מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Thesis for the degree
Doctor of Philosophy

עבודת גמר (תזה) לתואר
דוקטור לפילוסופיה

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

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

By
Dvir Schirman

מאת
דביר שירמן

שימוש בספריות גנטיות סינטטיות כדי לחקור את הבקרה והכלכלה של ביטוי גנים

Using large synthetic libraries to explore the regulation and economy of gene expression

Advisor:
Prof. Yitzhak Pilpel

מנחה:
פרופ' יצחק פלפל

October 2020

תשרי תשפ"א

1. Acknowledgment

First, I would like to thank my advisor and mentor Prof. Tzachi Pilpel for his dedicated guidance, for giving me inspiration during many conversations about my projects, and science in general, and for encouraging me to follow a career in science by presenting a role model for how enjoyable scientific career can be.

I thank Orna Dahan for her dedicated guidance in biological lab work, and for countless of hours discussing and solving all the technical aspects of my projects. Everything I know in lab work I owe to her dedicated and patient teaching.

I thank Ruth Towers for her help in all the logistic side of the lab, I could not have done anything in this work without her help.

I thank Prof. Zohar Yakhini for collaborating with us on the non-coding RNA library project, and for fruitful discussions.

I thank all the members of the Pilpel lab during the past five years for being a great team of scientists and creating a pleasant working environment. I wish to give special thanks to Idan Frumkin, Aviv Rotman, Ruthie Golomb and Shuting You for collaborating with me on two of the projects presented in this work.

I thank all the administrative stuff of the Molecular Genetics department for enabling all the work that is done in the department, including my own.

I thank the members of my PhD thesis committee, Igor Ulitsky and Schragi Schwartz for their advice during my research.

Finally, I wish to thank my spouse Naama, for providing me much support throughout this five-year journey.

2. Declaration

This thesis summarizes my independent research, yet some parts are the result of the following fruitful collaborations:

- The project described in section "6.1.1 Gene Architectures that Minimize Cost of Gene Expression in Bacteria" was done in collaboration with Idan Frumkin and Aviv Rotman. Idan and Aviv preformed the competition experiment and other lab work, Aviv devolved the pipeline to align sequencing results to the library, and Idan and I performed all the analysis.
- The project described in "6.3 Part III – The economy of expressing a shared public-good compound" was done in collaboration with Ruthie Golomb and Shuting You. Ruthie and I have designed the library, and performed the cloning and transformations of the library. Competition experiments and sequencing were performed by Ruthie, Shuting and I. Data analysis was performed by Ruthie, Shuting and I.

Table of Contents

3. List of Abbreviations

E. coli - *Escherichia coli* bacteria

S. cerevisiae – *Saccharomyces cerevisiae* yeast

GFP/YFP – Green fluorescent protein / Yellow fluorescent protein

FACS - Fluorescence-activated cell sorting

RBS – Ribosome binding site

SD / aSD – Shine-Dalgarno / anti Shine-Dalgarno sequence

ORF – Open reading frame

CDS – Coding sequence

UTR – Untranslated region

BC – Barcode

TF – Transcription factor

TFBS – Transcription factor binding site

CRE – *cis*-regulatory element

5'SS – 5' splice site

BS – Branch site

3'SS – 3' splice site

SECReTE – Secretion Enhancing *Cis* Regulatory Targeting Element

SP – Signal peptide

ER – Endoplasmic reticulum

SRP – Signal recognition particle

TM – Trans-membrane

4. Abstract

To flourish in nature cells must respond to changing environments via execution of different genetic programs. This is achieved by expressing genes in a regulated manner to reach the needed concentration of each gene. Gene expression affects the cell not only through production of needed proteins, but also by the burden it imposes on the cell.

In my research I aimed to disentangle regulatory elements that affect the regulation or economy of gene expression. I did so, using several synthetic oligo libraries.

In the first part of this work, I used a synthetic library in the bacteria *E. coli* to study the cost of gene expression. I achieved this goal by using a library with thousands of variations in a reporter gene and measured the relative fitness of each variant in this library using a competition assay. Since this gene is not needed by the cell, any differences in fitness are due to the cost of expressing it. By examining library variants that present higher or lower cost at a given expression level, I elucidated genetic features that contribute to efficient gene expression.

Next, I studied the regulation of mRNA splicing through variations in intron sequences. mRNA splicing is a crucial part of gene expression in eukaryotes, in which an intron is spliced out from a pre-mRNA to produce a mature mRNA. In this project I designed a new library of synthetic introns that were inserted into the yeast genome. Using this library, I revealed how different elements contribute to the efficiency of the splicing process. And I revealed how intron architecture have co-evolved with the splicing machinery.

In studying how sequence variations affect regulation of gene expression and the wellbeing of a unicellular organism, I assumed no interaction between strains. Lastly, I set out to challenge this assumption by studying the economy of expressing a public-good enzyme. Organisms in nature often cooperate within their community. Even the apparently simple unicellular organisms cooperate. They do so by secreting public-good compounds that are shared between neighboring cells. This relates to a classic question in evolutionary biology, which is how a costly cooperative strategy is maintained when a mutant cheater strain can take over the population. In the last part of this work, I studied producer-cheater dynamics in a complex population of different cooperation strategies in yeast. I used a synthetic library to introduce variations to a secreted public-good enzyme in yeast to create strains with varying levels of cooperation and measured their frequency dynamics in two growth conditions.

The three parts of this work combined demonstrate how genetic *cis*-regulatory elements have wide impact on the cell beyond their direct regulation of a specific gene. And how these systematic effects are under selection and has shaped the genome throughout evolution.

5. Introduction

To flourish in nature cells must respond to changing environments via execution of genetic programs that are stored in their genomes. This goal is achieved by expressing different genes in a regulated manner to reach the needed concentration of each gene. Gene expression affects the cell not only through production of needed proteins, but also by the burden it imposes on the cell. This burden is manifested through the usage of raw materials, energy and reducing power consumption[1–4], and allocation of common resources[5] involved in gene expression process. Hence during evolution, gene expression regulatory elements in-*cis* or in-*trans* are under selection to optimize the trade-off between the benefit of a gene[6] and the cost of producing it[7].

The relationship between cost and benefit as described above, generally describes the economy of expressing a gene which acts only intracellularly. However, even in unicellular organisms some of the proteins produced by the cell are acting extracellularly, either for communication[8], to process nutrients in the environment[9], or for extra-cellular stress response[10,11]. In such case more complex economic dynamics arises, since the cost of producing the protein is imposed only on cells producing it, while any cell in the environment can benefit from the common good. Such a scenario relates to a deep evolutionary question, on how cooperative strategies can arise in evolution[12,13], and on the relationship between cooperators and cheaters[14–16]. Moreover, studying the economy of gene expression through secreted common goods might provide a simplified model for the economy of gene expression in multi-cellular organisms.

One of the ways eukaryotic cells can regulate gene expression level is through regulation of mRNA splicing of constitutively spliced introns. During splicing introns are spliced out of a pre-mRNA molecule to produce a mature mRNA. Many of the research works on splicing deals with the regulation of alternative splicing. Yet, most introns are constitutively spliced[17,18], meaning the intron must be spliced out to create a functional mRNA. Therefore, efficient execution of this step is necessary for efficient expression of an intron-containing gene, and splicing regulation has a major effect on the cell. The principles that regulate the efficiency of this process are not yet fully understood.

5.1. Large synthetic libraries to study genetic *cis*-regulatory elements

In this research I aimed to disect *cis*-regulatory elements that affect the regulation or economy of gene expression, by utilizing several synthetic oligo libraries[19]. These libraries are a relatively new tool that enables a researcher to design thousands of short DNA oligos,

which enables a systematic exploration of regulatory sequence elements, and they have been used in several works to systematically study gene expression regulation[20–24].
In this work two such existing synthetic libraries that were designed to study gene expression regulation in bacteria[25] and yeast[26], were utilized to measure the cost of gene expression in two different systems.

5.2. Cost of gene expression

In the first part of this thesis I describe four different experiments in which I explored the effect of different regulatory elements on the cost of gene expression. While the beneficial contribution of a gene is specific to each gene according to its exact function, and changes in the gene's architecture will affect differently each gene[6], the rules that govern the cost of producing a gene are likely to be similar for all the genes in the same organism, as all genes are expressed using the same molecular machineries. Therefore, a study of ways in which the cell can minimize the production costs of a certain gene can help us reveal meaningful insights on how genes can be optimized for a more efficient expression.
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[1–4]. Even after their production, proteins might still impose costs, when degraded or by exerting toxicity, e.g. due to aggregation[27]. Understanding what molecular processes determine expression cost, its relation to 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[28–32], 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[1,7,33–36]. The production of such unneeded proteins diverts resources from synthesis of the cell's own proteins thus decreases cellular fitness[37–39]. Central to these studies 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 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. Large synthetic libraries can be harnessed to

disentangle expression level and expression costs, and reveal mechanisms that affect cost per protein molecule, since they can be used to create many different variants with different expression levels and associated costs. Thus, allowing to first characterize a relation between expression level and cost, and then study strains that consistently deviate from this predicted relationship achieving better or worse fitness than expected.

5.3. Sequence determinants and evolution of splicing in yeast species

In the second part of this thesis I describe how intron architecture affects the regulation of constitutive splicing in yeast, present evidence for co-evolution of intron architecture and the splicing machinery across yeast species, and examine the capacity of yeast to alternatively splice two-intron genes.

In human and other organisms, splicing is central to gene expression, as a typical human gene contains 8 introns[40], and these introns can be alternatively spliced to create different alternative splicing isoforms which increases the proteomic diversity of the human genome[41,42]. However, most introns or exons are constitutively spliced[17,18] and therefore their contribution to gene expression is not through increasing diversity. Nevertheless, when a pre-mRNA must undergo splicing to produce a functional mRNA, the efficiency of the splicing process directly affects the efficiency of the overall gene expression process. Hence, regulation of constitutive splicing in eukaryotes have an important role in gene expression regulation[43,44].

The budding yeast Saccharomyces cerevisiae, like other hemiascomycetous fungi, has a low number of intron-containing genes compared to other eukaryotes[45], and most of them have single intron which is constitutively spliced. Although they possess a small part of the genome, these intron-containing genes are highly expressed and are part of key cellular functions, for example, many of the ribosomal proteins in *S. cerevisiae* contain introns[46]. Hence, splicing regulation plays a major role in gene expression regulation in *S. cerevisiae*.

The evolution of introns and the splicing machinery in yeast is specifically interesting because of the fact that as opposed to other eukaryotes and even other fungi they have low number of introns[45]. The reason for this low number of introns remains an open question, and several works have tried to answer it using comparative genomics analysis[45,47,48], and the predominant explanation is that the last eukaryotic common ancestor (LECA) already had many introns[49], and that in the hemiascomycetous yeast there was a massive loss of introns. The most plausible mechanism suggested for this loss of introns is homologous

recombination between reverse-transcribed mature mRNA and the genomic locus[50] which explains a visible bias of intron positions in yeast towards the 5' end of genes.
Alternative splicing is the process in which a single intron-containing gene can produce several different isoforms that result in different proteins by inclusion or exclusion of specific exons[51]. It is considered to have major contribution to protein diversity in metazoan organisms. In higher eukaryotes this process involves many auxiliary factors[52] which were shown to be absent in yeast[53,54]. Yet, it has been shown that in *S. cerevisiae* there are cases of alternative 3' splice site selection[55], and there are rare examples of two-introns genes that are alternatively spliced[56–60]. Hence, this organism is an interesting model to study the minimal requirements for alternative splicing.

Previous works have studied constitutive introns' splicing efficiencies in *S. cerevisiae* by looking at existing natural introns. Either by analyzing RNA-seq data of intron containing genes[61–63], by creating a library of a reporter gene containing natural introns from *S.cerevisiae* genome[64] or by studying intron sequence features and looking at introns evolution and conservation[65,66].

These works have characterized principles of splicing regulation in budding yeast, but due to the low number of intron containing genes in this species there is not enough statistical power to study the effect of rare regulatory elements variants. Hence, by taking advantage of the power of large synthetic library the effect of different variants of regulatory elements, and their combination can be studied systematically. Other works have utilized large synthetic libraries to study how sequence features affect alternative splicing decisions in human cells [67–70]. In this work, *cis*-regulatory features that affect splicing efficiency of an intron, and the evolution of intron architecture and the yeast splicing machinery are studied using a designed large library of ~20,000 synthetic introns, each of them composed of different sequence features.

5.4. The economy of expressing a shared public-good compound

In the third part of this thesis I describe the design and initial results of a project aimed to decipher the complex evolutionary game theory dynamics in a population of yeast cells with varying levels of public-good production. In the parts described above, the economy of gene expression was studied by exploring how different architectures of genetic regulatory elements affect the cost of gene expression. A different general aspect of gene expression economy is the production and secretion of common goods which are consumed by a community of cells.

In unicellular organisms most of the cell's resources are devoted to intracellular functions which benefit only the cell itself. Nevertheless, microorganisms mostly live in colonies, and some of the cell's proteome are proteins which are secreted to the environment and benefit other cells in its environment. Those secreted proteins benefit cells in various manners, among them, communication through quorum sensing peptides[8], extra-cellular metabolism[9], or extra-cellular stress response[10,11].

Secreted proteins affect the cellular economy in a more complex manner since they break the direct relationship between cost and benefit. A cell might invest resources in producing a protein which is beneficial to the colony, but a neighboring cell can enjoy this common good, without investing resources of its own. In the simpler case of an intra-cellular gene, a cell that will optimize its cost-benefit ratio will prevail, but what would be the dynamics in the case of a secreted gene?

This question relates to a well-studied question of evolution of cooperation. Cooperation is a natural phenomenon that can be found at all levels of biological organization. However, the Darwinian theory of evolution is based on competition and survival of the fittest which should, therefore, reward selfish behavior. This raises a puzzling and fundamental question in evolutionary biology, how do systems of cooperation arise? There are various mechanisms suggested to enable the preservation of cooperation in nature, such as kin selection[71], reciprocity[72], and group selection[73].

Specifically, in the case of a unicellular organism that secrets a public-good compound, what protects the colony from invasion of a cheater strain that does not produce this compound? This question was studied either theoretically[74–77], or experimentally in microbial communities[14,16,78–81].

Secreted and membrane proteins in eukaryotes are translated by ribosomes bound to the endoplasmic reticulum (ER), to which they are directed via a hydrophobic ER-targeting sequences[82,83]. The predominant mechanism by which RNAs are transported to the ER is the signal recognition particle (SRP)[84–87] in which a hydrophobic peptide sequence at the N terminal of the gene, called the signal peptide (SP), is first translated in the cytoplasm, and as it emerges from exit channel of the ribosome it is recognized by the SRP, which is then recruited to its receptor on the ER membrane and translocation of ribosome-mRNA-nascent polypeptide chain complex from the cytoplasm to the ER occurs. The signal peptide is eventually cleaved from the final protein, but in some membrane protein the SRP interacts with instead with a transmembrane domain (TM) at the middle of the gene which is not cleaved, as this domain by nature is also composed of a stretch of hydrophobic amino acids.

While the SRP is the classical pathway for transporting RNA of secreted proteins to the ER other alternative mechanisms have been discovered and characterized in recent years[88–92]. Recently a new cis-regulatory RNA motif that enhance ER localization presumably in a translation independent manner was characterized. This motif is termed secretion-enhancing cis regulatory targeting element (SECReTE)[93], and it is composed of long three-way periodic repeats of pyrimidines (i.e. a periodic NNY template). It is presumed that this motif is recognized by some RNA binding protein (RBP) which participates in localization of the RNA to the ER. While this motif is characterized by a periodic repeat of pyrimidine, it was suggested that this pattern is a result of coding constraints within the coding region, and an optimal motif would be composed of a region with high pyrimidine content, as can be seen in SECReTE motifs which are found in the 3' UTR of secreted genes. This motif was only recently described in a single paper, and the strength of its effect is not yet known. Moreover it is not known whether the action of this motif is dependent on the SRP or other secretion pathways, or can it act independently.

One of the model systems to study cooperation through public-good secretion, is the SUC2 gene in the budding yeast *S. cerevisiae*. This gene codes for invertase which is an enzyme that hydrolyses the disaccharide sucrose into its two monosaccharide components of glucose and fructose[94]. *S. cerevisiae* cannot import sucrose into its cell, hence invertase is being secreted to the periplasmic space where it hydrolyses sucrose extracellularly and releases its degradation products to the surrounding media[9]. Hence it serves as public good, and several works have used this system to study public-good dynamics[14,15,95–97].

These previous works studied the dynamics of binary populations that included a mix of a producer strain that produces and secretes the public-good compound, and a cheater mutant that does not have the appropriate gene. This part is aimed to expand our understanding of these dynamics, by studying a heterologous population of cells that differentially secrete the common good protein *SUC2*. This continuous range of secretion is achieved using two designed synthetic libraries that modulate the secretion level of *SUC2* through modifications to two regulatory elements that control protein secretion. The first element is the classical regulatory element for protein secretion, the signal peptide. The second element is the recently characterized SECReTE motif[93]. Together, these two libraries create a population of yeast cell with continuous varying levels of production and secretion of a public-good enzyme.

6. Results

6.1. Part I – Cost of gene expression

6.1.1. Gene Architectures that Minimize Cost of Gene Expression in Bacteria

5’ gene-architecture affects cost of gene expression

My 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 [25] 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 (Fig 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 repeats. We then deep-sequenced the variable region of the pool of variants, and calculated relative fitness of each variant (Fig 1B). We regressed fitness values against GFP expression levels and observed a negative, linear correlation (Fig 1C, Pearson correlation r=-0.79, p-Value<10-200). The linear decline in fitness with expression is in agreement with previous studies [34,36]. 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 actual fitness we measured for the variant and the fitness expected for it according to its GFP expression level and the linear regression (Fig 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 (Fig 1C, blue and red dots). 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. This approach resulted in 975 positive and 815 negative variants. 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 (Fig 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 (Fig 1C, purple dots). 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.

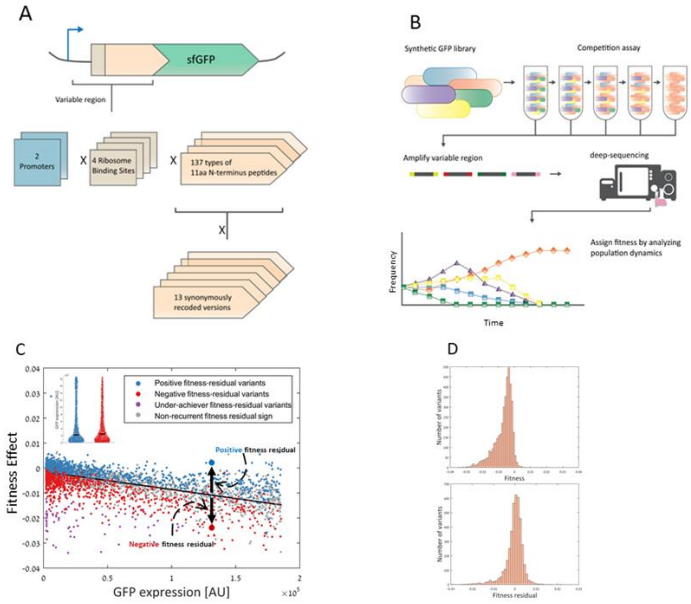

Figure 1 – 5’ gene architectures affect cost of gene expression at a given expression level
A. We utilized a synthetic library of ~14K E.coli strains, each expressing a GFP construct with a unique 5’ architecture that includes a promoter, ribosome binding site (RBS) and an 11-amino acid fused peptide. 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. 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. 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. **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.

Production of more proteins per mRNA molecule is an economic means to minimize expression costs

We first hypothesized that reaching 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 (Fig 1C, inset), we compared their GFP mRNA levels and found positive variants to have lower levels compared to negative variants (Fig 2A, Wilcoxon rank-sum p-Value=$1.6 \cdot 10^{-9}$, Effect size=58.26%). 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.

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 same GFP levels despite low mRNA levels in positive variants. We calculated initiation rates for all library variants using the "Ribosome Binding Site Calculator" [98], and observed that indeed positive variants had higher initiation rates (Fig 2B, Effect size=61.9%, Wilcoxon rank-sum p-Value=$3.7 \cdot 10^{-18}$). This observation holds true when examining mRNA level versus translation initiation rate at the individual variant level. 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 (Fig 2C,

Effect size=55.67%, Wilcoxon rank-sum p-Value=$3.4 \cdot 10^{-5}$). Moreover, we found that underachiever variants demonstrated even 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-Values=$9.6 \cdot 10^{-8}$ and $1.1 \cdot 10^{-4}$, respectively). Thus, by increasing translation efficiency, cells reduce transcription costs and hence also reduce the cost per protein.

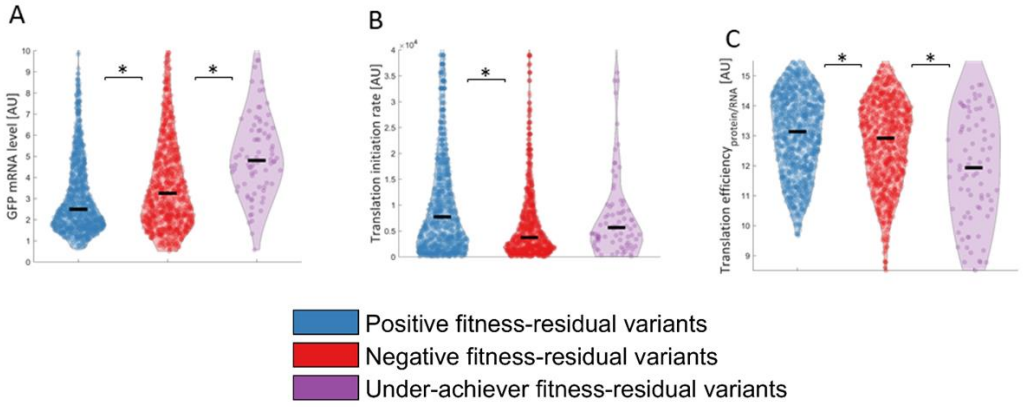

Figure 2 – Higher ratio of GFP protein/mRNA minimizes cost of gene expression
A. Although coming from the same distribution of GFP levels, positive variants (blue violin-plot) demonstrate lower mRNA levels of the GFP gene compared to negative variants (red violin-plot) (Effect size=58.26%, Wilcoxon rank-sum 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.

Translation speed at early elongation of coding region, achieved by diverse means, affects translation cost in opposing directions

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 [31,99–104]. Specifically, the prevalence of slowly translated codons at the 5' of ORFs has been suggested to support the efficiency of

gene translation [105]. 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 [106–111], 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 a measurement of ribosome elongation time per codon [111], as derived empirically from ribosome profiling data in E. coli (see Methods), to calculate translation elongation time for each library variant. We reasoned that if translational ramp is beneficial, then high elongation times, i.e. low ribosome speeds, should be more prevalent among the positive fitness residual variants. However, our results showed that positive variants demonstrate significantly longer elongation times at the N terminal fusion (Figure 3A, Effect size=59.55%, Wilcoxon rank-sum p-Value=$3 \cdot 10^{-12}$), and further for the underachievers (Effect size=64.79%, Wilcoxon rank-sum p-Value=$1.2 \cdot 10^{-5}$). Meaning that in terms of codon usage our experimental results contradict the theoretical predictions of the ramp model.
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 [25,112–114] and high affinity to the anti-Shine Dalgarno (aSD) motif of the ribosome[115]. We thus examined each of these factors separately and asked if 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 (Fig 3B). 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 up-stream 5' UTR (Fig 3C, Effect size=65.03%, Wilcoxon rank-sum p-Value=$5.4 \cdot 10^{-28}$). 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[25] 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 [115]. However, this observation has been recently questioned[116]. We next examined the effects of Shine Dalgarno (SD) 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 and found that positive fitness residual variants are characterized by low ribosome speed early in the ORF (Fig 3D, Effect size=63.82%, Wilcoxon rank-sum p-Value=$6.3 \cdot 10^{-24}$).

We thus notice three independent mechanisms that can affect the translation speed of the ribosome, and these mechanisms have opposing effects on fitness residual. According to the first mechanism which is translation elongation time as derived from codon usage, positive fitness residual variants are translated faster at their 5' end. However, according to mRNA structure and affinity to the anti-Shine Dalgarno motif positive fitness residual variants demonstrate slower translation at their 5' end. Thus, providing conflicting evidence regarding the theoretical predictions of the ramp model. We notice that effect size of the last two features (65.03% and 63.82% respectively) is stronger than the effect size of the first feature (59.55%). In addition, previous works have indicated that mRNA secondary structure has a stronger effect on ribosome elongation speed compared to codon usage[25,117].

Another means of reducing translation speed that was recently demonstrated (so far in yeast) is the incorporation of positively charged amino acids[118] or proline residues [119] 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. In summary, although the codon usage based means to ramp ribosome did not result in the expected effect on cost, two other measures, based on RNA energetics, propose two alternative ramp modes that do yield reduction in costs as a result of early ribosomal attenuation.

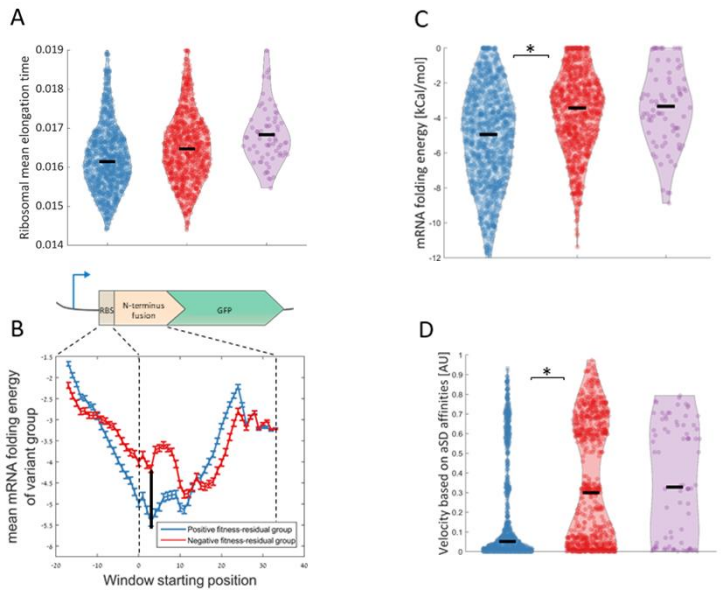

Figure 3 – Slow translation speed at early elongation, achieved by diverse molecular means, reduces expression cost

A+C+D. Positive variants show lower values of codon elongation time, 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.

Amino acid synthesis cost affects cost of gene expression

So far, we examined features that are based on the nucleotide sequence and how it associates with fitness residual. Next, we aimed at exploring 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[120]. 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 (Fig 4A, Effect size=72.74%, Wilcoxon rank-sum p-Value=$7.4 \cdot 10^{-62}$). Here, as well, underachiever variants show more expensive amino

acid usage compared to the negative group (Fig 4A, Effect size=72.75%, Wilcoxon rank-sum p-Value=$1.7 \cdot 10^{-11}$).

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. Remarkably, this frequency ratio was found to negatively correlate with the metabolic cost of each amino acid (Fig 4B, Pearson correlation r=-0.54, p-Value=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 [115] to calculate amino acid demands and utilized previously measured cellular concentrations as amino acid supplies[121]. Indeed, we found that amino acids with low demand-to-supply ratios are more prevalent in positive variants (Fig 4C, Pearson correlation r=0.82, p-Value=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 (Fig 4D, Pearson correlation r=-0.72, p-Value=$1.8 \cdot 10^{-3}$), we could not evaluate which mechanism – cost or availability - contributes more to fitness residual.

Amino acid hydrophobicity is associated with low fitness – a potential protein toxicity effect

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 [122] 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 (Fig 4E, Effect size=69.11%, Wilcoxon rank-sum p-Value=$3.2 \cdot 10^{-44}$). Underachievers showed an even more pronounced effect (Fig 4E, Effect size=81.67%, Wilcoxon rank-sum p-Value=$7.7 \cdot 10^{-21}$). 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.

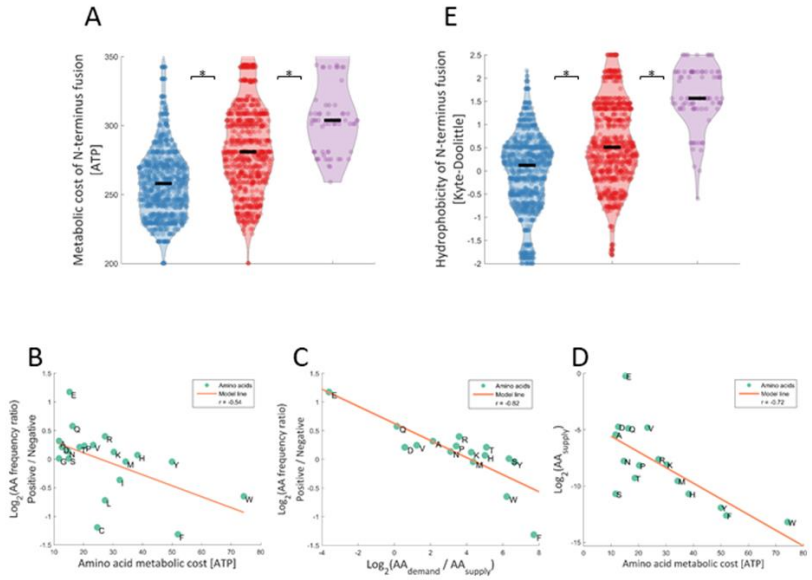

Figure 4 – Usage of expensive-to-synthetize, lowly available and hydrophobic amino acids decreases fitness residual

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-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). 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-Value=10^{-4}). Demand comes from occupancy of ribosomes on each transcript, and supply is the cellular concentration of each amino acid[121]. **D**. Amino acid availability and energetic cost are correlated (Pearson correlation r=-0.72, p-Value=$1.8 \cdot 10^{-3}$).

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

All sequence parameters contribute independently to fitness

We 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 may 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 (Fig 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 independently to the signal, even upon controlling for other factors as potential confounders.

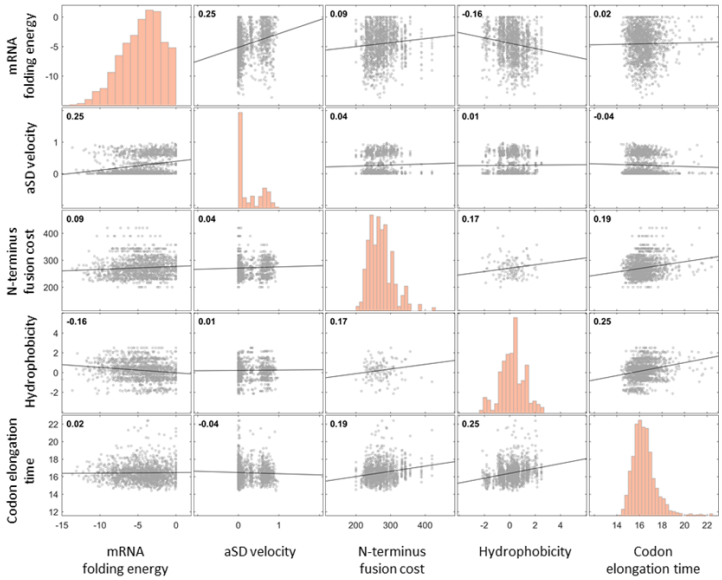

Figure 5 – Each feature affects fitness residual independently

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

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, I next asked if the mechanisms I 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, I compared the various architectural features between variants with above-average Δfitness-residual to variants with below-average Δfitness-residual.

