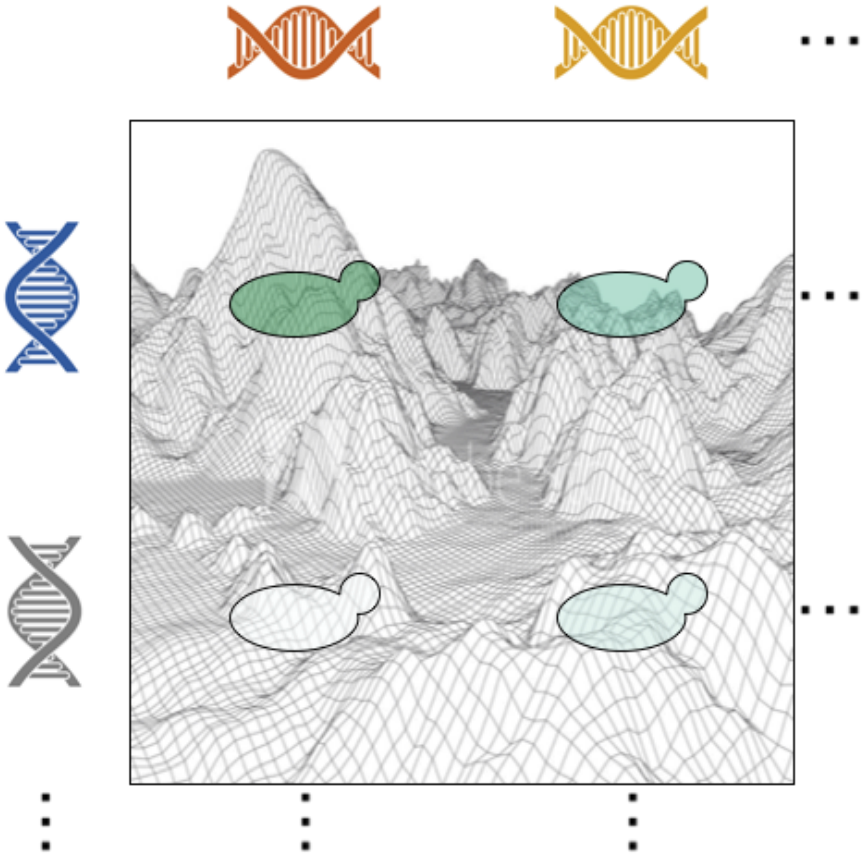
Methods for predictable and accelerated engineering of metabolism in eukaryotes

Petersen, Søren Dalsgård

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Methods for predictable and accelerated engineering of metabolism in eukaryotes



Søren Dalsgård Petersen

Ph.D. thesis, November 2019

The Novo Nordisk Foundations Center for Biosustainability

The Technical University of Denmark

Main supervisors:

Senior scientist Michael Krogh Jensen

Professor Jay D. Keasling

Preface

The work presented in this thesis was carried out in the period from December 1. 2016 to November 30. 2019. It was done mainly at the Novo Nordisk Foundation Center for Biosustainability (CFB) at the Technical University of Denmark, Kgs. Lyngby, Denmark. In the period from February 1. 2019 to May 1. 2019 I worked partly in the group for quantitative metabolic modeling at The Joint Bioenergy Institute (JBEI) and partly at the biotech company TeselaGen, both of which are located near or within San Francisco in California. My main supervisors were Michael Krogh Jensen at CFB and Jay D. Keasling at JBEI/CFB. The work was funded by The Technical University of Denmark and was associated with the Horizon 2020-funded international training network PAcMEN (Predictive and Accelerated Metabolic Engineering Network).

Søren Dalsgård Petersen

November 28, 2019

Copenhagen

Acknowledgments

There is a long list of people who deserve thanks and without whom this work would not have been possible. Still, among these, there is a shorter list of people who deserves a particular thank you for their involvement.

Of these, I would first like to thank my supervisors Michael Krogh Jensen and Jay D. Keasling for giving me the opportunity at this point in time to work within the field of synthetic biology together with world-leading experts. I also want to thank Jie Zhang for the invaluable help and for the collaboration that grew during my project. It has been a great joy working with you all, and I have benefitted so much from all the enthusiasm, expert mentoring and competent feedback. The projects presented in this thesis all have multiple co-authors as well as persons receiving acknowledgment and I would like to thank all for their contributions.

I would like to thank my collaborators abroad, in California, Chile, and Sweden. In California, at JBEI and TeselaGen, in particular, Mike Ferro, Eduardo Abeliuk, Sheila Somers, Tijana Radivojevic and Hector Martin Garcia for being great hosts during my secondment.

I would like to thank all the Ph.D. fellows and supervisors in the MSCA PAcMEN network for all the energy put into the network activities including conferences, meetings, courses, and social activities. It has been workwise inspiring and socially enjoyable and such a privilege to be part of the network. Also, I would like to thank my fellow Ph.D. students at the Novo Nordisk Center for Biosustainability and old study comrades for creating a nice community.

Working in environments full of dedicated colleges has been both fruitful and enjoyable to me. I am very grateful for the acquired inspiration and skills, and for the knowledge obtained about a variety of interesting research projects.

Finally, I would like to thank family and friends for their interest and support during the whole endeavor.

List of publications and manuscripts

1. **Petersen, S.D.**, Zhang, J., Lee, J.S., Jakočiūnas, T., Grav, L.M., Kildegaard, H.F., Keasling, J.D. and Jensen, M.K. (2018) Modular 5′ -UTR hexamers for context-independent tuning of protein expression in eukaryotes. Nucleic Acids Res., 10.1093/nar/gky734.

2. Jie Zhang[#], **Søren D. Petersen**[#], Tijana Radivojevic, Andrés Ramirez, Eduardo Abeliuk, Benjamin J. Sánchez, Zachary Costello, Yu Chen, Mike Fero, Hector Garcia Martin, Jens Nielsen, Jay D. Keasling, & Michael K. Jensen (2019). Predictive engineering and optimization of tryptophan metabolism in yeast through a combination of mechanistic and machine learning models. Submitted to Cell Systems. [#] = these authors contributed equally to this study.

Publications that I have contributed to during the time of my Ph.D. studies but are not included in the thesis.

3. Genee, H.J., Bali, A.P., **Petersen, S.D.**, Siedler, S., Bonde, M.T., Gronenberg, L.S., Kristensen, M., Harrison, S.J. and Sommer, M.O.A. (2016) Functional mining of transporters using synthetic selections. Nat. Chem. Biol., 12, 1015–1022.

4. Genee, H.J., Riesselman, A.J., **Petersen, S.D.**, Nath, S., Gronenberg, L.S., Salomonsen, B., Chan, L.J.G., Nhan, M., Baidoo, E.K., Wang, G., Oberortner, E., Hillson, N.J., Keasling, J.D., Marks, D.S., Petzold C.J., Deutsch, S., Sommer, M.O.A. (2019). Machine learning enables accurate prediction of complex topological features. Under revision for soon resubmission to Nature Biotechnology.

5. Dam, S.H., Friis, R.U.W., **Petersen, S.D.**, Esteban, A.M. and Laustsen, A.H. (2018) Venomics Display: An online toolbox for visualization of snake venomics data. Toxicon, 152, 60–64.

Abstract

Wealth increases and a growing world population necessitates sustainable methods for producing biomolecules within broad fields like fuels, chemicals, foods, feeds, and pharmaceuticals. The engineering of microorganisms to produce biomolecules from sustainable feedstock can be part of the solution. The development of organisms for bio-production can be considered as an iterative process with four major steps called Design, Build, Test, and Learn (DBTL). This process can be accelerated by the development of new and improved methodologies.

I have worked with all four steps in the DBTL cycle and have successfully contributed to developing new tools for regulating the production of selected metabolites. In the work presented in paper no. one (chapter no. 2 in this thesis), 8 hexameric sequences were identified in the untranslated part of the translation initiation site that leads to variable production of a green fluorescent protein (GFP) in yeast. It was shown that these hexameric sequences can be used for predictive and context-independent tuning of protein expression from yeast to CHO cells. Thus, the method has the potential to become a unified tool across a large group of organisms. The carotenoid production was effectively regulated by combining three of the eight hexamers and inserting them in front of the genes *crtE* and *erg9* which are known to code for key enzymes in the carotenoid synthesis pathway.

In paper no. 2 (chapter 3) the aim was to improve genotype-to-phenotype predictions of tryptophan synthesis rate in yeast as a tool for metabolic engineering. Genome scale modeling and literature was used in the identification of five gene targets for engineering, CRISPR/Cas9 facilitated genome editing for the construction of a cell library of combinatorial edits of promoters, high throughput methods for genotype and phenotype characterization, and state of the art machine learning (ML) models for genotype to phenotype prediction. The created library has at least a 5 fold variation in the tryptophan synthesis rate calculated as the ratio between the highest and lowest average rate for individual strains. It was possible to fit different ML models closely to the observed library strains, and these models could also be used to predict efficient strains not seen in the library. Thus, a strain was identified with a tryptophan synthesis rate that was 17 % larger than the highest rate found in the library and 106 % larger than that of an engineered, high

producing platform strain. In future studies, further increases in the tryptophan synthesis rate can be expected by combining relevant edits of new genetic targets with the five presented in this study.

The developed methods have the potential to become parts of more standardized workflows for leading metabolic engineering efforts in the routine development of cell factories.

Resume

Stigende befolkning og velstand nødvendiggør, at vi finder bæredygtige metoder til produktion af biomolekyler indenfor store områder som brændstoffer, kemikalier, fødevarer, foder og farmaceutiske produkter. Brugen af mikroorganismer til produktion af biomolekyler fra bæredygtigt råmateriale kan være en del af løsningen. Udviklingen af organismer til bioproduktion kan betragtes som en iterativ proces bestående af fire overordnede trin kaldet Design, Build, Test og Learn (DBTL). Denne proces kan gøres hurtigere og mere effektiv gennem udvikling af nye og forbedrede metoder inden for alle trin.

Vi har således arbejdet med alle fire trin i DBTL-cyklen og har med succes udviklet nye værktøjer til regulering af produktionen af udvalgte metabolitter. I arbejdet præsenteret i artikel nr. 1 (kapitel nr. 2 i denne afhandling), identificerede vi 8 sekvenser, hver bestående af seks basepar, i den ikke-translaterede del af translations-initierings-sitet, der fører til variabel produktion af et grønt fluorescerende protein (GFP) i gær. Vi viste, at disse hexamerer kan bruges til forudsigelig og kontekstuafhængig regulering af proteinekspression i gær og CHO-celler. Metoden har således potentiale til at blive et værktøj, der kan virke på tværs af en bred gruppe af organismer. Vi var i stand til at regulere carotenoid-produktion i gær ved at kombinere tre af de otte hexamerer og indsætte dem foran generne *crtE* og *erg9*, som er kendt for at kode for nøgle-enzymer i carotenoid-syntesevejen.

I artikel nr. 2 (kapitel 3) var vores mål at forbedre genotype-til-fænotype forudsigelser af tryptofan-syntese-hastighed i gær til brug som et værktøj til "metabolic engineering". Vi brugte genom-skala modellering og litteraturstudier til at identificere fem gen-mål, CRISPR/Cas9 faciliteret gen redigering til at konstruere et cellebibliotek med kombinatoriske forekomster af promotorer, højeffektive metoder til genotypisk og fænotypisk karakterisering, samt "state of the art machine learning" (ML-modeller) til at forudsige sammenhænge mellem genotyper og fænotyper.

Variationen i tryptofan-syntese-hastighed i det konstruerede bibliotek er mindst en faktor 5, beregnet som forholdet mellem den højeste og den laveste gennemsnitlige hastighed for

individuelle stammer. Det var muligt at tilpasse flere forskellige ML-modeller ganske godt til alle de observerede biblioteks-stammer. Modellerne kunne også bruges til at forudsige stammer (genotyper), der var mere effektive til at producere tryptophan end dem, der var indeholdt i biblioteket. Der blev således forudsagt en stamme med en tryptofan-syntese-hastighed, der var 17 % højere end den højeste hastighed fundet i biblioteket og 106 % større end hastigheden for en konstrueret, høj-produktiv referencestamme. Man kan forvente yderligere stigninger i tryptofan-syntese-hastighed, hvis man kombinerer relevante redigeringer af nye genetiske mål med de fem, der er præsenteret i nærværende undersøgelse.

De udviklede metoder til forudsigelig og acceleret redigering af stofskifteprocesser i eukaryoter har potentiale til at kunne indgå i standardiserede arbejdsgange i forbindelse med en rutinemæssig udvikling af cellefabrikker.

Table of contents

1. Introduction

1.1 Background

Population and wealth increases are increasing the pressure on Earth's resources. To meet this demand, we need to be able to produce consumer goods efficiently and sustainably. This also applies to essential biomolecules, eg fuels, chemicals, foods, feeds, and pharmaceuticals.

Metabolic engineering and the production of biomolecules in cell factories can be an important part of the solution. "Metabolic engineering is the science of rewiring the metabolism of cells to enhance the production of native metabolites or to endow cells with the ability to produce new products" (Nielsen and Keasling, 2016). There has recently been made excellent reviews covering the field (Keasling, 2010; Nielsen and Keasling, 2016; Davy et al., 2017; Liu and Nielsen, 2019). Metabolic engineering can potentially lead to the sustainable production of a large number of chemicals that today are derived from non-renewable resources (Nakamura and Whited, 2003). As a research field, metabolic engineering can be traced back to the end of the 1980s or early 1990s (Bailey, 1991), where there has been a sharp increase in knowledge about cell metabolism and bioinformatics. Since then, the field has led to the creation of a number of successful cell factories (Ro et al., 2006; Becker et al., 2011; Yim et al., 2011). But a number of limitations have also been identified.

