2424.

Taniguchi, Y., Choi, P.J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X.S. (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science *329*, 533–538.

Tholstrup, J., Oddershede, L.B., and Sørensen, M.A. (2012). mRNA pseudoknot structures can act as ribosomal roadblocks. Nucleic Acids Res. *40*, 303–313.

Tuller, T., and Zur, H. (2015). Multiple roles of the coding sequence 5′ end in gene expression regulation. Nucleic Acids Res. *43*, 13–28.

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

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

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

Vind, J., Sørensen, M.A., Rasmussen, M.D., and Pedersen, S. (1993). Synthesis of proteins in Escherichia coli is limited by the concentration of free ribosomes. Expression from reporter genes does not always reflect functional mRNA levels. J. Mol. Biol. *231*, 678–688.

Wagner, A. (2005). Energy constraints on the evolution of gene expression. Mol. Biol. Evol. *22*, 1365–1374.

Weinberg, D.E., Shah, P., Eichhorn, S.W., Hussmann, J.A., Plotkin, J.B., and Bartel, D.P. (2016). Improved ribosome footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. Cell Rep. *14*, 1787–1799.

Wen, J.-D., Lancaster, L., Hodges, C., Zeri, A.-C., Yoshimura, S.H., Noller, H.F., Bustamante, C., and Tinoco, I. (2008). Following translation by single ribosomes one codon at a time. Nature *452*, 598–603.

Yona, A.H., Bloom-Ackermann, Z., Frumkin, I., Hanson-Smith, V., Charpak-Amikam, Y., Feng, Q., Boeke, J.D., Dahan, O., and Pilpel, Y. (2013). tRNA genes rapidly change in evolution to meet novel translational demands. eLife *2*, e01339.

Optimizing gene expression by adapting splicing

Idan Frumkin[1,*,#], Ido Yofe[1,*], Raz Bar-Ziv[1], Yoav Voichek[1] & Yitzhak Pilpel[1,#]

1- Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel.

*- Co-first authors

#-correspondence: frumkin.idan@gmail.com; pilpel@weizmann.ac.il;

Abstract

Can splicing be used by cells to adapt to new environmental challenges? While various adaptation mechanisms for regulating gene expression have been revealed for transcription and translation, the role of splicing and how it evolves to optimize gene-expression patterns has not been thoroughly investigated. To tackle this question, we employed a lab-evolution experimental approach that challenged yeast cells to increase expression levels of a gene that carries an inefficiently-spliced intron. We followed the evolution of multiple lines and found independent routes by which cells adapted. Surprisingly, we did not observe an intron loss event, a mechanism believed to be common in intron evolution. Instead, we identified mutations in *cis* that improved the intron's splicing efficiency and increased the overall expression level of the entire gene. One of these *cis*-acting mutations occurred in an adjacent exon and hampered the functionality of the gene that was not under selection - demonstrating that adaptation of splicing efficiency may sometimes come at the expense of protein activity. Additionally, we observed adaptations in *trans*, which increased the cellular availability of the splicing machinery. These adaptations were achieved either by elevated expression levels of the splicing apparatus or, unexpectedly, by reduced expression levels of other intron-containing genes that are the natural consumers of this process. Ultimately, our work reveals novel molecular means by which the splicing machinery is changed by natural selection to optimize gene-expression patterns of cells.

Introduction

Changing environments can force cells to change their gene-expression programs, to better accommodate their surroundings. Throughout evolution, cells acquired regulatory mechanisms to tune gene expression, which have been the subject of intensive investigations – focusing mainly on transcription and translation. For example, when cells are challenged to increase protein expression levels, the DNA sequence of genes can change so as to increase transcription[1,2], support more efficient mRNA translation[3,4], or result in greater mRNA transcript stability[5,6]. Additionally, the transcription and translation machineries themselves have been shown to adapt to environmental challenges by altering the cellular pools of transcription factors[7] or tRNAs[8,9].

In evolving expression programs, adaptation often occurs either directly on the genes under pressure ("evolution in *cis*")[10], or indirectly, *e.g.* on the expression machineries, mostly transcription and translation ("evolution in *trans*")[11,12]. These two routes of evolution are profoundly different, as the first (*cis*) provides a localized solution that in principle can affect only a certain gene, while the later (*trans*) could be the method of choice if a coordinated change in many genes is needed.

Surprisingly, although the process of splicing is central to the maturation and regulation of mRNAs in eukaryotes[13–17], its role in adapting to novel demands on gene expression has not been thoroughly investigated. During mRNA splicing, precursor mRNAs are processed to remove introns while fusing exons together to create the mature transcript. This process provides an evolutionary means to diversify the proteome towards phenotypic novelty, as the choice of intron to be excluded, as well as the exons which are found in the mature transcript, can both be regulated based on the cell's needs[15,18,19]. One aspect of splicing evolution that has been extensively studied is gain and loss of intronic DNA, for which several molecular models have been proposed, mainly Reverse-Transcription and recombination-mediated intron loss, intron transposition and also exonization and intronization via mutations[20–23]. While intron loss and gain have been demonstrated experimentally[24,25], other forms of splicing evolution, such as alterations in splicing efficiency under changing conditions, have not.

Here, we set out to reveal whether introns or the splicing apparatus can evolve so as to alter the expression levels of genes in a timely and adaptive manner, and ask whether and how splicing evolves in *cis* and in *trans* to regulate gene expression. To this end, we generated a reporter construct in yeast cells that could simultaneously be read out and be selected for splicing efficiency. Namely, we introduced an inefficiently-spliced intron to a reporter gene that was fused to an antibiotic resistance gene. Using this approach, we could carry out a lab-evolution experimental setup to study the adaptation of splicing in the presence of the corresponding antibiotics.

Our results demonstrate two alternative adaptive routes for this evolutionary challenge. First, *cis*-acting solutions, in the form of adaptive mutations, occurred in the intron itself, but also surprisingly in an up-stream exon. These mutations resulted in increased splicing efficiency and higher expression levels of the antibiotic resistance gene. Remarkably, in some evolutionary lines there were no *cis* mutations, but rather *trans*-acting adaptations that have increased cellular availability of the splicing machinery - either by increased expression levels of the machinery's genes themselves, or surprisingly, by decreasing the expression levels of other intron-containing genes. Thus, this works unravels different layers at which the splicing machinery can be adapted to alter gene expression.

Results

Low splicing efficiency leads to stressed cells under restrictive conditions

We hypothesized that tuning splicing of genes could serve as a means to optimize their expression levels. To test this hypothesis, we used the yeast *Saccharomyces cerevisiae* in which ~30% of the transcriptome must be spliced, at a range of splicing efficiencies[17,26], to form mature mRNAs[27]. We built a synthetic gene construct that consists of two fused domains: A fluorescent reporter (YFP), which includes two alternative natural introns - with either high or low splicing efficiency - near the YFP's fluorescence site[26], fused to an antibiotics resistance gene (Kanamycin resistance gene). Specifically, we created three strains: (i) WT YFP strain without an intron; (ii) "Splicing^{High}" in which the YFP harbors the natural intron of *OSH7* and was previously reported to have high splicing efficiency within this YFP context[26]; and (iii) "Splicing^{Low}" in which the natural intron of *RPS26B*, with a low splicing efficiency[26], was inserted in the same location (Figure 1A).

We first hypothesized that cellular growth of each strain in the presence of the antibiotics, geneticin (G418), will associate with the YFP-Kan expression levels. We followed the growth of the three strains in the presence of the antibiotics and found that the WT strain had the highest fitness, Splicing^{High} grew slower, and Splicing^{Low} demonstrated a severe growth defect compared to the two other strains (Figure 1B+C). We then measured florescence intensity of the YFP-Kan reporter in the presence of the drug. In line with the growth measurements, we observed that WT cells demonstrated the highest fluorescence levels, followed by Splicing^{High}, and with Splicing^{Low} cells showing the lowest YFP-Kan levels (Figure 1D). These results demonstrate that the inefficiently-spliced intron in Splicing^{Low} reduces cellular levels of YFP-Kan and hence lead to a reduced fitness.