Figure 6A-E depicts for each of the analyzed features the difference in fitness residual 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 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=$6.2 \cdot 10^{-3}$, initiation rates=$7 \cdot 10^{-9}$, codon decoding speeds=$4.3 \cdot 10^{-2}$, mRNA folding=$3.5 \cdot 10^{-16}$ and aSD velocity =$7.6 \cdot 10^{-7}$). 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.

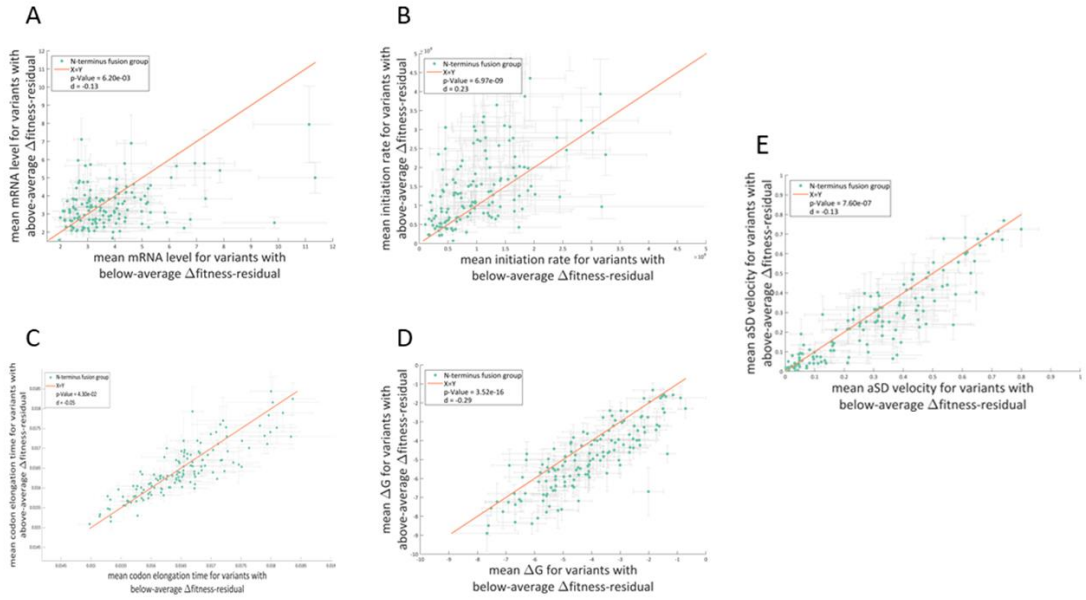

Figure 6 – Variant with same N-terminus amino acid fusion demonstrate a range of fitness residuals

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: mRNA levels $6.2 \cdot 10^{-3}$, initiation rates $7 \cdot 10^{-9}$, codon elongation time $4.3 \cdot 10^{-2}$, mRNA folding $3.5 \cdot 10^{-16}$ and aSD velocity $7.6 \cdot 10^{-7}$). d is Cohen's d that calculates the effect size.

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

So far, I examined fitness residual as a binary classification, namely categorizing variants with either positive or negative fitness residual. 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 (Fig 7A, r=0.53, p-Value<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. 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 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 (Fig 7B). This analysis revealed that the features contributing to fitness residual the most are hydrophobicity and metabolic cost of the N-terminus fusion, while ribosome elongation time contributing the least. To avoid over-fitting of our model on the library data, we performed feature selection using the Lasso algorithm. 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.

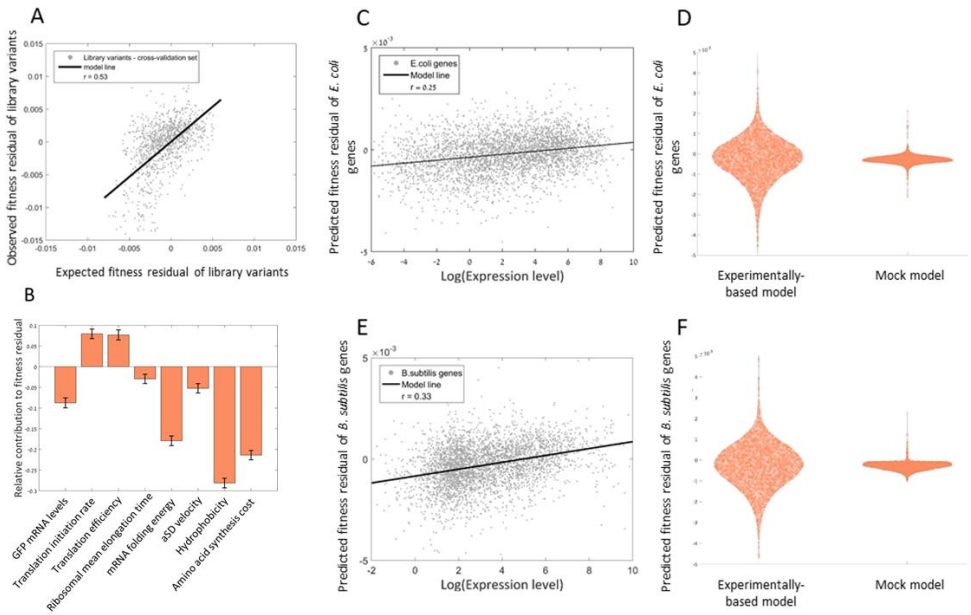

Figure 7 – A model that predicts fitness residual accurately reveals that fitness residual of natural *E.coli* genes is correlated with their expression level

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=$2.5 \cdot 10^{-12}$, TEGFP 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 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-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 residuals (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 predicted 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 next hypothesized that *E. coli* genes with higher expression levels are more likely to be endowed with cost reducing architectures. Indeed, I found a significant correlation between predicted fitness residual of *E.coli* genes and their protein expression levels (Fig 7C, r=0.25, p-Value=$2 \cdot 10^{-53}$), 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-Value=10^{-93}).
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 (Fig 7D+F, p-Value<10^{-5}). This observation suggests that the model we trained on the library data is able to expose the expression cost of natural 5' gene architectures.

6.1.2. Optimization of gene expression through promoter architecture

A synthetic promoter library is used to test how promoter architecture affects the cost of transcription

A key mechanism for regulating gene expression is through promoter architecture. Several studies have explored the effect of different promoter architectures on the expression level of a reporter gene[20,123,124]. In this project, I aim to elucidate how different promoter architectures impose different burdens on the cell. Specifically, I study how combination of different transcription factor binding sites (TFBS) contributes to the cost of gene expression independently of the cost of protein production.
I do so by utilizing a previously published synthetic promoter library expressed in *S. cerevisiae*[123]. This library includes ~2000 different synthetic promoters upstream of a YFP reporter gene. Each promoter is composed of a *cis*-regulatory element (CRE) which is built from random ligation of different TFBSs for four different transcription factors. Each transcription factor is represented by three possible sites, resulting in 12 different TFBSs

represented in the library. The four TFs represented in the library are: GCR1, MIG1, RAP1, & REB1, key transcription regulators in this organism, whose combined set of targets represent ~42% of yeast genes. In addition, all promoter variants have a basal TATA-box containing a minimal promoter downstream to the CRE. In order to recognize uniquely each variant both in DNA and RNA sequencing, a unique barcode was added to the 3' UTR of the reporter gene, and each barcode was paired to its appropriate CRE through sequencing (See Fig 8).

For the purpose of measuring the cost associated with each promoter in the library I applied the same method as in section 6.1.1, in which the relative fitness of each variant in a pooled library is measured using a competition assay. In addition to cellular fitness, I also measure the expression level of each variant using RNAseq of the barcoded region.

The competition assay was held in two growth conditions, a. optimal growth conditions on YPD media; b. YPD media with 1.7 mM DTT. The second condition was chosen since in an analysis of expression profiles of different stress conditions[125], I recognized that specifically under DTT, genes that are targets of both GCR1 & RAP1 demonstrated a coordinated expression reaction to the stress, which might imply that those TFs act coordinately as a response to this stress.

In order to support observations from this experiment, another dataset of fitness measurement in yeast was analyzed. In a recent work[6], the authors measured fitness values of 81 endogenous genes in *S.cerevisiae* under different expression levels. To drive the genes to different expression levels they have used a set of 120 synthetic promoters. While in that work the authors analyzed the data from a gene point of view, this dataset provides rich fitness measurements for 120 synthetic promoters in different genetic backgrounds. For the purpose of this project, these fitness measurements are analyzed from the promoter point of view.

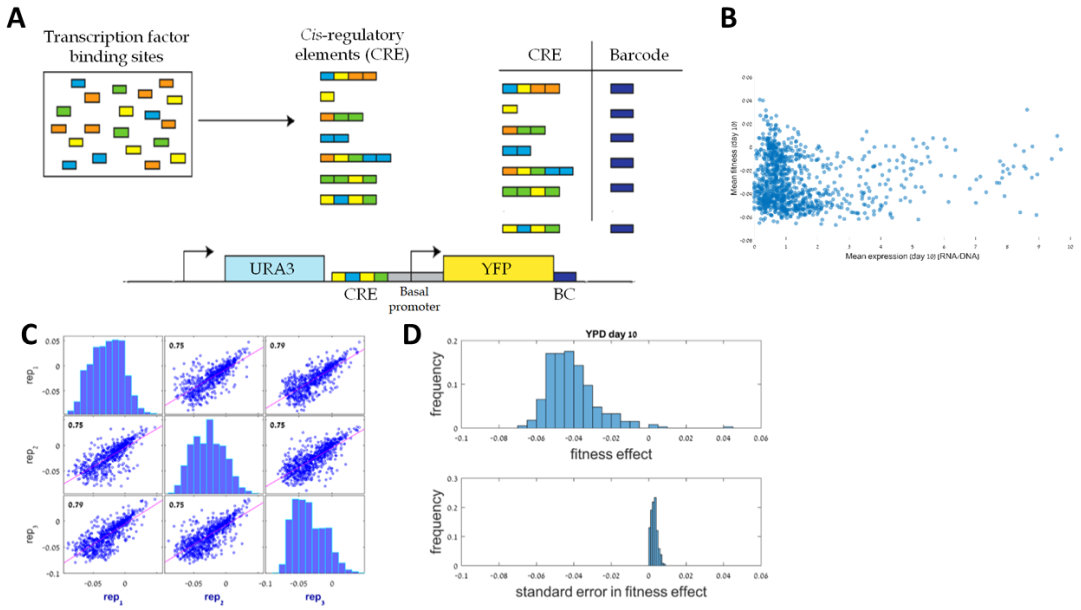

Figure 8 – Synthetic promoter library presents reproducible fitness measurements with no correlation to expression levels

A. Promoter library structure as was created by Mogno et al[123]. Double-stranded oligonucleotides encoding TFBS were mixed in a pool and ligated randomly to create a CRE library. After cloning CRE and barcode (BC) sequences into a reporter plasmid, the concordance between CREs and BCs was determined via sequencing, each BC identifies a single CRE. The cassette containing the library of CREs upstream of a basal promoter driving YFP and BC was integrated into the *S.cerevisiae* genome at the TRP1 locus. **B**. Each point in the plot represents a variant in the library. No significant correlation is observed between RNA expression level, and fitness effect. **C**. scatter plots and Pearson correlation of fitness effects from the first 3 repeats of the competition in YPD at day 10 (only 3 repeats are shown for the sake of clarity, mean Pearson correlation across all 45 pairwise correlations is 0.74). **D**. Distribution of mean fitness effect across repeats at YPD day 10 (top), and distribution of standard error (bottom). The error in mean is significantly lower than the mean fitness values.

Competition experiment

Ten independent repeats of a competition experiment were carried out in YPD media, another ten repeats were carried out in a YPD+DTT media. Each repeat was diluted in 1:120 ratio once a day (~ seven generations), and every two days a sample was frozen from each repeat. The experiment lasted for 16 days (~ 112 generations).

Following the competition experiment, genomic DNA was extracted from samples of the ancestral population, and days 6, and 10 of the competition experiment. The barcoded region

was amplified and deep-sequenced to measure the relative frequency of each variant in the library at any sample.

In addition, to measure expression level of each variant, at the start and at the end of the competition experiment, three samples from the ancestral population, and from day 10, were grown to mid-log phase and RNA and DNA were extracted.

Relative fitness of each variant is derived from the following equation:

$$f(t) = f_0 \cdot (1+s)^t \approx f_0 \cdot e^{st}$$

Where f is the variant frequency, t is the generation number and s is the fitness effect. Hence s is given by:

$$s = \frac{\ln\left(\frac{f(t)}{f_0}\right)}{t}$$

The per-cell expression level is calculated by dividing the observed frequency of the variant in the RNA sequencing, by the observed frequency in DNA sequencing.

The competition experiment in YPD and in YPD+DTT yielded similar results, and no significant promoter was enriched in one condition compared to the other. Therefore, in this report, only results from the YPD experiment are presented.

Consistent fitness values observed, but no correlation between fitness and expression levels

Following my results from the previous project[126] (Appendix 1), and previous studies[7,36,127], I expected that higher expression level in the library will manifest as lower fitness, due to the cost of expressing the unneeded reporter gene. However, the results show no correlation between expression level and cellular fitness (Fig 8B). A possible explanation for the lack of correlation is that the differences in expression loads spanned by this library are not big enough to result in a significant difference in burden in yeast cell. It should be noticed that previous studies, including ours[126], that demonstrated the cost of gene expression using such negative correlation when driving a single copy of a reporter gene to different expression levels where done in bacteria[7,127]. However, in yeast such negative correlation was observed only between strains having different copy numbers of a plasmid containing the reporter gene, and the reporter gene was expressed using a very strong promoter (TDH3)[36], resulting in a much wider range of expression levels. Hence, my present result might correspond to previous results on cost of gene expression in *S. cerevisiae* together indicating that a single gene copy expressed in yeast does not affect appreciably measured cost.

Nevertheless, fitness values are consistent across the independent repeats (Fig 8C). Meaning, the observed fitness values are a result of the genetic differences between the library variants. Hence, the synthetic promoter architecture contributes to fitness differences in the library.

Analysis of promoter features contributing to fitness

Despite the overall lack of correlation between fitness and expression I could still ask if there are features of the promoter architecture that affect fitness. First, I wish to check if the identity of the transcription factor associated with a binding site affects fitness. For that purpose, I examine the fitness distribution of all promoters with at least one binding site for a given transcription factor, compared to all promoter with no site for that transcription factor. Fig 9A, demonstrates that there is no observable effect in fitness for any specific transcription factor in the library.

I next turn to look for other promoter features that contribute to increased or reduced cost. For that purpose, I classify two groups of variants with fitness values from the extreme ends of the distribution. 337 out of 1835 of the original variants were extinct from the population until day 10, in at least 9 out of the 10 repeats and they are classified as “Extinct variants”. 289 variants have above average fitness values in at least 9 repeats and they are classified as “High-fitness variants” (Fig 9B). By comparing the two extreme groups I can examine features that are enriched in the high or low fitness group.

A significant effect is observed when comparing the fitness distributions of promoters with different total number of binding sites (irrespective of TFBS identity) (Fig 9C). Surprisingly, promoters with a single binding site have higher fitness compared to all other promoters. In addition, when comparing the distribution of number of binding sites between the high-fitness, the extinct groups, and the other promoters, it can be seen that higher fitness variants tend to have less binding sites (Fig 9D).

A potential reason for this effect is that higher number of sites in a promoter increases the demand for its associated transcription factor(s), which in turn decreases the supply of available transcription factors in the cell, for the rest of the transcriptome. However, such a model would predict that there will be a negative correlation between the number of binding sites in a promoter and its effect on fitness, and in Fig 9C no such correlation is observed for promoters with more than one binding site. This result suggests that there might be another explanation for the manner in which the number of binding sites affects cellular fitness.

Another feature which demonstrates significant difference in distribution between the high-fitness and the extinct group is predicted nucleosome occupancy (Fig 9E). Using a computational tool for predicting nucleosome occupancy[128], a predicted nucleosome occupancy for the reporter gene was calculated using the sequence as input, the sequence fragment that was taken for this calculation includes, the reporter gene, its synthetic promoter, and 1kb upstream to it. For each promoter a predicted nucleosome occupancy ratio score is calculated by dividing the number of occupied positions by the length of the promoter. In general I observe a significant (p-value = 10^{-6}; effect size (probability of superiority[129]) = 60%) trend – higher nucleosome occupancy ratio is typically associated with lower fitness. A previous work has classified natural yeast promoters into two distinct clusters, one with high nucleosome occupancy at the promoter, and the other with low nucleosome occupancy at the promoter[130]. They further demonstrate that they high occupancy promoters result in higher expression noise, and present higher histone turnover. These properties might explain why I observe higher cost for promoters with higher predicted occupancy.

Each of the TFs in this collection was represented by TFBSs with varying strengths, as gauged by the Position Weight matrix score. I next turned to examine potential correlation between this PWM score and fitness. Since promoters with a single binding site demonstrate different behavior than other promoters it is interesting to examine separately the 24 single-site variants. In Fig 9F it can be seen that there is no correlation between the fitness and the PWM score [131] of the transcription factor. Moreover, as seen for the entire library (Fig 9A), there is no observable contribution of transcription factor identity to fitness, and in addition there is no preference of one orientation over the other. Likewise, I could detect no correlation between the orientation of the motif (in forward or reverse strand) and fitness effect (Fig 9F).

In addition to the results presented in this section, other features were tested and yielded no significant contribution to the fitness differences between promoters. Those features are: Promoter's GC content; Shanon's entropy of the promoter; TFBS orientation (forward vs. reverse); Consensus sites vs. non-consensus sites. Further, analysis of the following features results in significant fitness differences, but the observed contribution can be explained by correlation of the feature to one of the features mentioned above: Number of unique transcription factors represented in a promoter (regardless of number of sites, this feature is highly correlated to the number of TFBS); Minimal edit distance to natural promoters in the genome, where each promoter is represented by its TFBS arrangement (this features is

correlated to the number of TFBS, promoters with low number of TFBS trivially, have lower edit distance to natural promoters). In conclusion, I mainly observe an intriguing association between promoters with a single binding site, and reduced cost. However, the lack of effect for the number of sites in the case of multiple site promoters, and the lack of plausible mechanism to explain this observation, leave this observation with convincing enough evidence.

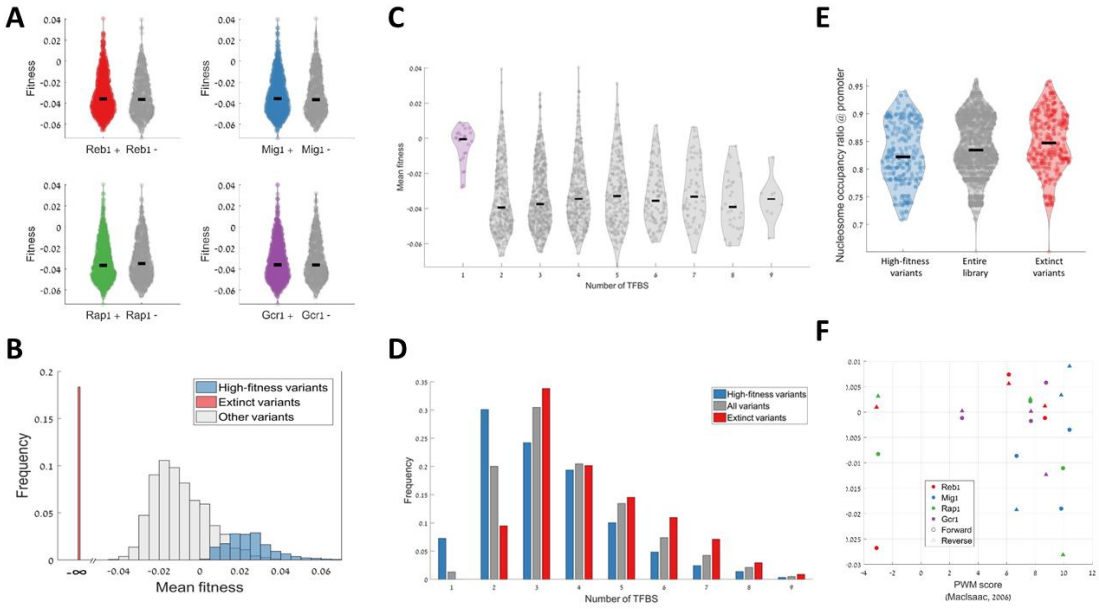

Figure 9 – number of TFBS in promoter affects fitness regardless of TF identity

A. In each plot, the fitness distribution of all promoters that include a site for the given TF (colored violin), is compared to the fitness distribution of all promoters without a site for that TF (grey violin). Black line represents the median of the distribution. **B.** Variants which are extinct from the population in at least 9 out of 10 repeats are classified as “Extinct variants” (red). Variants with fitness value above the mean fitness of a repeat in at least 9 out of 10 repeats are classified as “High-fitness variants” (blue). **C.** fitness distribution for promoters grouped by their number of binding sites. **D.** the distribution of number of binding sites for the High-fitness group (blue), extinct group (red), and the entire population (grey). **E.** Violin plots represent the predicted nucleosome occupancy ratio of the high-fitness variants (blue), extinct variants (red), and the entire library (grey). Black horizontal line represents the median of the distribution. **F.** Each dot in the plot is a promoter with a single TFBS. Each of the 4 transcription factors is represented by 6 binding sites, 3 possible sequences, each of them in a forward (circle) or reverse (triangle) orientation. The x-axis represents the PWM score as was calculated by [123] based on the PWM matrix in [131]. The y-axis represents the mean fitness effect of each promoter as was measured in the competition experiment.

A second dataset of promoter's contribution to fitness corroborates the effect of number of binding sites

To further examine the effect of promoter architecture on cellular fitness I turned to analyze an existing dataset of fitness measurements of synthetic promoters in *S. cerevisiae*. In a recent work[6] the authors measured the effect of endogenous gene expression level on cellular fitness for 81 genes. They did so by creating a synthetic construct in which an endogenous gene is transcribed using a set of 120 synthetic promoters, each with a known expression level from a previous research[20], following by a competition experiment of the entire library. This research provided a fitness Vs. expression curve for each endogenous gene (Fig 10), but naturally not every promoter variant fits perfectly on the gene's curve.

For the purpose of this project I utilize this dataset to analyze its fitness data from the promoter point of view. For each synthetic promoter 81 fitness values were measured when driving different endogenous genes, and for each of these genes an expected fitness at a given expression level can be extracted from the analysis done in the original paper[6]. Hence, for each promoter a fitness-residual score can be calculated for each of the measured genes by subtracting the expected fitness for this gene given its fitness-expression curve from the measured fitness of this promoter-gene variant (see Fig 10D for example). Using the fitness-residual score I can control for the endogenous gene expression contribution to cellular fitness and isolate the contribution of the synthetic promoter itself to fitness. After sorting all 120 promoters according to their median fitness-residual score (Fig 11A), I classify the bottom 23 promoters as negative fitness residual promoters, and the top 31 promoters as positive fitness-residual promoters (See Fig 11 legend for details).

Similarly, to Fig 9C,D I turn to look for the effect of the number transcription factor binding sites on the promoter's fitness residual scores. It can be seen in Fig 11B,C that as in my current experiment this dataset also provides evidence for a correlation between the number of binding sites, to the promoter's contribution to cellular fitness. As opposed to the current experiment, here promoters with more than 2 binding sites have even lower fitness-residual scores, however, there is a low number of many-TF variants in this library (n=12), so it would be hard to base a conclusion based on this observation.

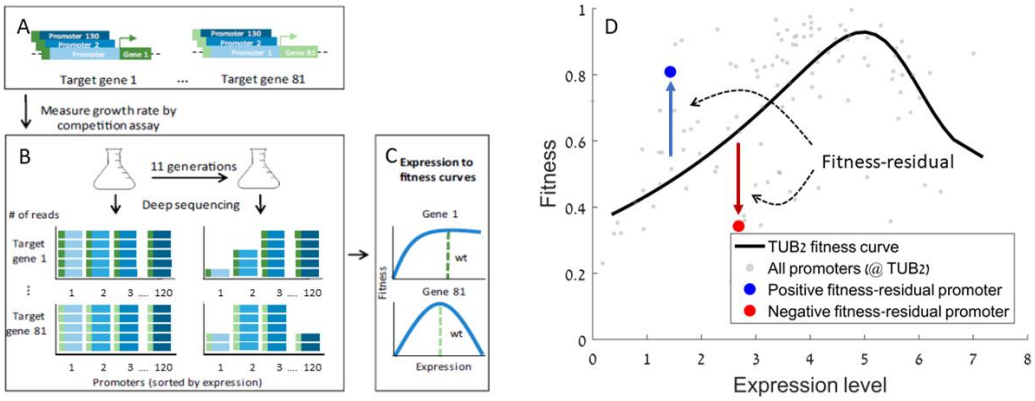

Figure 10 - Fitness measurements for 120 synthetic promoters driving 81 endogenous genes

A. A library of 120 synthetic promoters driving 81 endogenous genes was created by Keren et al[6]. When the expression level achieved by each synthetic promoter was previously measured[20]. **B.** The fitness of each promoter-gene in the library was measured using a pooled competition experiment. **C.** For each gene, a curve of fitness as a function of expression was created using the 120 fitness measurements for each gene and using some filtering techniques. **D.** An example of a fitness-expression curve for TUB2 gene. Each grey dot represents a fitness measurement for a variant with a different synthetic promoter driving this gene. For each promoter a fitness residual score is calculated by subtracting its expected fitness given the curve from the measured fitness value (blue, and red dots are positive, and negative fitness-residual examples).

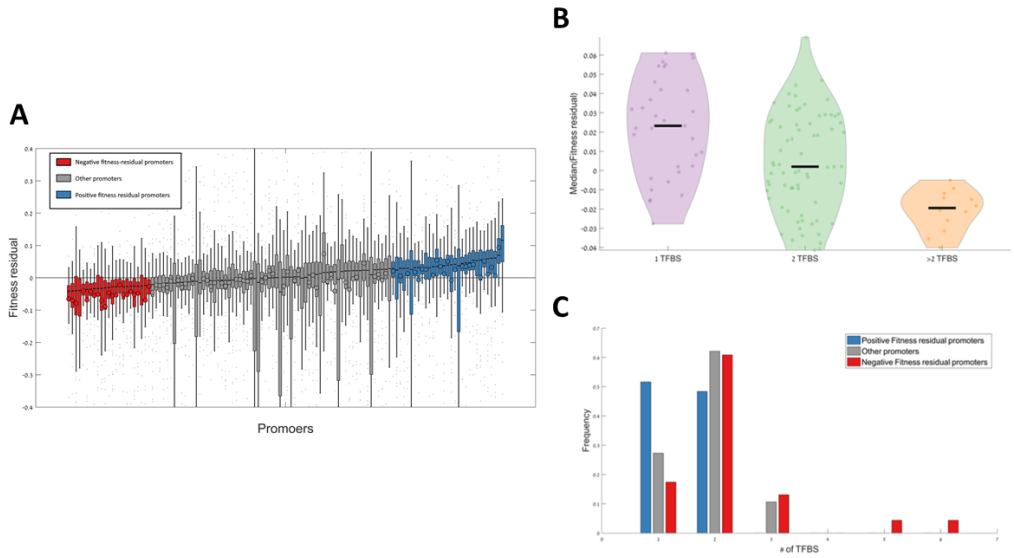

Figure 11 - In a second experiment, promoters with less binding sites demonstrate higher fitness

A. All 120 promoters sorted by their median fitness residual. Each box represents the distribution of fitness-residual scores for a specific promoter taken from 81 promoter-gene variants. In each box, vertical limits represent 1st and 3rd quartiles, horizontal line represents median value, and circle represents mean value. A group of negative fitness-residual promoters (red) is defined as all the

promoters with median value below the highest promoter for which the 3rd quartile value is negative. Respectively a group of positive fitness-residual promoters (blue) is defined by as all the promoters with median value above the lowest promoter for which the 1st quartile value is positive. **B.** Promoters are separated according to their number of transcription factor binding sites (1,2, or >2). For each group the distribution of median fitness-residual score is depicted with a violin plot. The single-TF group has significantly higher fitness-residual distribution compared to the 2-TF group. **C.** The distribution of the number of binding sites is plotted for the positive fitness-residual group (blue), negative fitness-residual group (red), and the others group (grey). It can be seen that the positive group is enriched with single-TF promoters.

6.1.3. Synthetic Intron library have no significant effect on other cellular functions in-*trans*

Another important aspect of gene expression regulation as described in Part II of this thesis is splicing regulation. Here I wish to test if introduction of a synthetic intron into the yeast genome affects cellular fitness and/or the splicing efficiency of other intron-containing genes in-*trans*, and if it is dependent on the intron architecture. For that purpose, I used the synthetic library described in part II, and applied to it two more high-throughput assays. I first performed a competition assay like I did to the other libraries to measure the relative fitness of each variant in the library. Next, in order to check the effect of intron architecture on other intron containing genes, I inserted into the library's yeast strain an intron-containing fluorescent reporter, with a fixed intron[64]. Then, after integration of the library, this reporter can serve as a reporter for the effect of the library's intron on splicing efficiency of this probe, which may reflect on splicing of the natural intron-containing genes in the genome. Introduction of a new intron to the genome might affect cellular fitness through affecting supply and demand economy of the shared splicing machinery[132–134], or through direct energetic costs involved in the spliceosome activity. In section 6.1.1, I demonstrated how subtle changes at the 5' UTR of a gene affect cellular fitness in bacteria. Hence, since in eukaryotic cells mRNA splicing involves a major molecular machine, I hypothesized that changes in intron architecture might impose a different burden on the cell. I measured cellular fitness of each variant in the library using a serial dilution competition assay of the entire library in a pool for 100 generations, followed by sequencing of the barcodes at 5 timepoints, then I calculated fitness using a log-linear regression of the relative frequencies of each variant at the measured timepoints. I performed the competition assay in three independent repeats, and additionally, each design was cloned with 4 different unique 8 nucleotides as an index, so each variant's fitness is measured independently 4 times within each repeat. Fitness

measurements between repeats or indexes have a positive weak correlation (mean pearson correlation between repeats r=0.12, and between indexes r=0.05). This suggests that any true fitness differences in the library are small and masked by the systematic noise in the system. Focusing on variants with consistent fitness measurement between repeats (~20% of the library), revealed no effect of splicing on cellular fitness (Fig 12B). Hence, I conclude that in my system introduction of one new synthetic intron imposes no apparent burden on the cell.