Cellular systems have been optimized by evolution for at least 3.7 billion years (Dodd et al., 2017). This includes systems that maintain a relatively stable internal environment, even when exposed to changing environmental conditions, i.e. homeostasis (Eelderink-Chen et al., 2010; Zhang et al., 2014). This homeostasis is achieved through extensive regulation of metabolic pathways (Chomvong et al., 2017). Thus changing the metabolism of cells requires reprogramming and/or disruption of the existing regulatory programming. Today we lack knowledge of the metabolic regulation, even in relatively simple model cells such as *E. coli* and *S. cerevisiae* (Gardner, 2013; Jeschek et al., 2016; Long and Antoniewicz, 2019). This is a major challenge in the design of cell factories.

Consequently, the development of cell factories is often described as an iterative process of Design, Build, Test, and Learn (The DBTL cycle; Figure 1; Culler, 2016; Nielsen and Keasling, 2016). This has become a popular framework originally adapted from more classical engineering sciences. The steps can be explained as follows, Design: planning of perturbations to a biological system. Build: implementing the perturbations. Test: assaying of the created variation. Learn: improve understanding, e.g. by modeling the variation as a function of initial perturbations. The updated understanding is then used to guide the next design phase.

The DBTL cycle can be accelerated by the development of more effective methods. An example is continuously improved capability of DNA synthesis and thereby the availability of still cheaper, custom-defined synthesized DNA (Kosuri and Church, 2014). Another example is the automation of the individual processes including robotics for high throughput building and testing of genetic designs (Carbonell et al., 2018). The output from the learn step should include hypotheses about relevant cell processes that can be tested in targeted experiments. In addition to improving process understanding, this can lead to more efficient designs and thus a further acceleration of the development of cell factories.

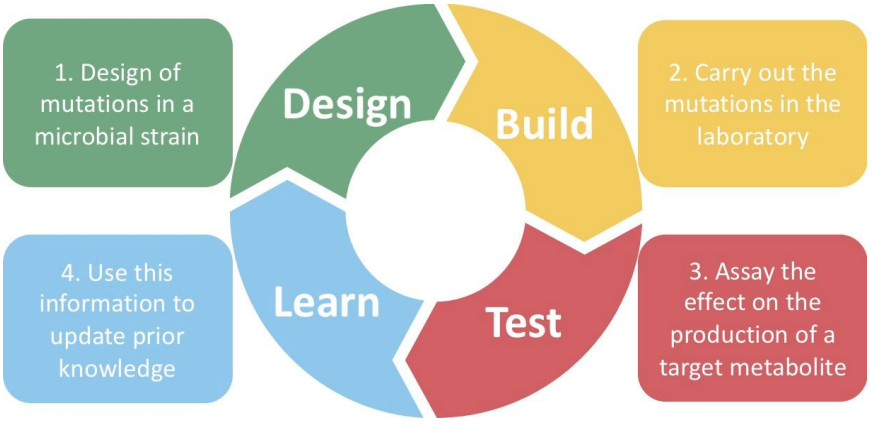


Figure 1. The Design-Build-Test-Learn cycle (the DBTL cycle) illustrating the development of cell factories as an iterative process with four major steps. An example of a single iteration in a particular application is explained in the boxes. Adapted from Culler (2016).

The DBTL cycle and my work

Design

The design step can be supported by the development of well-characterized gene regulatory genetic parts. Such characterization could involve e.g. accurate measurements of effects on mRNA or protein abundance. The availability of such parts would reduce the need for subsequent building and testing (Jeschek et al., 2016). An example of this could be the use of only meaningful genetic parts to eliminate metabolic flux imbalances. It is possible to regulate metabolic fluxes via the concentration of the corresponding enzymes. This can be achieved e.g. by changing genetic parts such as promoters and 5' UTRs (Rossell et al., 2011). It is an advantage to develop genetic parts that work across different organisms because different organisms have different capacities to produce different biomolecules (Martin et al., 2003). The main question raised in paper no. 1 relates directly to the characterization of 5'UTR subsequences for predictable tuning of protein expression in different genetic contexts. In paper no. 2, a set of characterized promoter parts was used to perturb the expression of five genes in the central carbon metabolism.

Build

Strain building generally consists of genomic editing of a base strain. It typically involves the introduction, the knockout, up- or downregulation, or the mutation of one or more genes. There are different methods that can be used for editing genomes. However, the preferred ones in many organisms are based on Clustered Regularly InterSpaced Palindromic Repeats (CRISPR) and their associated Cas proteins (CRISPR/Cas; Cong et al., 2013; Mali et al., 2013). For a more in-depth review on this topic the reader is referred to Jakočiūnas et al. (2016). In brief, for homologous recombination (HR) prone organisms such as yeast, the editing method works by using the CRISPR/Cas system to introduce a double-stranded DNA break and thereby increase the efficiency of HR for error-free DNA repair (Figure 2). The method is preferred because it is relatively quick to implement compared to other precision genome engineering techniques such as TALENs and ZFNs, it is efficient, and because a broad range of edit types can be targeted to almost anywhere in the

genome. Editing efficiencies of up to 100 % at a single locus by HR of multiple parts that are assembled in vivo has been reported using this method (Jakočiūnas et al., 2015). All yeast strains and cell libraries in paper nos. 1 and 2 were made by CRISPR/Cas9 facilitated genome editing. Most of these edits were gene introductions created by in vivo assembly and insertion. The CHO-S cell lines in paper no. 1 were all derived from a base strain edited by CRISPR/Cas to contain a Recombinase-Mediated Cassette Exchange (RMCE) landing pad. The CHO-S reporter cell lines were all constructed by RMCE through this landing pad to ensure high efficiency in the less HR prone CHO-S cell lines.

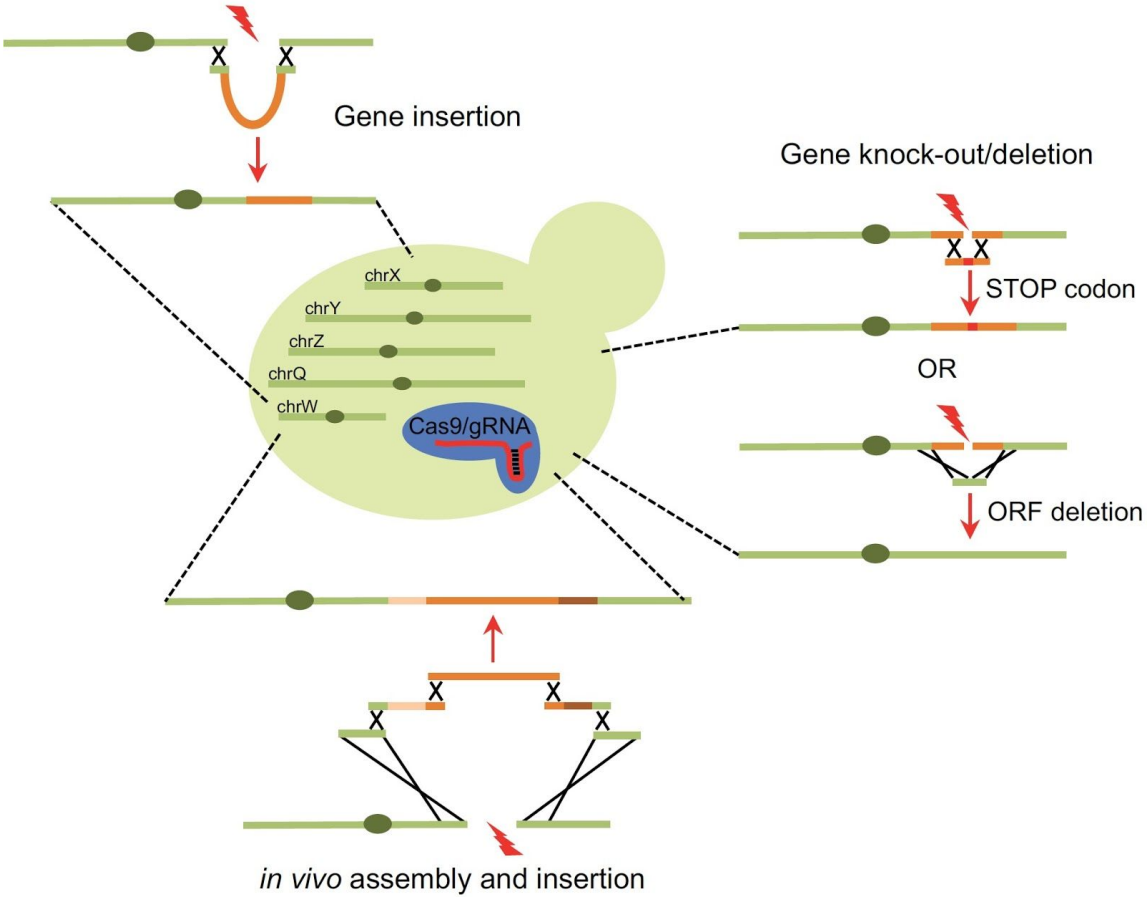


Figure 2. Common usages of CRISPR/Cas9 facilitated genome editing in S. cerevisiae. A CRISPR/Cas9 facilitated double-stranded break (red lightning symbol) can be repaired by homologous recombination (X symbols) via provided repair template (linear DNA fragments). This process can be utilized for *gene insertion*, *gene knockout*, and *in vivo assembly and insertion*. Figure from Jakočiūnas et al., (2016).

In general, all strain edits need to be validated by DNA sequencing, not only because the editing efficiency may be less than 100 %, but also because unintended events like untargeted or double integrations may have happened. The approximately 40 years of evolution of DNA sequencing technologies have recently been reviewed very well by Shendure et al. (2017). In 1976, the influential sequencing method commonly referred to as Sanger sequencing was invented by F. Sanger and A. R. Coulson (Sanger et al., 1977; Figure 3A). Still today, it is frequently used in many laboratories. Over the years the process has matured, it has been automated and it now fits well into many common day workflows. Today the method works well for sequencing reactions of about 1000 bp that can be sequenced relatively cheaply and with low error rates although these increase towards the end of the read (Stucky, 2012). From 2004, several new short-read sequencing technologies became widely available collectively called second/next-generation sequencing (SGS) or massively parallel sequencing (Figure 3B). These technologies overtook Sanger sequencing to some degree. The technique enabled millions of sequencing reactions to progress in parallel. Short read sequencing allowed millions of sequencing reads of lengths in the low hundreds to be sequenced with a very low error rate (Bentley et al., 2008). Thus multiple genomic loci or deep sequencing of a single locus could now be performed in a single experiment (Arsenic et al., 2015). Closely following the SGS technologies came the long-read sequencing technologies also called the third generation sequencing (TGS; van Dijk et al., 2018; Figure 3C exemplified by nanopore sequencing). TGS throughputs are generally lower than short reads sequencing but the longer reads make it easier to assemble the reads into contiguous sequences. The PacBio technology was one of the first examples (Eid et al., 2009). Today, this approach produces reads with an average length of more than 10 kb with reads longer than 100 kb (Wenger et al., 2019). The error rate is relatively high, but it is randomly distributed (i.e. unbiased) and lowered uncertainty can thus be obtained by repeated sequencing. Other important advantages of TGS technologies are that they generally do not require template amplification which can course copying errors and sequence-dependent biases. Also, the technologies have proved able to detect DNA modifications (Flusberg et al., 2010; Song et al., 2012). One remaining problem with the use of long-read sequencing for variant calling appears to be target enrichment (Kalinovski, 2019, personal communication). Hence, in both papers, most single strain edits were validated by Sanger sequencing. In paper no. 1, 5'UTR libraries were validated by Illumina short-read sequencing.