Since YFP-Kan expression level in Splicing^{Low} were significantly lower compared to the other strains, we hypothesized that Splicing^{Low} cells did not reach the needed concentration to sufficiently neutralize the antibiotics, and hence resulted in stressed

cells. To test this hypothesis, we performed mRNA sequencing of exponentially growing WT and Splicing^{Low} cells in an antibiotics containing medium, and analyzed their transcriptome profiles. Indeed, we observed that ribosomal genes were down-regulated in Splicing^{Low} compared to the control strain – a clear signature of stressed cells[28] (Figure 1E). Notably, the reduction in ribosomal expression levels (~8%) we observed here due to growth rate differences between WT and Splicing^{Low} cells is accurately predicted by a recent study, which calculated the linear correlation between growth rate and ribosomal expression levels in yeast cells[29]. In parallel, stress-related genes[30] were up-regulated in the Splicing^{Low} compared to the control strain (Figure 1E). We thus concluded that the general stress response was activated in Splicing^{Low} cells.

Rapid evolutionary adaptation increases expression level of the resistance gene

Our experimental system mimics an evolutionary scenario in which there is an immediate and continuous selection pressure to up-regulate the expression level of a gene of interest. How would the system evolve to better resist the antibiotics? Possible means to adapt include mutations in the gene's promoter to increase transcription, mutations that increase translation initiation, or mutations inside the gene itself that increase the functional efficiency of the protein. Additionally, the splicing machinery may also take part in adaptation of gene expression levels. To find which evolutionary track would be used by cells, we evolved the three strains by daily serial dilution on a medium supplemented with G418 for ~560 generations, in four independent cultures for each strain (Figure 2A). Interestingly, only the cultures of Splicing^{Low} cells demonstrated a significant improvement in fitness at the end of the experiment (Figure 2B+C). This observation implies that only Splicing^{Low} experienced a sufficiently strong selective pressure to adapt to the presence of the antibiotics in the medium, in contrast to the WT and Splicing^{High} strains which originally had much higher levels of the resistance gene.

Consistent with the fitness measurements, YFP measurements of the evolved cultures showed that expression levels of the resistance-YFP fusion gene increased in all four evolved cultures of Splicing^{Low} compared to the ancestral strain (Figure 2D). Conversely,

the increase in YFP-Kan expression levels in the evolved WT populations was smaller, and only one culture of the evolved Splicing^{High} cells demonstrated strong elevation of the YFP-Kan levels (Figure 2D). These results further support that Splicing^{Low} cells experienced the strongest selective pressure to adapt rapidly to the presence of the antibiotics in our experimental setup, and that they achieved this goal by increasing the levels of the resistance gene. We next moved to reveal the molecular mechanisms underlying this evolutionary process.

Adaptation in *cis* and *trans* leads to increased splicing efficiency

We hypothesized that improving the low splicing efficiency of the intron in Splicing^{Low} could be exploited by natural selection as an adaptation mechanism towards increasing the resistance gene levels. We therefore sequenced the YFP-Kan locus in 16 randomly chosen colonies from two evolved populations (termed here population A and population B) of Splicing^{Low}. Interestingly, we found that the colonies were split into two types – either with or without a mutation in the YFP-Kan locus. In population A, we found that the same mutation occurred in four out of eight colonies, changing adenine to cytosine inside the intron, 97 nucleotides up-stream to its 3' end (Figure 3A). In population B, we identified an exonic non-synonymous mutation that changed a valine at position 61 of the YFP protein into alanine (a thymine to cytosine 14 nucleotides up-stream of the intron) in three out of eight colonies. In the five other colonies from this population there were no mutations in the YFP-Kan locus.

Notably, none of the colonies demonstrated a mutation in the construct's promoter, terminator or in the sequence of the Kan resistance gene itself. These results propose that different mutations in the intron, or its vicinity, were adaptive and might affect splicing efficiency of the intron. Surprisingly, the observed mutations did not occur in the 5' donor, 3' acceptor, nor in the intron branch point – suggesting that other position of the intron can also be selected in evolution increase fitness by affecting splicing. While the intron- and exon-mutated colonies represent an evolutionary adaptation in *cis*, the

colonies that showed no mutation in the entire gene construct potentially found adaptive solutions in *trans* that may have occurred elsewhere in the genome.

We randomly chose six colonies: four colonies with a *cis* mutation and two colonies that showed no mutations in *cis*, for which we reasoned that such colonies may have adapted in *trans*. We termed these colonies according to the evolution lines from which they were derived: A-cis1, A-cis2, B-cis1, B-cis2, A-trans and B-trans. We followed the growth of these evolved colonies in the presence of G418 and found, as expected, that all grew faster than the Splicing^{Low} ancestor (Figure 3B). We then performed RNA-seq and transcriptome analysis of all colonies, which revealed relaxation of the stress response that was featured in the ancestor. Namely, the general stress response genes were reduced and ribosomal proteins were up- regulated in five evolved colonies (Figure 3C and Supplementary Figure 1). These observations suggest that the cells indeed adapted to the presence of the antibiotics in the environment and that the stress experienced by them was partially alleviated.

We next hypothesized that cellular fitness might correlate with mRNA levels of the YFP-Kan construct because increased transcript levels should result in higher concentrations of the Kan protein. Indeed, maximal growth rates of the control and Splicing^{Low} ancestors and for the six evolved colonies correlate with mRNA levels of the YFP-Kan construct, as deduced from the RNA-seq (Figure 3D) – supporting our conclusion that adaptation was based on increasing expression levels of the YFP-Kan gene. Since the observed *cis* mutations occurred at the vicinity of the intron, we hypothesized that they increased splicing efficiency of the YFP-Kan transcript. To test this possibility, we performed, for both *cis*- and *trans*-evolved colonies, a splicing efficiency assay with qPCR - targeting the un-spliced and spliced transcript versions. Interestingly, the ratio of spliced to un-spliced transcripts was higher in all evolved colonies compared to the Splicing^{Low} ancestor, suggesting that at least some of the mRNA level increase we observed in the evolved colonies results from increased splicing efficiency (Figure 3E).

To prove that adaptation of the colonies actually led to higher protein levels of the resistance gene, we measured fluorescence intensity using flow cytometry. We found that the two *cis*-colonies from population A (A-cis1 & A-cis2) and the two *trans*-colonies (A-trans & B-trans) showed higher YFP-Kan levels compared to the ancestor. However, the two *cis*-colonies from population B (B-cis1 & B-cis2) demonstrated decreased fluorescence intensity values (Figure 3F). These observations suggest that the non-synonymous, exon mutation reduced the fluorescence-per-protein value of the YFP-Kan construct in these colonies. Indeed, this position corresponds to a position that was recently reported to reduce florescence when mutated in the highly similar GFP[31]. Because YFP functionality was not selected for or against in our setup, it was free to mutate as long as it helps achieve a higher expression level of the entire construct by increasing the intron's splicing efficiency. It thus seems that modular domain-architecture of a protein may increase its evolvability under relevant conditions as it allows the optimization of each domain in isolation from the other.

It is possible that additional beneficial mutations exist in the genome of the *cis*-evolved colonies, which account for the phenotypes we observed. To directly assess the effects of the *cis* mutations, we generated two rescue strains, termed rescue-A and rescue-B, in which these *cis*-acting mutations were introduced individually to the ancestral Splicing^{Low} background. Notably, the two rescue strains grew better than Splicing^{Low} cells in the presence of the antibiotics (Figure 4A), though not as good as the wild-type, and the stress experienced by the Splicing^{Low} cells was relieved upon insertion of each individual *cis* mutation (Figure 4B). Finally, we measured splicing efficiencies and fluorescence intensity levels for both rescue strains, and found that they resembled the results of the evolved single colonies (Figure 4C-D, in comparison to Figure 3E-F). These observations strengthen our conclusions that the *cis*-acting mutations are sufficient to elevate YFP-Kan levels through an increased splicing efficiency, yet the non-synonymous mutation of population D also hampers the function of the YFP domain and reduces its florescence-per-protein ratio.