Next, I wanted to test if introduction of the library's synthetic intron influences splicing efficiency of other intron-containing genes. For this purpose, I inserted a reporter cassette to another location in the genome, with two fluorescent reporters. A YFP gene with an intron, such that only when spliced the reporter is expressed, and an mCherry to normalize for biological noise. For each variant in my library, I measured the YFP fluorescent levels that represent splicing efficiency of an intron-containing gene in-trans and check if they are influenced by its own library's intron architecture and splicing efficiency. YFP fluorescent levels were measured in a pool for the entire library by sorting the cells in a flow cytometer into 12 bins according to their YFP/mCherry ratio and sequencing the barcodes present in each bin. Fluorescent levels of each variant are inferred by reconstructing its distribution from its relative frequency in each bin.

By comparing each variant's own splicing efficiency to the YFP level I observe that there is hardly any effect of the variable synthetic intron on splicing efficiency in-trans. I do observe a significant but weak negative correlation (Pearson Correlation -0.085) between splicing efficiency and YFP expression (Fig 12A), which might suggest that high splicing efficiency might come at the expense of other intron-containing genes. But due to the low level of the correlation, I cannot conclude that the introduction of a new synthetic intron affects appreciably other intron-containing genes in my system.

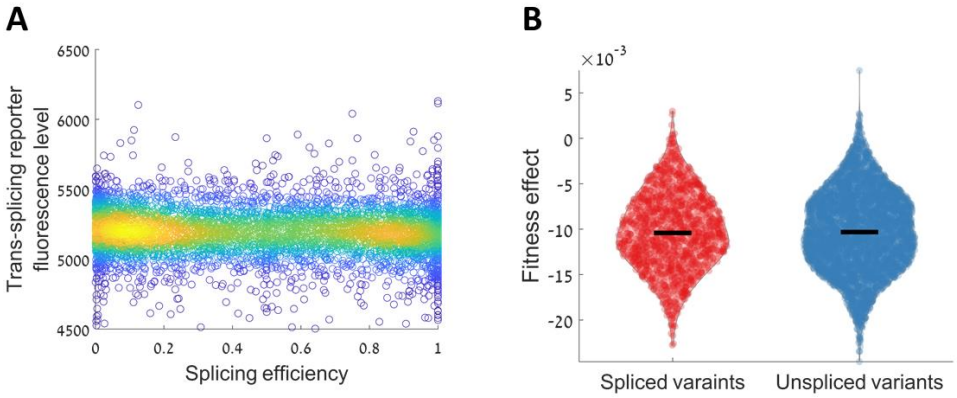

Figure 12 - Two high throughput assays for measuring effect of intron architecture on cellular functions

A. Library's intron splicing efficiency Vs. trans-splicing reporter fluorescent levels, colors represent density (Pearson r = -0.085, P<10-8). **B**. Fitness effect distribution for spliced variants (left), and unspliced variants (right). There's no significant difference between the two distributions (t-test, p-value=0.7)

6.1.4. T-rich elements contribute positively to fitness, and result in surprising connection between fitness and gene expression

In section 6.1.1 I have demonstrated how amino acid composition affects the cost of protein expression in *E. coli*. Next, when focusing only on mRNA production I wish to check if nucleotide composition has any effect on mRNA production costs in the yeast library. These costs might source from direct synthesis cost of the nucleotides, but it might also be affected from difference in the energy needed to separate double stranded DNA for GC-rich, or AT-rich sequences, as the melting temperature of GC pairs is higher than the melting temperature of AT pairs. In addition, nucleotide composition might affect properties of the mRNA itself like mRNA stability which also relates to expression costs per RNA mulecule, as a less stable RNA will result in a higher turnover rate of RNA molecules, to reach the same steady state levels.

For this purpose, I have designed a synthetic library of ~13,000 different variants spanning different nucleotide compositions. This library was synthesized together with the splicing library mentioned in Part I, and was cloned to the same vector, and integrated to the same genomic location.

As in research works aimed to measure the cost of gene expression, I measure the cost by expressing an unneeded gene, hence its contribution to fitness should be only through its

burden to the cell. Here, the unneeded gene is a synthetic non-coding RNA that was created by removing any ATG codons (at all possible frames), from an existing gene from the *S. cerevisiae* genome (*MUD1*). It is expected that the cost of gene expression will be positively correlated with expression level (i.e. the fitness will be negatively correlated), as was demonstrated in several previous works [7,36,135]. It is also possible that the fitness will not be correlated with expression level if the expression load is not high enough to affect the fitness as was demonstrated in section 6.1.2.

However, surprisingly I observe a positive correlation between mRNA levels and fitness (Fig 13A). It is highly unlikely that higher expression of a synthetic non-coding gene will cause directly to higher fitness, since I assume this gene is not beneficial. On the contrary, as explained above I predict a negative correlation between expression level and fitness. A possible explanation would be a causal effect of fitness on expression level, meaning when the cells are growing faster they also increase globally the rate of transcription which is observed through the expression levels of the library's gene. A second possible explanation would be a confounding factor that increases both the expression level of the gene, and its fitness. Such factor could be a feature that increases the stability of the mRNA, which will also increase steady state RNA levels, and might also result in reduced cost as it has lower demand from the RNA degradation machinery.

Looking at features that contribute to fitness, I see that fitness is strongly correlated with the variant's T content (Fig 13B), which also correlates with mRNA levels (Fig 13C). This correlation might be explained by the total T content in the variable region, or by the length of longest uninterrupted repeat of Ts (Fig 13D). It might be possible that T content serves as a confounding factor as it might influence both transcription rate, and fitness separately, thus creating apparent correlation between the two. However, partial correlation analysis ($r=0.27$, p-value$<10^{-35}$) rules out this possibility, since according to it fitness and expression level are correlated even whn the T content is held constant.

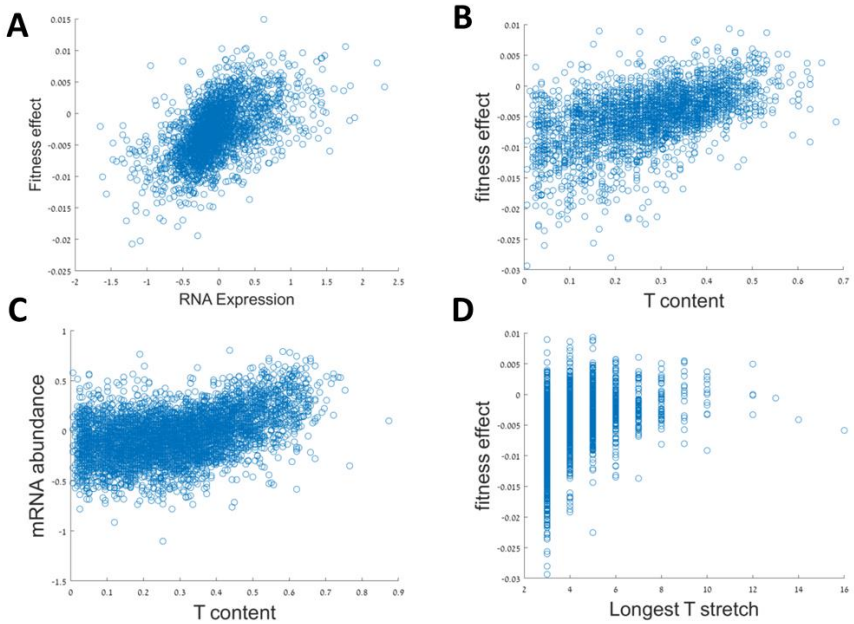

Figure 13 – T-rich elements in non-coding gene correlate with mRNA abundance and fitness

A. Scatter plot of mRNA abundance levels Vs. fitness effect for ~2,500 library variants with significant fitness measurements (Pearson r=0.42, p-value<10^{-100}). **B**. Scatter plot of the library's variable region's T content Vs. fitness effect (Pearson r=0.53, p-value<10^{-100}). **C**. Scatter plot of the library's variable region's T content Vs. mRNA abundance (Pearson r=0.44, p-value<10^{-100}) **D**. Scatter plot of the library's variable region's longest uninterrupted stretch of Ts Vs. fitness effect (Pearson r=0.37, p-value<10^{-85})

This observation of positive correlation between mRNA levels and fitness, which is related to T content poses an intriguing open question. I hypothesized that this effect might be connected to RNA stability, as a previous work demonstrated that U-rich elements at the 3' UTR stabilize the mRNA through base pairing with the polyA tail[136]. Such stabilization of the mRNA would naturally increase mRNA levels, and it might contribute to fitness relative to other variants in the library, since those other variants have higher RNA degradation costs. An alternative explanation to this correlation is the fact that in the growth media used for the competition assay (SD complete), Uracil is provided. Since high expression levels of high Thymine content variants will be manifested in higher demand for Uracil for production of the mRNA, this might explain why variants with high T have relative advantage over other variants. On the other hand, cells growing in this growth medium are also provided with Adenine, and I do not observe such relative advantage for high A content variants.

To test the effect of the growth media, I repeated the competition experiment for this library in two additional growth media: a rich medium (YPD), and a synthetic defined media without

addition of Uracil and Adenine (SD -UA). If the relative advantage of high T content variants was a result of the Uracil addition, I would expect to see this advantage decrease in the SD -UA media as no Uracil is provided for the cell. Moreover, in the rich medium I also expect the fitness advantage to decrease since in this case all the nucleic acids are provided in the media, and there should not be an advantage to high T variants over other variants. It can be seen in Fig 14 that the strong correlations between T content and fitness that were observed in SD complete media (Fig13) were not reproduced in the two other growth media. mRNA levels were not measured for the two additional growth media. This observation indicates that most probably the relative advantage for high T content variants was a result of supplying the cells with Uracil in the growth media, and therefore, I decided to not follow through and further investigate this result.

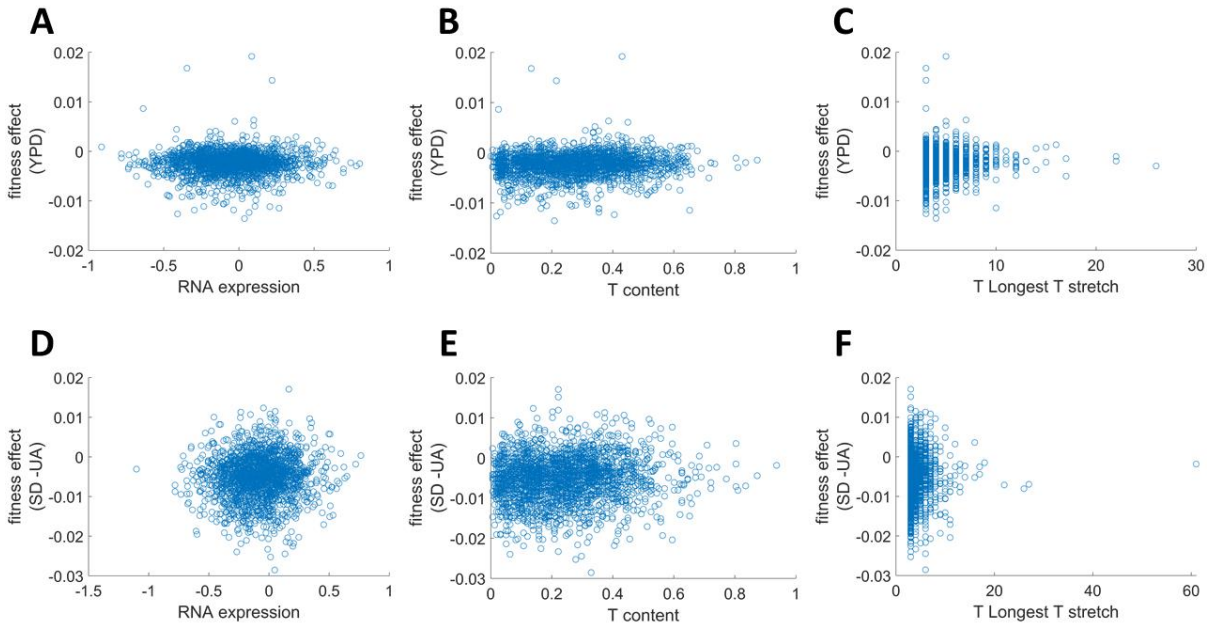

Figure 14 – Different growth media do not reproduce the fitness advantage of high T content variants

A-C. the same as Figures 13A,B,D respectively, but for rich media (YPD). Fitness vs. mRNA level correlation is not significant (p-value = 0.5). Fitness Vs. T content present a weak positive correlation (r=0.09, p-value<10^{-4}). Fitness Vs. longest T stretch present a weak positive correlation (r=0.06, p-value=0.006). **D-F**. the same as Figures 13A,B,D respectively, but for synthetically defined media lacking Uracil and Adenine (SD -UA). Fitness vs. mRNA level present weak correlation (r=0.06, p-value = 0.01). Fitness Vs. T content present a weak positive correlation (r=0.08, p-value<10^{-4}). Fitness Vs. longest T stretch correlation is not significant (p-value>0.05).

6.2. Part II – Sequence determinants and evolution of splicing in yeast species

High-throughput splicing efficiency measurements of thousands of synthetic introns

To explore how the intron architecture affects splicing efficiency, I designed a synthetic oligonucleotide library[19] of 18,061 variants. All the oligonucleotides were cloned into the same location inside a synthetic non-coding gene that was then integrated into the yeast genome. I cloned the library into a non-coding gene to avoid any differences between variants that might result from differences in translation.

Each oligo consists of fixed primers for amplification and cloning, a unique 12nt barcode, and a 158nt oligo that includes a unique intron sequence design within it. All synthetic intron sequences were introduced into a mutated version of the natural *S. cerevisiae* gene MUD1 lacking any ATG codon in all reading frames. The expression of the gene is driven by a synthetic promoter that was chosen from an existing promoter library [20,137] based on its high expression and low noise characteristics. The entire intron-containing gene library was integrated into the YBR209W dubious open reading frame in *S. cerevisiae* genome using a high-throughput Cre-Lox based method[138] (Fig 15).

Each intron design in the library is characterized by the sequence of its three functional sites (5'SS, BS, 3'SS), its length, the distance between its BS and 3'SS, and by the length of a short U-rich element upstream to the 3'SS.

Library subset	**Number of variants**
Synthetic Combinatorial designed introns	4,713
Synthetic introns with splice sites mutations	4,505
Synthetic introns with consensus splice sites and different background sequences	1,377
Natural introns' sequences from yeast species	1,173
Natural introns with mutated splice sites	1,328
Two-introns designs	823
Negative control	4,142

Table 1 - Summary of the different subsets composing the library and the number of variants in each of them

The library was composed of four major subsets (see Table 1): the first subset was created in a combinatorial design by introducing different splice site sequences with their exact sequence as observed in the genome, on the background of the *MUD1* derived gene. Introns were created with different lengths and different BS-to-3'SS length that represent the length characteristics of introns from non-ribosomal genes. In each oligo, an intron was created by replacing the background sequence at positions 6-11 with a 5'SS sequence, and then according to a choice of intron length, and BS-to-3'SS a BS sequence and a 3'SS were placed instead of the background sequence at corresponding positions. In addition, three versions of short poly uridine sequence were inserted upstream to the 3'SS. The second subset was based on perturbations to the former subset by introducing mutations to the genome-observed splice site sequences, in addition a set of synthetic variants was created by introducing only consensus splice sites sequences at varying length properties, within different background sequences. A third subset of the library was composed of introns that naturally occur in *S. cerevisiae* and in other yeast species, these introns' sequences were inserted into the 158 synthetic oligo, such that only the intron sequence was inserted instead of the background sequence at the 5' end of the oligo (introns longer than 158 nucleotides were not taken for this set). Lastly, the fourth subset of variants was composed of designs with two short introns, one next to the other, separated by a short exon. The introns' sequences used for this subset were natural introns taken from *S. cerevisiae* genome that are short enough to fit with another intron inside the 158 oligo. This last subset enables me to study the potential of the *S. cerevisiae* splicing machinery to process genes with multiple introns and to create alternative splice variants.

Splicing efficiency of each variant was measured using targeted RNA sequencing of the library's variable region. The sequence amplicon that was deep-sequenced included both the unique barcode of each intron design and its entire variable region, in either its unspliced or spliced forms. This allowed me to calculate splicing efficiency for each intron design. Shortly, each variant was identified by its unique barcode, and then the relative abundances of the spliced and unspliced isoforms were determined by aligning exon-intron and exon-exon junction sequences against the RNAseq reads. Splicing efficiency was defined as the ratio between the spliced isoform abundance and total RNA abundance. If for a specific RNA read, neither an unspliced isoform nor a spliced isoform of the designed intron were identified, I searched for an mRNA isoform that might have been created by a novel splicing

event. This was achieved by identifying uninterrupted gaps in the alignment of the full sequence to the RNAseq read.

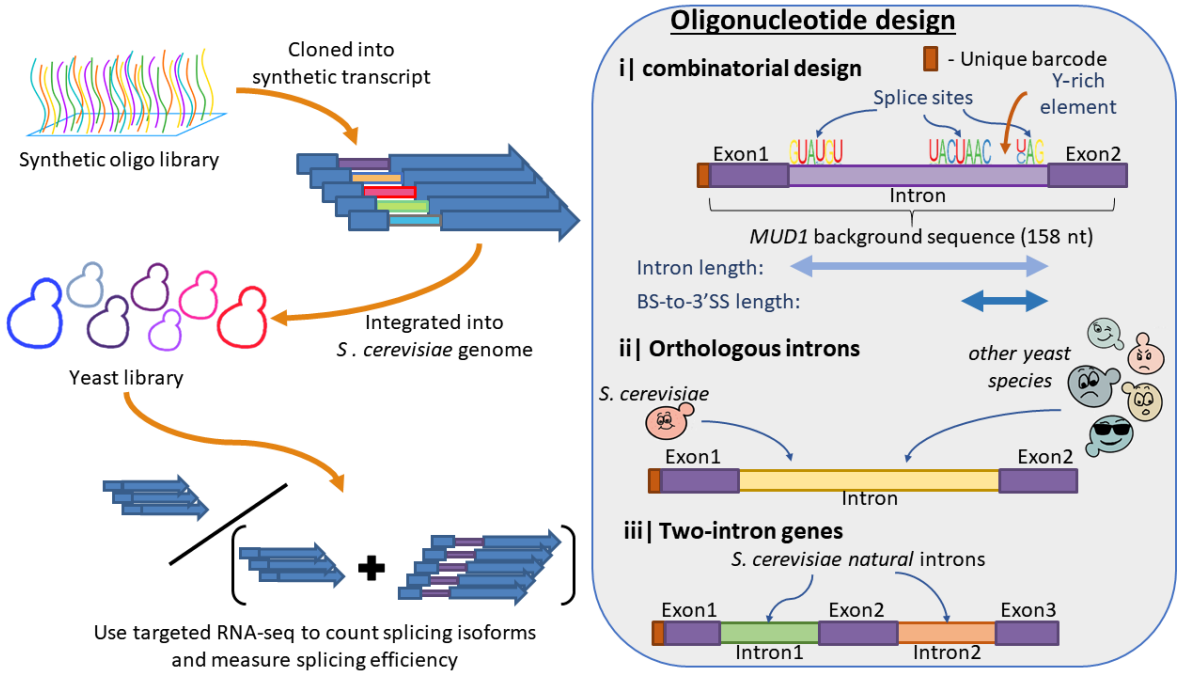

Figure 15 - A designed synthetic intron library in budding yeast

A large oligonucleotide library of designed introns was synthesized and cloned into a synthetic non-coding gene. The gene was then integrated into the budding yeast S. cerevisiae genome, to create a pooled yeast library. Then splicing efficiency was measured using targeted RNAseq of the intron's region, identification of RNAseq reads according to the barcode, and spliced isoform identification using alignment of exon-intron and exon-exon junctions. Inset: oligonucleotide design strategy - All oligos were identified using a unique 12nt random barcode at their 5' end; i) A set of synthetic introns was created by combinatorial design of different splice site sequences and other intronic features; ii) A set of natural introns from S. cerevisiae and other 10 yeast species was introduced into the library; iii) a set of synthetic two-intron genes was created by pairing together short intron sequences with an exon separating between them.

Synthetic introns are successfully spliced within the genomic construct

13,096 variants in the library were designed with a single intron (The remaining 4,695 are negative control variants or two-intron variants). For 33.5% of these single intron variants, I observed the designed splice isoform with median splicing efficiency of 0.428, where a value of 1 means that all RNA reads are from the spliced isoform. For comparison, 4,142 variants were designed as negative controls. These negative control variants are not expected to be spliced, as they were designed by introducing random sequences instead of splice site sequences. In this negative control set, only 1.9% yield a spliced isoform with a median

splicing efficiency of 0.039. In addition, when examining the natural introns of *S. cerevisiae* included in my library, I see that 84% of them yielded spliced isoforms, with median splicing efficiency of 0.675.

To verify that the introduction of a 12nt barcode upstream of the intron does not have a significant effect on splicing efficiency, I attached four different barcodes for a randomly chosen set of 517 designs. When comparing the variance in splicing efficiency of these quartets (considering only designs with non-zero splicing efficiency) to a set of randomly chosen quartets, I see that the mean variance of the multiple barcode set is much lower than the mean variance of each of 10^4 randomly shuffled variants quartets from the same set (data not shown), indicating that the barcode choice exerts at most a low effect on splicing efficiency.

Splicing efficiency is positively correlated with RNA abundance

When examining total RNA abundance, I see a positive correlation between splicing efficiency and total RNA abundance (i.e. summed level of unspliced and spliced isoforms) for the set of variants that are successfully spliced (Fig 16A). Since the calculation for splicing efficiency is dependent on total RNA abundance, I compare this result to a random set taken from the same distribution which does not present any correlation. This observation that total mRNA levels are correlated with splicing efficiency is interesting since in *S. cerevisiae* most genes do not contain an intron, and many genes have high expression despite having no intron, so one might expect that the act of splicing would not affect the levels of total RNA. Moreover, this experiment was done on synthetic genes that were not selected in evolution to regulate their gene expression through splicing. Thus, the correlation observed here between RNA expression level and splicing efficiency suggests a molecular mechanism at work. For example, the correlation might be explained by effects of splicing on nuclear export[139], or on RNA stability[140]. In particular, it may be suggested that variants for which RNA molecules are spliced are exported more efficiently and hence splicing can enhance steady-state RNA levels.

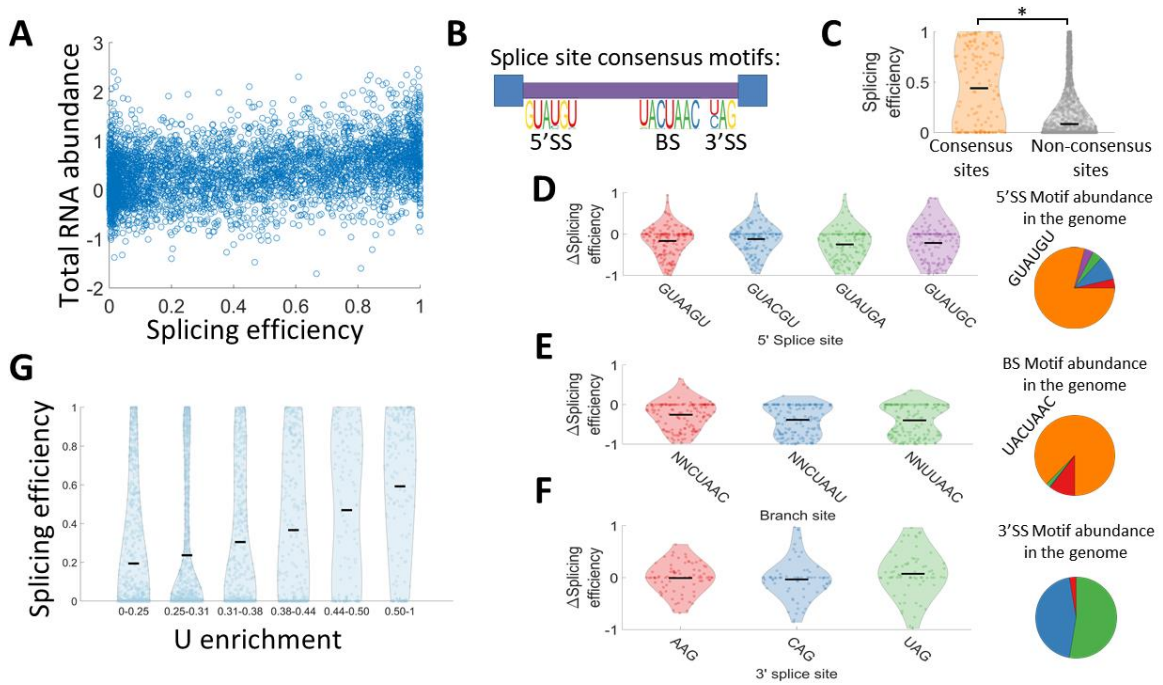

Figure 16 - A set of combinatorically designed synthetic introns elucidate contribution of different features to splicing efficiency

A. Scatter plot of the total RNA abundance and splicing efficiency for all the variants with splicing efficiency >0 shows a significant positive correlation between RNA level and splicing efficiency (Pearson r=0.4 p-value<10-72). **B.** Splice site motifs for S. cerevisiae introns, one can notice a dominant consensus sequence for the 5'SS and BS, and two consensus sequences for the 3'SS. **C.** Distribution of splicing efficiency for synthetic intron variants with consensus splice sites (orange), is significantly higher than splicing efficiency distribution for non-consensus splice sites (grey) (two sample t-test, p-value<10-80). Black horizontal line represents mean splicing efficiency. **D-F.** effect of non-consensus splice site sequences on splicing efficiency. Violin plots represent the distribution of the difference in splicing efficiency between a variant with a single non-consensus splice site, to a corresponding consensus sites variant which is identical in any other parameter. For the 3'SS since there are two consensus sequences, AAG variants were compared against the average of the two consensus variants, and CAG/UAG variants were compared against the other consensus variant. Pie charts show the relative abundance of each splice site in S. cerevisiae genome (orange slice represents the consensus site, and other colors correspond to the colors in the violin plots). In the case of the BS, NNCUAAC non-consensus variants represent all sequences that fit this template but different from the consensus sequence UACUAAC. **G.** Splicing efficiency distribution is binned according to poly uridine tract enrichment, which is calculated as the U content at a window of 20 nucleotides upstream to the 3'SS. I see a significant positive correlation, compared to a non-significant correlation for Y-rich elements (data not shown, Pearson r=0.95 p-value=4·10-3).

Combinatorial design of introns elucidates features contribution to splicing efficiency

I first analyzed the set of synthetic introns created by combinatorial design of different splice site sequences and length properties. I noticed that introns that contain the consensus splice site sequence in all three splice sites are better spliced than introns with at least one non-consensus splice site (Fig 16B,C). Next, for each of the non-consensus splice site variants I examined how it affects splicing efficiency by analyzing variants with a single non-consensus splice site, and comparing their splicing efficiency to the corresponding design with consensus splice sites and otherwise identical sequence (Fig 16D-F). I notice that almost all non-consensus branch site sequences result in much worse splicing, meaning that the consensus splice site is crucial for efficient splicing (Fig 16E). On the other hand, in the 5'SS while on average the non-consensus variants are spliced less efficiently I do observe a substantial number of variants that are spliced better than the corresponding variant with consensus site (Fig 16D). I also notice that for two of the splice site variants, lower splicing efficiency is observed only for longer introns (data not shown). For the 3'SS I see that there is no measurable difference in splicing efficiency between the three variants found in the genome, although two of them are significantly more abundant than the third (Fig 16F). The fact that the AAG 3'SS variants are spliced as well as the two YAG 3'SS variants is surprising due to the fact that ~95% of introns in all eukaryotes use a YAG 3'SS[141]. I used a set of variants with random mutations in their splice sites, to analyze the effect on splicing efficiency of all possible single nucleotide mutations in the three splice sites, and this analysis replicated the results observed for the splice sites variants found in the genome (data not shown).

Next, I examined how other intron features can affect splicing efficiency. In other eukaryotes, further than the splice sites, an intron is characterized by a poly-pyrimidine tract upstream to the 3'SS[142], However, it was noticed that in yeast a weaker feature is observed, and it is characterized by short uridine-rich sequence instead of pyrimidine (U or C)[53,143]. Using my library I examine the effect of uridine rich sequences upstream to the 3'SS, by binning variants according to their U content in a 20nt window upstream to the 3'SS, and comparing the splicing efficiency distribution in each bin (Fig 16G). I noticed that higher U content in this window is correlated with increased splicing efficiency. To demonstrate that the observed effect is due to specifically uridine enrichment and not generally pyrimidine, I repeat the same analysis by binning according to Y content, considering only variants with well balanced U and C composition, and I observe no correlation between Y enrichment and

splicing efficiency (Fig 17A). This result is the first experimental evidence that *S. cerevisiae* splicing machinery is specifically affected by a poly uridine tract, as opposed to other eukaryotes[142].

Previous studies have associated other intronic features with splicing efficiency: intron length[144,145], the distance between the branch site and the 3'SS (BS-to-3'SS length)[47,48,146], and intronic GC content[47,64,147,148]. The data from this library significantly supports the effect of intronic GC content (Fig 17B). As for intron length, I do not observe a specific length that is spliced more efficiently (Fig 17C). I note that introns taken for this library were bounded by a length of 158nt and the distribution of intron lengths represented in this library represents the length distribution of introns from non-ribosomal genes in *S. cerevisiae*. Introns from ribosomal genes are longer (mean intron length of ~400 nt), and these lengths are not represented in this work. I also did not observe a significant preference for a BS-to-3'SS length in terms of splicing efficiency.