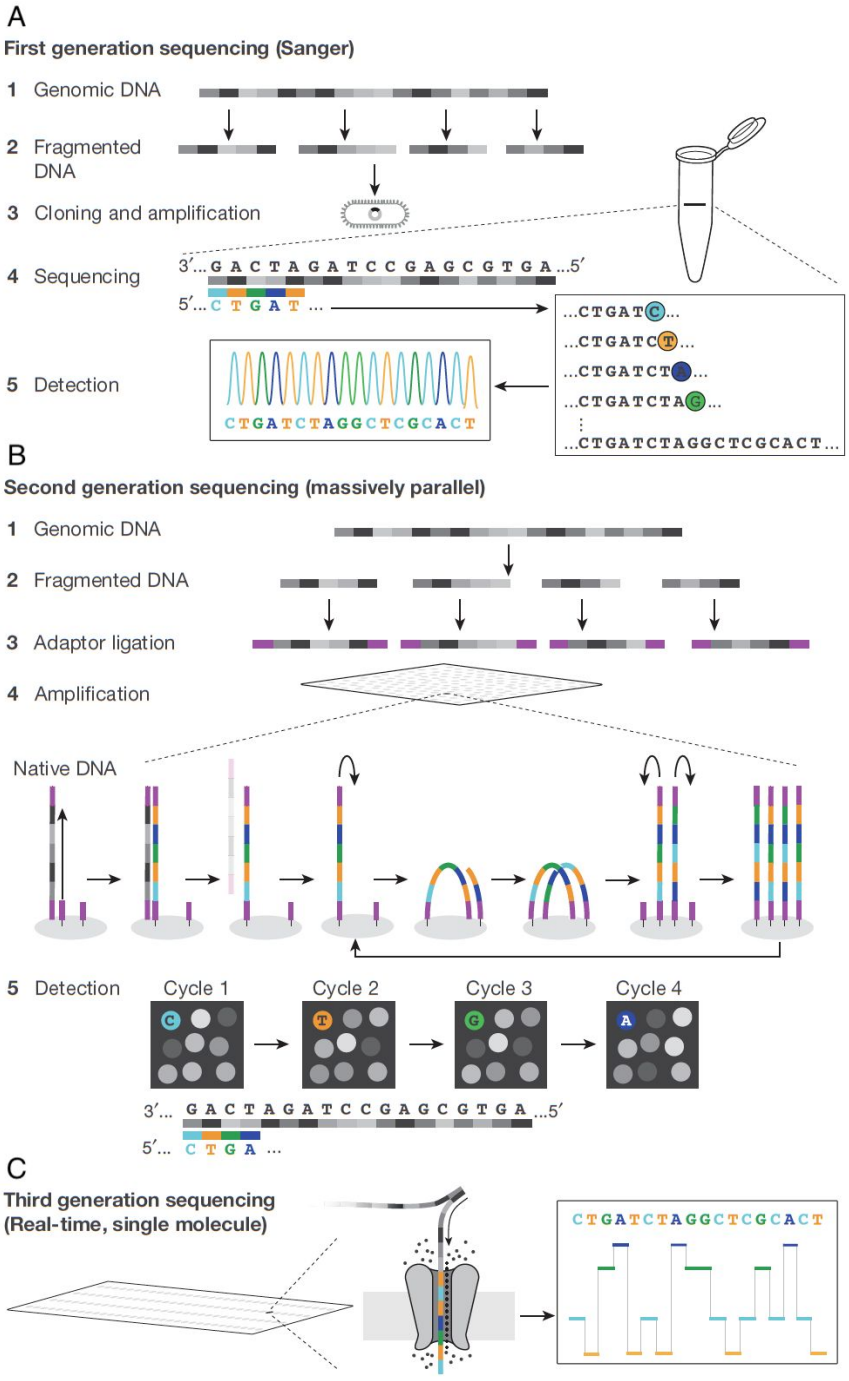


Figure 3. Underlying principles of three influential DNA sequencing technologies. Each of the technologies belong to its own technological generation and they are shown in chronological order. A: Sanger sequencing belonging to the first generation of sequencing technologies. B: Bridge amplification sequencing (second generation). C: Nanopore sequencing (third generation). Figure from Shendure et al. (2017).

The sequencing steps could also be considered as parts of the test step in the DBTL cycle.

Test

Once a strain has been 'build' and validated it is important to be able to evaluate the effect of the editing. Fluorescent proteins (FP) are core components in many powerful microbiology techniques (Cranfill et al., 2016; Figure 4). Among these techniques, multiple are used for single-molecule quantification (Bialecka-Fornal et al., 2016). These techniques rely on genetic systems that couple the amount of the single-molecule to the amount of FP, e.g. by genetically fusing a protein of interest with FP (a fluorescent fusion protein). Subsequently, the amount of FP and thus the amount of the protein of interest can be measured relatively easily by fluorescence spectroscopy techniques. In this way fluorescent proteins can be used to study the biology of proteins in vivo i.e. when, where and how much a protein is expressed (Snapp, 2005). FPs have been developed for decades and there are now FPs covering spectral regions from blue to red with optimized properties (Cranfill et al., 2016). No FP has been found to be the best concerning all properties. Thus it is important to find FP with the right properties for the given application. An example of this could be choosing FPs with similar maturation time when comparing fluorescence signals (Balleza et al., 2017). In paper no. 1 multiple FPs were used in yeast and CHO cells in combination with flow cytometry techniques to investigate the effects of 5'UTR subsequences on protein expression.

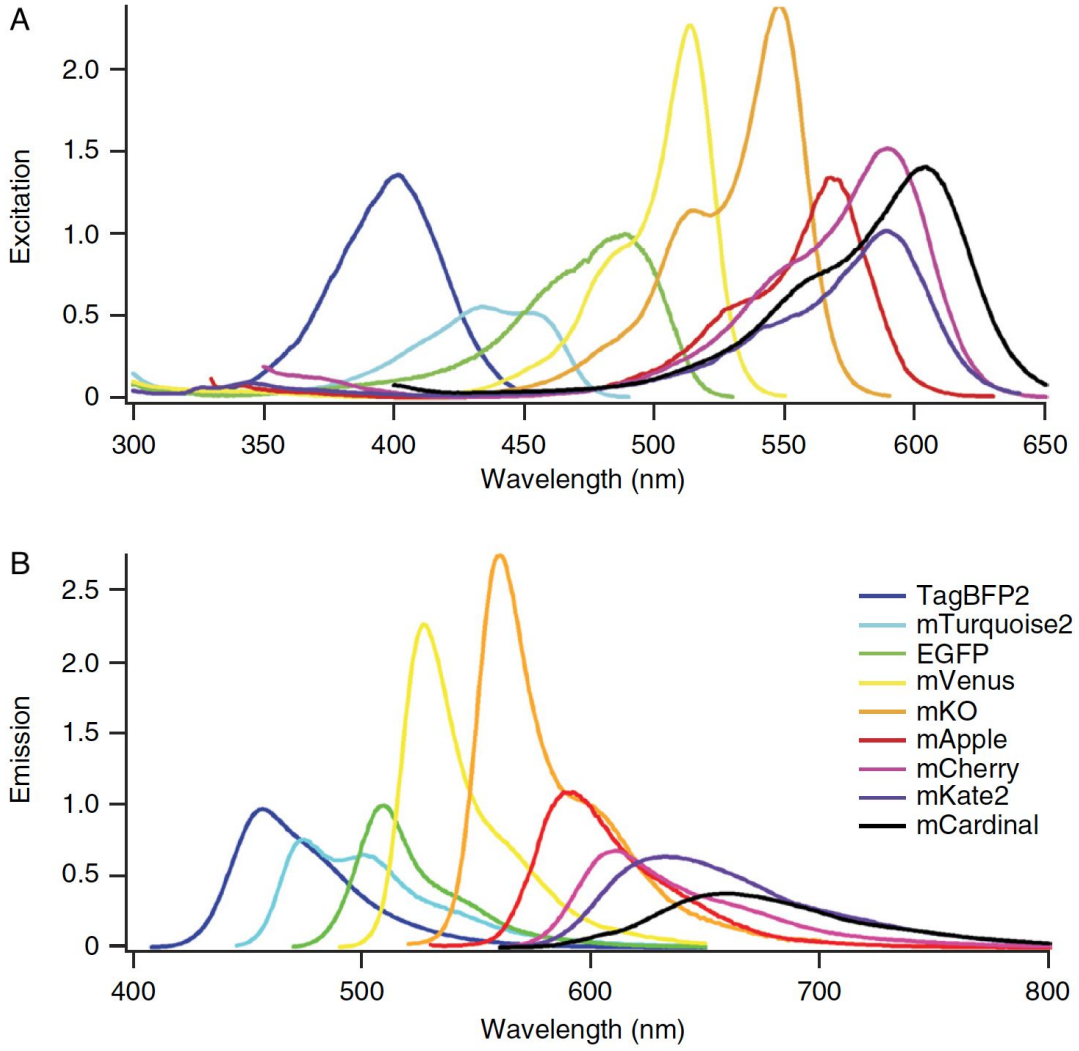


Figure 4. Excitation (A) and emission (B) spectra of 9 commonly used fluorescent proteins that covers the range of visible wavelengths. Figure from Cranfill et al., (2016).

Flow cytometry is a technique that is used to measure the characteristics of individual cells or particles in a population. The technique is often used with fluorescent molecules. Flow cytometry has been reviewed a number of times but an excellent review was made by Picot et al. (2012). In flow-cytometry, a sample of cells is focused to ideally flow one cell at a time through a laser beam (Figure 5). Subsequently, scattered light from the laser beam containing information about the cell is detected. This method enables accurate measurements of single cells instead of whole population aggregates. Today, throughputs can be more than a hundred thousand cells measured per second (Cossarizza et al., 2017). Flow cytometry is used for many applications in cell biology where any cell parameter that can be fluorescence-labeled can be measured in parallel.

Fluorescence-activated cell sorting (FACS) is a specialized type of flow cytometry. FACS allows the sorting of cells based on detected signals and thus the study of sub-populations at the single-cell level (Herzenberg et al., 1976). In paper no. 1 flow cytometry techniques were used for isoclonal variants at the population level. Furthermore, FACS was used to screen a 5' UTR cell library to identify UTR sequences that resulted in the differential expression of yeGFP.

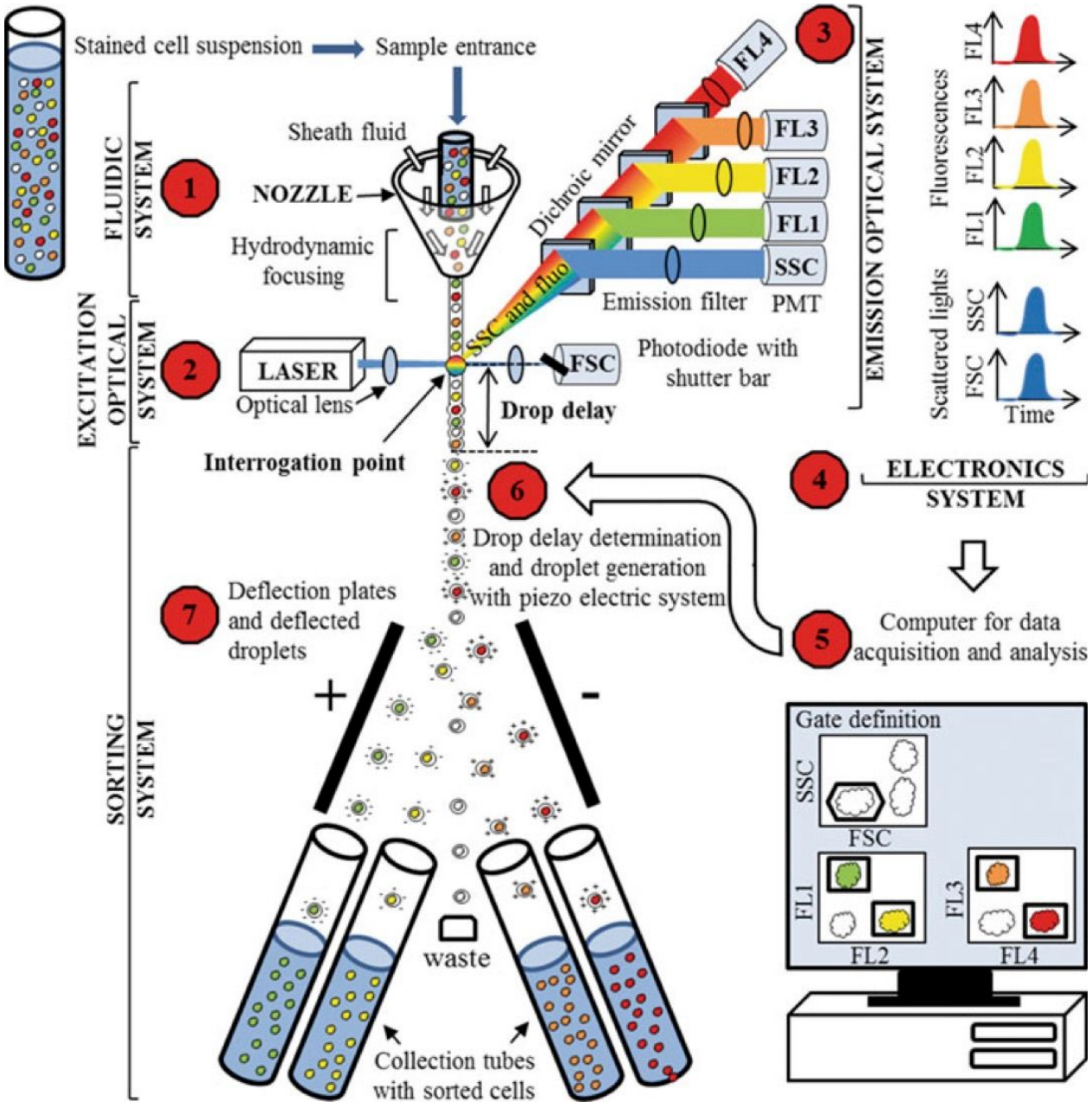


Figure 5. Principles of flow cytometry (systems 1-5) and Fluorescence Activated Cell Sorting (FACS; systems 1-7). Figure from Picot et al. (2012)

Our ability to create genetic edits by far exceeds our capacity to characterize these edits. This makes screening a bottleneck in the DBTL cycle. It has been shown that the capacity for screening diversity can be increased by using genetically encoded devices instead of classical analytical methods (Zhang et al., 2015). In nature, metabolites are in part detected by allosteric transcription factor proteins that upon metabolite binding controls the activity of corresponding promoters in order to regulate metabolism. Such transcription factors can be reengineered to monitor intracellular metabolites e.g. by linking the presence of a particular metabolite to the expression of a FP (Figure 6). Large cell libraries can be phenotypically characterized in high throughput by using such re-engineered transcription factors (i.e. biosensors). In paper no. 2, a biosensor was developed based on an engineered small molecule-binding transcriptional activator. It was used in combination with a microplate spectrophotometer for high throughput phenotypic characterization by real-time monitoring of fluorescence signal.