Our results thus far provide direct evidence that intron splicing takes part in the adaptation and optimization of gene expression patterns to environmental needs. Although intron sequences are much less conserved compared to exons, and are believed to be less functional, we demonstrate that their sequence can be used by natural selection as a molecular mechanism to regulate splicing efficiency and adjust gene expression patterns.

Increasing cellular availability of the splicing machinery can be adaptive

We finally aimed to decipher the mechanism behind the increased YFP-Kan levels in the *trans*-evolved colonies that showed no mutations in *cis*, i.e. within the reporter gene or in its vicinity. We reasoned that elevating availability of the splicing machinery as a global resource could be a means to increase splicing efficiency of the YFP-Kan transcript, and thus could be used as an adaptive mechanism to the antibiotics challenge. Increased splicing-availability could be achieved by increasing the expression of the splicing machinery genes. In addition, as with other cellular machineries whose functioning depends on supply-to-demand economy[4,8,32,33] reducing expression levels of the intron-containing genes, namely the "demand", could increase the availability of the machinery towards the intron under selection here.

To test if any of these evolutionary routes were indeed taken by the evolved cells, we calculated the expression level ratio of genes between the evolved colonies and their ancestor. In colony A-trans, we observed that while the average expression-ratio of the splicing machinery genes (the "supply") increased, that of the non-ribosomal intron-containing genes (the "demand") decreased (Figure 5A). This observation suggests that indeed the cellular availability of the splicing machinery was elevated in this evolved colony, which might have allowed for the observed increased splicing efficiency of the YFP-Kan gene. Next, we hypothesized that the *cis*-evolving colonies may have also adapted in *trans* and used this adaptation mechanism as well. Indeed, in all other evolved colonies we observed a similar trend, in which the overall supply-to-demand measurement of the splicing machinery was increased (Figure 5B). Notably, in some

colonies this phenotype was achieved by only increasing expression levels of splicing genes or by only reducing levels of the intron-containing genes (Supplementary Figure 2). Importantly, the two rescue strains, which did not evolve and only harbor our artificially introduced *cis*-acting mutation, did not show any change in splicing availability (Figure 5B), strongly supporting our conclusion that this phenotype was achieved by further adaptation of the cells during our lab-evolution experiment. Thus, we concluded that both *cis* and *trans* adaptation routes can co-occur in the same genome towards optimization of its gene expression patterns.

Discussion

Here we study the role of the splicing machinery in optimization of gene expression programs by placing selective pressure on cells to improve the splicing efficiency of a specific gene. Our results provide molecular evidence for the relevance of splicing as another instrument in the cellular toolbox towards adjusting its gene expression patterns. To the best of our knowledge, we demonstrate the first experimental evidence of splicing efficiency adaptation, confirming that this adaptation can occur in *cis* and *trans* similarly to adaptations of other means of gene regulation.

Two potential solutions to the burden we imposed on our ancestor lines were, surprisingly, not realized during our lab-evolution. First, considering previous studies of splicing evolution, one could have expected the intron to be lost by a genomic deletion or reverse transcription[34,35]. Such a solution could have been an ideal evolutionary adaptation to alleviate the burden, as we show that the intron-less strain has the highest fitness. The fact that we did not observe an intron-loss event suggests that nucleotide substitutions were more accessible solutions in this case, in agreement with previous evidence in yeast that nucleotide mutations are much more prevalent than deletion events, at a ratio of 33:1[36]. Another surprise was that the mutations we observed did not occur within any of the intron's three functional sites: the 5' donor, 3' acceptor, or the branch point of splicing. Surprisingly, one mutation was actually detected, and was verified here to affect splicing, in a region of the intron not known to exert a major effect on splicing, and another splicing-improving mutation happened in the up-stream exon - suggesting that various positions in the intron and its proximity may facilitate splicing rate and take part in the evolution of this process.

Notably, the fluorescence intensity per protein molecule of the YFP domain was decreased due to the non-synonymous mutation in the YFP first exon, suggesting that under certain evolutionary constrains selection may hamper superfluous functions of certain protein domains so as to increase availability of the entire protein. Why then, would the mutations we observed increase splicing efficiency? Past evidence showed that

mRNA secondary structures at the intron's edges influence splicing efficiency[26,37–39]. It is possible then that the mutations we observed in this study somehow open the structure of the intron under selection and make it more accessible for the splicing machinery.

Adaptive changes also occurred in *trans* to the YFP-Kan locus and increased availability of the splicing machinery. Recently, the competition of pre-mRNAs for the splicing machinery was shown to affect cellular function, as splicing efficiency of multiple introns was influenced by changes in the composition of the transcript pool[40]. While this mechanism was elegantly suggested to maintain the separation between meiotic and vegetative gene expression states, it is also possible that it can be utilized as an adaptive route available for cells to optimize expression levels of genes. More broadly, it has been shown[41] that yeast species that have a high content of intron-containing genes have adapted the codon usage of their splicing machinery genes more vigorously to their cellular tRNA pools compared to other species that have a lower number of introns in the genome. This evolutionary trend indeed shows that the splicing machinery adaptively responds to meet its own evolutionary demand.

Our findings demonstrate how availability of the splicing apparatus may have been beneficially increased both by elevating the expression level of the machinery and/or by reducing other intron-containing genes that compete with the antibiotic resistance un-spliced RNA for the spliceosome. Thus, increase in supply-to-demand ratio, analogous to the case in translation systems[8,42], appears to have evolved in this case.

Interestingly, we found that different adaptive means co-occurred in the evolved populations – independently in different cells or even simultaneously in the same genome. In particular, we saw that evolutionary lines that adapted in *cis* appear to also have had adaptations that are not encoded in the evolving gene, hence pointing to changes that must have occurred in *trans*. Further investigations will reveal which of these solutions, *cis* or *trans*, proves to be more evolutionarily stable - to fully reveal the dynamics of splicing adaptation when cells optimize their gene expression.

Materials and Methods

Yeast strains and plasmids

All *S. cerevisiae* strains in this study have the following genetic background: his3Δ1::TEF2-mCherry::URA3::RPS28Ap-YiFP-KAN::NAT; canΔ1::STE2pr-Sp_his5; lypΔ1::STE3pr-LEU2; leu2Δ0; ura3Δ0;

Strains of Y-intron-FP were taken from Yofe *et al*[26] and were introduced with a Kan resistance gene fused 3' terminally to the YFP. To reconstitute the mutations discovered after lab evolution (rescue strains), we amplified cassettes of Y-i_{mut}-FP-KAN and transformed these into the ancestor WT strain, selecting with KAN. Notably, all strains also carry an mCherry-fluorescent protein driven by an independent *TEF2* promoter that was used to normalize cell-to-cell variability for the YFP-Kan expression levels.

Media

Cultures were grown at 30°C in rich medium (1% bacto-yeast extract, 2% bacto-peptone and 2% dextrose [YPD]). Throughout all experiments, G418 was supplemented to the medium at a concentration of 3mg/ml, which is 10fold higher than the standard.

Evolution experiments

Lab-evolution experiments were carried out by daily serial dilution for 80 days. Cells were grown on 1.2ml of YPD+G418 at 30°C until reaching stationary phase and then diluted by a factor of 1:120 into fresh media (~7 generations per dilution, total of ~560 generations).

Liquid growth measurements

The cultures were grown at YPD+G418, and optical density (600nm) measurements were taken at 30min intervals. Growth comparisons were performed using 96-well plates, and the growth curve for each strain was obtained by averaging at least 15 wells.

FACS measurements of YFP-Kan levels

Cells were grown in YPD+G418 at 30°C until they reached the logarithmic growth phase at an optical density of ~0.4. Then, YFP and mCherry levels were measured for ~50,000 cells for each culture with flow cytometry. Gating was performed according to side and forward scatters, and YFP levels were normalized with the mCherry signal for each cell individually.

Quantitative PCR measurements of splicing efficiency

Cultures were grown in YPD+G418 at 30°C until cells reached the logarithmic growth phase at an optical density of ~0.4. Then, RNA was extracted using MasterPure kit (Epicentre), and were reverse-transcribed to cDNA using random primers. 2µl of cDNA were added to each reaction as template for qPCR using light cycler 480 SYBR I master kit and the LightCycler 480 system (Roche Applied Science), according to the manufacturer's instructions. For each strain, two qPCRs were performed with three biological repetitions and three technical repeats. A first qPCR was performed targeting the transcript spliced-version with a forward primer complementing the exon-exon junction and a downstream reverse primer. A second reaction targeted the un-spliced version of the transcript with a forward primer complementing the intron and the same reverse primer of the first reaction.