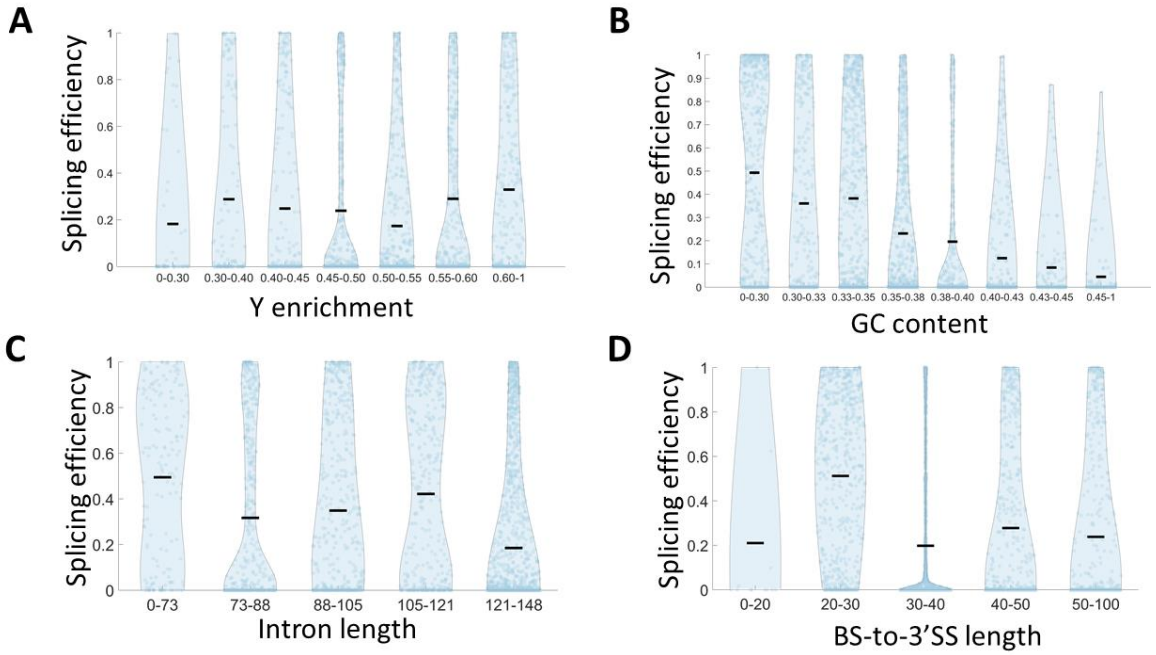

Figure 17 - A set of combinatorically designed synthetic introns elucidate contribution of different features to splicing efficiency

A. Splicing efficiency distribution is binned according to poly-pyrimidine tract strength , which is calculated as the Y (i.e. C or U) content at a window of 20 nucleotides upstream to the 3'SS. In order to specifically check elements that are not U-rich, only elements with at least 30% C out of their Y content are taken into account. Correlation is not significant (p-value=0.79), compared to the highly significant correlation for U-rich elements (Fig 3GD). **B.** Splicing efficiency distribution binned according to intronic GC content (Pearson r=-0.85 p-value<0.01). **C-D.** Splicing efficiency distribution of spliced variants for (C) different intron lengths, or (D) different BS-to-3'SS length.

Cryptic splicing events drive intron evolution

Up to this point, I have focused on designed splicing events. That is, successful splicing of the designed intron of each variant in the library. However, my designs might result in cryptic splicing isoforms, different than the ones I designed. To identify such splicing events, for each variant, I looked for cryptic spliced isoform by aligning the RNAseq reads to the full unspliced sequence, allowing large gaps in the alignment. A long uninterrupted gap in the RNA read is potentially a spliced intron, and if at least one of its ends is not found at the designed ends of this variant I label it as a cryptic intron.

I found cryptic splice isoforms in 25.2% of the variants designed to have a single intron with a median splicing efficiency of 0.038 for the cryptic splice isoforms, compared to the median splicing efficiency of 0.428 for the designed splice isoforms. As a comparison, only 3.1% of the negative control variants presented a cryptic spliced isoform, with a median splicing efficiency of 0.07. I then studied the location of the cryptic intron ends relative to the intended intron ends, I see that 87% of them have the same 5'SS as the designed intron and of course not the designed 3' SS, while only 1% of them have the designed 3'SS, but then not the designed 5'SS (Fig 18A). This observation suggests that a vast majority of cryptic splicing events are a result of selection of an alternative 3'SS during the splicing process. I acknowledge the fact that due to my amplicon sequencing based method, there is a low chance to detect upstream alternative 5'SS selection, and possibly selection of alternative 5'SS might result in unfinished splicing intermediant product[149] which also wouldn't be detected by my method. Previous work has also found that there are a significant number of alternative 3' splice site usage events in *S. cerevisiae* and suggested that the 3'SS choice can be explained by local RNA secondary structure at the original 3'SS[150]. To examine this observation, when analyzing my synthetic introns data, I considered the distribution of RNA free energy (**Δ**G) at a window of 30 nucleotides around the designed 3'SS. I notice that spliced variants with no cryptic splicing have more open structures at their 3'SS compared to spliced variants with cryptic splicing, and to a greater extent than unspliced variants with cryptic introns (Fig 18B). When studying the chosen alternative 3'SS I see that 70% of the alternative isoforms are spliced at one of the three sequence motifs found in the genome ([U/C/A]AG) (Fig 18C) and that 68.5% of them are spliced at the first downstream occurrence of this 3'SS motif after the designed 3'SS. Those isoforms that are spliced at the first downstream 3'SS motif, are more efficiently spliced than other cryptic splice isoforms (Fig 18D).

The observation that the *S. cerevisiae* splicing machinery can easily misidentify 3'SS leads me to hypothesize that mechanisms to avoid such cryptic splicing events must exist, as these events can result in frameshifts and appearance of a premature stop codon. Hence, I checked if I observe a selection against 3'SS motifs near the 3' end of natural introns in the genome. I registered all *S. cerevisiae* introns at their 3' end and calculated the frequency of the two dominant 3'SS motifs ([C/T]AG) around introns' end. Indeed, I found a depletion of these motifs at a window of -50 to +30 around introns' end compared to a null model based on 1000 random sets of genomic loci (Fig 18E).

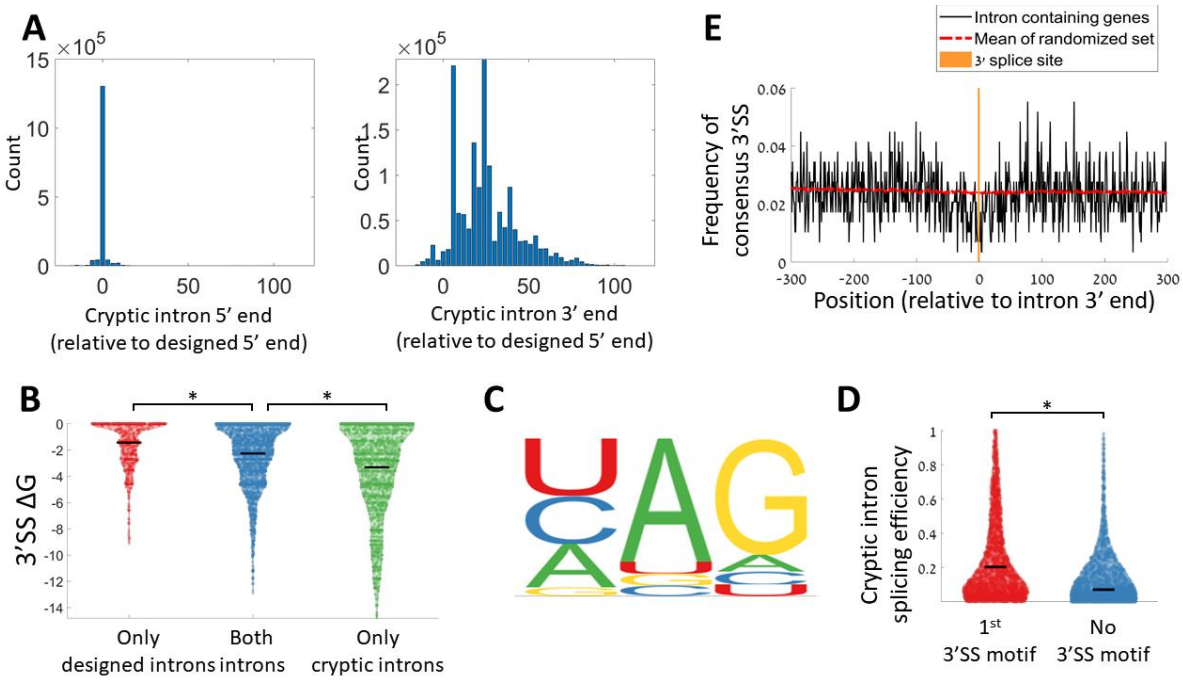

Figure 18 - Cryptic splice isoforms are produced due to selection of alternative 3'SS

A. Relative splice site position distribution for cryptic introns (relative to the designed splice site position), for the 5' splice site (left), and the 3' splice site (right). **B.** Distribution of **Δ**G values at a window of 30 nucleotides around the designed 3'SS for variants with designed splice isoform and no cryptic splice isoform (red), variants with both designed and cryptic splice isoform (blue), and variants with only cryptic splice isoform (green). The difference between all three distributions is significant (t-test, p-value < 10^{-18}). **C.** Sequence motif of the 3'SS for all cryptic intron isoforms detected. **D.** The distribution of splicing efficiency for cryptic splice isoforms, binned according to isoforms in which the cryptic intron is spliced at the first appearance of a 3'SS motif (red), or isoforms for which the last 3 nucleotides of the introns are not a 3'SS motif. (t-test, p-value<10^{-100}). E. 3'SS motif avoidance pattern around introns' 3' end in the *S. cerevisiae* genome. All *S. cerevisiae* intron-containing genes were registered according to the 3' end of their intron. The black line presents the frequency of TAG/CAG motif for each position. The red dashed line presents the expected frequency by averaging the motif frequencies over 1000 sets of sequences registered according to random positions inside coding genes.

Co-evolution of the splicing machinery and intron architecture across yeast species

Our system allows me to introduce any short intron sequence into the *S. cerevisiae* genome. This gives me the opportunity to study the evolution of intron architecture by introducing introns from other yeast species and observing how well they are spliced in my system. I first introduced all the naturally occurring introns from the *S. cerevisiae* genome that can fit in my oligonucleotide design length constraint. For each intron I inserted the full length of the intron flanked by 5 exonic nucleotides from each end, this sequence was inserted on the background of the standard *MUD1* derived background sequence of the library, at its 5' end. Hence, the length limit for an intron was 148 nucleotides, amounting to 149 introns out of 299 in this species. It should be noted that this limit on intron length forces me to use only introns from non-ribosomal genes in my library, as all the introns in ribosomal genes in *S. cerevisiae* are significantly longer (mean length of ~400 nucleotides).

Next, for each natural *S. cerevisiae* intron, I included in my library introns from orthologous genes of the *S. cerevisiae* intron-containing genes from a set of 10 other yeast species, with orthology identified by[48]. I found that most *S. cerevisiae* endogenous introns are spliced in my system (85.5%). Interestingly, introns from most of the other species are typically spliced at similar efficiencies (Fig 19A). Furthermore, for each species I compared each one of its introns to the intron of its orthologous gene in *S. cerevisiae* and define ΔSE as the difference in splicing efficiency for introns of the same gene. I found that many of the non *S. cerevisiae* introns are spliced better than their *S. cerevisiae* orthologs (Fig 19B), suggesting that *S. cerevisiae* introns are not specifically optimized for high splicing efficiency by their own splicing machinery.

Although I did not see a specific preference for the natural introns of *S. cerevisiae*, I still observe that introns from some species like *E. cymbalariae* or *K. thermotolerans,* are spliced in lower efficiency compared to introns from other species. I further note that these two species do not stand out phylogenetically from others (Fig 19A,B). Hence, I hypothesized that introns from these species might have been optimized to evolutionary changes in the splicing machinery. One such candidate could be the gene U2AF1, which is a splicing factor that is associated with the location of the branch site relative to the 3' end of the intron[47]. This gene is missing in 6 out of 11 of the yeast species I analyze here, including from *S. cerevisiae*, and in additional species (*T. blatae*) it is highly mutated and probably non-functional[48]. Indeed, the 11 yeast species I used here show a different distribution of BS-to-3'SS distances, that is concordant with the absence or presence of U2AF1 (Fig. S4A), while

other properties are not significantly different between the two groups (Fig. S4B-E) (intron length does seem to differ between the two groups, but this difference is solely ascribed to the BS-to-3'SS distance, as can be seen by the lack of difference in 5'SS-to-BS distance (Fig. S4B,C)). When comparing the distribution of splicing efficiencies between introns from species with or without a copy of U2AF1, I observed that introns that come from species lacking U2AF1 are better spliced in my *S. cerevisiae* system which also lacks U2AF1 (Fig 19C). Hence, I suggest that those introns were better optimized to splicing machinery lacking this factor, while introns that were adapted to machinery that use this factor are less suitable to *S. cerevisiae* splicing machinery.

In the previous section, I demonstrated that *S. cerevisiae* has a tendency to splice cryptic introns at alternative 3'SS downstream of the original site, which leads to a selection against 3'SS motifs near introns 3' end. Interestingly, when performing the same analysis for the other 10 yeast species I found that all but four species present significant *S. cerevisiae*-like avoidance of 3'SS sequence motifs near their introns 3'SS. Strikingly, the 7 species that show the avoidance signal are those lacking U2AF1, and the one species with highly mutated copy of this factor (Fig 19D, Fig. S5). This result suggests that loss of the U2AF1 gene results in a flexible recognition of the 3'SS, which in turn generates a selective pressure to avoid 3'SS motifs near the intended 3'SS in order to avoid cryptic splicing events. On the other hand, splicing machinery that includes U2AF1 results in a more stringent 3'SS recognition, possibly due to tight constraints on the BS-to-3'SS distance.

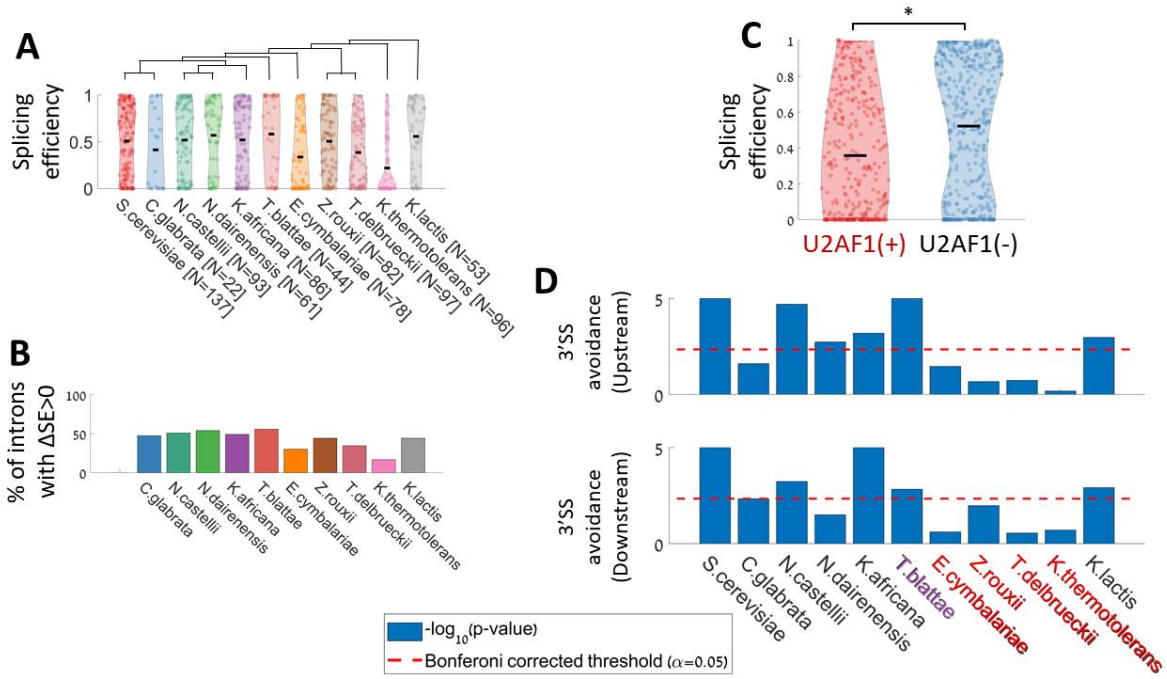

Figure 19 - Analysis of ortholog introns from other yeast species reveals intron architecture evolution

A. Splicing efficiency distribution of spliced variants for introns from orthologs of S. cerevisiae intron-containing genes. A species phylogenetic tree (created according to[63]) is presented above the corresponding violins. The number of introns included in the library from each species are indicated after each species name. **B.** Percent of introns that are spliced better than their S. cerevisiae ortholog intron for each species. **C.** Splicing efficiency distribution for introns that come from species that have a copy of U2AF1 splicing factor (left), and introns that come from species without U2AF1 splicing factor (t-test, p-value<10-9). **D.** Hypothesis test for the 3'SS motif avoidance for each of the 11 species upstream (top) or downstream (bottom) to the 3'SS. P-value was calculated by comparing the mean frequency of the 3'SS motif at a 30nt window upstream/downstream to the 3'SS against 105 sets of sequences each composed of coding genes sequences registered according to randomly chosen positions. Species with a copy of U2AF1 are marked in red, species with malfunctioned U2AF1 are marked in purple, and species without any copy of U2AF1 are marked in black.

A computational model elucidates important features that govern splicing efficiency

I created a large collection of single intron variants, with a systematic exploration of different intron design features. This wide collection of variants allows me to train a computational model that predicts splicing efficiency values from sequence features. For the purpose of this model I used a set of all single-intron variants including both synthetic and natural introns and excluding negative control variants (N=12,745). I trained a gradient boosting model[151,152] using a 5-fold averaging cross-validation technique[153] on randomly chosen 75% of the variants set (N=9,558). As an input to the model, I used a set of 39 features, comprising the

splice site sequences (as a categorical feature), intron length parameters, GC content, 3' u-rich element, and local secondary structure predictions at each splice site (see a full list of parameters at table S1). The model predictions were tested on the remaining 25% of the set of single-intron variants used for this model (N=3,187). Predicted splicing efficiency values for the test set are reasonably well correlated with the measured splicing efficiency values (Pearson r=0.75, Fig 20A).

The predictive model enables me to examine the contribution of each feature to a successful prediction of splicing efficiency. I used Shapley values[154] to infer individual features' importance. Meaning, I analyzed the global contribution of each feature to the predicted splicing efficiency value across all observations, moreover, in Figure 20B I present the individual feature contribution for each observation (i.e. library variant) of the 8 most important features according to this analysis, the distribution of Shapley values for each feature, and it's correspondence with the feature's values. I notice that the most important feature is the sequence of the BS which corresponds with the large difference in splicing efficiency I observed for non-consensus BS variants (Fig 16E). Next, I notice that intronic GC content has high contribution, as low GC content contributes to higher splicing efficiency, which corresponds to previous findings[64,147]. The 5'SS sequence also has a high contribution to efficient splicing. Interestingly, while the 3'SS sequence is considered one of the defining features of introns, it is only ranked 8th in terms of importance for the model predictions. In terms of local secondary structure, I see that only the local structure around the 3'SS is among the top 8 features, where I observe that open structure (high **Δ**G) around the 3'SS contributes positively to splicing efficiency.

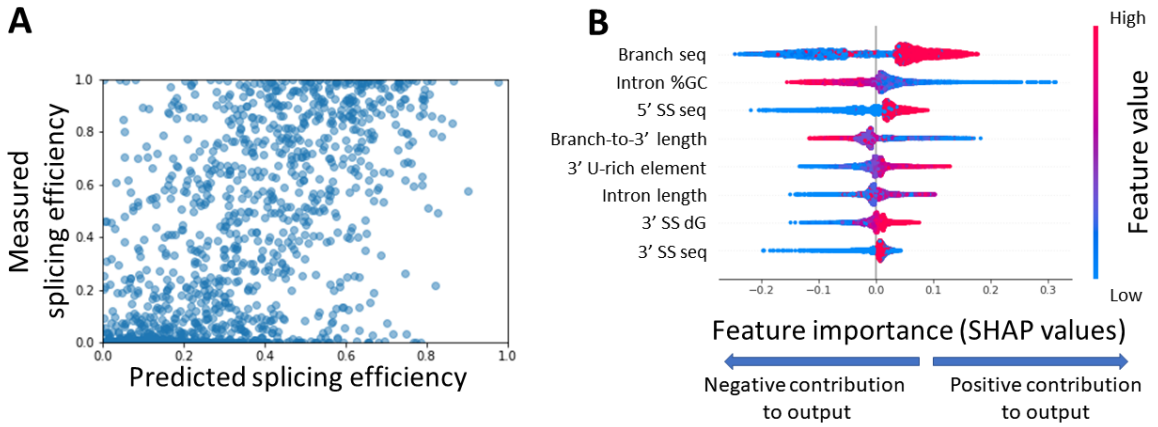

Figure 20 - Sequence features contribution to successful splicing are derived from a computational model

A. Measured splicing efficiency values versus predicted splicing efficiency values of the test set variants (N=3,187) as predicted by a gradient boosting model using 5-fold averaging cross-validation technique (Pearson r=0.75 p-value<10^{-100}). **B.** Distribution of Shapley values for the top 8 features when ranked according to the mean absolute value of the Shapley values. The x-axis represents the Shapley values, the higher the absolute value of the mean of the distribution, the higher its contribution to the model predictions. Positive values mean that the feature is predicted to improve splicing efficiency, and negative values mean that the feature is predicted to reduce splicing efficiency. Sample points are colored according to their feature's value for numerical features, and for splice site sequences that are treated as categorical features, they are colored by the splice site relative abundance in the S.cerevisiae genome (high values represent abundant sequence variants).

S. cerevisiae has the capacity to alternatively splice two tandem introns, thus generating alternative splice variants from the same RNA

Alternative splicing is not considered to have a major role in gene expression regulation in *S. cerevisiae*. There are 10 known genes with two tandem introns in the *S. cerevisiae* genome[155], and most of them are not known to be alternatively spliced. Previous works have examined alternative splicing of a two intron gene in *S. cerevisiae* by studying the spliced isoforms of the two genes that are known to be alternatively spliced (i.e. *DYN2* and *SUS1*)[56,59,156]. In these works, the regulation of alternative splicing of a specific gene was studied through chemical or genetic perturbations[56] or changing environmental conditions[59]. Here I use my library to provide a novel unbiased examination of the potential of *S. cerevisiae* splicing machinery to

alternatively splice a synthetic two intron gene. This system allows me to examine intronic features that may facilitate alternative splicing.

I created a subset of the library with two short introns separated by an exon. For this set I chose 25 short introns (<76 nucleotides), 10 of them are the 10 shortest natural introns in *S. cerevisiae*, additional 10 were randomly chosen from all the natural *S. pombe* introns that fit to the length limits and use splice sites that are found in *S. cerevisiae* as well, and lastly, I created 5 synthetic introns with consensus splice sites, a length of 56 nucleotides, and BS-to-3'SS distance of 20. Using these 25 short introns, I created a set of variants composed of all possible pairings of two introns, where the first intron was inserted at the 5' end of the variable region, and the second intron at the 3' end of the variable region, separated by an exon, the exon sequence was taken from the *MUD1* based background sequence used for other parts of the library.

Using this set of two-intron variants, I tested whether *S. cerevisiae* has the potential to alternatively splice and produce multiple spliced isoforms when given a two-intron gene. Such two-intron designs can result in 5 possible isoforms (Fig 21A). For each variant, I measured the relative frequency of each of the isoforms by aligning its predicted exon-exon junctions to the RNAseq reads. I observed all 4 spliced isoforms, as well as the unspliced, in my data (Fig 21B). Interestingly, out of 614 variants that are spliced, 130 variants have more than one spliced isoform observed for the same pre-mRNA sequence. This observation suggests that *S. cerevisiae* splicing machinery has the capacity to alternatively splice many possible two-intron sequences, and not only the two natural genes that are known to be alternatively spliced. This is despite the fact that its splicing machinery lacks some of the auxiliary factors involved in alternative splicing.

To decipher which intron properties contribute to multiple spliced isoforms, I analyzed all the possible 100 pairs assembled from natural *S. cerevisiae* introns. For each of the 10 introns in this analysis I counted the number of variants for which I observed an isoform in which this intron was spliced out. Then I ranked the 10 introns according to the number variants in which each of them was spliced (regardless of its position as the 1st or 2nd intron). I noticed that multiple isoforms are observed mainly when both introns are ranked high. A single isoform is observed when one intron is ranked low and the other high, and when a pre-mRNA consists of two introns that are ranked low, splicing is hardly observed (Fig 21D).

I next compared the abundance of splice variants that spliced either one of the two introns. There were two alternative hypotheses that I could test, namely that in each pair of introns

one of them will be better spliced than the other, or that the location of the two introns in the gene will dictate, so that either the up-stream or down-stream introns will be better spliced. I notice that isoforms with only the upstream or only downstream intron spliced, appear in similar numbers and have similar splicing efficiency distributions (Fig 21B). Furtner, when comparing the proportion of spliced variants for each intron sequence between variants in which it was placed as the up-stream intron, and variants in which it was placed as the down-stream intron, I see that those two measurements are in high agreement (Fig 21C), indicating that splicing efficiency of each intron is dependent mainly on its sequence, and less on the relative locations of the two introns in the gene.

To further elucidate which features allow production of several isoforms from the same sequence, I used a computational model as was used for the single intron variants. In this case, the input features for the model are the same 39 features as for the single intron case but multiplied by two, as I took the set of features for each of the introns, in addition, GC content and intron length were calculated also for the exon skipping isoform, which yields a total of 80 features. Similarly to the single intron case, I train a gradient boosting model, but in this case, I use it to predict a multi-class classification problem, where each combination of isoforms is considered as a different class, and a variant is assigned to a class according to the isoforms observed for it, ignoring their relative frequency (I exclude the “Both introns” isoform from this analysis as it wasn’t observed in enough samples). I use 5-fold cross-validation technique, where each time I take 20% of the variants set as a validation set and train the model on the remaining 80%, the model performances are inferred by examining all the variants predictions when each of them was in the validation set. The model manages to classify correctly in 68% of the cases, with a weighted average F1 statistic of 0.656. I infer the most important features for each class using a Shapley value analysis as was done for the single-intron model (Fig 21B). In figure 21D I summarize this analysis by looking at 6 features that had a significant contribution in at least one of the classes. First, I can see that a sequence is predicted to be unspliced mainly if its first 5’SS and second BS contribute negatively, thus preventing successful splicing of any of the 3 introns. Second, I notice that in order to get only “Intron 1”, “Intron 2”, or “Exon skipping” isoform, the features for the spliced intron should contribute positively, but also the features unique for the other introns should contribute negatively. And finally, for a combination of spliced isoforms I see that all features should contribute positively, but specifically, the short intron length contributes the most. I notice that due to my design approach, short intron lengths result also in a longer exon

between them, so it is possible that the length of the exon between the two introns is also important to allow alternative splicing.

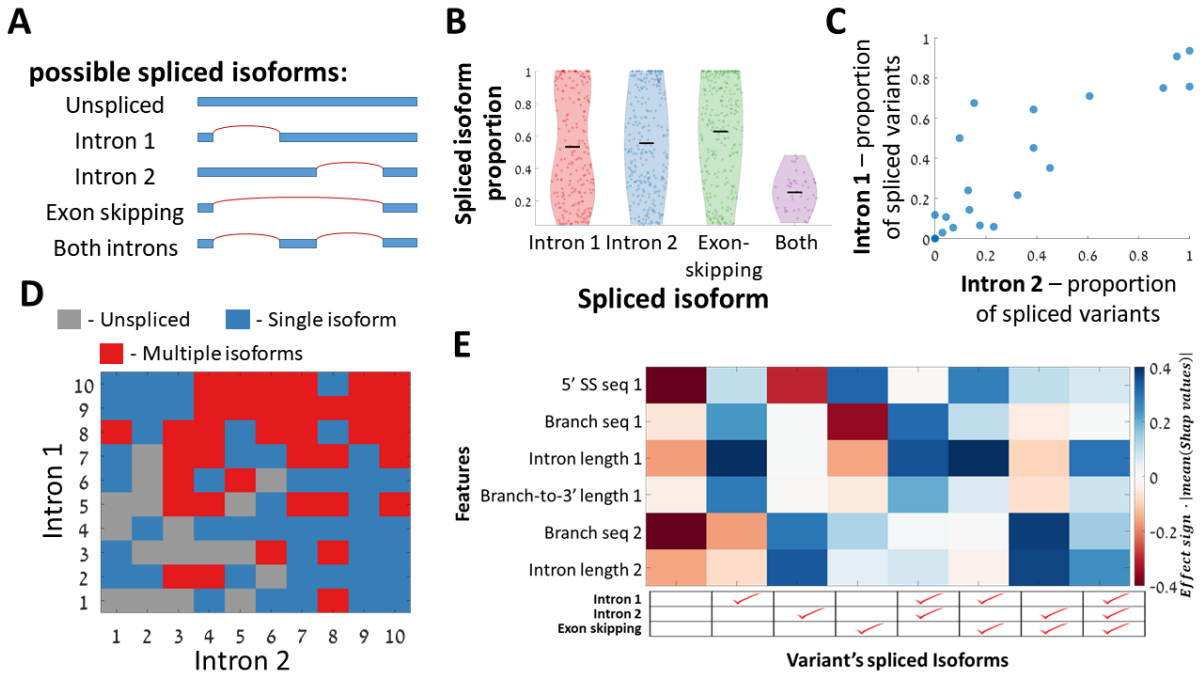

Figure 21 - Tandem two-intron designs demonstrate a capacity of *S. cerevisiae* to alternatively splice two introns.

A. Five possible isoforms can be observed for a two-introns design. **B.** Distribution of isoform's relative frequency for the 4 possible spliced isoforms. **C.** Intron performances when placed as the first intron Vs. placed as the second intron. Each dot represents a single intron sequence, the x-axis represents the proportion of variants with this intron spliced as the first intron, and y-axis axis represents the same for the second intron variants (Spearman correlation r=0.88 p-value=10^{-8}). **D.** The number of observed isoforms for each of the natural S. cerevisiae intron pair variants. Each number represents a different intron sequence, and they are ordered according to the number of variants in which the intron was spliced. **E.** Feature importance (Shapley values) derived from a gradient boosting model trained on a multi-class classification problem aiming to predict which spliced isoforms are observed for each variant. A set of Shapley values is calculated for each class (i.e. set of isoforms), and I present features that have a mean absolute value >3 standard deviations in at least one of the classes. For each feature and each class, I present the mean absolute value multiplied by an "effect sign" to indicate if the feature contributes positively/negatively to produce this isoform (for details on "effect sign" calculation, see Materials and Methods)

6.3. Part III – The economy of expressing a shared public-good compound

A synthetic library creates a variation in the production levels of a public-good enzyme in yeast

A general aspect of gene expression economy which can be studied in a systematic manner is the economy of production and secretion of common goods which are consumed by a community of cells. In unicellular organisms most of the cell's resources are devoted to intracellular functions which benefit only the cell itself. Nevertheless, microorganisms mostly live in colonies, and some of the cell's proteome are proteins which are secreted to the environment and benefit other cells in its environment.

Secreted proteins affect the cellular economy in a more complex manner since they break the direct relationship between cost and benefit. A cell might invest resources in producing a protein which is beneficial to the colony, but a neighboring cell can enjoy this common good, without investing resources of its own. In the simpler case of an intra-cellular gene, a cell that will optimize its cost-benefit ratio will prevail, but what would be the dynamics in the case of a secreted gene?