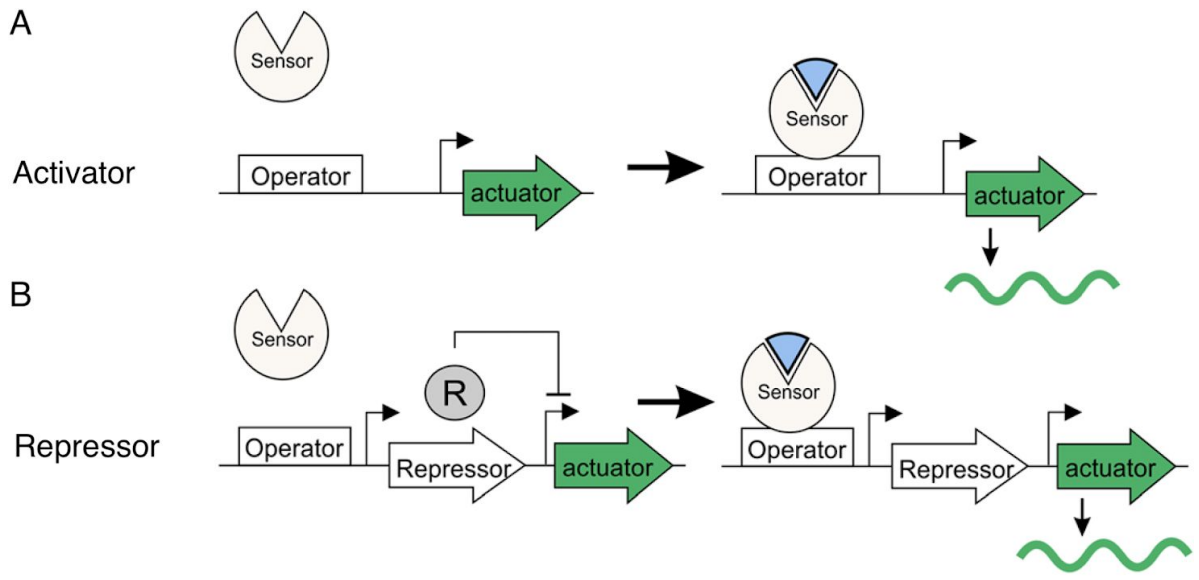


Figure 6. Principles for biosensor development based on allosterically regulated transcription factors. A: A transcriptional activator system can be reengineered to activate the expression of an actuator gene in the presence of a ligand. B: A similar response can be obtained from a transcriptional repressor system by reengineered it to relieve the repression of an actuator gene in response to the presence of a ligand. Figure from Mahr and Frunzke (2016).

Learn

Nielsen and Keasling (2016) state that "Learning is possibly the most weakly supported step in current metabolic engineering practice, yet perhaps the most important to increasing the rate of success". There are two general approaches to learning: reductionist and holistic. The reductionist approach is to study individual processes before combining them to finally reach an overall mechanistic understanding of a system. This is generally regarded as the traditional approach to (biological) science. An example of this can be found in the influential review "Foundations of Engineering Biology" where it is described how we can improve our ability to engineer biological systems via the concepts of standardization, decoupling, and abstraction (Endy, 2005). The holistic approach is to look at the process as a whole and try to understand the overall general principles and secondly decompose these principles in order to understand subprocesses. This approach is much like the approach of systems biologists. So which approach should one choose? Presnell and Alper (2019) hypothesize that the most effective strategy would be to integrate aspects of both approaches. In paper no. 2 holistic methods were used to describe the combined effect of elements that have been studied in a reductionist manner.

Both the reductionist and holistic approaches can be well supported by modeling, i.e. by hypothesis and data-driven modeling, respectively. Data-driven or statistical models are interesting because we generally do not have enough process understanding to build perfect hypothesis-driven models of entire biological systems. Today, a popular type of statistical model is based on machine learning (ML). An excellent review on the application of machine learning in the field of systems metabolic engineering was recently made by Presnell and Alper (2019). According to this review, the applications can so far be fit into one of four categories called product maximization, de novo pathway design, phenotypic profiling, or robust systems modeling. Even so, ML models are still used relatively little in metabolic engineering. Today, it is not clear how far you can get towards modeling of biological systems with machine learning based on high throughput characterization. In particular, it is not known how much and what types of data are needed. Still, machine learning has been used to extract patterns in the data that can be used to improve the design of new biological systems (Bonde et al., 2016; Costello and Martin, 2018; Zhou et al., 2018; Jervis et al., 2019).

There are many different machine learning approaches. Programming packages such as the open-source python library "Scikit-learn" contains a long list of implementations of popular ML algorithms (> about 50 depending on how you define an individual model) that are ready to use (Pedregosa et al. 2011). Machine learning models can be divided into supervised and unsupervised models (Dey, 2016). Contrary to the unsupervised methods, the supervised models need supervision in the form of prior knowledge about the target variable. However, in this thesis, I will only address the supervised models. The supervised models are generally good at making predictions based on complex relationships (Ohler et al., 2002; Segal et al., 2006). The ML models vary in success depending on input data and application. This is due to the fact that every model has built-in specific assumptions about data. In general, it is complicated to tell in advance which model will fit your data and application the best (Wolpert, 1996). Thus, model selection requires experimentation.

It may be a combination of contributions from different models (a model ensemble) that is optimal (Dietterich, 2000; Rokach, 2010). There are three reasons for this as illustrated in Figure 7. First, the training of an ML-model can be viewed as searching a space of hypotheses (H) in order to identify the best. A statistical problem arises if the amount of training data available is too small compared to the size of the hypothesis space. The model will find many hypotheses (h1, h2, ...) that provide the same good accuracy (Figure 7, shown in blue) but not necessarily the right hypothesis (i.e. the true function, f). By constructing an ensemble out of all of these accurate models, that is by "averaging" their votes, the chance of choosing the right hypothesis increases (Figure 7A). Second, ML-models usually work by performing some sort of search for a better fit that can get stuck in a local optimum (see figure on the frontpage). Hence, in situations where there is enough training data (so that the statistical problem is absent), it may still be very difficult computationally for the ML-model to find the best hypothesis. An ensemble constructed by running the local search from many different starting points may provide a better approximation to the true unknown function than any of the individual models (Figure 7B). Third, in most ML-applications, the true function f cannot be represented by any of the hypotheses in H. By forming a weighted sum of hypotheses from the space of hypotheses (H), it may be possible to represent functions outside H (Figure 7C).

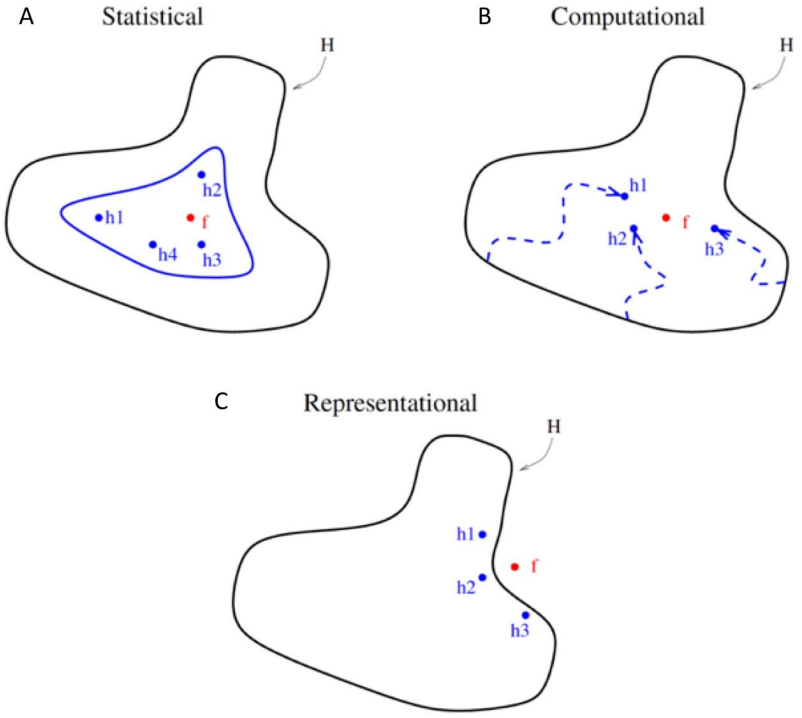


Figure 7. Three reasons (Statistical, Computational, and Representational) why an ML-model ensemble may work better than a single ML-model. Figure from Dietterich, (2000).

A sophisticated method to automate this model selection process called the Automated Recommendation Tool (ART; Radivojević et al., 2019) has recently been developed. ART creates an ensemble model by combining up to 30 different ML models contained in the Scikit-learn library. In paper no. 2 ART was used to select the best ensemble of ML models for modeling genotype to phenotype relationships and for predicting unseen genotypes with improved performance.

The training of ML algorithms requires large datasets of high quality (Camacho et al., 2018). This requirement still needs to be met even though laboratories are continuously developing new experimental workflows that are getting more efficient and accurate (Carbonell et al., 2019). One aspect of data quality is the variation created by manipulating relevant genetic targets. Finding such relevant targets is not trivial. The combination of a large genetic space and limited experimental capacity necessitates a certain mechanistic understanding. In paper no. 2 a combination of genome scale modeling for genetic target selection and ML modeling was used to identify strains with increased tryptophan synthesis rate.

1.2 Outline of work done during the Ph.D. study

All work that has been done in the timeframe of this thesis, except for the work related to paper no. 5, has been on methods for lowering the turn-around time of the metabolic engineering cycle and/or reducing the needed number of iterations, and can, therefore, be linked to one or more steps in the DBTL cycle (Figure 8). Paper titles and numbers are shown on the cycle according to their main contribution although there may also be significant contributions elsewhere. For instance, paper no. 1 indicates a main contribution in the Design step although the context-independent hexameric TIS sequences can also be seen as a significant help in building smart libraries and therefore a contribution to the Build step. And paper no. 2 is placed next to the Learn step even though it also relates to the other steps: to the Build step because 13-part and 20 kb library construction, to the Test step because the development of a new biosensor was part of the study, and to the Design step because the design of new strains was supported by genotype to phenotype modeling.

Papers no. 3 to 5 in the list of publications are not included in the thesis because the majority of the work was done before the start of my Ph.D. studies. Moreover, papers no. 3 and 4 concerns bacteria, not eukaryotes. My main contributions to these three papers are:

No. 3: to develop a functional selection system for thiamine pyrophosphate in *E. coli*. Proofreading was done during the PhD study.

No. 4: to construct and characterize a library of refactored thiamine pyrophosphate pathways in *E. coli*. Recommended strains was constructed during the PhD study.

No. 5: to search the literature and collect published snake venom proteomics data. Paper writing and proofreading was done during the Phd study.

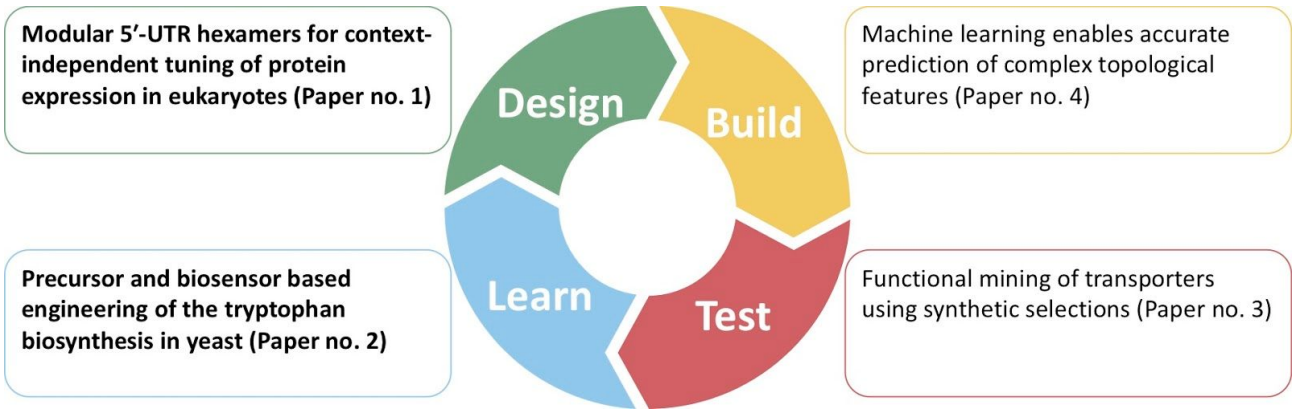


Figure 8. Work related to the DBTL cycle done during the Ph.D. study. Paper titles and numbers are shown on the cycle according to their main contribution although there may also be significant contributions elsewhere. Bold text indicates work included in the thesis which will be described in the following sections.

1.3 Research aims and specific questions addressed in the thesis

The overall aim of this thesis is to develop methods for accelerated and predictable engineering of metabolism in eukaryotes. More specifically the aim is to construct and characterize libraries of regulatory genetic parts, to combine such regulatory parts with pathway genes thereby creating combinatorial pathway libraries, to characterize such libraries by linking the individual genetic combination to productivity, and to use machine learning approaches to discover unseen pathway combinations with superior performance. More specific research aims and key research questions addressed in the two papers are stated below.

In paper no 1, the aim was to contribute to the development of tools for predictable tuning of protein expression in eukaryotes. Here, the following research questions were addressed:

1. Can the untranslated part of the translation initiation site (TIS) sequence be used for predictable tuning of protein activity in a broad variety of eukaryotes?
    1.1. How much variance in yeGFP fluorescence can be achieved by randomizing the TIS sequence bases at positions -6 to -1 relative to the open reading frame?
    1.2. How much difference in median yeGFP fluorescence can be obtained between single clone populations sorted from a TIS sequence cell library?