$F_{exon\text{-}exon}$ = 5'-CACTACTTTAGGTTATGGTTT-3'

F_{intron} = 5'-CTTCAATTTACTGAATTTGTATG-3'

R_{both} = 5'-GTCTTGTAGTTACCGTCA-3'

Splicing efficiency is reported as the average Cp of the spliced transcript minus the average Cp of the un-splice version.

mRNA deep sequencing

Cultures were grown in YPD+G418 at 30°C until cells reached the logarithmic growth phase at an optical density of ~0.4. Cells were then harvested by centrifugation and flash-frozen in liquid nitrogen. RNA was extracted using a modified protocol of nucleospin® 96

RNA kit (Machery-Nagel). Specifically, cell lysis was done in a 96 deep-well plate by adding for each well 450µl of lysis buffer containing 1M sorbitol, 100mM EDTA and 0.45µl lyticase (10IU/µl). The plate was incubated at 30°C for 30min to break cell wall, centrifuged for 10min at 3000rpm, followed by the removal of the supernatant. Then, extraction continued as in the protocol of nucleospin® 96 RNA kit, only using β-mercaptoethanol instead of DTT. Poly(A)-selected RNA extracts of size ~200bps were reverse-transcribed to cDNA using poly(T) primers that were barcoded with a unique molecular identifier (UMI). cDNA was then amplified and sequenced with an Illumina HiSeq 2500.

Analysis of mRNA deep sequencing

Processing of RNA-seq data was performed as described in Voichek *et al*[43]. Shortly, reads were aligned using Bowtie[44] (parameters: --best –a –m 2 –strata -5 10) to the genome of *S. Cerevisiae* (R64 from SGD) with an additional chromosome containing the sequence of the YFP-Kan construct. For each sequence, we normalized for PCR bias using UMIs as describe in Kivioja *et al*[45]. Next, reads for each gene end (400bp upstream to 200bp downstream of the ORF's 3' end) were summed-up to estimate the gene's expression level. Genes with coverage lower than 10 reads were excluded. To normalize for differences in coverage among samples, we divided each gene expression by the total read count of each sample and then multiplied by 10^6. Then, expression ratio was calculated between an evolved/rescue colony to the ancestor and a log2 operation was performed on that ratio. These values were used to compare expression levels of gene groups (ribosomal genes, general stress response genes, splicing machinery genes, intron-containing genes) and of the YFP-Kan mRNA levels as described in the manuscript. When calculating the expression levels of splicing machinery and intron-containing gene groups, the ribosomal and general stress response genes were excluded from the analysis in order to avoid bias from cellular regulation due to changes in physiology and growth rate of the cells.

Acknowledgments

We thank Martin Mikl for help with the qPCR experiments. We also greatly appreciate comments to the text from Tslil Ast, Orna Dahan, Dvir Schirman, and Avihu Yona, and the entire Pilpel lab for discussions of the project. IF thanks the Azrieli Foundation for an Azrieli PhD Fellowship. This study was supported by the Minerva Foundation who funded the Minerva Center for Live Emulation of Evolution in the Lab.

Bibliography

1. Weingarten-Gabbay, S. & Segal, E. The grammar of transcriptional regulation. *Hum. Genet.* (2014). doi:10.1007/s00439-013-1413-1

2. Amorós-Moya, D., Bedhomme, S., Hermann, M. & Bravo, I. G. Evolution in regulatory regions rapidly compensates the cost of nonoptimal codon usage. *Mol. Biol. Evol.* **27,** 2141–51 (2010).

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

4. Gingold, H. & Pilpel, Y. Determinants of translation efficiency and accuracy. *Mol. Syst. Biol.* **7,** 481 (2011).

5. Presnyak, V. *et al.* Codon optimality is a major determinant of mRNA stability. *Cell* **160,** 1111–24 (2015).

6. Bazzini, A. A. *et al.* Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *EMBO J.* **35,** 2087–2103 (2016).

7. Babu, M. & Teichmann, S. A. Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res.* **31,** 1234–44 (2003).

8. Yona, A. H. *et al.* tRNA genes rapidly change in evolution to meet novel translational demands. *Elife* **2,** e01339 (2013).

9. Kirchner, S. & Ignatova, Z. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. *Nat. Rev. Genet.* (2014). doi:10.1038/nrg3861

10. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8,** 206–16 (2007).

11. Lemos, B., Araripe, L. O., Fontanillas, P. & Hartl, D. L. Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 14471–6 (2008).

12. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in cis and trans gene regulation. *Nature* **430,** 85–8 (2004).

13. Matera, a. G. & Wang, Z. A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.* **15,** 108–121 (2014).

14. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338,** 1593–9 (2012).

15. Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338,** 1587–93 (2012).

16. Petibon, C., Parenteau, J., Catala, M. & Elela, S. A. Introns regulate the production of ribosomal proteins by modulating splicing of duplicated ribosomal protein genes. *Nucleic Acids Res.* **44,** gkw140 (2016).

17. Parenteau, J. *et al.* Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. *Cell* **147,** 320–31 (2011).

18. Reyes, a. *et al.* Drift and conservation of differential exon usage across tissues in primate species. *Proc. Natl. Acad. Sci.* 2–7 (2013). doi:10.1073/pnas.1307202110

19. Bush, S. J., Chen, L., Tovar-Corona, J. M. & Urrutia, A. O. Alternative splicing and the evolution of phenotypic novelty. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **372,** (2017).

20. Irimia, M. & Roy, S. W. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb. Perspect. Biol.* **6,** (2014).

21. Roy, S. W. & Irimia, M. Mystery of intron gain: new data and new models. *Trends Genet.* **25,** 67–73 (2009).

22. Hooks, K. B., Delneri, D. & Griffiths-Jones, S. Intron evolution in Saccharomycetaceae. *Genome Biol. Evol.* (2014). doi:10.1093/gbe/evu196

23. Shabalina, S. a *et al.* Distinct patterns of expression and evolution of intronless and intron-containing mammalian genes. *Mol. Biol. Evol.* **27,** 1745–9 (2010).

24. Lee, S. & Stevens, S. W. Spliceosomal intronogenesis. *Proc. Natl. Acad. Sci. U. S. A.* 201605113 (2016). doi:10.1073/pnas.1605113113

25. Derr, L. K., Strathern, J. N. & Garfinkel, D. J. RNA-mediated recombination in S. cerevisiae. *Cell* **67,** 355–64 (1991).

26. Yofe, I. *et al.* Accurate, Model-Based Tuning of Synthetic Gene Expression Using Introns in S. cerevisiae. *PLoS Genet.* **10,** e1004407 (2014).

27. Ares, M., Grate, L. & Pauling, M. H. A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA* **5,** 1138–9 (1999).

28. de Nadal, E., Ammerer, G. & Posas, F. Controlling gene expression in response to stress. *Nat. Rev. Genet.* **12,** 833–45 (2011).

29. Metzl-Raz, E. *et al.* Principles of cellular resource allocation revealed by condition-dependent proteome profiling. *Elife* **6,** (2017).

30. Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11,** 4241–57 (2000).

31. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533,** 397–401 (2016).

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

33. Brackley, C. A., Romano, M. C. & Thiel, M. The dynamics of supply and demand in mRNA translation. *PLoS Comput. Biol.* **7,** e1002203 (2011).

34. Sverdlov, A. V, Babenko, V. N., Rogozin, I. B. & Koonin, E. V. Preferential loss and gain of introns in 3' portions of genes suggests a reverse-transcription mechanism

of intron insertion. *Gene* **338,** 85–91 (2004).

35. Cohen, N. E., Shen, R. & Carmel, L. The role of reverse transcriptase in intron gain and loss mechanisms. *Mol. Biol. Evol.* **29,** 179–86 (2012).