A known model system for public-good cooperation dynamics in budding yeast is sucrose metabolism. Yeast cells cannot intake the disaccharide sucrose into the cell, however when the preferred monosaccharaides glucose and fructose are not available, it can metabolize sucrose, by converting to glucose and fructose extra-cellularly[9,157]. The gene which is responsible for that reaction is *SUC2* that expresses the enzyme Invertase, which is secreted to the cell's periplasm[158] where it can hydrolyze the sucrose in the media, and release the monosaccharide products back to the media, where they can be digested through the standard glucose and fructose pathways.

In this work we wish to study how perturbations to regulatory elements that control protein secretion affect population dynamics in a complex population that includes strains with varying invertase secretion or production levels. The classic regulatory element that controls secretion is the signal peptide (SP)[87] which is a short sequence at the N terminד of the protein, enriched with hydrophobic amino acids. This sequence targets the nascent protein to the endoplasmic reticulum (ER) while it is being translated (Fig 22A). In addition, a recent study from our lab has described a new RNA regulatory motif that is presumed to modulate protein secretion by targeting their RNA molecules to the ER. This motif is termed Secretion-Enhancing *Cis* Regulatory Targeting Element (SECReTE)[93], and is defined as long repeats of

three way periodicity of pyrimidines (i.e. long NNY repeats) (Fig 22B). SUC2 gene has both of these regulatory elements.

To create a standing variation of yeast cells with varying invertase secretion levels we designed and created two synthetic oligo libraries[19], the first introduces variations to *SUC2*'s signal peptide, and the second introduces variations to *SUC2*'s SECReTE motif. The two libraries were designed to systematically introduce different types of mutations (i.e. synonymous, non-synonymous, enhancing/decreasing hydrophobicity, enhancing/decreasing Y content etc). Each of these libraries was cloned separately to *S. cerevisiae* cells to create two SUC2 secretion libraries, through perturbations of two different mechanisms and regulatory elements. The SP library was designed with 4,500 variants, and the SECReTE library was designed with 4,800 variants.

The two libraries were cloned separately into a yeast integrative plasmid[159] that includes the wild-type sequence of *S. cerevisiae SUC2* gene. Each library replaced the relevant region in the WT sequence. The plasmid library was then inserted into an *S. cerevisiae* strain for which the SUC2 in its original location was deleted, the library's *SUC2* gene was inserted to chromosome X, (Fig 22C).

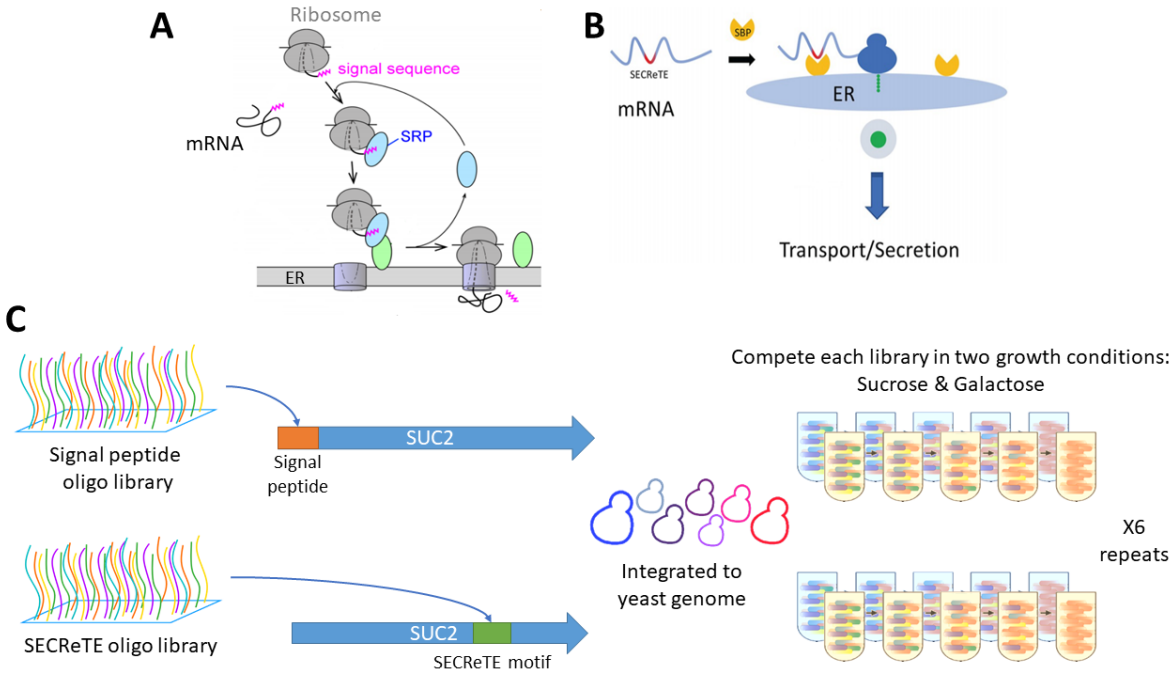

Figure 22 – Perturbations to SUC2 secretion levels using two synthetic libraries

A. The translated signal sequence is recognized by the SRP protein that co-translationally targets the translating ribosome to the ER. Adapted from[87] **B.** The mRNA transcript containing the SECReTE motif is recognized by a putative SECReTE binding protein that facilitates translocation of the mRNA to the ER. Presumably, the enhanced association of SECReTE containing transcripts to the ER promotes secretion[93]. **C.** Two synthetic oligo libraries were designed and synthesized, one for the

SUC2 SP, and one for the *SUC2* SECReTE motif. Each of the libraries was separately cloned and inserted into *S. cerevisiae* genome. The relative fitness of library's variants was measured in two different conditions, galactose to measure differences in cost of gene expression, and sucrose to measure differences in both cost and invertase activity.

Measuring cost of gene expression and public-good associated fitness using two competition assays

To study the population dynamics of the different library strains we wished to separate the cost component from the benefit component of expressing the different *SUC2* variants. As we described in section 6.1, cost of gene expression can be measured by measuring fitness differences between variants of unneeded gene. In the case of *SUC2*, it is known that the gene is activated by glucose repression[160], but it is needed for the cell's metabolism only when grown on sucrose or raffinose. Hence, if we grow the cells on galactose, they will express *SUC2*, but in this condition the gene is unneeded. So, by preforming a competition assay with galactose as the carbon source we could measure differences in the cost of expressing different *SUC2* variants. Next, we wished to measure the effect of the *SUC2* variation on the benefit from the gene, and on the population dynamics. For that purpose, we performed a competition assay on sucrose containing media (Fig 22C).

Before starting the competition assay on the entire library we ran preliminary competition assays with three existing variants of *SUC2* that differ in their SECReTE motif[93]. In this preliminary competition we noticed that the main differences in fitness between the strains are observed during exponential growth (data not shown). Therefore, for the SECReTE library we chose to run the competition experiment while maintaining the cells in exponential growth. This was done by diluting the cells culture every 12 hours while the culture was in mid-log phase (OD600 ~ 0.5).

On the other hand, as a preliminary experiment for the SP library, we isolated specific strains from the library by plating and choosing single colonies. We then compared between the strains by measuring their growth curve during overnight growth and noticed that differences between those variants are manifested also in the stationary phase. Therefore, for the SP library the competition was done by diluting the cell culture every 24 hours when the culture had reached stationary phase.

The SECReTE library competitions on both media was run for 32 generations and samples were taken for deep sequencing of the library's variable region at generations {0,4,8,12,16,24,32}. The SP library competitions on both media was run for 90 generations

and samples were taken for deep sequencing of the variable region at generations {0, 11, 22, 56, 90}. Competition of each library in each growth condition was done in six independent repeats.

SECReTE library demonstrates combination of public-good related fast dynamics, and slower cost related dynamics

We first infer relative fitness of each variant in the library using log-linear regression of the frequency of each variant as a function of number of generations. As a first assessment for effect of true biological differences in the library as opposed to random noise, we analyze the correlation between independent repeats. In the SECReTE library competitions we notice that the mean pairwise correlation coefficient between repeats in the sucrose competition experiment (Pearson, $\bar{r}$=0.52) is appreciably larger than the mean pairwise correlation coefficient in the galactose competition experiment (Pearson, $\bar{r}$=0.29) (Fig 23A, for visual clarity only a single pair of repeats is presented). This result indicates that when the library is grown on sucrose the true biological differences between variants have stronger effect compared to when grown on galactose, suggesting that the modifications to the SECReTE motif introduced in this library indeed affect the functionality of invertase.

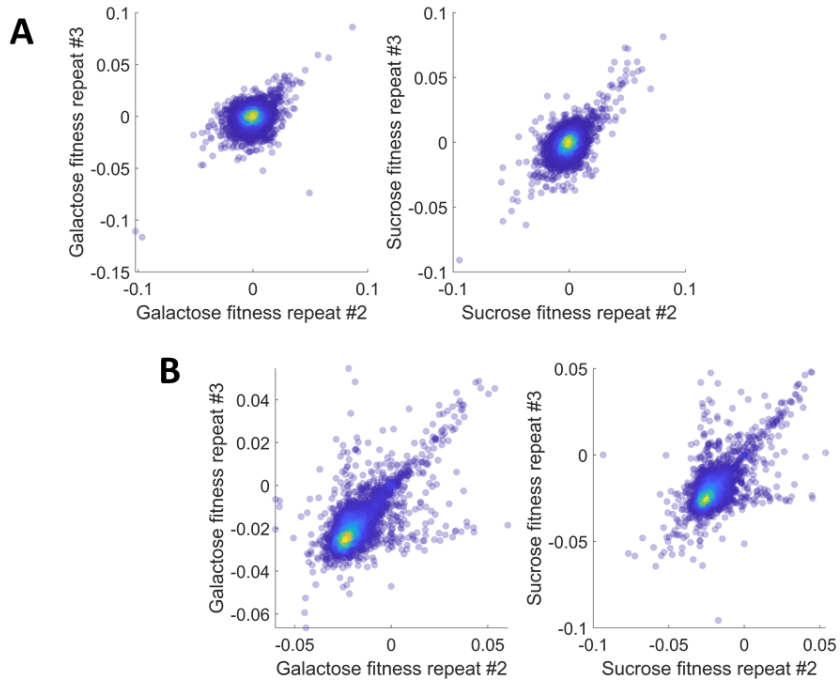

Figure 23 – Pairwise correlations between repeats

A. scatter plot of the SECReTE library's variants fitness as inferred from repeat #2 Vs. repeat #3, in the galactose competition experiment (left, Pearson r=0.27, mean Pearson pairwise correlation $\bar{r}$=0.29), compared to the sucrose competition experiment (right, Pearson r=0.49, mean Pearson

pairwise correlation $\bar{r}$=0.52). **B.** scatter plot of the SP library's variants fitness as inferred from repeat #2 Vs. repeat #3, in the galactose competition experiment (left, Pearson r=0.69, mean Pearson pairwise correlation $\bar{r}$=0.66), compared to the sucrose competition experiment (right, Pearson r=0.64, mean Pearson pairwise correlation $\bar{r}$=0.65).

Next, when comparing the variant's fitness between the two growth conditions we notice a significant positive correlation between fitness on sucrose and fitness on galactose (Pearson, r=0.35), Figure 23A. Nevertheless, like the analysis presented in section 6.1.1, we are interested more in the residuals from the linear regression line between the fitness in the two conditions. Here, the fitness component that is common to both conditions would be the cost of gene expression, and differences in fitness when growing on sucrose compared to galactose would represent functional differences in sucrose metabolism and invertase activity. Therefore, the residuals of this regression line should correspond with the fitness derived from *SUC2* benefit. Similarly to the analysis presented in 6.1.1 we define positive and negative fitness residual variants, as variants who consistently demonstrate positive or negative fitness residual across all pairwise comparisons (Fig 24A, see materials and methods for detailed description on positive/negative residual variants definitions). We observed 88 positive fitness residual variants, and 89 negative fitness residual variants, in the SECReTE library.

Since the variants in the library differ in their public-good production, we predict that fitness values of variants grown on the sucrose media will not be constant and will be dependent on the population composition. Therefore, we turn to look also on close-to instantaneous fitness measurement, in which we infer fitness from the log ratio of frequencies between consecutive generations. We then infer fitness for different timepoints along the competition, and plot the mean fitness on sucrose versus the mean fitness on galactose for these timepoints. When observing the scatter plot of fitness on sucrose and fitness on galactose, and the location of the positive/negative variants as defined above using the regression based fitness (Fig 24B), we observe that the positive and negative fitness residual groups present different dynamics. We notice that the negative group present extreme negative fitness values only at the start of the competition, suggesting that these variants suffer in a sucrose limiting environment but only until the population reaches some equilibrium. However, the positive group are a group of variants that increase their relative fitness as the competition proceeded, but although they were classified as variants with significant residuals from the regression line, we notice that when analyzing this group alone a significant correlation between sucrose and galactose is

observed (Pearson, r=0.79) suggesting that the dominant mechanism contributing to fitness in these variants is nevertheless cost of gene expression.

To further test what differentiates the negative fitness residual group from the rest of the library we compared the pyrimidine (Y) content of this group to the positive residual group and the entire SECReTE library. We chose to look at Y content because the SECReTE motif is characterized by enrichment in Ys. We notice that the negative group has significantly lower Y content (Wilcoxon ranksum test, p-value < 10^{-4}) (Fig 24C), while the positive group has the same distribution of Y content as the rest of the library. According to our current understanding of the SECReTE motif, this suggests that the negative residual variants have lower secretion of SUC2 gene, hence we hypothesize that these variants suffer from reduced level of sucrose degradation product at the start of the competition because of reduced selfish production, until the population stabilizes in subsequent generations and these variants can enjoy degradation products from neighboring cells.

Moreover we looked at enrichment or depletion of R-to-Y mutations in both the negative residual group and the positive residual group. We analyzed the frequency of SNPs at different sites both for the positive and negative group and compared them to the frequency of SNPs in the entire library. Position in which there was a significant enrichment or depletion in mutation rate are presented in Fig 24D. We notice that the negative group has less SECReTE enhancing mutations near the natural SECReTE motif site compared to the rest of the library, suggesting again that these variants have lower secretion levels.

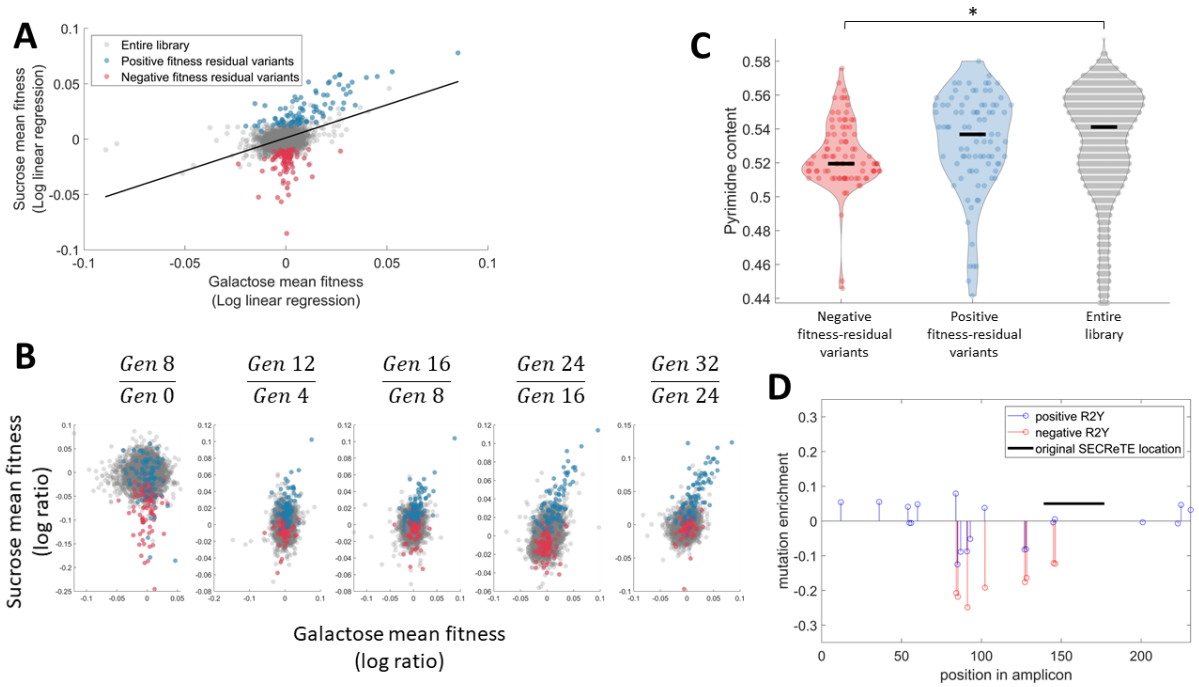

Figure 24 – population dynamics of the SECReTE library

A. Scatter plot of the mean fitness as inferred from all timepoints using log-linear regression of relative frequency data. Positive fitness residual variants are marked in blue and negative fitness residual variants are marked in red (see materials and methods for how these groups were defined). Fitness in the two condition is weakly correlated (Pearson, r=0.35). **B.** Fitness on sucrose Vs. fitness on galactose in presented in (A), only here in each plot the fitness is derived by the log ratio of frequency between two timepoints separated by 8 generations. The timepoints taken for each panel are indicated on top of the panel. Blue and red variants are the positive or negative fitness residual variants as defined using the regression based fitness, exactly as described in (A). **C.** Distribution of pyrimidine (Y) content for negative/positive fitness residual variants, and for the entire library. Negative fitness residual variants have significantly lower Y content compared to the entire library (Wilcoxon ranksum test, p-value < 10^{-4}). **D.** Mutations enrichment per location for both the negative and positive fitness residual groups. At each position an enrichment score was calculated by the fraction of variants containing an R-to-Y mutation in this position in each fitness residual group, subtracted by the ratio of this mutation in the entire library. Positive value means that the mutation was enriched in this group, a negative score means that the mutation is depleted from this group. Only positions with statistically significant mutation enrichment score are presented (Two proportion Z test, corrected for multiple hypothesis using FDR value of 0.05[161]). The horizontal black line represents the location of the natural SECReTE motif of SUC2 within the library's variable region (138-177).

Signal peptide library also demonstrates a combination of fast and slow dynamics

When analyzing the SP library competition experiments, we first notice that in both sugars there is a good correlation between repeats (mean pairwise Pearson correlation $\bar{r}$=0.66 for

galactose, and $\bar{r}$=0.65 for sucrose, Fig 23B). Moreover, when comparing the mean fitness inferred from the entire competition experiment using log-linear regression in both conditions we notice a high correlation between fitness on galactose and fitness on sucrose (Fig 25A, Pearson r=0.7). This high correlation might indicate that the dominant factor contributing to fitness differences in this library is the cost of gene expression. However since we predict that fitness differences due to invertase activity will change along time as the population structure changes, we also analyzed this correlation as a function of time by calculating "instantaneous fitness" i.e. the log-ratio of frequencies between subsequent timepoints. By comparing the correlation between instantaneous fitness on both sugars in the first timepoint and in the last timepoint we see that indeed the correlation is specifically high in the last timepoint (Fig 25B, Pearson r=0.3 and, 0.76 in generation 11 and 90 respectively). Now we wish to analyze the temporal dynamic of this correlation along the competition, however, since the time difference between subsequent time-points was not constant, it is not trivial to compare between correlation values (since during longer time period we expect the signal to noise ratio to increase). Therefore, in order to properly compare the correlation in instantaneous fitness between the two sugars across time, we decided to compare in each timepoint the "between-condition mean pairwise correlation", and "within conditions mean pairwise correlation". In order to compare between the two values we take the difference between the Fisher Z transformations[162] of each correlation value, as this transformation transforms the correlation coefficient to approximately normal distribution. We expect to see a negative value for this score, as we expect the between-condition correlation to be lower, the more negative the score, it means that the within-condition correlation is more distinct than the between-conditions correlation which suggests a stronger effect for the *SUC2* functionality on fitness. We observe lower between-conditions correlation for the first timepoint compared to later timepoint (Fig 25C), which like in the SECReTE library case suggests that differences in invertase activity are more dominant at the start of the competition.

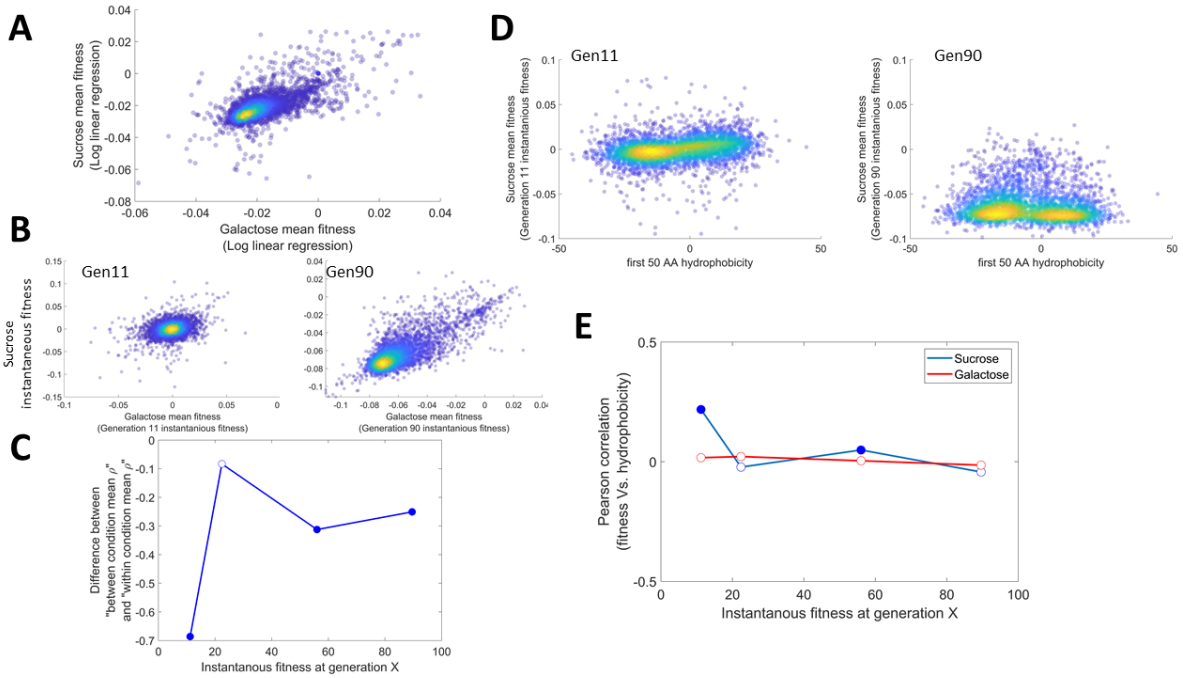

Figure 25 - population dynamics of the signal peptide library

A. Scatter plot between mean fitness on sucrose and mean fitness on galactose. Fitness values are calculated by log-linear regression across all timepoints (Pearson, r=0.7, p-value<10^{-100}). **B.** Scatter plot between conditions of instantaneous fitness calculated as the log ratio of frequencies between subsequent timepoints normalized by the difference in generations between the two timepoints. The left panel presents the instantaneous fitness for the time period between generation 0 and generation 11 (Pearson, r=0.3, p-value<10^{-100}), and the right panel presents the instantaneous fitness for the time period between generation 56 and generation 90 (Pearson, r=0.76, p-value<10^{-100}). **C.** The difference between the Fisher Z transformation of “between conditions mean pairwise correlation” and the mean of the two “within condition pairwise correlation”. Filled circles represent scores that are significantly lower than zero (Z-test). **D.** Scatter plot of instantaneous fitness Vs. total hydrophobicity score of the 50 amino acids variable region (Kyte-Doolitle). Left panel presents the instantaneous fitness for the time period between generation 0 and generation 11 (Pearson, r=0.27, p-value<10^{-80}), and the right panel presents the instantaneous fitness for the time period between generation 56 and generation 90 (Pearson, r=-0.02, p-value=0.13). **E.** Pearson correlation coefficient between hydrophobicity score and instantaneous fitness on each condition. Filled circles represent statistically significant values (corrected for multiple hypotheses using Bonferoni correction).

Signal peptide activity is associated with its hydrophobicity, therefore we decided to test a possible correlation between the variable region hydrophobicity score calculated using the Kyte-Doolitle score[122], and instantaneous fitness on sucrose at different timepoints. At the first timepoint we observe a significant correlation between fitness and hydrophobicity (Pearson r=0.27, p-value < 10^{-80}), while in the last timepoint no such correlation is observed (Pearson, p-value =0.13) (Fig 25D). When examining this correlation across all timepoints,

and also for the galactose fitness measurements, we notice that a significant correlation is observed only in sucrose and only in the first timepoint (Fig 25E).
Following this result, we hypothesize that in the signal peptide library, at the start of the competition there is an advantage for variants that secrete higher levels of invertase, but when the population reaches a steady state in terms of public-good production and enough degradation products are shared in the media, the cost component becomes more dominant. This conclusion goes in line with what we have seen with low Y content variants in the SECReTE library.

7. Materials and methods

7.1. Gene architecture that minimize the cost of gene expression

Library architecture

The synthetic library was provided to me by Goodman *et al.* [25] 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 Sites (RBS) and an N'-terminus amino acid fusion of 11 amino acids followed by a sfGFP gene. The library as a whole includes: two promoters with either high or low transcription rate. Three synthetic RBSs with strong, medium, or low translation initiation rates, as well as 137 different genomic RBSs that were defined as the 20bps upstream to the ORF of 137 *E. coli* genes. And finally, 137 coding sequences (CDS) 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 1.2ml of LB + 50μg/ml kanamycin at 30°C, the exact same conditions as was used in Goodman *et al.* 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 four days (~27 generations), mixed with glycerol and kept at -80°C.

Library preparation and sequencing

Plasmids from time zero (library "ancestor") and all other samples were purified with a QIAgene mini-prep kit and used as templates for PCR to amplify specifically the variable region of all variants in the population. To minimize PCR and sampling biases, we used a large amount of template, ~500ng of DNA, and a relatively short PCR of 26 rounds. The forward primer (sequence: CAGCTCTTCGCCTTTACGCATATG) was paired with 5 different reverse primers that are one bp shifted from each other to insure that library complexity was high enough for Illumina sequencing:

R1: GACAATGAAAAGCTTAGTCATGGCG ; R2:
ACAATGAAAAGCTTAGTCATGGCG

R3: CAATGAAAAGCTTAGTCATGGCG ; R4: AATGAAAAGCTTAGTCATGGCG

R5: ATGAAAAGCTTAGTCATGGCG

PCR products were then run on BluePipin to capture the correct amplicon size of ~140 bps and remove any un-specific amplicons. Then, DNA buffer was exchanged using Agencourt AmPure SPRI bead cleanup protocol. Hiseq library was prepared next using the sequencing library module from *Blecher-Gonen. et al.* 2013 [163]. In short, blunt ends were repaired, Adenine bases were added to the 3' end of the fragments, barcode adapters containing a T overhang were ligated, and finally the adapted fragments were amplified. The process was repeated for each sample with a different Illumina DNA barcode for multiplexing, and then all samples were pooled in equal amounts and sequenced. We performed a 125bp paired end high output run on HiSeq 2500 PE Cluster Kit v4. Base calling is performed by RTA v. 1.18.64, and de-multiplexing is carried out with Casava v. 1.8.2, outputting results in FASTQ format.

Data processing

De-multiplexed data was received in the form of FASTQ files split into samples. First, SeqPrep (https://github.com/jstjohn/SeqPrep) was used to merge paired reads into a single contig, to increase sequence fidelity over regions of dual coverage. The size of each contig was then compared to the theoretical combined length of the forward primer, the reverse primer and the variable region of the variants. Next, the forward and reverse primers were found on each contig (allowing for 2 mismatches) and trimmed out. This step was performed for both the forward and reverse complement sequences of the contig, to account for non-directional ligation of the adaptors during library preparation. Then, the reverse primer was searched at the last 5 nucleotides of the contig to account for different primer lengths. After primers were trimmed, the contig was tested again for its length to ensure no indels had occurred. Contigs were then compared sequentially to the entire library, comparing the sequence of each contig to the sequence of each variant. Any contig without a matching variant within two mismatches or less was discarded. Contigs with more than a single matching variant with the same reliability were also discarded due to ambiguity. Each contig that passed these filters was counted in key-value data structure, storing all variants in the library and their frequency in each sample. These data were then used for all downstream analyses.

Fitness estimation

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 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 both for generation ~84 and generation ~56

GFP expression level estimation

GFP expression levels were taken from *Goodman et al.* [25] data, in which it was calculated using the method described in *Kosuri et al.*[21]. In short, cells were sorted into 12 expression bins using FACS, and in each bin the relative frequency of each variant was measured using deep-sequencing. The estimated expression level of each variant was then calculated by computing the weighted geometric mean of the bins' median expression level, using the relative frequency of each variant as the bin's weight.

In order to validate this data, we estimated the GFP expression level from the raw data in *Goodman et al.* by fitting gamma distribution parameters (suggested before as a model to capture noise, or spread of expression values of a gene within an isogenic population [164]) to the histogram of each variant's frequencies in all bins. This gamma distribution follows this equation: $P(x) = \frac{x^{a-1}e^{-\frac{x}{b}}}{b^a\Gamma(a)}$ where Γ denotes the gamma function.

These two estimation methods are highly correlated (Supplementary Figure 5C, r=0.94, p-Value<10^{-200}). However, we noticed that ~600 variants showed high expression levels according to the gamma fit method, while coming from the entire range of expression level using the geometric mean method. When closely examining these cases, we noticed that the source for the disagreement between the two methods is that these variants were observed only in two bins, with one of them being the highest bin, and the other not being the second highest. Therefore, we decided that the expression estimation for these variants is unreliable and excluded them from our analyses.

Calculation of fitness residuals and classifying variants according to their positive or negative fitness residual sign

We 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. To calculate fitness residual, we performed the following steps:

First, we filtered out variants that demonstrate a lower GFP level than 2^{11}[AU], since below this threshold the GFP measurement method is not sensitive to accurately measure GFP. We also excluded variants with a GFP level above $2^{17.5}$[AU], as above this threshold the

measurement method saturates. Notably, only variants with the “high promoter” were included in the analysis, since almost all “low promoter” variants did not pass the protein level filter. This decision was essential as the few “low promoter” variants that did pass this threshold show biased values of sequence features, such as a very low GC-content, which could mask real signals.

Next, we fitted a linear regression model between fitness and GFP expression levels for each of the six independent FitSeq repeats separately at each of the last two time points (generations ~56 and ~84). Then, variants for which fitness residual was in the top or bottom 5% were excluded and a new regression line was fitted in order to reach a better fit. These outliers were excluded only for the sake of fitting a regression line and were still included in our downstream analyses. Then, a fitness residual score was calculated for each variant at each repeat of the experiment and on each of the two time points.

We then split the variants into two groups: positive or negative fitness residual variants. To account for random processes (experimental errors and drift), "positive" or "negative" class was assigned for a given variant only if it showed a positive/negative fitness residual sign in at least 5 lineages in both time points. The set of all the above filters resulted in 975 variants in the positive variant group and 815 in the negative variant group.