1.3. Are the relative difference in yeGFP fluorescence due to TIS sequences independent of a promoter sequence, reporter sequence, carbon source, and chassis?

1.4. How does the expression from the strongest selected hexamer (i.e. TCGGTC) compare with the strength of the most frequently used hexamer in CHO cells (i.e. GCCACC)?

1.5. How much variation in growth rate and carotenoid production can be achieved by using the hexamers to balance the Erg9 and crtE protein expression?

In paper no. 2, the aim was to improve methods for genotype-to-phenotype predictions as a tool for metabolic engineering. Here, the following research questions were addressed:

1. How much can the tryptophan synthesis rate in yeast be increased in one iteration of the DBTL cycle when using a GSM to assist the selection of genetic targets and ML models for prediction?
    1.1. How much variation in the tryptophan synthesis rate can be achieved when using a GSM in the identification of gene targets?
    1.2. How much genetic variation can be achieved in a combinatorial library when combining 5 target genes with 30 promoters using CRISPR/Cas9 facilitated genome editing methods?
    1.3. Can a biosensor be used to overcome the measurement of tryptophan concentration more than 144000 times?
    1.4. How big a dataset is needed to train the ML models?
    1.5. Can machine learning approaches be used to capture significant patterns in the measured data with a small mean absolute error (MAE)?
    1.6. Can ML models be used to predict and thereby identify strains with a higher tryptophan synthesis rate than any in the library?

Published online 14 August 2018

Nucleic Acids Research, 2018, Vol. 46, No. 21 ***e127***
doi: 10.1093/nar/gky734

2. Modular 5′-UTR hexamers for context-independent tuning of protein expression in eukaryotes

Søren D. Petersen[1], Jie Zhang[1], Jae S. Lee[1], Tadas Jakočiūnas[1], Lise M. Grav[1], Helene F. Kildegaard[1], Jay D. Keasling[1,2,3,4,5,6] and Michael K. Jensen[1,*]

[1]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark, [2]Joint BioEnergy Institute, Emeryville, CA 94608, USA, [3]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [4]Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA 94720, USA, [5]Department of Bioengineering, University of California, Berkeley, CA 94720, USA and [6]Center for Synthetic Biochemistry, Institute for Synthetic Biology, Shenzhen Institutes of Advanced Technologies, Shenzhen 518055, China

Received March 09, 2018; Revised July 24, 2018; Editorial Decision July 30, 2018; Accepted August 01, 2018

ABSTRACT

Functional characterization of regulatory DNA elements in broad genetic contexts is a prerequisite for forward engineering of biological systems. Translation initiation site (TIS) sequences are attractive to use for regulating gene activity and metabolic pathway fluxes because the genetic changes are minimal. However, limited knowledge is available on tuning gene outputs by varying TISs in different genetic and environmental contexts. Here, we created TIS hexamer libraries in baker's yeast *Saccharomyces cerevisiae* directly 5′ end of a reporter gene in various promoter contexts and measured gene activity distributions for each library. Next, selected TIS sequences, resulted in almost 10-fold changes in reporter outputs, were experimentally characterized in various environmental and genetic contexts in both yeast and mammalian cells. From our analyses, we observed strong linear correlations (R^2 = 0.75–0.98) between all pairwise combinations of TIS order and gene activity. Finally, our analysis enabled the identification of a TIS with almost 50% stronger output than a commonly used TIS for protein expression in mammalian cells, and selected TISs were also used to tune gene activities in yeast at a metabolic branch point in order to prototype fitness and carotenoid production landscapes. Taken together, the characterized TISs support reliable context-independent forward engineering of translation initiation in eukaryotes.

INTRODUCTION

Control of protein expression is critical for cellular development, differentiation and adequate response to intra- and extracellular conditions (1). From simple bacteria to multicellular eukaryotes, control of protein expression involves the sequence composition of the 5′-untranslated regions (5′-UTRs) of existing messenger RNAs (mRNAs). Specifically, translation initiation, where the AUG start codon is identified by ribosomes and decoded by methionyl-(transfer RNAs) tRNAs (met-tRNAs), is recognized as one of the most crucial steps in translation (1–4). As a consequence, a large number of studies have been performed to deduce the relationship between 5′-UTR sequences and protein expression (5–11).

In bacteria, simple base-pairing between the 6-nt Shine–Dalgarno (SD) sequence located immediately 5′ end of the start codon and the anti-SD sequence in the peptidyl decoding site of the 16S ribosomal subunit controls translation initiation from 50 to 70% transcripts by modulating the ribosomal accessibility to the SD sequence (12,13). Moreover, deep sequence–function characterization of SD libraries has enabled the development of predictive algorithms for tuning of protein expression over several orders of magnitude by simple modulation of the SD sequence (4–6).

In eukaryotes, translation is initiated at the 5′ end of mRNA by the recruitment of the 40S ribosomal subunit, auxiliary initiation polypeptides and the met-tRNA, collectively the 43S pre-initiation complex (PIC) (14). Different from translation initiation in bacteria, once recruited, PIC scans along a much larger sequence space of eukaryotic 5′-UTRs, often several hundred nucleotides in length, until encountering an AUG codon (14–16). During scanning, a number of 5′-UTR sequence features are known to affect translation, including mRNA secondary structures,

*To whom correspondence should be addressed. Tel: +45 6128 4850; Fax: +45 4525 8001; Email: mije@biosustain.dtu.dk
Present address: Jae S. Lee, Department of Molecular Science and Technology, Ajou University, Suwon 16499, Republic of Korea.

decoy AUG codons, PIC stalling at upstream open reading frames (uORFs) and the sequence context surrounding the cognate AUG for translation, commonly referred to as the Kozak sequence (17–29).

Similar to the algorithms established in bacteria for tuning protein expression, major efforts have been performed to mine the causal sequence elements of native eukaryotic 5′-UTRs, in order to attempt to model and forward engineer 5′-UTRs with predictive protein expression outputs (7,9). Initially, Kozak *et al*. reported GCCRCCAUGGG (R = A/G, start codon underlined) to effectively control ribosomal recognition of AUG and thereby initiation of translation (30–32). In particular the positioning of a purine at position -3 and a guanine at position +4 from the AUG codon has later been adopted for efficient translation initiation (33–35). Expanding on this, in mammalian cell lines, Noderer *et al*. have systematically probed the efficiency of start codon recognition for all possible translation initiation sites (TISs) flanking the AUG start codon at positions -6 to -1, and +4 and +5, totaling ∼65 000 TIS sequences, concluding that the motif RYMRMVAUGGC (Y = U or C, M = A or C, R = A or G and V = A, C, or G, start codon underscored) enhanced start codon recognition and GFP translation efficiency (8). Likewise, in yeast, recent studies have attempted to accurately estimate TIS efficiencies on reporter protein expression by randomizing 5′-UTR elements up to 50 nt upstream AUG (uAUG) and training computational models on smaller subfractions (4×10^{-26} – 0.2%) of these libraries (7,9). Here, both studies pointed out that in addition to uORFs and mRNA secondary structure, positions -3 to -1 from the AUG start codon are the most important parameters for tuning protein expression, ultimately enabling the construction of an algorithm explaining up to 70% of the observed variation in protein levels (7). More recently, the computational model by Dvir *et al*. has been further validated with new experimental data, again investigating the interactions with polymorphism in nucleotides at positions -10 to -1 relative to the start codon, but this time also placing the TIS in genomic contexts of two different reporter proteins and two different promoters (10). Here, similar correlations between experimental data and model predictions were observed ($R^2 = 0.36$–0.73), ultimately suggesting that the ∼30% variation observed for which the current models cannot account for arises from experimental noise and yet-uncharacterized biological factors (7). Likewise, though Noderer *et al*. showed strong linear relationships between GFP expression in different mammalian cell lines and cultivation media, comparisons of TIS efficiencies of GFP expression compared to other reporter genes suggested some context-dependence of TISs and open reading frame (ORF; $R^2 = 0.39$–0.76), in line with the model being trained on ORF-specific library sequences including the +4 and +5 positions (8). Taken together, the above studies indicate that systematic characterization of the impact of short TISs on protein expression in broad contexts still remains to be elucidated before TISs can be used as a tool for predictable tuning of protein expression.

In this study, we sought to establish a robust, simple and experimentally validated workflow to assess the sequence–function relationship of TISs in diverse genomic and environmental contexts (Figure 1). To do so, we created three TIS libraries spanning more than 4500 designs for nucleotide positions -6 to -1 directly upstream of an ORF of GFP controlled by three different promoters in yeast. Based on fluorescence-activated cell sorting (FACS) and single clone validations, a diverse sample of TIS hexamers with a robust output range of ∼10-fold was selected for further characterization in the context of different ORFs, promoters, host chassis, growth medium and cell densities (Figure 1). In general, the linear relationship between relative fluorescence output from selected TIS sequences obtained in different contexts was high, with correlation coefficients ranging from 0.77 to 0.98. Moreover, testing TISs derived from yeast in mammalian cell lines, we specifically uncovered a TIS sequence stronger than the Kozak element commonly used to drive protein production in mammalian cell factories. In addition, we used selected TIS hexamers to investigate the carotenoid production landscape in yeast by tuning dual protein activities at an essential metabolic branch point, thereby prototyping the fitness and carotenoid production landscape in a simple and cost-effective manner. Our detailed, experimental analyses allow us to put forward a list of short sequence-validated TISs to be used as a method for predictable tuning of protein expression in diverse genomic and environmental contexts.

MATERIALS AND METHODS

Strains, cell lines and growth media

Baker's yeast *Saccharomyces cerevisiae* strains were derived from CEN.PK2-1C (EUROSCARF, Germany). Yeast strains were cultured in yeast synthetic drop-out media (Sigma-Aldrich) at 30°C. CHO-S cells (ThermoFisher) and derivative cells were maintained in CD CHO medium (Gibco Cat. #10743-029) supplemented with 8 mM L-glutamine (Lonza Cat. #BE17-605F) and 2 ml/L anti-clumping agent (Gibco Cat.#0010057AE) in 125 ml Erlenmeyer shake flasks (Corning Inc., Acton, MA), incubated at 37°C, 5% CO_2 at 120 rpm and passaged every 2–3 days. *Escherichia coli* DH5α were cultured in Luria-Bertani (LB) medium containing 100 mg/l ampicillin (Sigma-Aldrich) at 37°C.

Plasmid and strain construction

Yeast integrative plasmids were created by USER cloning (36) and propagated in *E. coli* DH5a. Yeast transformations were performed by LiAc/SS carrier DNA/PEG method (37). Plasmids and polymerase chain reaction (PCR) products were purified using kits from Macherey-Nagel. Biobricks for USER assembly were amplified using Phusion U Hot Start PCR Master Mix (ThermoFisher), parts for transformation by Phusion High-Fidelity PCR Master Mix with HF Buffer (ThermoFisher), whereas colony PCRs were performed using 2xOneTaq Quick-Load Master Mix with Standard Buffer (New England Biolabs). Oligos, duplex oligos and gBlocks were purchased from Integrated DNA Technologies (IDT). Sequencing was performed by Eurofins. All primers, plasmids, yeast strains and CHO cell lines are listed in Supplementary Tables S1, S2, S3 and S4, respectively.

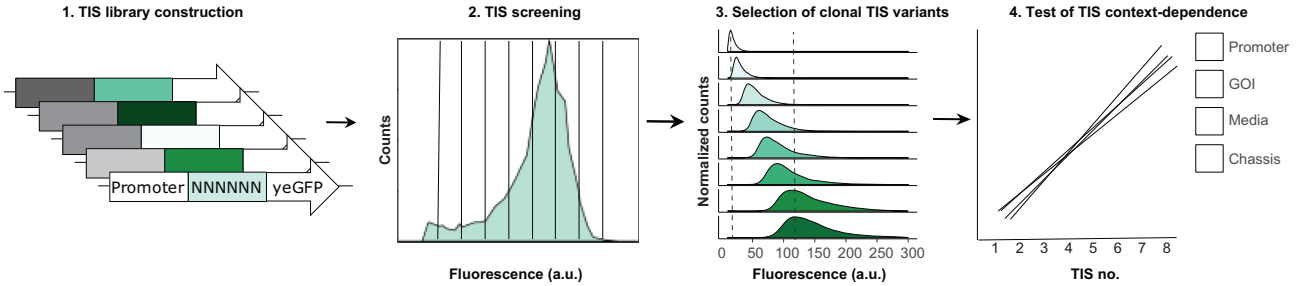


Figure 1. Workflow for TIS library construction and characterization.

Constructions of TIS libraries

Three TIS libraries were constructed using the EasyClone method (38) with slight modifications. Briefly, 0.1 pmol of promoter REV1 (389 bp upstream position +1 of YOR346W; Supplementary Table S5), RPL18B (700 bp; upstream position +1 of YNL301C) and TEF1 (420 bp; upstream position +1 of YPR080W) were cloned together with yeast-enhanced GFP (yeGFP; Supplementary Table S6) (39) with six randomized nucleotides upstream the start codon, into 0.03 pmol linearized EasyClone vector pCfB261 amplified from a vector excluding a *ccdB* cassette (suicide gene (40); pCFB8168) to counterselect for non-linear vector template. Correctly assembled EasyClone plasmids (0.5 pmol) containing the yeGFP cassette under either REV1, RPL18B or TEF1 promoter was linearized by NotI (Fermentas) and transformed into TC-3 cells for recombination at the EasyClone site XII-5 facilitated by Cas9 (41). *Escherichia coli* and *Saccharomyces cerevisiae* library colonies were scraped and pooled from five individual transformation plates. In *E. coli*, TIS library sizes in the context of REV1, RPL18B and TEF1 promoters were estimated by colony count to ∼1180, 4600 and 4520, respectively. In *S. cerevisiae*, REV1, RPL18B and TEF1 library sizes were estimated to ∼845, 3300 and 2750, respectively.