36. Zhu, Y. O., Siegal, M. L., Hall, D. W. & Petrov, D. a. Precise estimates of mutation rate and spectrum in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E2310-8 (2014).

37. Soemedi, R. *et al.* The effects of structure on pre-mRNA processing and stability. *Methods* **125,** 36–44 (2017).

38. Gahura, O., Hammann, C., Valentová, A., Půta, F. & Folk, P. Secondary structure is required for 3' splice site recognition in yeast. *Nucleic Acids Res.* **39,** 9759–67 (2011).

39. Warf, M. B. & Berglund, J. A. Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem. Sci.* **35,** 169–78 (2010).

40. Munding, E. M., Shiue, L., Katzman, S., Donohue, J. P. & Ares, M. Competition between Pre-mRNAs for the Splicing Machinery Drives Global Regulation of Splicing. *Mol. Cell* **51,** 338–348 (2013).

41. Man, O. & Pilpel, Y. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat. Genet.* **39,** 415–21 (2007).

42. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324,** 255–8 (2009).

43. Voichek, Y., Bar-Ziv, R. & Barkai, N. Expression homeostasis during DNA replication. *Science* **351,** 1087–90 (2016).

44. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

45. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9,** 72–4 (2011).

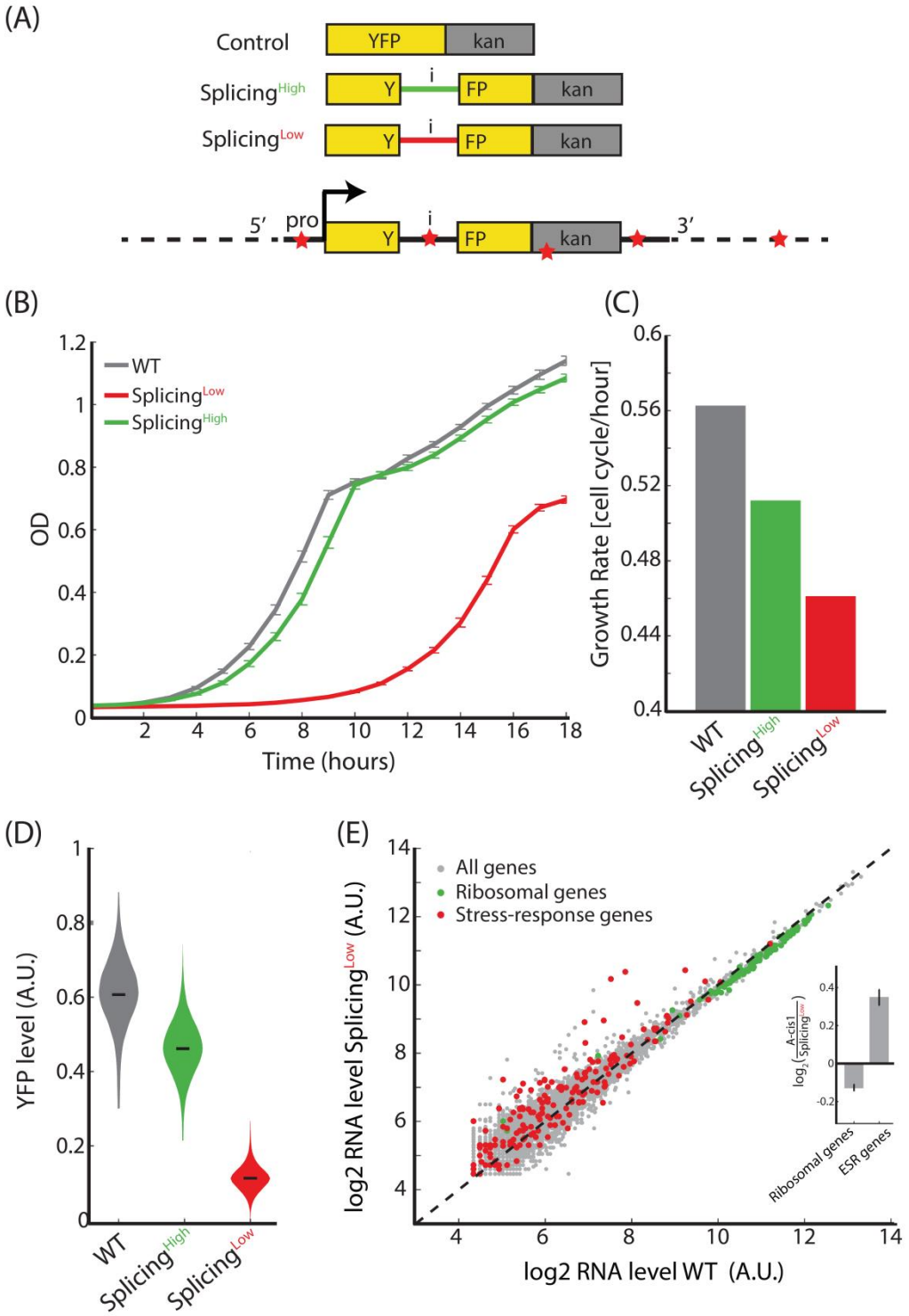


Figure 1 – Inefficient intron splicing leads to lower gene expression levels and compromised antibiotics resistance.

A| We introduced two alternative introns into a YFP domain that was fused to a kanamycin resistance domain - to generate three strains: (i) WT without an intron; (2) Splicing^{High} with an efficiently spliced intron; and (iii) Splicing^{Low} with an inefficiently spliced intron. Evolving cells at the presence of the antibiotics could adapt by mutating different parts of the YFP-Kan construct (evolution in cis) or other loci, evolution in *trans* (red stars represent potential locations of such putative mutation sites).

B+C| Splicing^{Low} suffers from a severe growth defect compared to WT or Splicing^{High} cells when the antibiotic is supplemented to the medium. The growth defect is manifested as a longer lag phase and a lower maximal growth rate.

D| Florescence intensity of the YFP-Kan reporter for all three strains shows that Splicing^{Low} cells have lower expression levels of YFP-Kan. This observation links between YFP-Kan expression levels and cellular fitness.

E| Transcriptome profiling shows that ribosomal genes were down-regulated (green dots, p-Value=4.62x10^{-26}, paired t-test) and stress-related genes were up-regulated (red dots, p-Value=3.40x10^{-5}, paired t-test) in Splicing^{Low} compared to WT cells. This observation suggests that Splicing^{Low} cells are stressed because of compromised resistance to the antibiotics and that the general stress response was activated in them. **Inset|** Mean $\log_2$ ratio of ribosomal and ESR gene groups.

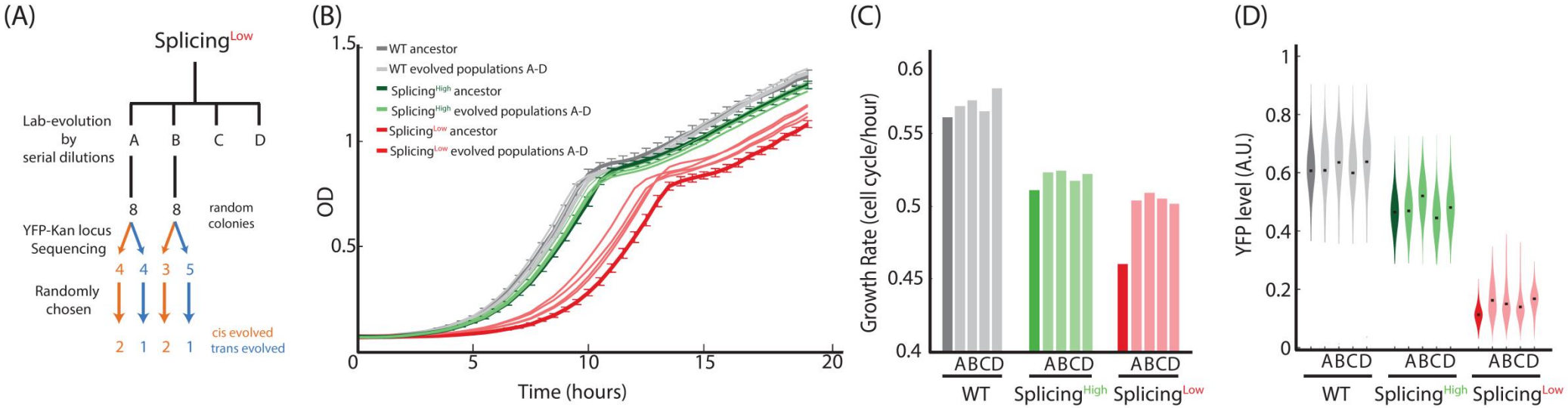


Figure 2 – Rapid adaptation to the presence of the antibiotics is observed only for Splicing^{Low} cells.