Since we noticed that some of the negative variants have extreme negative fitness residual values, we further classified them as “underachievers”. Underachiever variants were defined as variants that repeatedly showed fitness residual values in the bottom 5% of the entire library. Similar to the positive/negative classification, a variant is assigned as “underachiever” only if it is found in the bottom 5% in at least 5 out of the 6 linages in both time points, which resulted in 80 variants.

Parameter comparison between two fitness residual groups

A one-sided Wilcoxon rank-sum test was used to compare the distributions of different sequence parameters between the positive and negative fitness residual groups. We also tested the effect size of each parameter using the “Probability of superiority” method [129] that calculates the probability to randomly choose a member from group A with a higher value than a random member from group B.

To compare between effect sizes according to GFP expression levels, we split the positive and negative variant groups into three quantiles according to GFP expression levels. Then, the effect size for hydrophobicity or amino acid synthesis cost between positive and negative variants were calculated for each quantile. We then performed an empirical p-Value

estimation by randomly choosing three data sets with the same number of variants, and computed the effect size at each set. This sampling was performed 10^4 times, and p-Value was estimated by counting the number of times the difference in effect sizes between the first and second sets and between the first and third sets were lower in the real data than the difference in effect sizes of the random groups.

Calculating translation initiation rate per variant

We estimated the translation initiation rate of each variant with the "RBS calculator" [165,166], which simulates initiation rates given a UTR and a coding sequence. This calculation is achieved by using a bio-mechanic model combining the affinity to the anti-Shine Dalgarno sequence of the ribosome, mRNA secondary structure of the UTR and coding sequence, and steric interference of the ribosome and the mRNA.

Ribosomal mean elongation time estimation

To evaluate codon-decoding times by the ribosome we used a published values [111], which were derived from ribosome profiling data [115]. Mean elongation times for each of the 61 sense codons are driven from measured ribosome density, when the ribosome A site is on a codon, averaged over all the appearances of the codon within mRNAs. This measurement estimates the translation elongation time of each codon, and its inverse which represents translation speed correlates significantly (r=0.46 for *E. coli)* with tRNA availability. The final score given to each variant was the harmonic mean of its elongation time values over the first 11 amino acids.

Folding energy estimation of mRNA secondary structure

We calculated folding energy of mRNA secondary structure for each variant by using the ViennaRNA package algorithm [167]. Each sequence was computed by a sliding window, whose starting position ranged from position -18 to position 32. The calculation was repeated with different window sizes (20-60bps). All calculations were done assuming a temperature of 30ºC.

Model for estimating translation velocity based on anti-Shine Dalgarno affinity

The Shine-Dalgarno affinity was calculated identically to Li *et al.* [120]. In short, for each position we calculated the affinity of 8-11bps 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 10mer sequences for that position using the RNA annealing function from the

ViennaRNA package algorithm [167], 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 by:

$$K = e^{-\frac{\Delta G}{RT}}$$

Where ΔG denotes the SD affinity, R denotes the gas constant and T denotes the temperature. This equilibrium constant, at the nth codon along a sequence, is defined in turn, given the association reaction rate (k_f) which represents the association to the current site (n), and a dissociation reaction (k_b) that represents the dissociation to the current site as:

$$K = \frac{k_{f_n}}{k_{b_n}}$$

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 $n+1$:

$$\frac{1}{v_{n\to n+1}} = \frac{1}{k_{b_n}} + \frac{1}{k_{f_{n+1}}}$$

$$v_{n\to n+1} = \frac{k_{b_n} k_{f_{n+1}}}{k_{b_n} + k_{f_{n+1}}}$$

We further assume that the association reaction rate is not dependent on the sequence, therefore for every n, $k_{f_n} = k_f$. Introducing equations (i)-(ii) to the equation (iv), results in a term for the ribosomal velocity at a specific position by the anti-Shine Dalgarno affinity:

$$v_{n\to n+1} = \frac{k_f \cdot k_f K^{-1}}{k_f(1 + K^{-1})} = k_f \frac{e^{\frac{\Delta G}{RT}}}{1 + e^{\frac{\Delta G}{RT}}}$$

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. The analysis was performed also at a codon resolution, taking into account only positions of the sequence that are the first nucleotide of codons, which yielded similar results to the nucleotide-based analysis.

<u>Amino acid property estimation of N-terminus fusion amino acids</u>

Hydrophobicity of each 11-amino acid N-terminus peptide was calculated according to its score on the Kyte-Doolittle scale [122]. Amino acid cost was derived from Akashi and Gojobori

[120] in the form of the amount of energy consumed for its production in high energy ATP or GTP bonds. Cost was either evaluated per amino acid or summed for the whole peptide. Supply of amino acids were derived from *Bennet et al.* [121], which measured cellular concentrations of amino acid in exponnentially grown *E. coli*. Notably, two amino acids are missing from this table (Gly & Cys), and two amino acids are indistinguishable (Lys & Ile). Therefore, those 4 amino acids were excluded from the this analysis. The demand per amino acid was calculated by multiplying the frequency of each amino acid in each *E. coli* gene by the median ribosome profiling score of the gene [115]. The sum of all genes was defined as the total amino acid demand.

Amino acid enrichment in positive and negative variant groups

To calculate the frequency of the various amino acids in the collective proteome in either the positive or the negative fitness residual group, we counted the occurrences of each amino acid in each variant. We then summed this number for each amino acid across all variants in each group and divided the sum by the number of variants in each group multiplied by 11. To quantify the relationship between amino acid enrichment and energetic-cost or availability we calculated the frequency ratio of each amino acid by dividing the amino acid frequency of the positive fitness residual group by the frequency of the negative group. We then calculated the Pearson correlation between the log2 amino acid enrichment ratio and the amino acid energetic-cost or their availability.

Comparing fitness residual among variants with the same N-terminus fusion by Δfitness-residual

We defined Δfitness-residual as the difference between the fitness residual of a given variant with the average fitness residual of the variant group with the same N-terminus amino acid fusion. Therefore, Δfitness-residual measures the expression cost of a variant normalized to its GFP expression level and its N-terminus amino acid sequence. We then spilt each variant group of the same N-terminus fusion to above-average and below-average variants and calculated for each sub group the mean value for six features (RNA levels, translation initiation rates, translation efficiency, codon decoding speed (MTDR), mRNA secondary structure, and anti-Shine Dalgarno affinity). For each feature, the mean value of the below-average (x-axis) and above-average (y-axis) Δfitness-residual groups were depicted as a scatter plot, in which each point represents a different N-terminus fusion. Then, the deviance of all dots from the identity (X=Y) line was calculated and tested for significance with a one-tailed Student's t-test. To compute an effect size for this enrichment, we used Cohen's d: $d =$

$\frac{\bar{x}_{high}-\bar{x}_{low}}{S}$ where $\bar{x}_{high\backslash low}$ represents the mean of the feature in the above- or below-average group, and S represents the standard deviation of the feature in the entire set of library variants that was used in this study.

A multiple linear regression model to predict fitness residual

We performed a multiple linear regression using all eight features as independent variables (RNA levels, translation initiation rate, translation efficiency (GFP protein/mRNA), mRNA secondary structure, codon-decoding speed, aSD affinities, amino acid metabolic cost and hydrophobicity) and the mean fitness residual across six repeats of FitSeq as the dependent variable. The regression yielded a coefficient for each feature, which were all used in order to predict fitness residual of a given variant.

As a negative control for this model we randomly shuffled each of the features in the library, trained a mock model on this shuffled library, and computed the Pearson correlation coefficient between the observed fitness residual and the expected fitness residual according to the mock model. In order to compute a p-Value we repeated this process 10^5 times, and counted the number of times the correlation coefficient from the mock model was higher than the correlation coefficient from the real model.

To predict fitness residual of natural *E. coli* and *B. subtilis* genes, a second regression model was performed, in which we excluded translation efficiency (due to lack of data for the entire ~4000 *E. coli* genes) and hydrophobicity (due to the fact that hydrophobic motifs in membrane proteins are functional, hence including this feature might lead to wrong estimation of membrane proteins). We also used Lasso regularization and feature selection method [168] with Matlab's "lassoglm" function from the "Statistics and Machine Learning" toolbox to avoid overfitting of the model. The λ value was chosen as the value for which the deviance was one standard deviation higher than the minimum deviance achieved in a 1000-fold cross validation. Out of the six features used for this model, none were excluded by Lasso method. This model performed well in predicting fitness residual of the library variants and a cross validation test resulted in correlation of r=0.3 (p-Value=10^{-10}).

This model was then used to predict fitness residual scores for natural *E. coli* (strain MG1655) and *B. subtilis* (strain 168) genes. For each gene of these species, we computed a score for each feature of the model. We used RNA levels for *E. coli* from a previous RNA-seq experiment in which cells ware grown in LB and were harvested at the logarithmic growth phase. We used published RNA data for *B. subtilis* [169]. Translation initiation rates was computed with the same initiation rate model as was used for the library variants [165,166].

mRNA secondary structure, codon-decoding speed and aSD affinities were calculated as explained for the library variants. MTDR values for both species were taken from published data [111]. Amino acid metabolic cost was calculated as the mean value for the entire protein, and for both species the same cost was assigned for each amino acid [120]. Protein expression levels for both species were taken from the integrated datasets in Pax-Db [170].
As a negative control for the prediction of fitness residual for natural *E. coli* genes, we generated a mock model by randomly shuffling each of the features in the library, training a linear regression model on this shuffled library and using it to predict fitness residual for all *E. coli* genes. We then compared the standard deviation of the fitness residual predictions by the real model to the one of the mock model. This analysis was repeated 10^5 times to compute a p-Value for the chance of the real model to show a higher standard deviation than the mock model.

Cross validation sets

Cross validation tests of the regression model were performed by randomly choosing training and test sets, in proportions of 70% and 30% of the entire library variants, respectively. In order to account for the fact that some of the information lays in the amino acid sequence, the training/test sets were also separated by the N-terminus amino acid peptide sequences with 41 peptide sequences (~30%) chosen as test set, and the rest as training sets. 10-fold cross validation was performed by randomly generating ten different pairs of training and test sets. The results are based on the average across these 10 repeats.

RNA fitness residual calculation

To evaluate mRNA fitness residuals, we repeated the same calculation as described for fitness residual only with the mRNA levels instead of protein levels placed on the x-axis.

7.2. Optimization of gene expression through promoter architecture

Library design

This library includes ~2000 different synthetic promoters upstream of a YFP reporter gene. Each promoter is composed of a *cis*-regulatory element (CRE) which is built from random ligation of different TFBSs for four different transcription factors. Each transcription factor is represented by three possible sites, resulting in 12 different TFBSs represented in the library. The four TFs represented in the library are: GCR1, MIG1, RAP1, & REB1. In addition, all promoter variants have a basal TATA-box containing a minimal promoter downstream to the

CRE. In order to recognize uniquely each variant both in DNA and RNA sequencing, a unique barcode was added to the 3' UTR of the reporter gene, and each barcode was paired to its appropriate CRE through sequencing (See Fig 8).
I got the library as 3 separate yeast cells pools frozen in glycerol. The 3 sets represented 3 different parts of the library, and they were separated due to the synthesis technique employed in the Cohen lab when creating it. After receiving the pooled libraries, I mixed them into a single pool. Each of the sets had different number of unique variants, an I wished to mix such that eventually all variants will be represented equally. Hence, before mixing I counted the cell density in each subpool, and mixed them in appropriate ratios according to the number of unique variants in each subpool.

Growth media

Growth media used for this part:

- YPD - 10g/L yeast extract, 20 g/L peptone, 20 g/L glucose
- YPD + DTT - 10g/L yeast extract, 20 g/L peptone, 20 g/L glucose, 1.7mM DTT

Competition experiment

Ten independent repeats of a competition experiment were carried out in YPD media, another ten repeats were carried out in a YPD+DTT media. Each repeat was diluted in 1:120 ratio once a day (~ seven generations), and every two days a sample was frozen from each repeat. The experiment lasted for 16 days (~ 112 generations).
Following the competition experiment, genomic DNA was extracted from samples of the ancestral population, and days 6, and 10 of the competition experiment.

Fitness estimation

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, I 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 both for generation ~42 and generation ~70.

RNA extraction, genomic DNA extraction, and next generation sequencing

RNA extraction, cDNA synthesis, genomic DNA extraction, library preparation for next generation sequencing, and mapping of sequencing results were done in the same manner as was done for the non-coding RNA library (See section 7.3) .
Primers used for next generation sequencing library preparation are prDSO1-7 (see table S2)

7.3. Non-coding RNA library in yeast

This section relates to results sections 6.1.3, 6.1.4, and 6.2

Synthetic library - general design notes

I used Agilent's oligo library synthesis technology[19] (Agilent Technologies) to produce a pool of 45,000 designed single-stranded DNA oligos at a length of 230 nucleotides. Each oligo includes two 30 nucleotides fixed homology regions at their 5' and 3' end for amplification and cloning, and a 12 nucleotide unique barcode downstream to the 5' homology. This leaves an effective variable region of 158 nucleotides for each variant. The entire synthesized library was composed of 3 sub-libraries that are separated in the initial amplification stage using different homology sequences.
The experiments and data reported in sections 6.1.3 and 6.2 are based on one sublibrary with 24,510 variants (termed SplicingLib1). The experiments and data reported in section 6.1.4 are based on a second sublibrary with 13,257 variants (termed NucleotideCompLib).
Barcodes were chosen such that the minimal edit distance between any two barcodes will be greater than 3 to allow for single error correction for all types of errors including insertion/deletion which are the common error types in oligo synthesis.
The library was designed as a non-coding RNA library to avoid possible differences between variants that result from translation. Hence, for each variant, any occurrence of ATG triplet at any frame was mutated to avoid occurrences of a start codon. Except for cases where a 5'SS includes an ATG triplet, in which case, a stop codon was introduced 2 codons downstream of the ATG.

Synthetic library – splicing library variants design

The synthetic introns library is composed of several sets of variants. A first set is based on a combinatorial assembly of intron features. Six features were chosen to represent an intron, and all the possible combinations of features were combined to create a set of 5,331 synthetic introns. The features used for this set are: the three splice sites, 5'SS, BS, 3'SS, intron length, BS-to-3'SS length, and a 3' U-enriched sequence element. For each feature, a set of few values was chosen, the 5'SS and 3'SS sets included all the splice sites variants that are found

in *S. cerevisiae* genome (5 sequence variants for the 5'SS, and 3 for the 3'SS). The BS included the consensus BS sequence (TACTAAC), and three template sequences with two random nucleotides at the first two positions, since non-consensus BS sequences differ greatly in these positions (NNCTAAC, NNCTAAT, NNTTAAC). For the intron length feature, 5 representing lengths were chosen (73, 89, 105, 121, and 137 nucleotides), and for the BS-to-3'SS length 4 representing lengths were chosen (20, 30, 40, and 50 nucleotides). For the 3' U-enriched sequence element, 3 sequences at different lengths were used (ATTTTTAA, TTTAA, TAA). In addition, For each of the splice sites, a random control sequence was created and a set of control variants was created by assembling the three control sites with all combinations of the other features.

Full oligo sequences were based on a background sequence that was derived from the intron-containing region of *MUD1* gene from *S. cerevisiae* genome (positions 4-161 in MUD1 open reading frame), followed by randomization of its three splice sites. Each oligo sequence was created by placing a 5'SS 5 nucleotides downstream of the effective variable region instead of the background sequence in this position. Then a BS, U-rich element, and 3'SS sequences were placed in a similar manner according to the chosen length parameters of each variant. In addition, a set of 2,094 variants was created by taking only the consensus splice site sequences at different lengths and incorporating them within 9 additional background sequences, the first two from UBC9, and SNC1 genes in a similar manner, and the remaining 7 based on random sequences.

A second set is based on mutating consensus sites' variants from the previous set. 3,607 variants were created by introducing random mutations to splice site sequences. 1,344 additional variants were created by mutating positions adjacent to splice sites with the aim to create a stem-loop RNA structure at the splice sites. This aim was achieved by introducing random mutations and selections *in-silico* of variants for which RNA secondary structure tool[167] predicts that the splice site will be base-paired within a stem-loop structure.

A third set was based on 1,297 naturally occurring introns from 11 yeast species. I first took all the endogenous intron sequences from *S. cerevisiae*[155] that fit into my 158 nucleotides effective variable region (149 introns). Each intron was inserted with a flanking region of 5 nucleotides from each side on the background of the *MUD1* derived background sequence described above. Next, I took intron sequences from orthologs of these intron-containing genes from a set of 10 other yeast species and added them to the library in the same manner. Intron annotations were taken from[48]. In addition, I also incorporated into the library a set of 200 randomly chosen introns from *S. pombe* that have *S. cerevisiae*-like splice sites. For the

S. cerevisiae and *S. pombe* sets, I also created a set of 3,151 variants with random mutations in introns' splice sites.

Finally, a fourth set of 1,467 variants was created by combining two intronic sequences to create synthetic two-intron variants. For this set, I chose all the introns from *S. cerevisiae* genome shorter than 76 nucleotides (10 introns), plus 10 randomly chosen short introns from *S. pombe* genome and an additional 5 synthetic introns based on combining consensus splice sites on the background of a random sequence. Each variant sequence was created by placing two introns on the background of the *MUD1* sequence, the first intron at the 5' end of the variable region, and the second at the 3' of the variable region. All possible pairs of intronic sequences were created and introduced to the library. In addition, for each two-intron variant, a corresponding variant of a joined long intron was created by removing the BS and 3'SS of the first intron, and the 5'SS of the second intron. These sequences were replaced by the corresponding sequence in the *MUD1* background sequence.

Synthetic library – nucleotide composition variants design

The nucleotide composition library was designed by systematically creating random sequences with a defined nucleotide composition. I defined 405 sets of quartets representing the relative frequency of each of the four nucleotides. For each such set I randomly chose 25 158 nucleotides sequences taken from the distribution defined by the composition set. In addition, when assigning the 12 nucleotide barcode to this library I chose barcodes with similar nucleotide composition to be assigned to each variant.

Construction of master plasmid

In order to integrate the library into *S. cerevisiae* genome, I used a Cre-Lox based method[138]. I built a master plasmid to clone the library into, which is compatible with this method. The master plasmid was based on pBAR3[138]. A Lox71 site was cloned into pBAR3 to allow Cre-Lox recombination using restriction-free cloning method[171] (primers prDS20, prDS21) to create pDS101. Then I cloned into the plasmid a background sequence that will serve as the library's non-coding gene. A non-coding sequence was designed by taking the sequence of *MUD1* intron-containing gene from its transcription start site to its 3' UTR(-70 to 1106, relative to the start codon), excluding a region around the intron into which the oligo library would be cloned (-45 to 211, relative to start codon). The background sequence was then mutated at any occurrence of ATG to avoid start codons, and additional 27 sites were mutated to reduce homology to the endogenous copy of *MUD1* in the genome. In the cloning

site of the oligo library two 20 nucleotides sequences were added, to be used as homology sequences during library cloning. Upstream to the background gene I added a synthetic promoter taken from a published promoter library that was chosen for its high expression level and low noise (Promoter id #2659, from Supp table 3 in Sharon et al.[20]). Downstream to the background gene I added ADH1 terminator sequence. The entire promoter+background gene+terminator construct (total length of 1,397 nucleotides) was synthesized as a Gene Fragment (Twist Bioscience). The synthesized background gene was cloned into pDS101 using NEBuilder HiFi DNA Assembly (New England Biolabs) to create pDS102 (primers prDS22, prDS23).

Synthetic library - cloning and amplification of plasmid library

Synthetic oligos were first amplified according to Agilent's recommendations[172]. Library oligos were amplified using sublibrary specific homology plus 4 different 8 nucleotide sequences that were inserted to serve as an index for control purposes, such that every unique variant could be measured independently 4 times, and a homology sequence to the master plasmid for cloning.

The library was amplified in 4 PCR reactions, Each PCR reaction included:

- 25 ul - KAPA HiFi HotStart ReadyMix (Roche)
- 1.5 ul - 10uM forward primer (prDS45-48 for SplicingLib2, prDS55-58 for SplicingLib1)
- 1.5 ul - 10uM reverse primer (prDS49 for SplicingLib2, prDS59 for SplicingLib1)
- 200 pM of DNA oligo library
- H_2O to complete volume to 50ul

PCR program:

- 95°C 3 min
- 98°C 20 sec
- 58°C 15 sec
- 72°C 15 sec
- Repeat steps 2-4 for 15 cycles
- 72°C 1 min

After amplification, the PCR product was cut from gel, purified using Wizard SV Gel and PCR Clean-Up System (Promega), and all 4 reactions were pooled together. The master plasmid pDS102 was linearized using PCR reaction (primers prDS62, prDS63). Then a

plasmid library was assembled using 4 independent reactions of NEBuilder HiFi DNA Assembly (New England Biolabs) to avoid biases in assembly that might affect the library's distribution.

From this stage, I followed Agilent's library cloning kit protocol[173] steps 2-7. In short, the plasmid library was purified using AMPure XP beads (Beckman Coulter), then inserted to electrocompetent *E.coli* cells (ElectroTen-Blue, Agilent Technologies) using electroporation. Then bacterial cells were inoculated into two 1 liter low gelling agarose LB bottles, in order to grow isolated colonies in 1-liter volume. After 48 hours of growth in 37◦C bacterial cells were harvested using centrifugation, and cells were grown overnight on liquid LB media in 37◦C. Finally, the amplified plasmid library was extracted from bacterial cells using 4 reactions of Midiprep kit (Macherey-Nage NucleoBon Xtra Midi Plus).

Growth media

Growth media used in this work:

- YPG - 10g/L yeast extract, 20 g/L peptone, 20 g/L galactose
- YPD - 10g/L yeast extract, 20 g/L peptone, 20 g/L glucose
- SC complete - 6.7 g/L nitrogen base without aminoacids, 20 g/L glucose, 1.5 g/L amino acid mix
- SC -URA - 6.7 g/L nitrogen base without aminoacids, 20 g/L glucose, 1.5 g/L URA drop-out mix

Yeast strain construction

The yeast strain for which the library was inserted was based on SHA185, *S. cerevisiae* strain with Cre-Lox landing pad (derived from BY4709, ura3Δ ybr209w::GalCre-KanMX-1/2URA3 -lox66) kindly supplied to me by Sasha F. Levy's lab. In order to measure the effect of the library's synthetic intron on splicing efficiency of other intron-containing genes *in-trans* I inserted into the genome a fluorescent reporter cassette. A cassette of mCherry, and intron-containing YFP was taken from an existing reporter library that includes all natural introns from *S. cerevisiae* genome[64]. From this library I chose an intron with medium splicing efficiency (YDL108W intron) such that it will have potential to increase or decrease its efficiency.

As part of this work I measured the relative fitness of each variant in the library using a competition assay. I did not want the intron-containing reporter to be expressed during the competition assay to avoid any feedback effect due to differential expression of the reporter.

Therefore, I replaced the promoter of the intron-containing YFP with an inducible *CUP1* promoter.

The modified reporter cassette was assembled from three fragments using NEBuilder HiFi DNA Assembly (New England Biolabs). The mCherry fragment, and YFP-NAT fragment were amplified from the intron-containing reporter strain kindly supplied to me by Maya Schuldiner's lab. *CUP1* promoter fragment was amplified from BY4741 *S. cerevisiae* strain (primers prDS1, prDS6-10). The assembled cassette was transformed to the HO locus of SHA185 *S. cerevisiae* strain to create yDS101 strain, and was plated on YPD+NAT agar plates for selection.

Synthetic library - integration into yeast genome

Transformation of the plasmid library to yeast cells was done using a Cre-Lox based high throughput genomic integration method[138], that inserts the plasmid sequence into the YBR209W dubious open reading frame. yDS101 yeast cells were transformed with 500ug plasmid library and grown overnight in YPG media to induce Cre expression. Then cells were plated on selective media (SC-Ura) approximately 50 plates per transformation. I counted the number of colony-forming units by plating diluted samples and got 1.5106 CFUs for SPlicingLib1, and 0.9105 CFUs for SplicingLib2 which are ~60 and ~50 times the number of unique variants in the library accordingly.

Flow cytometry sorting to determine YFP fluorescence levels of library variants

In order to measure YFP fluorescence levels of each variant in the library I used a FACS sorting assay followed by deep-sequencing of each bin. Yeast library cells were grown for 24 hours in SC complete media, and then diluted to fresh SC complete + 100uM Copper(II) to induce YFP expression, and grown overnight, dilution ratio was calculated such that the culture will reach a cell density of 10^7 cells/ml (mid-log phase) the morning after.

Cells were sorted using BD FACSAria FUSION in the Weizmann institute flow cytometry unit. Cells were sorted into 12 equally populated bins according to their YFP/mCherry ratio. YFP fluorescence level represents the in-trans splicing efficiency, and mCherry is used to normalize for variation between cells. In each bin I sorted 3105 cells, and then grown them overnight before harvesting the cells for genomic DNA extraction.

YFP fluorescence levels were measured only for SplicingLib1.

Pooled competition assay to measure relative fitness

To measure the relative fitness of each variant in the library I used a pooled competition assay, followed by deep sequencing of the library's barcode. The competition was done in 3 independent repeats. The library was grown in ~3ml of SC complete media + Doxycycline to prevent bacterial contamination in 15 ml falcon tubes at 30◦C. Once every 24 hours, 96ul of the culture were transferred to 2.98ml of fresh media. Then 1.5ml of the grown culture from the previous day, was centrifuged, and the pellet was frozen for DNA extraction. I conducted the competition for 20 days, which accounts for 100 generations.

RNA extraction, cDNA synthesis, and genomic DNA extraction

Total RNA of the library cells was extracted in two independent repeats. Library cells were grown overnight in SC complete media, and then diluted to a fresh media by 1:100 and grown for an additional 6 hours until they reached OD_{600} of 0.5, such cells are harvested in mid-log phase. The cell culture was centrifuged for 45 seconds at 4,000g and the pellet was immediately frozen in liquid nitrogen.

RNA was extracted using MasterPure Yeast RNA Purification Kit (Lucigen), and treated with TURBO DNase (ThermoFischer) to remove any residues of genomic DNA. I then synthesized cDNA using reverse transcription with random primers using qScript Flex cDNA Synthesis Kit (QuantaBio).

To normalize RNA levels by the relative frequency of each variant in the sample, I extracted genomic DNA from the same samples used for RNA extraction. Cells were harvested and frozen at mid-log the same as for RNA extraction. In addition, genomic DNA was also extracted from frozen samples taken from the FACS sorting, and competition experiments. DNA was extracted from all samples using MasterPure Yeast DNA Purification Kit (Lucigen).

Next-generation sequencing - library preparation

Both cDNA and genomic DNA samples were prepared for sequencing in the same manner. I used a two-step PCR protocol to amplify the library's variable region and link it to Illumina's adaptors with indexes.

The first PCR reaction was used to amplify the variable region and link homology sequences to Illumina's adaptors, I performed 8 parallel reactions to each sample to reduce PCR biases. I used 6 different forward primers each with one extra nucleotide, to create shifts of the amplicon sequence to avoid low complexity library.

Each reaction included:

- 25 ul - KAPA HiFi HotStart ReadyMix (Roche)

- 1.5 ul - 10uM forward primer (prDS137-142)
- 1.5 ul - 10uM reverse primer (prDS143)
- 100ng DNA
- H_2O to complete volume to 50ul

PCR program:

- 95°C 3 min
- 98°C 20 sec
- 58°C 15 sec
- 72°C 15 sec
- Repeat steps 2-4 for 20 cycles
- 72°C 1 min

Next, I pooled all 8 reactions for each sample and purified the PCR product using AMPure XP beads (Beckman Coulter). The second PCR was used to link specific indexes to each sample so I can multiplex several samples in a single sequencing run.
Each reaction included:

- 25 ul - Phusion High-Fidelity PCR Master Mix with HF Buffer (New England Biolabs)
- 2.5 ul - 10uM forward primer (prDS144)
- 2.5 ul - 10uM reverse primer (prDS145)
- 1-5ng DNA
- H_2O to complete volume to 50ul

PCR program:

- 98°C 30 sec
- 98°C 10 sec
- 62°C 20 sec
- 72°C 15 sec
- Repeat steps 2-4 for 15 cycles
- 72°C 5 min

Next, I purified the PCR product using AMPure XP beads (Beckman Coulter), quantified final concentration using Qubit dsDNA HS (ThermoFischer), diluted all samples to 4nM, and pooled together all the samples. NGS library was sequenced in Illumina NextSeq 500 system,

using 150x2 paired-end sequencing. I obtained a total of 13.9, 12 million reads for the two RNA samples of SplicingLib1, and 1.9, 1.9 million reads for the two corresponding DNA samples, and 1.6, 2 million reads for the two RNA samples of SplicingLib2, and 0.25, 0.2 million reads for the two corresponding DNA samples. Samples from the competition assay had average coverage of 10 million reads per sample, and samples from the FACS sorting assay had average coverage of 4 million reads per sample.

Mapping sequencing reads to the library's variants

Sequencing reads from all samples were processed the following way: I first merged paired-end reads using PEAR[174], next I trimmed homology sequences and demultiplexed the reads according to the 4 control indexes using Cutadapt[175], then I clustered all unique reads using 'vsearch --derep_prefix'[176].

All the unique reads were mapped to a library variant according to the first 12 nucleotides in the read, which are the designed barcode. A read was mapped to one of the library's variants by searching the barcode with minimal edit distance to the read's barcode. If this minimal distance was <3, and only a single library barcode is found in this distance, the read was aligned to this variant.

Data analysis

All data analysis except the gradient boosting model were done in Matlab (R2018b). Gradient boosting modeling was done in Python 3.7.

Computing splicing efficiencies

For each variant, the mapped reads obtained from the RNA sequencing were first classified into three possible types: unspliced, intended spliced isoform, and undetermined. A read was classified into one of these types using an alignment of 40 nucleotide sequences representing *exon-intron*, and *exon-exon* junction sequences. I aligned each read to the reference junction sequences using local Smith-Waterman alignment (swalign function in Matlab), and a normalized alignment score was defined the following:

$$Junction\ alignment\ score = \frac{SW(junction, read)}{SW(junction, junction)}$$

If the normalized score was >0.8 I infer the junction is positively aligned to the RNA read. A read was classified as unspliced if it was aligned to the *exon-exon* junction, and not aligned to the two *exon-intron* junctions. A read was classified as 'intended spliced isoform' if it was aligned to the *exon-exon* junction and not to the two reference *exon-intron* junction. All other reads were classified as undetermined.

Intended splicing efficiency for each variant was then calculated for each index according to:

$$SE = \frac{spliced\ isoform\ abundance}{total\ RNA\ abundance}$$

A final splicing efficiency value for each variant was then set by taking the median between indexes in each repeat and then taking the mean between the two repeats.