Construction of promoter and reporter strains

Thirty-two yeast strains were constructed similarly to the TEF1 promoter TIS library, by combinatorial assembly of promoter, TIS and reporter sequences using the CasEMBLR method with assemblies integrated into EasyClone site XII-5 (38,42). For these strains, the TIS was defined as one of eight sequences TGATAT, CGACTT, ACGTTC, GGGGGT, TAGGTT, AGGACA, TGTGAA or TCGGTC. Eight of the 32 strains were constructed by transformation of the alcohol dehydrogenase II (ADH2) promoter fragment (700 bp upstream position +1 of YMR303C), and one of eight TIS-yeGFP into strain TC-3 (43) for recombination at the XII-5 site. Homology between fragments was 30 bp, with up and down fragments of at least 450 bp for homologous recombination into EasyClone site XII-5. The remaining 24 strains were constructed similarly by transformations of the previously used TEF1 promoter fragment, with each of 24 fragments made by combination of the eight TIS sequences with the three reporter genes encoding ymUkG1 (44), yeGFP and mKate2 (45).

Construction of carotenoid strains

A background carotenoid expression strain was created by first, amplifying genes *crtI* and *crtYB* from plasmid YIplac211-YB/I/E* (46). Second, genes were USER cloned together with bidirectional promoter (pTDH3_pTEF1) into linearized vector pCfB390 (38) to create a plasmid pTAJAK-11. Third, linearized pTAJAK-11 was integrated into CEN.PK2-1C strain, XI-3 site as described in Jensen *et al.* (38), and the strain was named TC-9. Fourth, TC-9 was transformed with the Cas9 expression plasmid pCfB176 to create strain TC-10. Further, carotenoid strains were constructed by transforming (i) linear 90-bp DNA donor fragments spanning *Erg9* promoter and coding sequence introducing either TIS no. 1 (TGATAT), no. 5 (TAGGTT) or no. 8 (TCGGTC) directly upstream the start codon, (ii) a pCfB261 upstream part with phosphoglycerate kinase gene (PGK1) promoter (YCR012W; 984 bp) and (iii) either TIS no. 1 (TGATAT), no. 5 (TAGGTT) or no. 8 (TCGGTC) directly upstream the *crtE* start codon with a pCfB261 downstream homology part into EasyClone site XII-5 of strain TC-10 by the CasEMBLR method (Erg9 gRNA sequence: CACATATCACACACACACAA; XII-5 gRNA sequence: TTGTCACAGTGTCACATCAG) (42).

Construction of the CHO reporter cell pools

CHO reporter cell pools were derived from a master cell line harboring a recombinase-mediated cassette exchange (RMCE) landing pad. The master cell line was made by CRISPR-mediated homology directed targeted integration of CHO-S cells as previously described (47), with minor changes in the homology-directed repair (HDR) donor plasmid. The mCherry coding sequence in the HDR-donor plasmid (pCFB8173) has been flanked by a *loxP* sequence at the 5′ end and a *lox2272* sequence at the 3′ end (pEF1α-*loxP*-mCherry-*lox2272*-BGHpA), and the 5′ and 3′ homology arms target a non-coding region. Promoterless and polyAless RMCE vectors were constructed by assembly of PCR fragments containing TIS sequences no. 1 (TGATAT), no. 5 (TAGGTT), no. 8 (TCGGTC) or the mammalian consensus TIS (GCCACC) in combination with mammalian-enhanced GFP (meGFP) (48) or ZsGreen1 (Clontech #632428) that were flanked by *loxP* and *lox2272* sequences. The CHO master cell line at a concentration of 1×10^6 cells/ml was transfected with TIS-GFP or TIS-ZsGreen1 RMCE reporter plasmids and Cre-

recombinase vector in 3:1 ratio (w:w) in six-well plates using FreeStyle™ MAX transfection reagent to exchange mCherry coding sequence with TIS-GFP or TIS-ZsGreen1 cassettes. For Cre recombinase expression, PSF-CMV-CRE recombinase vector (OGS591, Sigma-Aldrich) was used. Transfected cell pools were passaged two times after transfection. After 7 days, cell pools were analyzed by flow cytometry. Flow cytometry revealed that 1–3% of the cells in all cell pools were changed from mCherry to GFP positive.

Next-generation sequencing of TIS libraries

Genomic DNA was extracted from over night cultures using PureLink Genomic DNA Purification Kit (Invitrogen). Genomic DNA extracts were used as template in PCR amplifying ∼300 bp overlapping the TIS sequence within the first 50 bp. Purified PCR products were indexed with Nextera XT indexing. The indexed amplicons were quantified using Qubit 2.0 Fluorometer (Life Technologies), pooled in equimolar quantities and sequenced on Illumina MiSeq using 75-bp reads. TIS sequences were extracted from sequencing reads using the cutadapt command line tool (49) treating the TIS sequence flanking regions as anchored adapters. Reads shorter than 70 bp was removed and up to 5 bp mismatches in the flanking regions were allowed in total. From 4.1 million usable reads from the three sequenced libraries, we identified 4037, 4093 and 4037 (or 98.6–99.9%) of the 4096 possible hexameric TIS sequences for each library. Of these 1721, 2174 and 844 TISs exceeded our cutoff of 100 reads for reliable quantification.

Flow cytometry and TIS library sorting by FACS

Yeast cells were grown in 96-well microtiter plates ON to saturation, diluted to OD_{600} 0.025 (measured by reading the absorbance at 600 nm on Microplate Reader, BioTek) and incubated for 4–6 h (until OD_{600} reached 0.1–0.2) before being measured by flow cytometry using a MACSquant VYB (Miltenyi) or BD Fortessa (BD Biosciences) flow cytometer. CHO reporter cell pools in exponential phase were analyzed by flow cytometry using a BD FACSJazz cell sorter (BD Biosciences). Fluorescence of yeGFP, ymUkG1, meGFP and ZsGreen1 was measured after excitation by 488 nm laser and detected through 525/50 nm bandpass filters. Fluorescence of mKate2 and mCherry was measured after excitation by 561 nm laser and detected through a 615/20 nm bandpass filter. Cells were gated based on FSC-A and FSC-H (singlets) as well as FSC-A and SSC-A profiles for robust measurements. All fluorescence data presented are median values for at least 10 000 or 5000 cells from yeast and CHO cells, respectively. Flow cytometry data were analyzed using FlowLogic version 700.2A (Inivai Technologies). Fluorescence measurements from each fluorescent protein were mean normalized (each measurement divided by the mean and multiplied by 100).

The TIS library in the context of the TEF1 promoter was divided into 10 equal gates based on yeGFP signal, and 48 cells were sorted out from each gate using BD FACS ARIA II (BD Biosciences). In total 480 cells were spotted onto agar plates, grown in liquid cultures and validated by flow cytometry (BD Fortessa, BD Biosciences). Eight colonies spanning the range of fluorescence were selected and sequenced to reveal the corresponding TISs.

Characterization of carotenoid strains

Pre-cultures were inoculated from glycerol stock and incubated for 48 h before 2 μl culture was spotted onto SD agar. Pictures of colonies on agar plates were taken after incubation for 48 h at 30°C and 72 h at 5°C. Maximum specific growth rates (μ_{max}) were calculated using the Easylinear function from the growth rates R package (50) and setting the number of consecutive data points to 10 ($h = 10$). Growth were measured in 200 μl cultivations (Growth Profiler 960, EnzyScreen).

β-Carotene extraction and quantification by HPLC

Measurements were performed using a method described by (51) with a few modifications. β-Carotene was extracted from 2 ml culture broth. The pelleted cells were lysed with 250 μl glass beads and in 500 μl ethyl acetate supplemented with 0.01% 3,5-di-tert-4-butylhydroxy toluene (BHT). Finally, 300 μl ethyl acetate was evaporated from cell extracts and the pellet was redissolved in 1.5 ml ethanol with 0.01% BHT for high pressure liquid chromatography (HPLC) measurements.

Data analysis

Data analysis were mainly done using the *R* statistical environment (version 3.4.1). Additional analysis using RNAfold from the Vienna RNA package (version 2.4.6), the yUTR-calculator by Decoene *et al.* (10) and cutadapt version 1.13 was performed in a Python 3.5 environment. Systematic names of yeast genes using one of TISs 1 to 8 were found using the find Motif search tool in CLC Main Workbench (version 7.7.2) on the CEN.PK113-7d genome (Genbank ID: AEHG01000000).

RESULTS

Construction and characterization of TIS libraries

Our first aim was to characterize in high-throughput how short TIS sequences affect protein expression in eukaryotes. In yeast and mammalian cells, earlier studies have characterized TIS libraries ranging from -50 to -1, -10 to -1, -6 to +4 and -6 to +5 positions relative to the AUG start codon (7–9,33,52). From those studies, it has been inferred that (i) positions -3 to -1 from the AUG start codon, (ii) a purine (R) in position -3 or lack of G in position +4, (iii) mRNA secondary structure and (iv) out-of-frame uAUGs are the most important parameters for tuning protein expression. In order to experimentally investigate in greater detail the potential for identifying a list of hexameric TISs spanning the -6 to -1 position that can be used to predictably tune protein expression in a context-independent manner (i.e. promoter or gene of interest (GOI) proximal sequences), we initially constructed three TIS libraries, each containing a yeGFP expression cassette controlled by either a weak (REV1), medium (RPL18B) or strong (TEF1) constitutive

promoter, which span approximately three orders of magnitude in expression level (53) and have <54% similarity in the -50 to -7 position (Supplementary Figure S1). Between promoter and gene, we cloned randomized hexameric TISs and genomically integrated the reporter expression cassettes (Figure 2A). For each of the three libraries, we also constructed a corresponding control strain by substituting the randomized nucleotides with -6 to -1 positioned nucleotides, AAAACA, from the strong PGK1 promoter (54) (Figure 2A). From the 4096 possible hexamer variants, DNA sequencing revealed library coverages of 42% (1721), 53% (2174) and 21% (844) from the REV1, RPL18B and TEF1 TIS libraries, respectively (Supplementary Figure S2). Both the per base position and overall frequency of nucleotides A, C, G and T were ∼20%, 10%, 45% and 25%, respectively (Supplementary Figure S3). Flow cytometry analysis of the three TIS libraries revealed variances in yeGFP fluorescence of ∼2, >3 and >4 orders of magnitude for the REV1, RPL18B and TEF1 promoters, respectively (Figure 2B–D). Also, none of the three libraries included TIS variants that exceeded the fluorescence measurements observed in cells expressing the PGK1 TIS AAAACA (Figure 2B–D, red).

In order to identify, and further characterize, TISs showing a high degree of protein expression tunability, we sorted the TIS library in the context of the TEF1 promoter, which showed the highest detectable variance of the three libraries. Briefly, we selected 480 single cells based on gating (Supplementary Figure S4), thereby covering a large fraction of the yeGFP expression range (Figures 1 and 3A). Following single clonal validation of fluorescence, we selected eight strains uniformly covering the maximum 10-fold range in yeGFP expression observed in our TIS library and determined the TISs by sequencing (Figure 3A). Importantly, as the chromophore formation is an O_2-dependent autocatalytic process (55), it is critical to consider cell density of the small cultivation volumes (150 μl) used when comparing fluorescence intensities of cell populations expressing individual TIS variants. Accordingly, as we observed a rapid decrease in maximum per cell fluorescence with increasing cell densities, we analyzed yeGFP expression at OD_{600} = 0.1–0.2 as also reported from studies in bacteria (Supplementary Figure S5) (56).

Modular TISs show context-independent tuning of protein expression in yeast

In eukaryotes, earlier studies of TIS libraries, spanning larger sequence spaces (e.g positions -50 to -1 or -6 to +5), have focused on building computational models to enable forward engineering of protein expression levels (7–10). Though these efforts have enabled high-throughput enumeration of sequence parameters of importance for predictive tuning of protein expression, the predicted protein expression levels from placing TISs in new genomic contexts (i.e. promoter or gene of interest) have so far not been able to explain >30% of the variation observed between experimentally deduced behavior compared to the genomic context in which the TIS algorithms were originally designed (7,10).

To investigate if shorter TIS variants selected from our FACS analysis (Figure 3A) would have context-dependent effects in protein expression, we placed the TIS sequences in the context of a different promoter, different genes of interest (GOI) and in another cultivation medium, and used standard flow cytometry to analyze protein expression from a total of 32 genomic designs (Figure 3B and Supplementary Figure S9). Specifically, in addition to the strong constitutive TEF1 promoter, we also tested the eight selected TISs in the context of the glucose-repressed ADH2 promoter, and for GOI we included two other fluorescent reporters: ymUKG1 and mKate2 (44). The selection of GOIs and promoter contexts was based upon maximal sequence diversity and carbon source dependent expression, respectively (57) (Supplementary Figure S1). Next, from the 32 designs we experimentally validated, fluorescence measurements revealed from 2- to 10-fold variation, with glucose and TEF1 contexts displaying the largest fold changes between the weakest (TIS 1, TGATAT) and the strongest (TIS 8, TCGGTC) TISs (Figure 3C–E). The five strongest TISs all had a purine in position -3, which is in line with earlier reports (7,8,10,52). Furthermore, the TIS sequence dictates the fluorescence in a similar manner across all tested genomic and environmental conditions, as evidenced by the linear correlations between mean fluorescence values between individual promoter or reporter contexts (Pearson's, R^2 = 0.75–0.98) (Figure 4). This range overlaps with previous studies based on larger number of TISs (8).