A| We evolved WT, Splicing^{High}, and Splicing^{Low} cells for ~560 generations with the presence of the antibiotics in four independent cultures for each strain. We measured fitness and YFP-Kan expression levels for all evolved lines (see below), and also randomly chose 16 colonies from two evolved lines of Splicing^{Low}. We sequenced the YFP-Kan locus of those colonies and observed that around half showed mutations in the YFP-Kan construct (indication of evolution in *cis*) and the other half did not (indication of evolution in *trans*). Of those colonies, we randomly chose two *cis*-evolved and one *trans*-evolved colonies from each evolved population for further examinations (see figure 3 onwards).

B+C| Growth of evolved populations compared to the three ancestors. Only evolved Splicing^{Low} cells demonstrate significant improvement in growth for all four independent evolution lines. This observation suggests that the inefficiently spliced intron led to a rapid adaptation of Splicing^{Low} cells.

D| Florescence intensity of the YFP-Kan reporter for all evolved cultures show that expression levels were much increased in all four evolved cultures of Splicing^{Low} compared to the ancestral strain (effect sizes = 78.67, 79.54, 75.17, 83.19). Conversely, the increase in expression levels in the evolved WT and of Splicing^{High} populations were smaller (WT effect sizes = 64.66, 68.44, 63.51, 67.74; Splicing^{High} effect sizes = 54.33, 70.66, 52.43 and 58.27). This observation suggests that adaptation of Splicing^{Low} cells was based on their ability to increase expression levels of the resistance proteins.

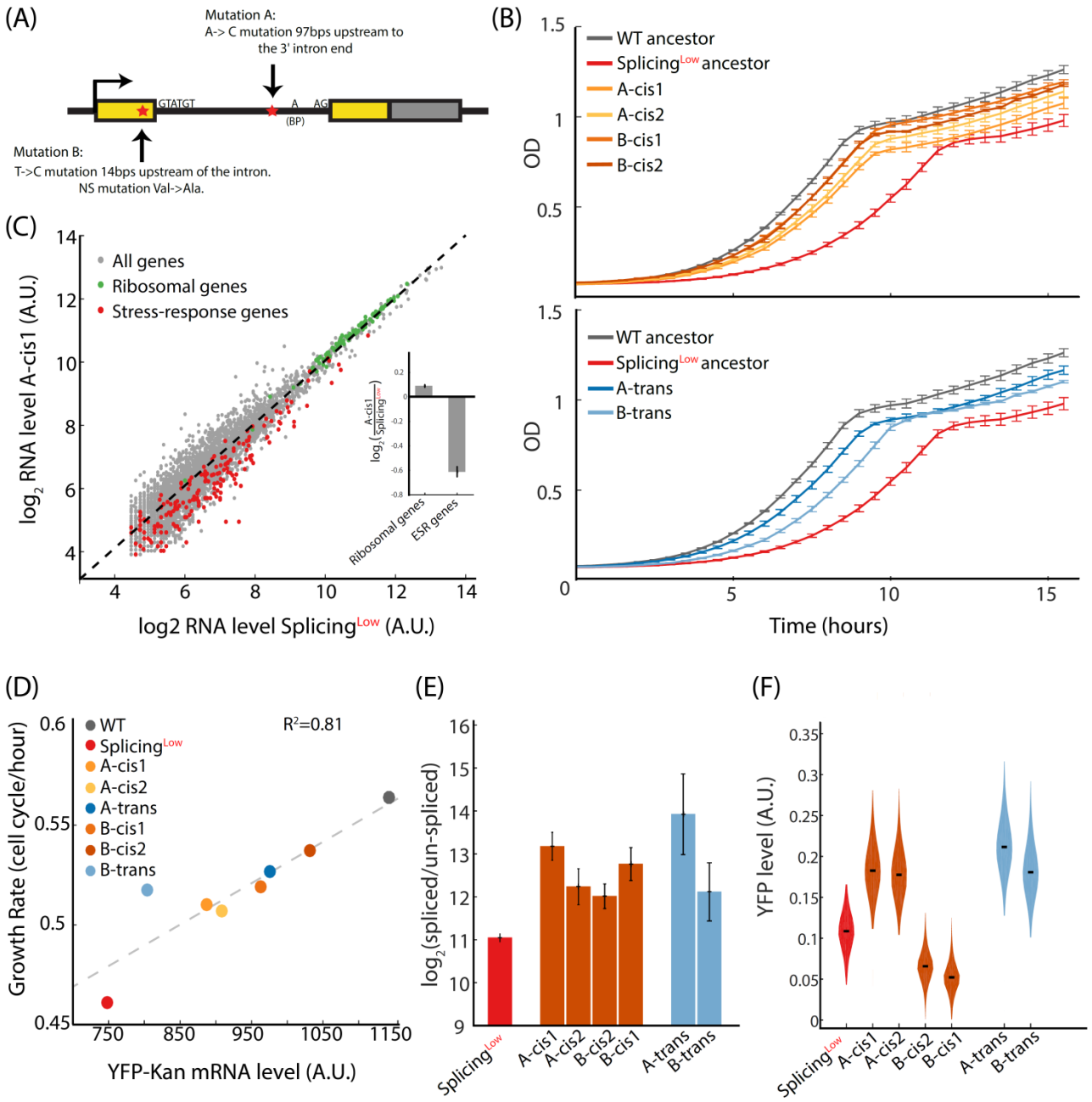


Figure 3 – Evolved colonies demonstrate increased splicing efficiency that results in higher transcript levels and relieved stress.

A| Sequencing of the YFP-Kan construct in the evolved colonies revealed two mutation types: (i) in the intron itself and (ii) in the up-stream exon – see text for full description. These mutations did not occur in the intron 5' donor, 3' acceptor, or the branching point – suggesting that other positions of the intron and its vicinity are phenotypically functional and may affect splicing efficiency.

B| All *cis*-evolved colonies (upper graph) and *trans*-evolved colonies (lower graph) show increased fitness compared to the Splicing^{Low} ancestor, yet still lower than the WT ancestor.

C| Transcriptome profiling reveals that ribosomal genes were up-regulated (green dots, p-Value=4.94 x10^{-18}, paired t-test) and stress-related genes were down-regulated (red dots, p-Value=3.64 x10^{-15}, paired t-test) in the evolved colony A-cis1 compared to the Splicing^{Low} ancestor. This trend was observed in 5 out of 6 evolved colonies (Supplementary Figure 1). **Inset|** Mean $\log_2$ ratio of ribosomal and ESR gene groups.

D| mRNA levels of YFP-Kan transcripts correlate with growth rate – suggesting that cellular fitness in our set-up is indeed determined by the availability of Kanamycin-resistance proteins to overcome the antibiotics.

E| All *cis*- and *trans*-evolved colonies demonstrate increased splicing efficiency of the YFP-Kan mRNA compared to the Splicing^{Low} ancestor. This result suggests that all adaptation trajectories led to the adaptation of the splicing process to better mature the un-spliced YFP-Kan transcript.

F| Florescence intensity of the YFP-Kan reporter show increased levels for the two *cis*-evolved colonies with the mutation in the intron and for the two *trans*-evolved colonies. In contrast, the two *cis*-evolved colonies with the non-synonymous mutation in the exon demonstrate decreased YFP-Kan levels. This observation suggests that the non-synonymous mutation hampered the ability of the YFP domain to florescent and reduced the Florescence intensity per protein molecule (see text for full explanation).