The undetermined reads were further analyzed to search for unintended spliced isoforms, meaning, isoforms that result from splicing of an intron different than the designed intron, hence no *exon-intron* reference junction could be defined. Each read was aligned against the full reference design with the following parameters to the Smith-Waterman algorithm (Gapopen = 100, ExtendGap=1) in order to allow for alignment with long uninterrupted gaps, if the normalized alignment score was <0.7 and the number of mismatches in the alignment was <6, a read was set as unintended spliced isoform. Then the 5' and 3' end of the intron were set according to the ends of the uninterrupted gap.

For each variant, unintended isoforms were clustered according to their 3' and 5' intron ends, and for each cluster, I calculated the splicing efficiency as described above for the intended spliced isoforms. Unintended spliced isoforms were counted only for isoforms with splicing efficiency higher than 0.01.

Computing two-intron spliced isoforms ratio

For the set of two-intron variants, I needed to classify each read to one of five possible isoforms: unspliced, intron 1, intron 2, exon-skipping, or 'both introns spliced'. Reads were classified into one of these isoforms according to junction's alignment as described above. A read was classified to an isoform according to the following conditions:

- Intron 1 - positive alignment to the *exon1-exon2* junction, and negative to the *exon1-intron1* and *intron1-exon2* junctions.
- Intron 2 - positive alignment to the *exon2-exon3* junction, and negative to the *exon2-intron2* and *intron2-exon3* junctions.
- Exon-skipping - positive alignment to the *exon1-exon3* junction, and negative to the *exon1-intron1* and *intron2-exon3* junctions.
- Both introns - positive alignment to the concatenated *exon1-exon2-exon3* junction, and negative to all the 4 *exon-intron* junctions.
- Unspliced - negative alignment to both *exon1-exon2* and *exon2-exon3* junctions, and positive alignment to all 4 *exon-intron* junctions.

Then the splicing ratio of each isoform was determined by the ratio of its spliced isoform abundance and the total RNA abundance.

Total RNA abundance, and data filtering

Genomic DNA levels were used to determine total RNA abundance, and to filter outlier spliced isoforms.

To determine total RNA abundance I wish to normalize by the variant's frequency in the population. Hence, Total RNA abundance of variant x was determined according to:

$$Total\ RNA\ abundance(x) = \log_{10}\left(\frac{RNA\ frequency(x)}{DNA\ frequency(x)}\right)$$

Some RNA read alignments might be inferred as spliced isoforms due to errors in synthesis, or systematic errors in alignment. Therefore, the splicing efficiency calculation was done also on the DNA samples, and if a variant had an intended or unintended splicing efficiency higher than 0.05 in the DNA samples the corresponding value was set to zero.

Inferring YFP fluorescence levels of each variant

Fluorescence levels of each variant were inferred using the same method described in (Kosuri et al. 2013). In short, I mapped each read to one of the library's variants using barcodes as described above. Then I normalized the read count in each bin by the total number of reads in the bin to get the relative frequency of each variant within the bin f_{ij} fij.

$$f_{ij} = \frac{reads_{ij}}{\sum_i reads_{ij}}$$

Then for each variant I normalized the read frequency in a bin, by the sum of frequencies across all bins, to get its relative representation in each bin a_{ij}.

$$a_{ij} = \frac{f_{ij}}{\sum_j f_{ij}}$$

Then, the fluorescence level for each variant Y_j was calculated by taking the weighted geometric mean of fluorescence levels across bins, where a_{ij} is the weight of bin j, and m_j is the median YFP fluorescence level of bin j.

$$Y_j = e^{\sum_j a_{ij} \log(m_j)}$$

Inferring relative fitness from competition assay

Relative fitness of each variant in the library was inferred from sequencing 5 time points along a 20 days competition assay. I sequenced samples from days: 0,1,2,17,20 which correspond to generations 0,5,10,85,100. I mapped each read to one of the library's variants using barcodes as described above.

For each sample I calculated the relative frequency of each variant, by normalizing the read count of each variant by the total read count of this sample. Then for each variant I calculated fitness by performing log-linear regression between the vector of frequencies for each variant, and the vector of generations. The fitness is then taken as the slope of this regression. The competition was done in 3 independent repeats, and in addition each variant was cloned with four 8 nucleotides sequence which serves as an index, to allow me to detect outliers that might result from beneficial mutations *in-trans*. When looking at correlations between repeats and indexes I see weak positive correlations. So, I wanted to examine only variants that present reproducible fitness measurements. For each variant I infer 12 fitness measurements (3 repeats times 4 indexes). I calculate the mean and standard deviation of this fitness vector for each variant, and consider only variants for which the coefficient of variation is smaller than 1 (<1), and the standard deviation is smaller than 0.01 (<0.01). This method was used both for the synthetic intron library (section 6.1.3) and the nucleotide composition library (section 6.1.4).

3'SS avoidance calculation

For each of the 11 yeast species, I examine the frequency of the 3'SS sequence motif, to check if it is avoided near introns' 3' end. First, I calculate the frequency of the two major 3'SS sequences ([C/T]AG) at positions relative to the introns' 3' end. For that purpose, in each species, I register the sequences of all the intron-containing genes at their intron's 3' end and set the end of the intron as position 0. Then, at every position downstream or upstream to the intron end, I count the number of occurrences of the two 3'SS sequences and divide it by the number of introns in each species.

Then I test if there is a statistically significant depletion of this motif at a window of 30 nucleotides upstream or downstream of the intron end. I perform the statistical using sampling of random control sets. Each control set includes N random positions from coding regions in the same genome, where N is the number of introns in a species. Those positions are set as the reference positions at which I register N sequences and measure the 3'SS motif frequency around them. I randomly sample 10^5 such control sets, and then I count the number

of sets for which the mean frequency within a window of 30 nucleotides is lower than the mean frequency in the true introns set. p-value is defined according to: ($f_{motif}^{introns}$ and $f_{motif}^{control}$ are the frequency of the 3'SS motif at the true introns set, or the control set accordingly)

$$p-value(upstream) = \left[\sum_{i=-32}^{-3} f_{motif}^{control} < \sum_{i=-32}^{-3} f_{motif}^{introns}\right] \cdot 10^{-5}$$

$$p-value(downstream) = \left[\sum_{i=1}^{30} f_{motif}^{control} < \sum_{i=1}^{20} f_{motif}^{introns}\right] \cdot 10^{-5}$$

A computational model for predicting single-intron splicing efficiency

I used a gradient boosting regression model to predict splicing efficiencies of library variants. The gradient boosting implementation is based on LightGbm[152] library for Python, and the feature importance inference is based on SHAP[154] library for python.
Each variant is characterized by a set of 39 features (see table S1). I took a set of 15,516 variants that includes all the designed single intron variants, excluding negative controls and the 1,704 variants set with identical intron features. This set was randomly divided into a training set composed of 75% of the variants and a test set with the remaining 25%. I then trained the model on the training set using 5-fold averaging cross-validation technique[153], meaning, I divided the training set to 5 subsets, each time training the model on 4 of them, using the fifth as a validation set, and predicting the splicing efficiency value for the test set. Thus, creating 5 different predictions for the test set, which I next averaged to create a single prediction.
The parameters given to the model are the following:

- Number of leaves - 50
- Learning rate - 0.1
- Feature fraction - 0.8
- Bagging fraction - 0.8
- Bagging frequency - 5
- Number of boost rounds - 500
- Number of early stopping rounds - 5

Feature importance was inferred by running Shapley value analysis[154] on the training set for each of the 5 k-fold iterations, followed by averaging the Shapley values over the 5 iterations.

A computational model for predicting two-intron spliced isoforms

I used gradient boosting classification to predict which set of isoforms will be produced from a two-intron variant. The gradient boosting implementation is based on LightGbm library for Python, and the feature importance inference is based on SHAP library for python.

I took a set of 1,414 two-intron variants, and each variant was characterized by the 39 features of each of its two introns (the same features as for the single intron variants, see table S1), in addition, the GC content and intron length of the exon-skipping intron were also added as features to create a set of 80 features (39 features for each of the two introns, plus 2 extra features for the "exon-skipping" intron). I then classified the variants according to the splice isoforms seen for each variant in the data (excluding the 'both introns spliced' isoform because there were not enough cases for it). Each class is characterized by a set of isoforms that appear together for the same variant, the isoforms are namely, intron 1, intron 2, and exon-skipping.

In this classification task, I are interested specifically in the feature importance, so I trained a model on the entire set using 5-fold averaging cross-validation technique as described above. The Shapley values analysis outputs a set of Shapley values for each feature, for each of the 8 classes. I are looking for features that have a high contribution in at least one of the classes. So, I decided to look only at features that have an absolute mean Shapley value greater than 3 times the standard deviation of all Shapley values, in at least one of the features. I find 6 such features (Fig 21D). Then for each of these features in each of the classes, I present the absolute mean Shapley value which presents the total importance of this feature for this class, multiplied by the feature importance sign. A positive sign means that a variant that has a value in this feature which contributes positively to splicing efficiency, is more likely to be found in this class, while a negative sign, means that a variant that has a value in this feature that contributes negatively to splicing efficiency is likely to be found in this class. Feature importance sign is calculated by computing the Pearson correlation between the feature values and the Shapley values both for the single-intron set described above, and the two-intron set for each class described here. If the two correlations have the same sign (meaning, the feature has the same effect as in the single intron case), the feature importance sign is set to '1', otherwise, if the two correlations have opposing signs, the feature importance sign is set to '-1'.

7.4. Yeast SUC2 secretion libraries

Strains and plasmids

All *Saccharomyces cerevisiae* strains used for this part are based on the BY4741 (MATa his3Δ1, leu2 Δ 0, met15 Δ 0, ura3 Δ 0) genetic background. The library was inserted on a background of deletion of the natural copy of SUC2 gene, so the strain used for the library construction was BY4741 ΔSUC2::NAT.

pGS2223 – Plasmid created for the construction of the Suc2 secretion variant library. It is based on the integrative plasmid pCfB2223 from the EasyClone 2.0 Yeast Toolkit[159], which includes a Kan resistance gene and homologous regions for an integrative site on chromosome X.33 The WT *SUC2* gene was cloned into this plasmid to create pGS2223 using the NEBuilder HiFi DNA Assembly according to its reaction protocol.

Growth media

- YPD - 10 g/L yeast extract, 20 g/L peptone and 20 g/L glucose.
- YP-Gly - 10 g/L yeast extract, 20 g/L peptone, 30 g/L glycerol.
- SD - 6.7 g/L nitrogen base, 1.5 g/L mix of all amino acids and 20 g/L glucose.
- SC-sucrose - 6.7 g/L nitrogen base, 1.5 g/L mix of all amino acids and 20 g/L sucrose.
- SC-low sucrose- 6.7 g/L nitrogen base, 1.5 g/L mix of all amino acids, 0.5 g/L sucrose, and 0.05 g/L glucose.
- SC-low galactose- 6.7 g/L nitrogen base, 1.5 g/L mix of all amino acids, 1 g/L galactose
- LB – Media composed of tryptone 10 g/L, yeast extract 5 g/L, Nacl 10 g/L.

Cloning of plasmid sub-libraries

~18,000 DNA fragments of 300bp each, that contain the designed sequences, were ordered from Twist Bioscience. From these fragments, 200bp-long oligos were amplified into sub-libraries based on their location in the *SUC2* gene, with the middle 150bp being the variable region. The 100bp that were not amplified are not part of the *SUC2* gene and were relevant for a previous method that was tried for cloning the secretion variant library, which proved not to work for our needs.

For the SECReTE sub-library each PCR reaction included:

- 25ul Phusion High-Fidelity PCR Master Mix

- 7.5 ul 10uM F primer (primers prGS3, prGS5 table S2)
- 7.5 ul 10uM R primer (primers prGS4, prGS6 table S2)
- 0.75 ul library DNA
- ddW: complete volume to 50ul

PCR program:

- 98°C - 30 sec
- 98°C - 10 sec
- 51°C - 20 sec
- 72°C - 30 sec
- Repeat steps 2-4 for 20 cycles
- 72°C - 10 min

For the SP sub-library each PCR reaction included:

- 25ul Phusion High-Fidelity PCR Master Mix
- 11.5 ul 10uM F primer (prGS1, table S2)
- 11.5 ul 10uM R primer (prGS2, table S2)
- 0.2 ul library DNA
- ddW: complete volume to 50ul

PCR program:

- 98°C - 30 sec
- 98°C - 10 sec
- 47°C - 20 sec
- 72°C - 30 sec
- Repeat steps 2-4 for 20 cycles
- 72°C - 10 min

In parallel, the backbone plasmid pGS2223 (that the oligos will later be cloned into) was created by cloning the WT *SUC2* gene into the linearized integrative plasmid pCfB2223 from the EasyClone 2.0 Yeast Toolkit[159], using NEBuilder HiFi DNA Assembly cloning kit.
The linearization PCR reaction included:

- 25ul KAPA HiFi HotStart
- 1.5 ul 10uM F primer (prGS13, table S2)
- 1.5 ul 10uM R primer (prGS14, table S2)
- 25 ng plasmid DNA
- ddW: complete volume to 50ul

PCR program:

- 95°C - 3 min
- 98°C - 20 sec
- 56°C - 15 sec
- 72°C - 5 min
- Repeat steps 2-4 for 25 cycles
- 72°C - 9 min

pGS2223 was then linearized, specific to each sub-library location, by performing a PCR reaction that amplified the vector using primers in opposite directions to each other, which then removed the relevant variable region.

Each such vector linearization reaction included:

- 25ul KAPA HiFi HotStart
- 1.5 ul 10uM F primer (primers prGS15, prGS17, prGS19, table S2)
- 1.5 ul 10uM R primer (primers prGS16, prGS18, prGS20, table S2)
- 25 ng plasmid DNA
- ddW: complete volume to 50ul

PCR program (for SECReTE library):

- 95°C - 3 min
- 98°C - 20 sec
- 50°C - 15 sec
- 72°C - 5 min
- Repeat steps 2-4 for 25 cycles
- 72°C - 9 min

PCR program (for SP library):

- 95°C - 3 min
- 98°C - 20 sec
- 49°C - 15 sec
- 72°C - 1 min
- Repeat steps 2-4 for 25 cycles
- 72°C - 9 min

The amplified oligo sub-libraries were cloned into the relevant linearized plasmids using NEBuilder HiFi DNA Assembly cloning kit with a 1:2 vector:insert ratio, replacing the removed region. The product was then cleaned with 1.5X AMPure XP beads according to the one-sided manufacture protocol.

Bacterial transformation of sub-libraries

The clean product of the plasmid sub-libraries was transformed by electroporation into ElectroTen-Blue Electroporation Competent Cells (Agilent). For each sub-library, between 4-8 electroporation reactions were performed (depending on the resulting efficiency) and 5.5ul of the purified plasmids was added per reaction. The electroporation protocol was used from the Agilent's library cloning protocol[173]. In this protocol, the plasmid DNA was added to thawed ElectroTen-Blue Competent Cells and moved to a chilled cuvette which was then placed in the electroporator. After the cells were electroporated, they were immediately resuspended in a rich medium and moved to a fresh tube to recuperate in a shaking incubator set at 37˚c for one hour. Next, the cells were plated on 20-30 LB+amp plates for selection. The plates were then scraped and all colonies from the same sub-library were pooled together

reaching an estimated 1012 cells suspended in ~40-60 ml of DDW (2ml was used to scrape each plate). The estimated cell count of 1012 is based on the ~106 single colonies grown per sub-library and estimating about 106 cells in each single colony. From the pooled culture, some glycerol stocks were made by suspending the culture to a final concentration of 30% glycerol, then being frozen and stored at -80˚c. The plasmid sub-library was extracted directly from the rest of the pooled culture using Promega midiprep kit.

Integration of plasmid libraries into yeast cells

The purified plasmid sub-libraries were linearized using the restriction enzyme NotI, according to the manufacturer protocol. Sixteen such reactions were performed per sub-library to ensure sufficient material for the subsequent transformation.
The linearized plasmid product was then cleaned with 1.5X AMPure XP beads and transformed into Δ*suc2* yeast strain using heat shock transformation.
20 transformation reactions were performed per sub-library to ensure sufficient representation of the library. Transformed cells were plated on YPD + G418 + NAT selection plates. G418 selection was done to select for successful transformation of the *SUC2* gene, and NAT selection was done to select for the Δ*suc2* strain. The plates were then scraped and all colonies from the same sub-library were pooled together in ~80 ml of DDW. From this pool, some glycerol stocks were made by suspending the culture to a final concentration of 30% glycerol, then being frozen and stored at -80˚c
Due to the transformation resulting in not only the expected colonies (with library insert in the correct location) but also in petite yeast cells, we grew the pooled sub-libraries on glycerol for one day to exclude the petite cells from the pool. For the glycerol growth, 200ul from the pooled yeast sub-library (which contains ~$1.5x10^9$ cells) was added to 100ml of YP-glycerol and grown in a shaking incubator at 30˚c for ~ 24 hours. Then, glycerol stocks were frozen as previously described and stored at -80˚c. Colony PCR directed to insertion junctions was used to verify integration of the library in the correct location (See primers prGS27-30, table S2).

SECReTE sub-library Competition assay

Competition assays on two growth media was done for the SECReTE sub-library to measure the variants fitness on sucrose and galactose. Competitions experiments were done on SC-low sucrose and on SC-low galactose, both with added Doxycycline, at 30˚c for 4 days. Before starting the competition, the pooled sub-library was grown in 5ml of SC-low sucrose media and in 5ml of SC-low galactose media for 24 hours until reaching stationary phase. For

each medium, the OD600 of the culture was measured and then diluted to reach mid-log (in this case, OD600 of ~0.6) after 12 hours. This mid-log culture was the starter for the competition experiment since we did not want to include the lag of exiting stationary phase in our experiment and was diluted to reach OD600 of ~0.6 over the next 12 hours.
Throughout the experiment, for each medium, the cells were grown in 5ml cultures in glass 25ml falcon tubes in a roller. Every 12 hours, the OD600 of the culture was measured and based on the OD600, the culture was diluted to reach the OD600 of 0.6 at the following time point. This dilution factor was typically approximately 1:16 which allowed for ~4 generations per time point. At each dilution, cells were sampled for sequencing.
Competition in each media was done in 6 independent repeats, and samples from the following generations were used for sequencing {0, 4, 8, 12, 16, 24, 32}.

Signal peptide sub-library competition

Competition assays on two growth media was done for the signal peptide sub-library to measure the variants fitness on sucrose and galactose. Competitions experiments were done on SC-low sucrose and on SC-low galactose, both with added Doxycycline, at 30˚c for 16 days. The cell culture was diluted once every 24 hours when the cells had reached stationary phase at dilution rate of 1:50. At each dilution a cell pellet was frozen for subsequent sequencing. Bot competition experiments were done in 6 independent repeats. Samples were taken for sequencing at days {0,2,4,10,16} which represent approximately generations {0,11,22,56,90}.

Genomic DNA extraction, and next generation sequencing

Genomic DNA extraction, and library preparation for next generation sequencing was done in the same manner as for the non-coding RNA library (see section 7.3)
SECReTE library sequencing library preparation was done with primers prGS33-38, SP library sequencing library preparation was done with primers prGS43-48 (table S2).

Mapping sequencing reads to the library's variants

Sequencing reads from all samples were processed the following way: We first merged paired-end reads using PEAR[174], we trimmed homology sequences using Cutadapt[175]. Next, only for ancestor samples of each sub-library we clustered all unique reads using 'vsearch --derep_prefix'[176]. The purpose of this step was to detect if there are variants in the library that were not part of the original design, that could come from synthesis errors. Detected undesigned variants with >100 reads in the ancestor sample were treated as a a new

library variant. In the SECReTE library 5,474 undesigned variants were detected and added to the original 4,800 variants. In the SP library 759 undesigned variants were detected and added to the original 4,500 variants.

For mapping each read to its library variants we used bowtie2 alignment[177]. We construct an artificial genome with all the library variants (including the newly found undesigned variants) as "chromosomes" and aligned the next generations sequencing reads to this genome. Then we counted the number of aligned reads per each variant.

Definition of positive and negative fitness residual variants

The purpose of this analysis is to isolate variant groups that differ significantly in their fitness on sucrose compared to their measured fitness on galactose. We looked for variants that repeatedly demonstrate such behavior across all pairwise comparisons between repeats, hence we did the following analysis for all 36 pairwise comparisons of repeats on sucrose and galactose:

We first calculated a regression line which represents the best fit for agreement between the two conditions. For the purpose of this regression analysis we filtered out variants with low fitness values $\left(|fitness - \mu(fitness)| < 2\sigma(fitness)\right)$, as this variants' fitness measurement is more influenced by noise and by bias the regression analysis. We did a linear regression analysis with 95% confidence interval, then a variant was considered as positive fitness residual variant for this pairwise comparison, if its fitness value was larger by 0.005 than the higher confidence interval line, or negative if lower by 0.005 than the lower confidence interval line. Eventually variants were classified as positive or negative if they were classified as such in at least 34 of 36 pairwise comparisons.

8. Discussion

Cells in nature must execute different genetic programs at different conditions. To achieve this goal, the expression of each gene is regulated at various levels and at every stage in its life cycle. The synthesis of a gene is regulated both at the transcription level and at the translation level, and the steady state level of the gene is also regulated by regulating the degradation of both its mRNA copies, and the protein. Moreover, the delivery of the protein to its cellular destination is regulated by several additional mechanisms. In addition, in many cases, to produce a final protein product, mRNAs and proteins also undergo several processing steps like RNA editing, splicing, or post translation modifications. Each of these steps is regulated through complex cellular systems. The operation and regulation of these systems comes with a cost to the cells. As these systems are composed of limited resources and require energy to operate, they must obey economic rules of supply and demand and their activity can be optimized according to cost-benefit tradeoffs. Therefore, I hypothesized that different molecular machineries and their associated *cis*-regulatory elements have evolved to optimize these tradeoffs.

Some genes participate in a more complex economic setting, as they encode for a secreted protein which serves as a public-good compound. Meaning, a cell that express this protein invests the costs in expressing it, but the benefits from this protein are shared with neighboring cells. Through these genes, unicellular organisms apply cooperative strategies, and the economic principles that act on them involve complex game theoretic models of interaction between different actors.

In this study I set out to elucidate how the genetic architecture of different *cis*-regulatory elements contribute to the economy of gene expression. First within the cell, studying the economy of expressing genes that act only intracellularly, and second using a model public-good enzyme to study the economic principles that govern cooperative population dynamics. Understanding mechanistic features that contribute to optimizing gene expression, helps us to better understand how the genome was shaped through evolution through selection acting to better optimize gene expression.

I approached this question by utilizing and designing a set of synthetic oligo libraries. Through combination of massively parallel reporter assay (MPRA), pooled competition assay, and targeted deep sequencing, I used these libraries to systematically test how different regulatory elements contribute to gene expression regulation, and its economy.

In the first part of my research I demonstrated how the cost of gene expression in bacteria can be regulated even at a specific and desired expression level. I further demonstrated that this cost is affected by different features at all levels of the Central Dogma. I showed how the cost of producing a protein at a given expression level is affected by mRNA properties such as secondary structure, codon usage, and affinity to the anti-Shine Dalgarno (aSD) motif of the ribosome. I suggest that these features affect the optimization of gene expression through regulation of translation elongation speed. However, these features have opposing predicted effects on elongation speed, and therefore the contribution of translation regulation to cost optimization is yet undecided. I also demonstrated that properties of the amino acids that compose the final protein contribute to the cost of producing it. Either directly through the synthesis cost of the amino acids, or indirectly through possible toxic effects of the protein. I further demonstrated 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, this study suggests design elements that could be both utilized for better heterologous gene expression or by natural selection for the optimization of natural genes.

In addition, I showed that regulating initiation and mRNA levels affects expression cost, as increasing the number of proteins that are produced per mRNA is associated with reduced cost per expression level. This architecture could be beneficial because it reduces energy and resource consumption that are devoted to mRNA production. If cost-reducing, why genomes do not utilize even further the strategy of low transcription and mRNA abundance, combined with high translation initiation? One potential reason is that too low mRNA levels might lead increased response time to an environmental signal[125] or to increased expression noise[178], although the data from our experiments did not show a relationship between gene expression noise and cost. 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[179].

Following this observation, I decided to focus specifically on the cost of producing mature mRNA, by studying the cost of expressing a synthetic non-coding gene. For this purpose, I chose to work with the buddying yeast *S. cerevisiae* as a model system, since RNA processing is richer in eukaryotes and I predicted that genetic variation in a non-coding gene will result in higher variation in terms of cost.

First, I utilized a synthetic promoter library[26] to study the effect of promoter architecture on gene expression. While significant differences in fitness were observed in this library, I did not observe the predicted correlation to expression level. Presumably because in yeast the

differences in expression level of a single copy gene do not create sufficient differences in burden. I notice, that in other work that studied the cost of gene expression in *S. cerevisiae*[36], the range of expression load was much higher due to differences in copy numbers of the reporter gene. As for differences in cost between different promoter architecture, I observed an association between the number of transcription factor binding sites and cost, but the results were not conclusive enough.

Next, I designed a large oligo library of different designs of a synthetic non-coding gene to study the effect of transcribed genetic features on the cost of producing mature mRNA. Using a subset of this library I observed a surprising relationship between thymidine content, reduced cost, and increased expression level, suggesting a surprising phenomenon of negative correlation between cost of gene expression and expression levels. However, there are no sufficient evidence to support this potential conclusion.

I further used this library to study the effect of intron architecture on the cost of gene expression by introducing about 20,000 different synthetic introns into *S. cerevisiae* genome. I measured the cost of introducing an heterologous intron both through competition assay, and by an intron-containing reporter gene *in-trans*, to test how introduction of another intron imposes burden on the cell by affecting the splicing regulation of other intron-containing genes. However, in this study I did not detect measurable effects for introduction of a synthetic intron on the wellbeing of the cell. This result suggests that the operation of the splicing machinery on a single gene is not a limiting factor in terms of resource allocation in the cell. In contrast, a different work from our lab has shown that under a strong selective pressure cells can adapt by accumulating mutations in introns that increase the expression level of a resistance gene[180].

In a second part of this research I utilized the synthetic intron library mentioned above to decipher how intronic architecture affects splicing regulation in budding yeast. Since *S. cerevisiae* genome includes a low number of introns, the natural variation that exists in the genome is not sufficient to understand all the principles that govern splicing regulation in this organism. Therefore, this large synthetic library created a unique opportunity to decipher such principles.

A subset of this library was used to study the evolution of introns architecture by introducing into it natural introns from 11 yeast species. I tested if introns that come from different hosts that use modified splicing machineries are spliced differently when processed by the *S. cerevisiae* machinery. I suggest that a loss of a splicing factor (U2AF1) that occurred in some of the yeast species examined here, affects intron architecture through the distance between

the branch site and the intron's 3' end. I then observe that introns coming from hosts including this factor are spliced less efficiently in this system lacking the U2AF1 factor. Interestingly, in humans, a mutation in U2AF1 was associated with hematopoietic stem cell disorders that can progress into acute myeloid leukemia, and this mutation was shown to cause misplacing events which are presumed to be related with the disease[181–183].

I suggest that introns evolved to adapt to the presence or absence of U2AF1 splicing factor. Additionally, I found that the loss of this splicing factor is accompanied by a genomic depletion of the 3'SS motifs near introns' 3' ends. I hypothesize that in species that lack U2AF1 splicing factor the splicing machinery misidentifies the intended 3'SS due to similar sequence motifs near that site. This imposes a burden on the cell by creating unintended transcripts and proteins that need to be processed by the degradation machineries. This in turn creates a selective pressure to avoid such cryptic 3'SS motifs near the intended 3'SS. This hypothesis is supported by the fact that many cases of alternative 3'SS isoforms are observed both in this study and in previous findings in *S. cerevisiae*[60,150]. This is a demonstration of a biological machinery that is permissive, where even a few random mutations in the sequence can create a new recognition site for it to act on. Similar permissive behavior was recently shown in bacteria for the RNA polymerase and the probability of a random sequence to create a promoter[184].

Regulated alternative splicing is the focus of many studies on splicing because of its effect on increasing the proteomic diversity of the genome[67,68,70]. In yeast, there are very few known examples of functional alternative splicing[56–59]. However, using the relative simplicity of the yeast splicing machinery, I aimed to decipher novel ways by which alternative splicing is carried out in biological systems. The hallmark of such effort will be to create a synthetic gene that is alternatively spliced to different isoforms in reaction to different environmental conditions. Here I provide the first step towards this aim by creating a set of multiple two-intron designs and by characterizing which of them give rise to multiple RNA isoforms. I found that the potential for creating multiple isoforms is mainly dictated by the individual splicing capacity of each of the two introns irrespective of their relative location. If both introns are designed for efficient splicing, their corresponding two-intron design will have the capacity to create multiple isoforms. This result suggests that alternative splicing might be an easily evolvable trait whose evolution might have not necessarily required sophisticated evolution in *trans* as an initial evolutionary step.

Lastly, I presented here initial results of an ongoing project in which I study the population dynamics in two complex communities that include varying levels of production and\or

secretion of a public-good enzyme. The results presented here indicate that during a competition experiment of these populations different dynamics are observed at different time scale. I further hypothesize that these different temporal dynamics are associated with the differences between effects of expression cost on fitness, as opposed to shared enzymatic activity by the public-good protein. In a future work I will further characterize the variants in this library by measuring the enrichment of the variant's mRNA in the ER.

To summarize, by using the powerful tool of large synthetic libraries I was able to systematically explore the function and associated cost of different cis-regulatory elements in a manner that is not possible by studying merely the natural variation in the genome. Finally, these experiments enabled me to infer how economic principles of optimizing cost while maintaining desired expression levels have shaped the genome throughout evolution.