Benchmarking measured TIS efficiencies

When comparing the measured relative expression values of all 32 combinatorial designs with existing computational algorithms for predicting translation initiation (8,10), we observe substantial correlation coefficients (R^2 = 0.44–0.86) between measured and predicted values (Supplementary Figures S6 and S7), with the model inferred by Noderer *et al.* generally displaying stronger correlations compared to the model generated by Decoene *et al.* However, as reported in these previous modeling studies, translation efficiency for some TISs (in our case TIS no. 4) would require additional experimental validation (incl. context) to further refine predictive tuning of protein expression (8).

Complementary to these observations, we performed a genome-wide search for the occurence of the eight different TISs and compared the result with the translation efficiency of the native gene products as reported by Lathvee *et al*. (58). Briefly, this analysis identified TISs 2, 3, 5, 6 and 7 in the -6 to -1 position in a total of 11 genes of which four genes with TISs 5, 6 and 7 had translation efficiencies reported (58). From this small number of hits, the translation efficiencies of the two genes with TIS 6 were higher than the efficiency reported for the gene with TIS 5, whereas the gene with TIS 7 had the lowest translation efficiency reported among the four genes (Supplementary Table S6).

Moreover, one critical parameter known to influence translation initiation efficiency is the folding propensity of 5′-UTRs (17,20,30). To further benchmark translation initiation efficiencies of TIS 1–8, we calculated the minimum free energy (kcal/mol) for bases at positions -15 to +50 by RNAfold as a function of normalized mean fluorescence

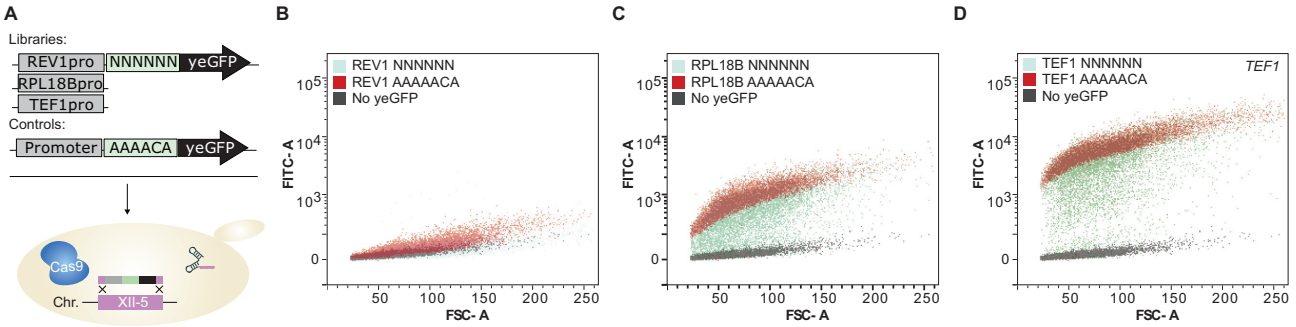


Figure 2. Distribution of reporter gene activities of three TIS libraries. (**A**) Schematic outline of the TIS library designs in the context of three different promoters (REV1, REPL18B and TEF1) controlling the expression of yeGFP. Negative control strain was without yeGFP expression. Fluorescence outputs for all three libraries were compared to yeGFP expression under the control of the TIS AAAACA from the strong PGK1 promoter. Libraries and control designs were integrated into yeast chromosome XII, EasyClone site 5 by CRISPR-mediated double-strand breaking and homologous recombination. (**B**) Fluorescence (FITC-A) as a function of forward scatter (FSC-A) of the TIS library in the context of REV1 promoter. (**C**) Fluorescence (FITC-A) as a function of forward scatter (FSC-A) of the TIS library in the context of RPL18b promoter. (**D**) Fluorescence (FITC-A) as a function of forward scatter (FSC-A) of the TIS library in the context of TEF1 promoter. Scatter plots in (B–D) are displayed together with control populations having TIS AAAACA (red) and wild-type cells without yeGFP expressed (gray).

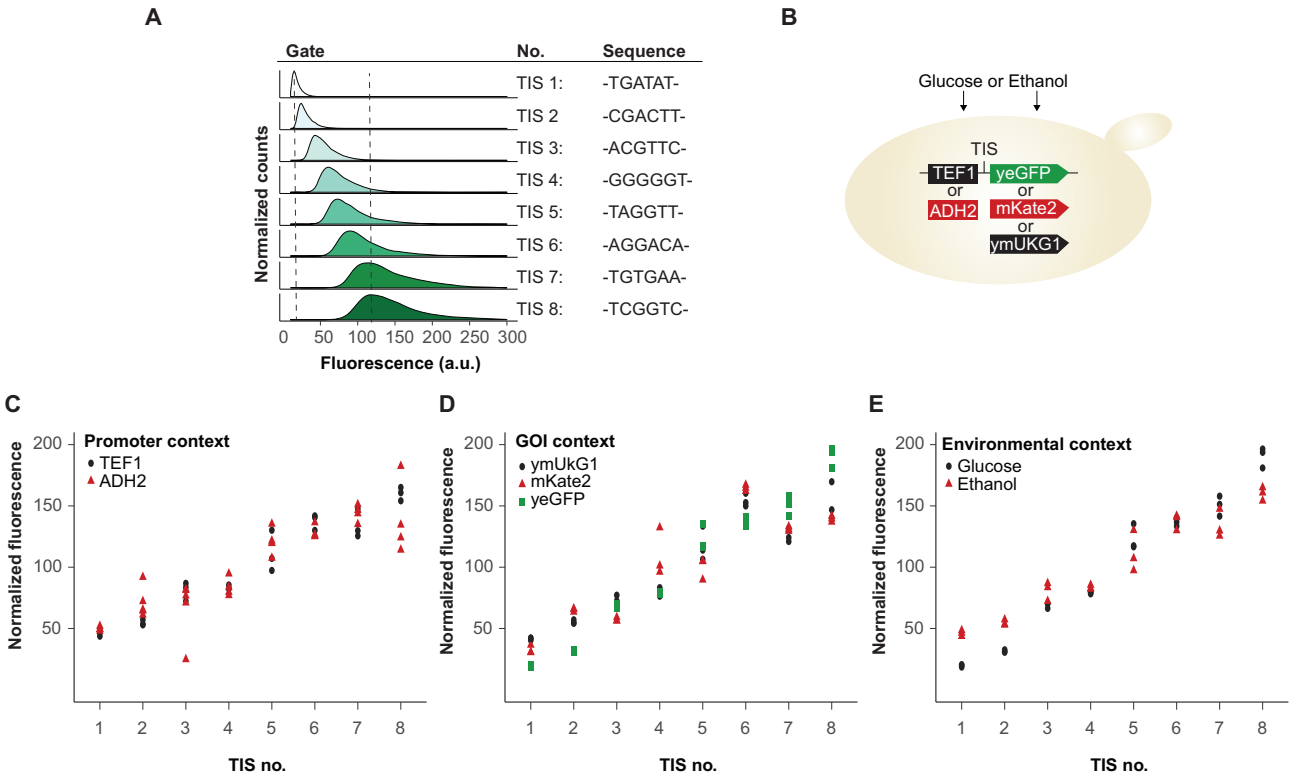


Figure 3. Investigation of interaction effects between TIS sequences and promoter, gene of interest (GOI) or growth culture condition. (**A**) Fluorescence histograms of gated populations of single clonal variants from the TIS library in the context of the TEF1 promoter. To the right, selected TISs from each of the populations are represented. (**B**) Schematic outline of the experimental design used to investigate interaction effects. (**C**) Normalized median fluorescence measured for the eight different selected TIS sequences in the context of a constitutive promoter (TEF1) and a glyconeogenic promoter (ADH2). (**D**) Normalized median fluorescence measured for the eight different selected TIS sequences in the context of three different GOI; ymUkG1, mKate2 and yeGFP. (**E**) Normalized median fluorescence measured for the eight different selected TIS sequences in the context of two different carbon sources present in the growth medium; glucose or ethanol. In plots (D–E), the TIS sequence order is selected from the TIS library controlling yeGFP expression under the control of the TEF1 promoter. In plot (C–E), median fluorescence values are shown for at least three biological replicates each based on at least 5000 single cell measurements.

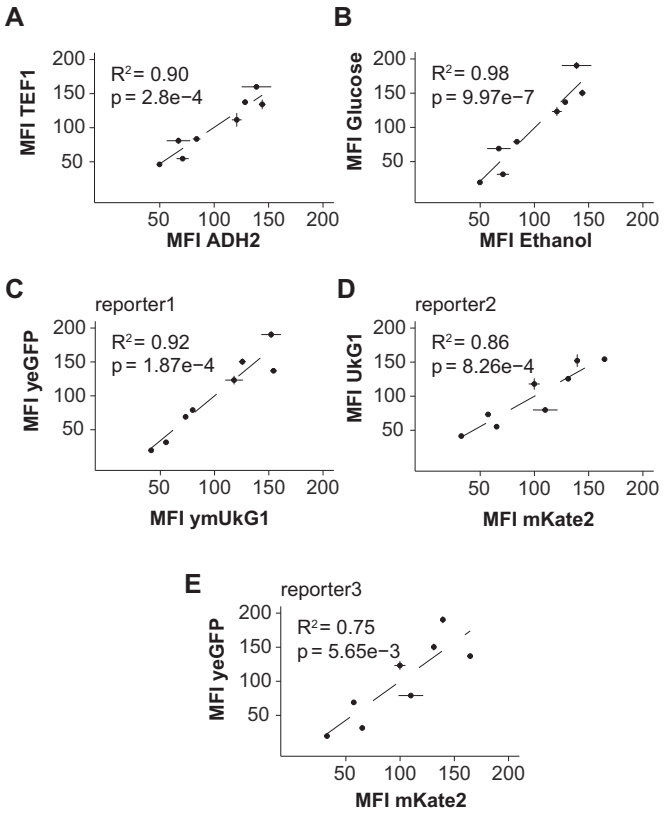


Figure 4. Linear correlations between normalized fluorescence measurements for eight TIS sequences in diverse genomic and environmental contexts. (**A**) Correlation between yeGFP fluorescence measurements for the two tested promoters (TEF1 and ADH2). (**B**) Correlation between yeGFP fluorescence measurements when using the two different carbon sources, glucose or ethanol. (**C–E**) Correlation between fluorescence measurements between each pair of the three different fluorescent reporters tested, ymUkG1, mKate2 and yeGFP. The fluorescence values are means of normalized median fluorescence values from at least three biological replicates each based on at least 5000 single cell measurements.

values for all 32 strain designs in this study. From this analysis, we observed no significant positive correlation ($R^2 = 3.0 \times 10^{-3}$, $P = 0.77$) (Supplementary Figure S8), which could indicate that the effect of varying the relative small hexameric TISs reported in this study, only have a modest effect on the minimum free energy observed for the sequence space analyzed (-15 to +50) (21).

Taken together, benchmarking the eight TISs with existing translation initiation prediction tools (8,10,21) and experimentally measured translation efficiencies (58) reveals that the model output overall correlated with our measured TIS efficiencies. However, when investigating the correlation between genome-wide occurrences of the identified TISs and their translation efficiency, the numbers are too low to infer statistical significance. Finally, we observe no significant correlation between the TIS strength of the 32 different designs studied and the folding propensity of their -15 to +50 regions (21), indicating that the hexameric TISs only modestly affect folding propensity of the 5′-UTRs.

TISs show context-independent tuning of protein expression in mammalian cells

In mammalian cells, the TIS sequence RYMRMVAUGGC (Y = U or C, M = A or C, R = A or G and V = A, C or G, start codon underscored) has been reported as a high-efficiency TIS, with positions -4, -3, -2, +4 and +5 as the most critical for efficient translation initiation (8,33). This consensus dictates the use of the CGx anticodon of ala-tRNA following incorporation of the AUG start codon for efficient translation initiation. Our best sequence TCGGTC (motif YYRRYVAUG–) is not fully in accordance with the earlier reported high-efficiency TIS motif RYMRMVAUGGC, as it deviates at position no. -6, -4 and -2. Moreover, neither does our TIS toolkit take into consideration the use of specific codons following the AUG start codon. Still, to further investigate if yeast-derived TIS variants spanning only positions -6 to -1 could also tune context-independent protein expression in mammalian cells, we decided to engineer CHO cells, the biotechnology workhorse for recombinant therapeutic protein production (59). Here, we constructed six CHO cell pools containing three different TISs derived from our FACS-based selection (Figures 1 and 5A) in combination with either meGFP or ZsGreen1, selected for their low sequence similarity, and optimized fluorescence intensity for CHO cells (Supplementary Figure S1) (48). Additionally, we created a cell pool containing the mammalian consensus TIS GCCACC (32) in combination with meGFP (Figure 5A and B; Supplementary Figure S10). First, testing meGFP expression in both yeast and CHO cells showed that the TIS strength was maintained between the two chassis and revealed fluorescence measurements with almost 10-fold variation between weakest and strongest TISs (Figure 5B). Importantly, considering the inherent efficiency of CHO cells for protein production, we observed that meGFP expression in combination with TIS TCGGTC was almost 50% stronger than the mammalian consensus TIS GCCACC (TIS no. 11) (Figure 5B). Moreover, just as was observed in yeast, TISs dictated the fluorescence of each of the reporter genes tested (Figure 5C). Finally, we observed strong linear correlations between fluorescence outputs from yeast versus CHO cells ($R^2 = 0.98$) as well as for mean meGFP versus ZsGreen fluorescence values across biological duplicates ($R^2 = 0.91$) (Figure 5D–E).