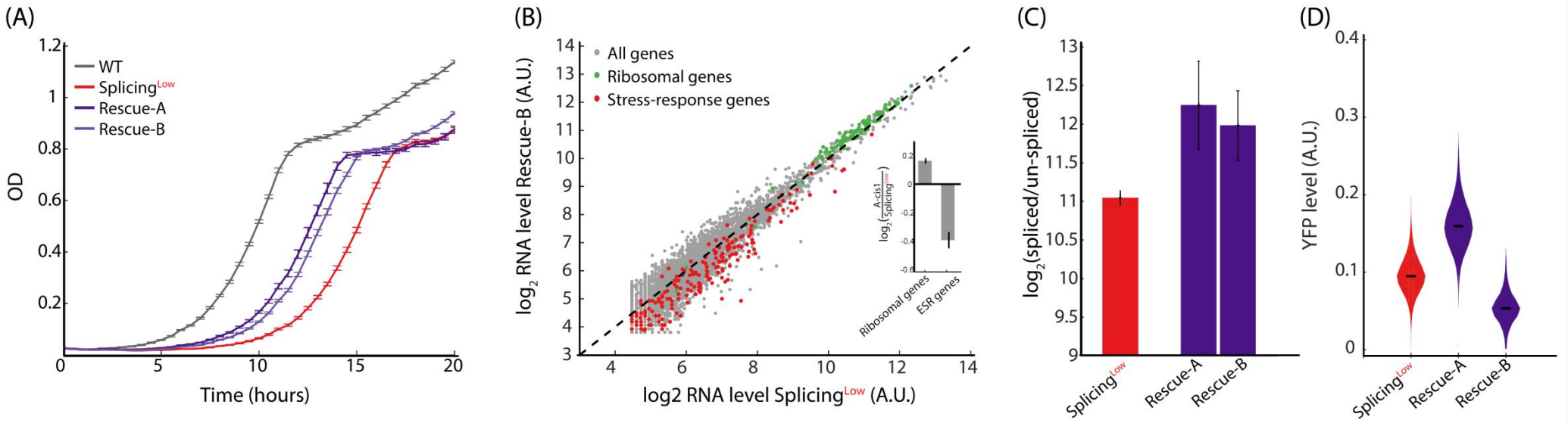


Figure 4 – cis-acting mutations are sufficient to increase fitness by elevating splicing efficiency.

A| We created two rescue strains, each harboring one of the mutations that appeared spontaneously in the evolved populations. Growth of the two rescue strains show that a single mutation in the YFP-Kan construct is sufficient to increase fitness compared to Splicing^{Low}.

B| The exonic mutation is also sufficient to alleviate stress, as ribosomal genes were up-regulated (green dots, p-Value=1.02x10^{-18}, paired t-test) and stress-related genes were down-regulated (red dots, p-Value=9.02x10^{-12}, paired t-test) in Rescue-B compared to Splicing^{Low}. The same trend was also observed for the intronic mutation for Rescue-A cells. **Inset|** Mean $\log_2$ ratio of ribosomal and ESR gene groups.

C| The two rescue strains demonstrate higher splicing efficiency of the YFP-Kan mRNA compared to the Splicing^{Low} ancestor. This result suggests that a single mutation is sufficient to improve splicing efficiency.

D| Florescence intensity of the YFP-Kan reporter for the Rescue-A and Rescue-B strains show similar trends as the colonies in Figure 3D - supporting earlier conclusions.

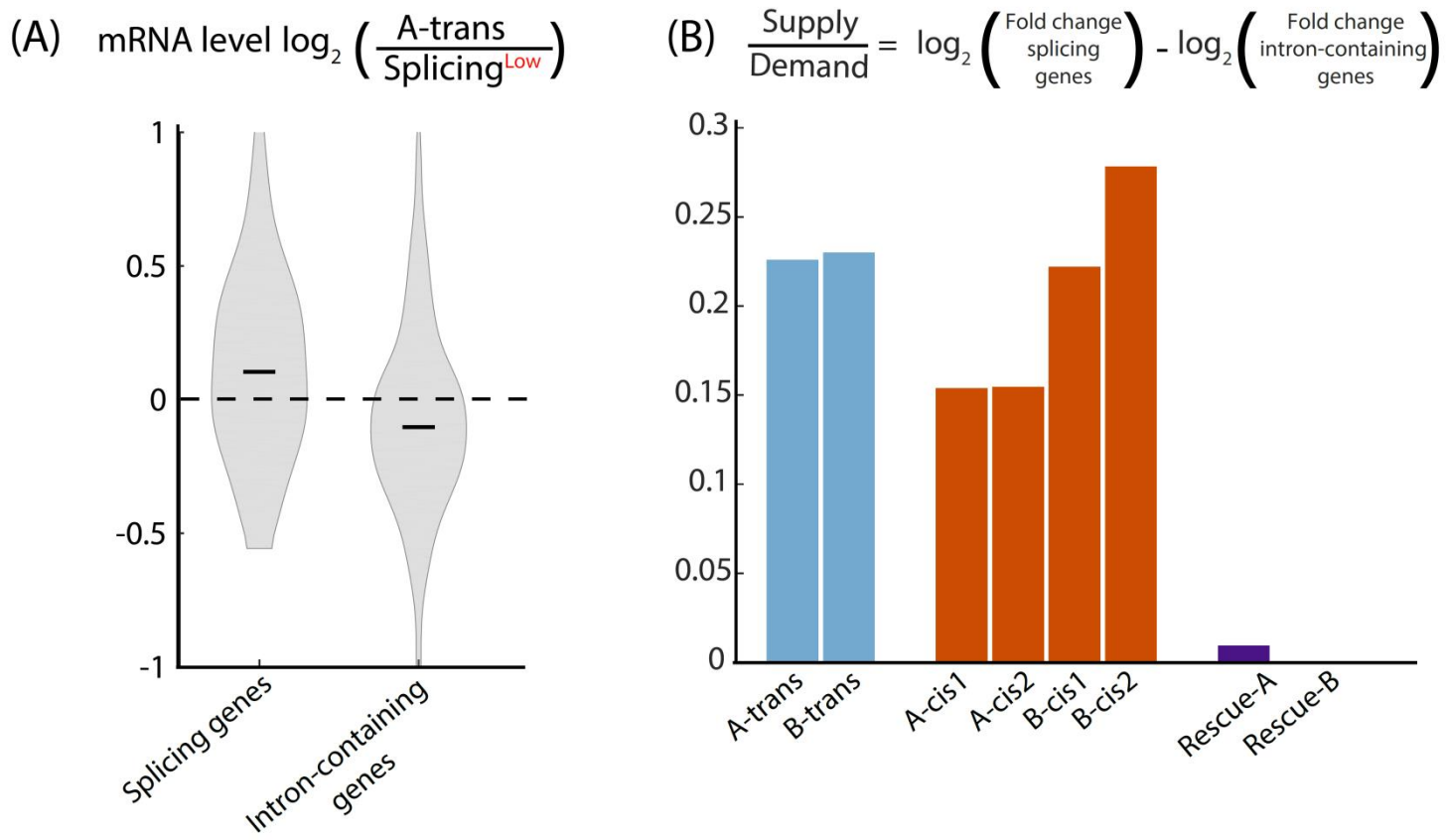
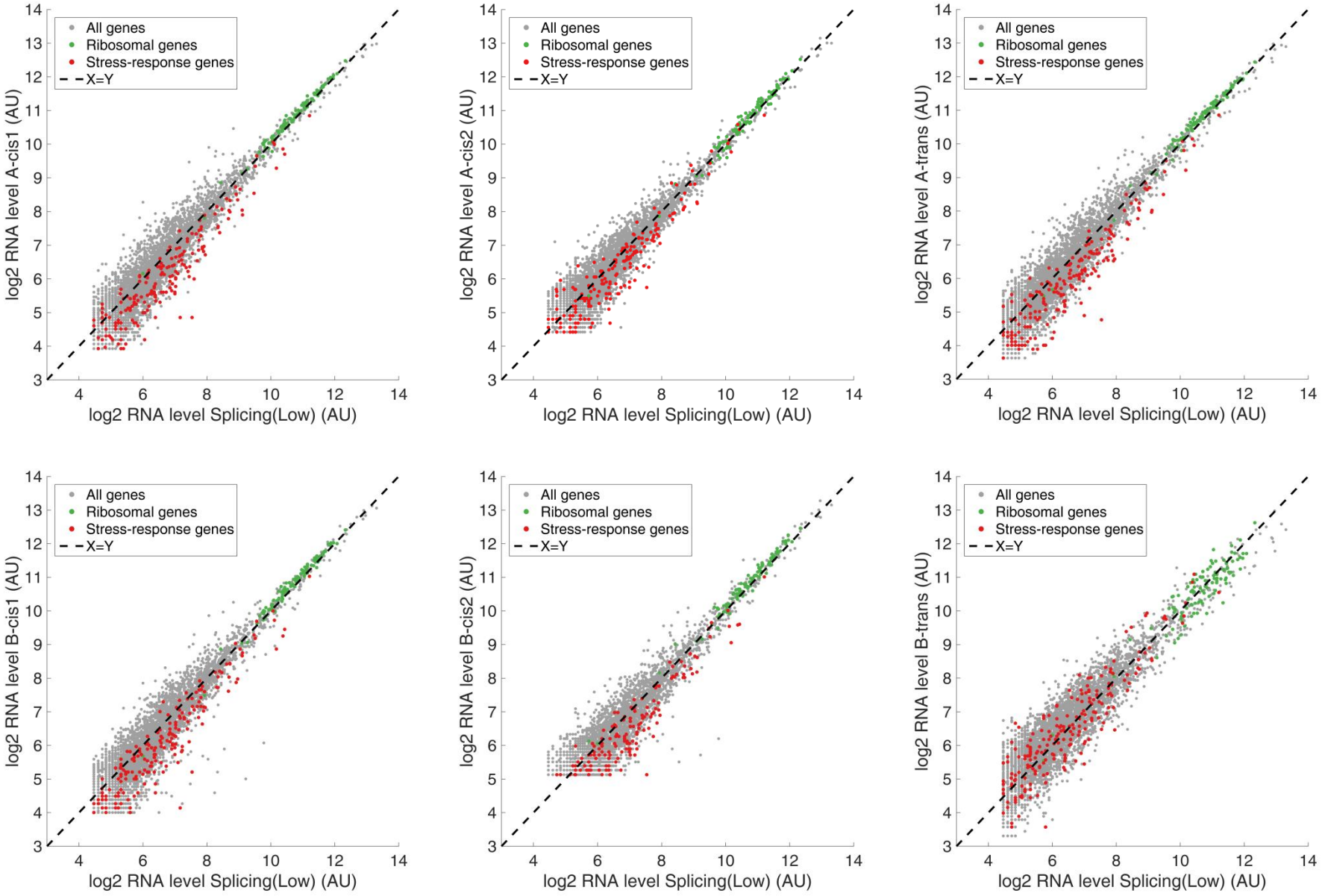