9. References

1. 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).
2. Bienick, M. S. *et al.* The interrelationship between promoter strength, gene expression, and growth rate. *PLoS One* **9**, e109105 (2014).
3. 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).
4. Glick, B. R. Metabolic load and heterologous gene expression. *Biotechnol. Adv.* **13**, 247–61 (1995).
5. Scott, M., Mateescu, E. M., Zhang, Z. & Hwa, T. Interdependence of Cell Growth Origins and Consequences. *Science (80-.).* **330**, 1099–1102 (2010).
6. Keren, L. *et al.* Massively Parallel Interrogation of the Effects of Gene Expression Levels on Fitness. *Cell* **166**, 1282-1294.e18 (2016).
7. Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**, 588–592 (2005).
8. Miller, M. B. & Bassler, B. L. Quorum Sensing in Bacteria. *Annu. Rev. Microbiol.* **55**, 165–199 (2001).
9. Carlson, M. & Botstein, D. Two differentially regulated mRNAs with different 5′ ends encode secreted and intracellular forms of yeast invertase. *Cell* **28**, 145–154 (1982).
10. Kapteyn, J. C. *et al.* The contribution of the O-glycosylated protein Pir2p/Hsp150 to the construction of the yeast cell wall in wild-type cells and beta1,6-glucan-deficient mutants. *Mol. Microbiol.* **31**, 1835–1844 (1999).
11. Moukadiri, I. & Zueco, J. Evidence for the attachment of Hsp150/Pir2 to the cell wall of *Saccharomyces cerevisiae* through disulfide bridges. *FEMS Yeast Res.* **1**, 241–245 (2001).
12. Nowak, M. A. Five rules for the evolution of cooperation, Science 314. *Science (80-.).* **314**, 1560–1563 (2007).
13. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–6 (1981).
14. Gore, J., Youk, H. & Van Oudenaarden, A. Snowdrift game dynamics and facultative cheating in yeast. *Nature* **459**, 253–256 (2009).
15. Greig, D. & Travisano, M. The Prisoner's Dilemma and polymorphism in yeast SUC genes. *Proc. R. Soc. B Biol. Sci.* **271**, 25–26 (2004).
16. Fiegna, F., Yu, Y. T. N., Kadam, S. V. & Velicer, G. J. Evolution of an obligate social cheater to a superior cooperator. *Nature* **441**, 310–314 (2006).
17. Chen, F.-C., Wang, S.-S., Chen, C.-J., Li, W.-H. & Chuang, T.-J. Alternatively and constitutively spliced exons are subject to different evolutionary forces. *Mol. Biol. Evol.* **23**, 675–682 (2006).
18. Ding, F. & Elowitz, M. B. Constitutive splicing and economies of scale in gene

expression. *Nat. Struct. Mol. Biol.* 1 (2019) doi:10.1038/s41594-019-0226-x.

19. LeProust, E. M. *et al.* Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38**, 2522–2540 (2010).

20. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* (2012) doi:10.1038/nbt.2205.

21. Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc. Natl. Acad. Sci.* **110**, 14024–14029 (2013).

22. Shalem, O. *et al.* Systematic dissection of the sequence determinants of gene 3' end mediated expression control. *PLoS Genet.* **11**, e1005147 (2015).

23. Weingarten-Gabbay, S. *et al.* Comparative genetics. Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science* **351**, (2016).

24. Maricque, B. B., Chaudhari, H. G. & Cohen, B. A. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4285.

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

26. Mogno, I., Kwasnieski, J. C. & Cohen, B. A. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res.* **23**, 1908–1915 (2013).

27. Geiler-Samerotte, K. A. *et al.* Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 680–5 (2011).

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

29. 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).

30. 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).

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

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

33. 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).

34. 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).

35. 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).

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

37. 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).

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

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

40. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

41. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457–463 (2010).

42. Kalsotra, A. & Cooper, T. A. Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* **12**, 715–729 (2011).

43. Köhler, U., Donath, M., Mendel, R. R., Cerff, R. & Hehl, R. Intron-specific stimulation of anaerobic gene expression and splicing efficiency in maize cells. *Mol. Gen. Genet.* **251**, 252–258 (1996).

44. Gotic, I. *et al.* Temperature regulates splicing efficiency of the cold-inducible RNA-binding protein gene Cirbp. *Genes Dev.* **30**, 2005–2017 (2016).

45. Stajich, J. E., Dietrich, F. S. & Roy, S. W. Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. *Genome Biol.* **8**, R223 (2007).

46. Parenteau, J. *et al.* Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. *Cell* **147**, 320–331 (2011).

47. Neuvéglise, C., Marck, C. & Gaillardin, C. The intronome of budding yeasts. *C. R. Biol.* **334**, 662–670 (2011).

48. Hooks, K. B., Delneri, D. & Griffiths-Jones, S. Intron evolution in Saccharomycetaceae. *Genome Biol. Evol.* **6**, 2543–2556 (2014).

49. Rogozin, I. B., Carmel, L., Csuros, M. & Koonin, E. V. Origin and evolution of spliceosomal introns. *Biol. Direct* **7**, 1 (2012).

50. Mourier, T. & Jeffares, D. C. Eukaryotic intron loss. *Science (80-.).* **300**, 1393 (2003).

51. Black, D. L. Mechanisms of alternative pre-messenger {RNA} splicing. *Annu. Rev. Biochem.* **72**, 291–336 (2003).

52. Jurica, M. S. & Moore, M. J. Pre-mRNA splicing: Awash in a sea of proteins. *Molecular Cell* vol. 12 5–14 (2003).

53. Schwartz, S. H. *et al.* Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res.* **18**, 88–103 (2008).

54. Plass, M., Agirre, E., Reyes, D., Camara, F. & Eyras, E. Co-evolution of the branch site and SR proteins in eukaryotes. *Trends in Genetics* vol. 24 590–594 (2008).

55. Plass, M., Codony-Servat, C., Ferreira, P. G., Vilardell, J. & Eyras, E. RNA secondary structure mediates alternative 3′ss selection in Saccharomyces cerevisiae. *RNA* **18**, 1103–1115 (2012).

56. Howe, K. J., Kane, C. M. & Ares, M. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. *RNA* **9**, 993–1006 (2003).

57. Grund, S. E. *et al.* The inner nuclear membrane protein Src1 associates with subtelomeric genes and alters their regulated gene expression. *J. Cell Biol.* **182**, 897–910 (2008).

58. Juneau, K., Nislow, C. & Davis, R. W. Alternative splicing of PTC7 in Saccharomyces cerevisiae determines protein localization. *Genetics* **183**, 185–194 (2009).

59. Hossain, M. A., Rodriguez, C. M. & Johnson, T. L. Key features of the two-intron Saccharomyces cerevisiae gene {SUS1} contribute to its alternative splicing. *Nucleic Acids Res.* **39**, 8612–8627 (2011).

60. Meyer, M., Plass, M., Pérez-Valle, J., Eyras, E. & Vilardell, J. Deciphering 3'ss Selection in the Yeast Genome Reveals an RNA Thermosensor that Mediates Alternative Splicing. *Mol. Cell* **43**, 1033–1039 (2011).

61. Carrillo Oesterreich, F. *et al.* Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell* (2016) doi:10.1016/j.cell.2016.02.045.

62. Douglass, S. M., Leung, C. S. & Johnson, T. L. Extensive splicing across the Saccharomyces cerevisiae genome. *bioRxiv* 515163 (2019) doi:10.1101/515163.

63. Feng, B. *et al.* Reconstructing yeasts phylogenies and ancestors from whole genome data. *Sci. Rep.* **7**, (2017).

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

65. Schwartz, S. H. *et al.* Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res.* **18**, 88–103 (2008).

66. Hooks, K. B., Delneri, D. & Griffiths-Jones, S. Intron Evolution in Saccharomycetaceae. *Genome Biol. Evol.* **6**, 2543–2556 (2014).

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

68. Baeza-Centurion, P., Miñana, B., Schmiedel, J. M., Valcárcel, J. & Lehner, B. Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. *Cell* **176**, 549-563.e23 (2019).

69. Cheung, R. *et al.* A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing

Disruptions. *Mol. Cell* **73**, 183-194.e8 (2019).

70. Mikl, M., Hamburg, A., Pilpel, Y. & Segal, E. Dissecting splicing decisions and cell-to-cell variability with designed sequence libraries. *Nat. Commun.* (2019) doi:10.1038/s41467-019-12642-3.

71. Hamilton, W. D. The genetical evolution of social behaviour. II. *J. Theor. Biol.* **7**, 17–52 (1964).

72. Smukalla, S. *et al.* FLO1 Is a Variable Green Beard Gene that Drives Biofilm-like Cooperation in Budding Yeast. *Cell* **135**, 726–737 (2008).

73. Wilson, D. S. A theory of group selection. *Proc. Natl. Acad. Sci.* **72**, 143–146 (1975).

74. Nowak, M. A. & Sigmund, K. Evolutionary dynamics of biological games. *Science* **303**, 793–9 (2004).

75. Hardin, G. The tragedy of the commons. Hardin and Baden (eds) Managing the Commons. (1968).

76. Damore, J. A. & Gore, J. Understanding microbial cooperation. *J. Theor. Biol.* **299**, 31–41 (2012).

77. MacLean, R. C., Fuentes-Hernandez, A., Greig, D., Hurst, L. D. & Gudelj, I. A Mixture of "Cheats" and "Co-Operators" Can Enable Maximal Group Benefit. *PLoS Biol.* **8**, e1000486 (2010).

78. Velicer, G. J., Kroos, L. & Lenski, R. E. Developmental cheating in the social bacterium Myxococcus xanthus. *Nature* **404**, 598–601 (2000).

79. Diard, M. *et al.* Stabilization of cooperative virulence by the expression of an avirulent phenotype. *Nature* **494**, 353–356 (2013).

80. Schiessl, K. T. *et al.* Individual- versus group-optimality in the production of secreted bacterial compounds. *Evolution (N. Y).* **73**, 675–688 (2019).

81. Drescher, K., Nadell, C. D., Stone, H. A., Wingreen, N. S. & Bassler, B. L. Solutions to the Public Goods Dilemma in Bacterial Biofilms. *Curr. Biol.* **24**, 50–55 (2014).

82. Cross, B. C. S., Sinning, I., Luirink, J. & High, S. Delivering proteins for export from the cytosol. *Nature Reviews Molecular Cell Biology* vol. 10 255–264 (2009).

83. Rapoport, T. A. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature* vol. 450 663–669 (2007).

84. Blobel, G. & Dobberstein, B. Transfer of proteins across membranes: I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. *J. Cell Biol.* **67**, 835–851 (1975).

85. Walter, P. & Blobel, G. Translocation of proteins across the endoplasmic reticulum. III. Signal recognition protein (SRP) causes signal sequence-dependent and site-specific arrest of chain elongation that is released by microsomal membranes. *J. Cell Biol.* **91**, 557–561 (1981).

86. Gilmore, R., Blobel, G. & Walter, P. Protein translocation across the endoplasmic reticulum. I. Detection in the microsomal membrane of a receptor for the signal recognition particle. *J. Cell Biol.* **95**, 463–469 (1982).

87. Akopian, D., Shen, K., Zhang, X. & Shan, S. O. Signal recognition particle: An essential protein-targeting machine. *Annual Review of Biochemistry* vol. 82 693–721 (2013).

88. Jonikas, M. C. *et al.* Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science (80-.).* **323**, 1693–1697 (2009).

89. Schuldiner, M. *et al.* The GET Complex Mediates Insertion of Tail-Anchored Proteins into the ER Membrane. *Cell* **134**, 634–645 (2008).

90. Stefanovic, S. & Hegde, R. S. Identification of a Targeting Factor for Posttranslational Membrane Protein Insertion into the ER. *Cell* **128**, 1147–1159 (2007).

91. Ast, T., Cohen, G. & Schuldiner, M. A network of cytosolic factors targets SRP-independent proteins to the endoplasmic reticulum. *Cell* **152**, 1134–45 (2013).

92. Aviram, N. *et al.* The SND proteins constitute an alternative targeting route to the endoplasmic reticulum. *Nature* **540**, 134–138 (2016).

93. Cohen-Zontag, O. *et al.* A secretion-enhancing cis regulatory targeting element (SECReTE) involved in mRNA localization and protein synthesis. *PLoS Genet.* **15**, e1008248 (2019).

94. Boyer, P. D., Lardy, H., Myrbäck, K. & others. The enzymes: volume 7. *Enzym. Vol. 7.* (1963).

95. H. Koschwanez, J., R. Foster, K. & W. Murray, A. Sucrose Utilization in Budding Yeast as a Model for the Origin of Undifferentiated Multicellularity. *PLoS Biol.* **9**, e1001122 (2011).

96. Koschwanez, J. H., Foster, K. R. & Murray, A. W. Improved use of a public good selects for the evolution of undifferentiated multicellularity. *Elife* **2**, e00367 (2013).

97. CRAIG MACLEAN, R. & BRANDON, C. Stable public goods cooperation and dynamic social interactions in yeast. *J. Evol. Biol.* **21**, 1836–1843 (2008).

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

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

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

101. 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).

102. 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).

103. Weinberg, D. E. *et al.* Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Rep.* **14**, 1787–99 (2016).

104. Goodarzi, H. *et al.* Modulated Expression of Specific tRNAs Drives Gene Expression

and Cancer Progression. *Cell* **165**, 1416–27 (2016).

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

106. 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).

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

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

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

110. 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).

111. 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.

112. 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).

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

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

115. 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).

116. Mohammad, F., Woolstenhulme, C. J., Green, R. & Buskirk, A. R. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Rep.* **14**, 686–94 (2016).

117. Allert, M., Cox, J. C. & Hellinga, H. W. Multifactorial Determinants of Protein Expression in Prokaryotic Open Reading Frames. *J. Mol. Biol.* **402**, 905–918 (2010).

118. Charneski, C. a. & Hurst, L. D. Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. *PLoS Biol.* **11**, (2013).

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

120. 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).

121. Bennett, B. D. *et al.* Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. *Nat Chem Biol* **5**, 593–599 (2009).

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

123. Mogno, I., Kwasnieski, J. C. & Cohen, B. A. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res.* **23**, 1908–15 (2013).

124. Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc. Natl. Acad. Sci.* **110**, 14024–14029 (2013).

125. Gasch, A. P. & Werner-Washburne, M. The genomics of yeast responses to environmental stress and starvation. *Funct. Integr. Genomics* **2**, 181–92 (2002).

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

127. 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).

128. Xi, L. *et al.* Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics* **4**, e1000175 (2010).

129. Ruscio, J. A probability-based measure of effect size: robustness to base rates and other factors. *Psychol. Methods* **13**, 19–30 (2008).

130. Tirosh, I. & Barkai, N. Two strategies for gene regulation by promoter nucleosomes. *Genome Res.* **18**, 1084–91 (2008).

131. MacIsaac, K. D. *et al.* An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* **7**, 113 (2006).

132. Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V & Kondrashov, F. A. Selection for short introns in highly expressed genes. *Nat. Genet.* **31**, 415–418 (2002).

133. Pozzoli, U. *et al.* Intron size in mammals: complexity comes to terms with economy. *Trends in Genetics* vol. 23 20–24 (2007).

134. Seoighe, C., Gehring, C. & Hurst, L. D. Gametophytic Selection in Arabidopsis thaliana Supports the Selective Model of Intron Length Reduction. *PLoS Genet.* **1**, e13 (2005).

135. Scott, M., Mateescu, E. M., Zhang, Z. & Hwa, T. Interdependence of Cell Growth. **330**, 1099–1102 (2010).

136. Geisberg, J. V, Moqtaderi, Z., Fan, X., Ozsolak, F. & Struhl, K. Global Analysis of mRNA Isoform Half-Lives Reveals Stabilizing and Destabilizing Elements in Yeast. (2014) doi:10.1016/j.cell.2013.12.026.

137. Sharon, E. *et al.* Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Res.* **24**, 1698–706 (2014).

138. Levy, S. F. *et al.* Quantitative evolutionary dynamics using high-resolution lineage tracking. (2015) doi:10.1038/nature14279.

139. Zhou, Z. *et al.* The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. *Nature* **407**, 401–405 (2000).

140. Wang, H.-F., Feng, L. & Niu, D.-K. Relationship between mRNA stability and intron presence. *Biochem. Biophys. Res. Commun.* **354**, 203–8 (2007).

141. Wilkinson, M. E., Charenton, C. & Nagai, K. RNA Splicing by the Spliceosome. *Annu. Rev. Biochem.* **89**, 359–388 (2020).

142. Coolidge, C. J., Seely, R. J. & Patton, J. G. Functional analysis of the polypyrimidine tract in pre-{mRNA} splicing. *Nucleic Acids Res.* **25**, 888–896 (1997).

143. Patterson, B. & Guthrie, C. A U-rich tract enhances usage of an alternative 3' splice site in yeast. *Cell* **64**, 181–187 (1991).

144. Wieringa, B., Hofer, E. & Weissmann, C. A minimal intron length but no specific internal sequence is required for splicing the large rabbit β-globin intron. *Cell* **37**, 915–925 (1984).

145. Dewey, C. N., Rogozin, I. B. & Koonin, E. V. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *{BMC} Genomics* **7**, 311 (2006).

146. Luukkonen, B. G. & Séraphin, B. The role of branchpoint-3' splice site spacing and interaction between intron terminal nucleotides in 3' splice site selection in Saccharomyces cerevisiae. *{EMBO} J.* **16**, 779–792 (1997).

147. Wong, J. J.-L. *et al.* Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154**, 583–95 (2013).

148. Mordstein, C. *et al.* Codon Usage and Splicing Jointly Influence mRNA Localization. *Cell Syst.* **10**, 351-362.e8 (2020).

149. Harigaya, Y. & Parker, R. Global analysis of mRNA decay intermediates in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 11764–9 (2012).

150. Plass, M., Codony-Servat, C., Ferreira, P. G., Vilardell, J. & Eyras, E. RNA secondary structure mediates alternative 3′ss selection in Saccharomyces cerevisiae. *RNA* **18**, 1103–1115 (2012).

151. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).

152. Ke, G., Meng, Q., Finley, T., Wang, T. & Chen, W. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* 3146 (2017).

153. Jung, Y. & Hu, J. A K-fold Averaging Cross-validation Procedure. *J. Nonparametr. Stat.* **27**, 167–179 (2015).

154. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable {AI} for trees. *Nat. Mach. Intell.* (2020) doi:10.1038/s42256-019-0138-9.

155. Clark, T. A., Sugnet, C. W. & Ares, M. Genomewide analysis of {mRNA} processing in yeast using splicing-specific microarrays. *Science (80-.).* **296**, 907–910 (2002).

156. Li, X. *et al.* A unified mechanism for intron and exon definition and back-splicing. *Nature* (2019) doi:10.1038/s41586-019-1523-6.

157. Dodyk, F. & Rothstein, A. Factors influencing the appearance of invertase in Saccharomyces cerevisiae. *Arch. Biochem. Biophys.* **104**, 478–486 (1964).

158. Esmon, P. C., Esmon, B. E., Schauer, I. E., Taylor, A. & Schekman, R. Structure, assembly, and secretion of octameric invertase. *J. Biol. Chem.* **262**, 4387–4394 (1987).

159. Stovicek, V., Borja, G. M., Forster, J. & Borodina, I. EasyClone 2.0: expanded toolkit of integrative vectors for stable gene expression in industrial Saccharomyces cerevisiae strains. *J. Ind. Microbiol. Biotechnol.* **42**, 1519–1531 (2015).

160. Trumbly, R. J. Glucose repression in the yeast Saccharomyces cerevisiae. *Molecular Microbiology* vol. 6 15–21 (1992).

161. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

162. Fisher, R. A. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika* **10**, 507 (1915).

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

164. Friedman, N., Cai, L. & Xie, X. S. Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Phys. Rev. Lett.* **97**, 1–4 (2006).

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

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

167. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).

168. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc.* **58**, 267–288 (1996).

169. Cohen, O. *et al.* Comparative transcriptomics across the prokaryotic tree of life. *Nucleic Acids Res.* gkw394 (2016) doi:10.1093/nar/gkw394.

170. Wang, M. *et al.* PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell. Proteomics* **11**, 492–500 (2012).

171. Bond, S. R. & Naus, C. C. RF-Cloning.org: An online tool for the design of restriction-free cloning projects. *Nucleic Acids Res.* **40**, W209–W213 (2012).

172. Technologies, A. G7555-90000 SureGuide Custom CRISPR Guide Library - Guidelines for Amplification and Cloning Assembly. 1–2 (2016).

173. Technologies, A. G7556-90000 SureVector CRISPR Library Cloning Kit - Protocol. (2017).

174. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).

175. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).

176. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile

open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

177. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

178. Taniguchi, Y. *et al.* Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–8 (2010).

179. Hausser, J., Mayo, A., Keren, L. & Alon, U. Central dogma rates and the trade-off between precision and economy in gene expression. *Nat. Commun.* **10**, 1–15 (2019).

180. Frumkin, I. *et al.* Evolution of intron splicing towards optimized gene expression is based on various Cis- and Trans-molecular mechanisms. *{PLoS} Biol.* **17**, e3000423 (2019).

181. Graubert, T. A. *et al.* Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat. Genet.* **44**, 53–57 (2012).

182. Przychodzen, B. *et al.* Patterns of missplicing due to somatic U2AF1 mutations in myeloid neoplasms. *Blood* **122**, 999–1006 (2013).

183. Ilagan, J. O. *et al.* U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Res.* **25**, 14–26 (2015).

184. Yona, A. H., Alm, E. J. & Gore, J. Random sequences rapidly evolve into de novo promoters. *Nat. Commun.* **9**, 1–10 (2018).

10. Appendix - Supplementary tables

Table S1 - gradient boosting model features

Feature name	Feature type
5’SS sequence	categorical
BS sequence	categorical
3’SS sequence	categorical
Intron GC%	numeric
U-enrichment @ 3’ end	numeric
Intron length	numeric
BS-to-3’SS length	numeric
5’SS ΔG (30nt window)	numeric
BS ΔG (30nt window)	numeric
3’SS ΔG (30nt window)	numeric
3’SS GC% (30nt window)	numeric
5’SS stem length	numeric
BS stem length	numeric
3’SS stem length	numeric
5’SS stem arm	numeric
BS stem arm	numeric
3’SS stem arm	numeric
5’SS fraction of nucleotides based paired	numeric
BS fraction of nucleotides based paired	numeric
3’SS fraction of nucleotides based paired	numeric
5’SS - is 1st nucleotide paired	categorical
5’SS - is 2nd nucleotide paired	categorical
5’SS - is 3rd nucleotide paired	categorical
5’SS - is 4th nucleotide paired	categorical

5′SS - is 5th nucleotide paired	categorical
5′SS - is 6th nucleotide paired	categorical
BS - is 1st nucleotide paired	categorical
BS - is 2nd nucleotide paired	categorical
BS - is 3rd nucleotide paired	categorical
BS - is 4th nucleotide paired	categorical
BS - is 5th nucleotide paired	categorical
BS - is 6th nucleotide paired	categorical
BS - is 7th nucleotide paired	categorical
3′SS - is 1st nucleotide paired	categorical
3′SS - is 2nd nucleotide paired	categorical
3′SS - is 3rd nucleotide paired	categorical

Table S2 - List of primers

Name	Used for	Sequence
prDS1	NEBuilder assembly of reporter cassette	CTCATAAGCAGCAATCAATTCTATCTATACTTTAAAATGCTTTCTGCATCTATATTACCCTGTTATCCC
prDS6	NEBuilder assembly of reporter cassette	GATCGGCTTACTAATATGGGGCCGTATACTTAC
prDS7	NEBuilder assembly of reporter cassette	ACGGCCCCATATTAGTAAGCCGATCCCATTAC
prDS8	NEBuilder assembly of reporter cassette	TCACCTTTAGACATTTTATGTGATGATTGATTGATTG
prDS9	NEBuilder assembly of reporter cassette	AATCATCACATAAAATGTCTAAAGGTGAAGAATTATTCACTGGTGT
prDS10	NEBuilder assembly of reporter cassette	CTGGTTGAAACAAATCAGTGCCGGTAACGCTTTTTGTATCTTGAGTCGACACTGGATGGCGGC
prDS20	RF cloning pBAR3	CCTTCGTTCTTCCTTCTGTTCGGAGGGGACCAGGTGCCGTAAG
prDS21	RF cloning pBAR3	CCGGGTGACCGATTCGGTAATCCCGGTAGAGGTGTGGTCAATAAG

prDS22	Linearize pDS101	TCCGAACAGAAGGAAGAAC
prDS23	Linearize pDS101	GATTACCGAATCGGTCAC
prDS55	Amplification & cloning SplicingLib1 index1	AAAAGTGGAAGTCAGGGTGTTGGTGTAAAGAACATCTAAATACGAGGCACTTACTCCG
prDS56	Amplification & cloning SplicingLib1 index2	AAAAGTGGAAGTCAGGGTGTTGGTGTAAAGTGTTGGGAAATACGAGGCACTTACTCCG
prDS57	Amplification & cloning SplicingLib1 index3	AAAAGTGGAAGTCAGGGTGTTGGTGTAAAGAAGCCATGAATACGAGGCACTTACTCCG
prDS58	Amplification & cloning SplicingLib1 index4	AAAAGTGGAAGTCAGGGTGTTGGTGTAAAGGCTAAAGAAATACGAGGCACTTACTCCG
prDS59	Amplification & cloning SplicingLib1 R	ATTGTGGGGAGTGGAACGCAGTCACATTGATAGGAATAGCGAACTCCAGG
prDS62	Linearize pDS102	CTTTACACCAACACCCTGAC
prDS63	Linearize pDS102	TCAATGTGACTGCGTTCCAC
prDS137	NGS library preparation shift0	ACGACGCTCTTCCGATCTGTCAGGGTGTTGGTGTAAAG
prDS138	NGS library preparation shift1	ACGACGCTCTTCCGATCTAGTCAGGGTGTTGGTGTAAAG
prDS139	NGS library preparation shift2	ACGACGCTCTTCCGATCTTCGTCAGGGTGTTGGTGTAAAG
prDS140	NGS library preparation shift3	ACGACGCTCTTCCGATCTCATGTCAGGGTGTTGGTGTAAAG
prDS141	NGS library preparation shift4	ACGACGCTCTTCCGATCTACTAGTCAGGGTGTTGGTGTAAAG
prDS142	NGS library preparation shift5	ACGACGCTCTTCCGATCTTAGCCGTCAGGGTGTTGGTGTAAAG
prDS143	NGS library preparation R	AGACGTGTGCTCTTCCGATCTGTGGAACGCAGTCACATTGA
prDS144	NGS library preparation PCR2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
prDS145	NGS library preparation PCR2 with index	CAAGCAGAAGACGGCATACGAGAT [index]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
prDS50	Amplification & cloning NucleotideCompLib index1	AAAAGTGGAAGTCAGGGTGTTGGTGTAAAGAACATCTAACAACGCTTTCTGTGTCGTG

prDS51	Amplification & cloning NucleotideCompLib index2	AAAAGTGGAAGTCAGGGTGTTGGTGTAAAGTGTTGGGAACAACGCTTTCTGTGTCGTG
prDS52	Amplification & cloning NucleotideCompLib index3	AAAAGTGGAAGTCAGGGTGTTGGTGTAAAGAAGCCATGACAACGCTTTCTGTGTCGTG
prDS53	Amplification & cloning NucleotideCompLib index4	AAAAGTGGAAGTCAGGGTGTTGGTGTAAAGGCTAAAGAACAACGCTTTCTGTGTCGTG
prDS54	Amplification & cloning NucleotideCompLib R	ATTGTGGGGAGTGGAACGCAGTCACATTGAAAGTAGGGTTTCTGACCTCG
prGS1	Oligo Amplification of SP lib F	GCTTTTCTTTTCACTAACG
prGS2	Oligo Amplification of SP lib R	GAAAGTACAGATGCCATTTG
prGS3	Oligo Amplification of SECReTE lib ins1 F	GAGTGGGTTTTTCAATGATAC
prGS4	Oligo Amplification of SECReTE lib ins1 R	GAGAAGGTTCATACCAGAACAC
prGS5	Oligo Amplification of SECReTE lib ins2 F	AGATCAACCCATTGCTATCG
prGS6	Oligo Amplification of SECReTE lib ins2 R	TAAAAGTGTAACCACCATCAAG
prGS13	pCfB2223 linearization	AAGCTATCTGTAACAGGAGCATCGCGTGCATTCATCCG
prGS14	pCfB2223 linearization	GATTGTGACGTGTGGATGCTATCGCACGCATTCCGTTG
prGS15	pGS2223 linearization SP lib	CAAATGGCATCTGTACTTTC
prGS16	pGS2223 linearization SP lib	CATATACGTTAGTGAAAAGAAAAGC
prGS17	pGS2223 linearization SECReTE lib ins1	AAGGTGTTCTGGTATGAACC
prGS18	pGS2223 linearization SECReTE lib ins1	ATCAATAGTATCATTGAAAAACCC
prGS19	pGS2223 linearization SECReTE lib ins2	TCTCTTGATGGTGGTTACAC

prGS20	pGS2223 linearization SECReTE lib ins2	CTTGGGAGCGATAGCAATG
prGS27	Verification of yeast sub-libraries (down junction)	CCTGCAGGACTAGTGCTGAG
prGS28	Verification of yeast sub-libraries (down junction)	CCGTGCAATACCAAAATCG
prGS29	Verification of yeast sub-libraries (up junction)	GTTGACACTTCTAAATAAGCGAATTTC
prGS30	Verification of yeast sub-libraries (up junction)	TGACGAATCGTTAGGCACAG
prGS33	SECReTE NGS library preparation shift0	ACGACGCTCTTCCGATCTAGATCAACCCATTGCTATCGC
prGS34	SECReTE NGS library preparation shift1	ACGACGCTCTTCCGATCTTAGATCAACCCATTGCTATCGC
prGS35	SECReTE NGS library preparation shift2	ACGACGCTCTTCCGATCTACAGATCAACCCATTGCTATCGC
prGS36	SECReTE NGS library preparation shift3	ACGACGCTCTTCCGATCTCTCAGATCAACCCATTGCTATCGC
prGS37	SECReTE NGS library preparation shift4	ACGACGCTCTTCCGATCTTTAGAGATCAACCCATTGCTATCGC
prGS38	SECReTE NGS library preparation R	AGACGTGTGCTCTTCCGATCTGAGAAGGTTCATACCAGAACAC
prGS43	SP NGS library preparation shift0	ACGACGCTCTTCCGATCTGCTTTTCTTTTCACTAACG
prGS44	SP NGS library preparation shift1	ACGACGCTCTTCCGATCTTGCTTTTCTTTTCACTAACG
prGS45	SP NGS library preparation shift2	ACGACGCTCTTCCGATCTACGCTTTTCTTTTCACTAACG
prGS46	SP NGS library preparation shift3	ACGACGCTCTTCCGATCTCTCGCTTTTCTTTTCACTAACG
prGS47	SP NGS library preparation shift4	ACGACGCTCTTCCGATCTTTAGGCTTTTCTTTTCACTAACG

prGS48	SP NGS library preparation R	AGACGTGTGCTCTTCCGATCTACGGTGTCATTTGGGTTGTATTG
prDSO1	PromoterLib NGS library preparation shift0	ACGACGCTCTTCCGATCTTGAATTGTACAAATAACGGCCGAA
prDSO2	PromoterLib NGS library preparation shift1	ACGACGCTCTTCCGATCTCATGAATTGTACAAATAACGGCCGAA
prDSO3	PromoterLib NGS library preparation shift2	ACGACGCTCTTCCGATCTTCGTGAATTGTACAAATAACGGCCGAA
prDSO4	PromoterLib NGS library preparation shift3	ACGACGCTCTTCCGATCTATCGTGAATTGTACAAATAACGGCCGAA
prDSO5	PromoterLib NGS library preparation shift4	ACGACGCTCTTCCGATCTGCCATTGAATTGTACAAATAACGGCCGAA
prDSO6	PromoterLib NGS library preparation shift5	ACGACGCTCTTCCGATCTTACCGATGAATTGTACAAATAACGGCCGAA
prDSO7	PromoterLib NGS library preparation R	AGACGTGTGCTCTTCCGATCTGATAAAGTCAGTGCTTAAACAC