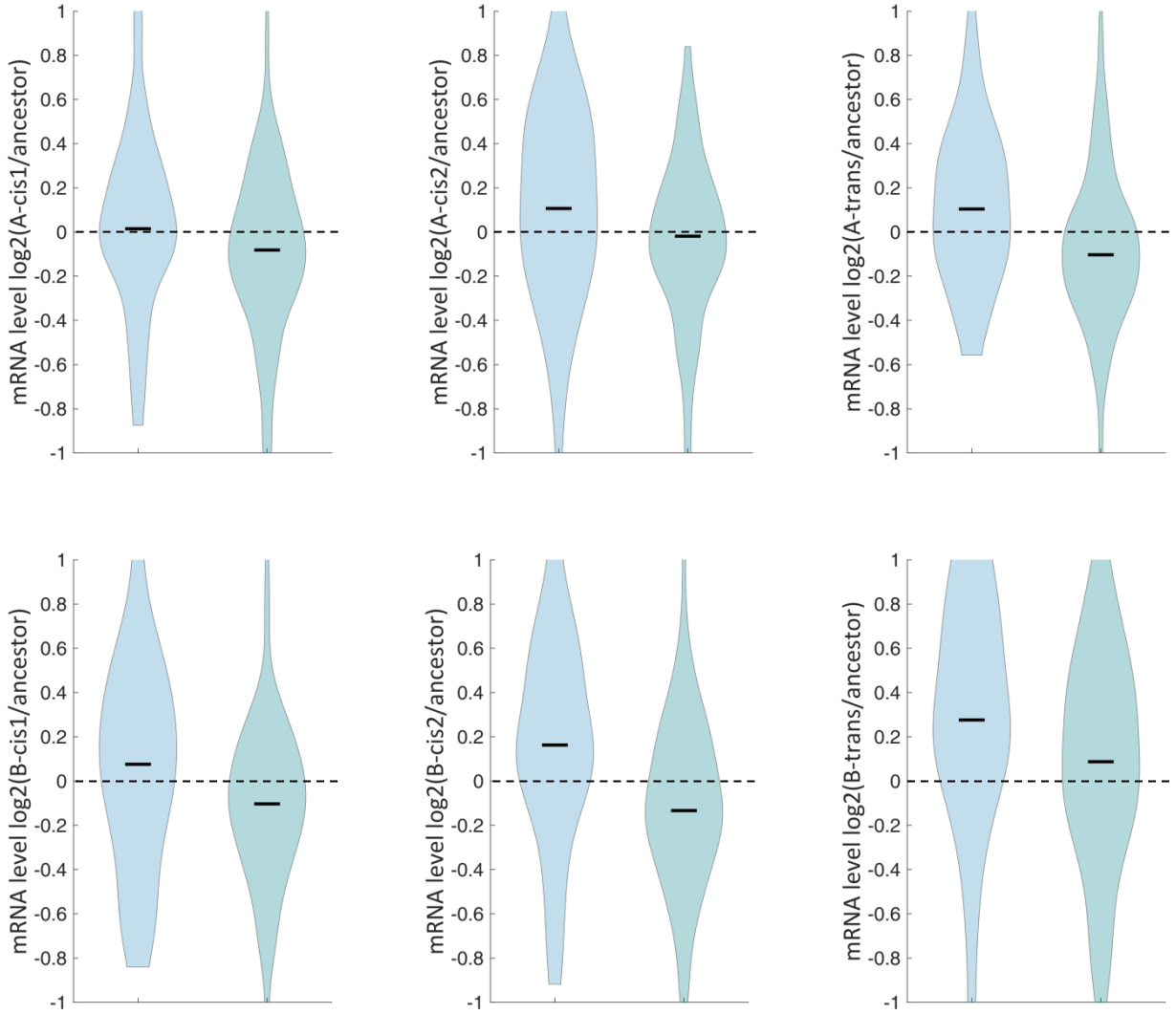
Figure 5 – Increasing cellular availability of the splicing machinery is an adaptive mechanism of splicing.

A| The groups of splicing genes and intron-containing genes were increased (p-Value=1.36x10^{-3}, paired t-test) and decreased (p-Value=1.67x10^{-2}, paired t-test), respectively, in the trans-evolved colony A-trans compared to Splicing^{Low} ancestor. This observation suggests that the supply-to-demand ratio of the splicing machinery was increased in A-trans colony, which allowed its increased splicing efficiency of the YFP-Kan transcript.

B| Supply-to-demand ratios for the splicing machinery were calculated to all evolved colonies and to the rescue strains as the difference between the mean fold-change of splicing genes to the mean fold-change of intron-containing genes. While supply-to-demand ratios were increased in all evolved colonies, they remained the same for the two rescue strains. These results suggest that indeed the cellular availability of the splicing machinery was elevated in the evolved colonies – a *trans*-adaptation mechanism to optimize gene expression using the splicing process.



Supplementary Figure 1 | Transcriptome profiling reveals that ribosomal genes were up-regulated and that stress-related genes were down-regulated in 5 out of 6 of the evolved colonies compared to the Splicing^{Low} ancestor.



Supplementary Figure 2 | Supply-to-demand ratio in each of the evolved colonies. Folding change ratios in log2 are shown for splicing genes (left) and intron-containing genes (right). Black line represents the median of the distribution.