Tuning metabolic fluxes using modular TISs

The short length and context-independence of TISs make them particularly useful for engineering regulatory branch points of cellular metabolism, similar to earlier reports from MAGE-derived replacement of SD sequences in bacteria (60). To demonstrate an application of simple tuning of a metabolic branch point by the use of hexameric TIS variants, we aimed to tune metabolic fluxes through the native mevalonate biosynthesis pathway, and the *Xanthophyllomyces dendrorhous* 4-step β-carotene pathway (Figure 6A) (46). Specifically, this included targeting the genes encoding squalene synthase (*ERG9*) and the heterologous geranylgeranyl diphosphate (GGPP) synthase (*crtE*) at the farnesyl diphosphate (FPP) branch point (Figure 6A), both have

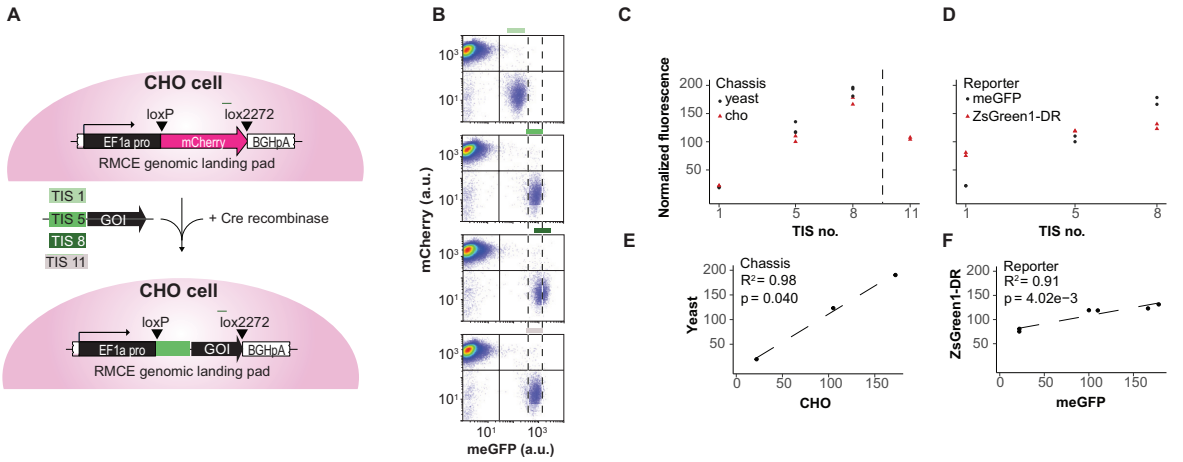


Figure 5. Comparison of interaction effects between TIS sequences and eukaryote chassis. (**A**) Schematic outline of the RMCE methodology used for introducing GOI with varying TISs into the genomic landing pad. In this study we compared hexameric TIS no. 1 (TGATAT), 5 (TAGGTT) and 8 (TCGGTC), with the mammalian consensus TIS no. 11 (GCCACC). (**B**) Scatter plots of CHO singlets with genomically integrated reporter meGFP in the context of TIS no. 1, 5, 8 or 11. Dashed lines indicate the distribution of the gated population from meGFP expressed in the context of TIS no. 11. (**C**) Comparison of yeast and CHO cells expressing meGFP under the control of varying TISs. (**D**) Comparison of two reporters (meGFP and ZsGreen1) expressed under the control of varying TISs. (**E**) Correlation between fluorescence measurements for three TIS sequences in yeast and CHO. (**F**) Correlation between fluorescence measurements for two reporters (meGFP and ZsGreen1) in CHO cells. In (E), the data shown are means of the normalized median fluorescence for each of the three TIS sequence with paired measurements from (C). In (C–F), the fluorescence values are means of normalized median fluorescence values from at least two biological replicates each based on at least 5000 single cell measurements.

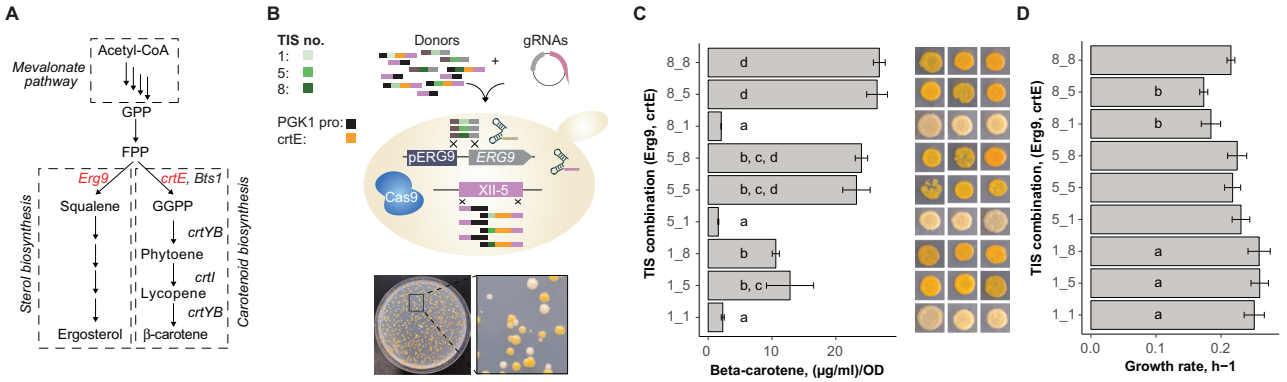


Figure 6. Effect of balancing Erg9 and crtE protein expression on carotenoid production. (**A**) Map of the carotenoid pathway and its connection to the native metabolism in yeast. The dashed lined boxes indicate sectioned native and heterologous metabolic pathways, with the branch point genes, *Erg9* and *crtE*, colored in red. (**B**) Schematic outline of the TISs tested and the genome engineering approach (top), and the resulting phenotypic landscape observed from multiplex TIS targeting of crtE (on Chr. XII-5) and *Erg9* loci on the yeast genome (bottom). (**C** and **D**) β-Carotene quantification and phenotype of carotenoid strains with variable duplex TISs controlling the expression of Erg9 and CrtE. Mean β-carotene content and maximum specific growth rate are shown (with standard errors, $n = 6$). Mean values with different lettering are significantly different according to pairwise Tukey HSD test ($P < 0.05$). Phenotypes are depicting three biological replicates.

earlier been reported to impact isoprenoid production in yeast (43,61,62).

Here, starting from a baseline strain with Cas9 and the 4-step β-carotene pathway genes integrated, we performed a one-pot transformation of a double guide RNA (gRNA) construct and repair templates introducing TISs (no. 1, 5 and 8) controlling the expression of ERG9 at the native site, as well as the *crtE* under the control of the PGK1 promoter genomically integrated at EasyClone site XII-5 (Figure 6B).

Following library transformation, colonies stably displayed diverse carotenoid-associated orange coloring and colony sizes (Figure 6B). Having observed the wide phenotypic distribution offered from combinatorial perturbation of TISs, we next re-constructed nine defined designs by duplex integration of TISs no. 1, 5 and 8 to control Erg9 and CrtE protein expression in all combinations in our background strain (Figure 6B) (46), and then quantified β-carotene levels as well as measured growth rates for all designs. From this analysis we observed up to 16-fold differences in β-carotene levels (Figure 6C), as well as up to 50% differences in growth rate (Figure 6D, and Supplementary Figure S11, $R^2 = 0.79$–0.99 with a mean $R^2 = 0.97$). Though no linear effect between fitness and production was observed, it is evident that stronger TISs (TISs 5 or 8) are needed to drive the expression of crtE in order to direct flux toward carotenoid production (Figure 6C). Interestingly, the strains with TIS 1 controlling ERG9 showed the highest growth rates. This finding is surprising in light of Erg9p

being an essential enzymatic step for conversion of FPP to squalene. However, acknowledging the intricate transcriptional and product-inhibited regulation of Erg9, low translation initiation efficiency of Erg9 could relieve ergosterol feedback inhibition and lead to upregulation of transcription (63,64). Alternatively, we could imagine the accumulation of toxic intermediates causing a reduction in growth rates for some of the strains with TISs 5 and 8 engineered to control Erg9 translation initiation (65).

Taken together, this example corroborates the simple design, rapid construction and testing of intricately regulated production and fitness landscapes offered from library transformations of hexameric TIS variants.

DISCUSSION

In this study, we have characterized and classified hexameric TISs according to their impact on protein expression in yeast and mammalian cells. Starting from three TIS libraries collectively covering 4739 TIS variants in yeast, we identified TISs that can tune protein expression up to 10-fold irrespective of the diverse genomic (38–54% similarity of the -54 to +13 positions, Supplementary Figure S1) and environmental (cell density or growth medium) conditions. Importantly, in terms of applicability, we showed that TIS TCGGTC was stronger than the mammalian consensus TIS GCCACC frequently used for protein production in CHO cells, and that a multiplex transformation of TIS variants targeting an essential metabolic branch point could be used to probe the production and fitness landscape of yeast cell factory designs. Though the combined use of large *de novo* synthesized TIS libraries and FACS screens to deduce sequence to function relationships has recently been reported in both bacteria and eukaryotes (6–9,66), the sequence space (positions -6 to -1) covered in this study is to our knowledge the smallest space systematically studied in broad genomic and environmental contexts, yet the dynamic range covered is similar to variants selected from larger TIS sequence spaces (7). Moreover, as the TISs characterized in this study only cover positions upstream the AUG start codon, protein expression of any ORF should technically be possible by a simple hexameric 5′-end primer extension using said ORF as a template. As such, both scalability and cost-effectiveness in both design and construction of engineered cells are ensured.

In the further positioning of our findings in relation to earlier studies, we find the five strongest TISs identified in our study have a purine at position -3 (Figure 3), consistent with earlier studies (33,34). Also, the degree of tunability observed in this study is similar to the ∼7-fold changes in protein expression observed from studies characterizing larger 5′-UTR sequence space (e.g positions -50 to -1 or -6 to +5) (7,8), underscoring the potential to use hexameric TISs for efficient protein expression tuning. Interestingly, among the five different fluorescent reporter genes tested in this study, the ones displaying the largest tunability in the context of varying TISs are the yeGFP and mammalian GFP (Figures 3 and 5; Supplementary Figure S1). The ORFs of these two genes are the only ones not having a guanine at position +4, otherwise reported to be important for efficient translation initiation (8,33–34), suggesting that the TISs identified in this study could be recalcitrant to ORF sequence diversity at this exact position.

More generally speaking, one immediate observation from studies of TISs in eukaryotes is that even though small TIS sequence spaces can robustly tune protein expression in a predictable manner, the degree of tuning is several orders of magnitude lower than the tuning offered by TIS variants in bacteria (6,67). This is largely due to more intricate regulatory mechanisms associated with translation initiation in eukaryotes compared to bacteria, including ribosome scanning mode-of-action, longer 5′-UTRs, 5′-end capping of mRNA, assembly of eukaryotic initiation factors, internal ribosome entry sites, uAUGs and uORFs observed in eukaryotes (10). Yet, engineering excessively short 5′-UTR (≤20 nt) may not provide TISs with higher tunability, as genome-wide mapping of yeast 5′-UTRs with such short 5′-UTRs has been observed to be detrimental to translation initiation control and exhibit below-average translational efficiency (68), and hence is not considered a viable route to dereplicate the impact combinations of native 5′-UTR elements would have on translation initiation.

Furthermore, looking ahead, it is important to consider system-level limitations of protein expression (69), and continue to improve current and new models for predicting TIS strengths in broad genomic contexts and, cellular and environmental conditions (5–10). Also, from the range of fluorescent outputs observed in our TIS libraries in three promoter contexts (Figure 2), it is evident that the native transcriptional regulation, conferred by promoter usage, controls the absolute quantitative impact TISs will have on protein expression, as was recently reported from genome-wide studies in yeast (58). As such, we envision that in order to engineer synthetic translation initiation elements with higher dynamic output ranges based on the existing features of the translation machinery, a more detailed understanding of both *cis* and *trans* initiation mechanisms is expected to enhance our ability to predictably control larger spans of protein expression levels.

Finally, when quantifying changes in protein expression within an order of magnitude as observed from varying short TISs, mitigating experimental noise and conforming to standardized experimental procedures become essential for deducting sequence–function relationships (7,70). With the ongoing development of advanced genome engineering technologies, especially in relation to <100-bp edits (71,72), and the drop in DNA synthesis costs, we expect that TISs will be particularly useful baits for multiplex targeting, tuning and optimization of protein expression levels in robust genomic contexts, thereby expectedly improving signal-to-noise ratios, and ultimately enabling predictable and rational tuning of genetic circuits and cellular behavior.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Saranya Nallapareddy for help with CHO transfections, Nachon Charanyanonda Petersen for help with flow cytometry analyses, and Larissa

Tramontin and Kanchana Kildegaard for help with HPLC. Also, a warm thanks to colleagues at the Novo Nordisk Foundation Center for Biosustainability for fruitful discussions and comments.

FUNDING

Novo Nordisk Foundation; European Commission Horizon 2020 programme (PAcMEN, No. 722287). Funding for open access charge: Novo Nordisk Foundation and European Commission HZ2020 Programme.
Conflict of interest statement. J.D.K. has a financial interest in Amyris, Lygos, Demetrix, Constructive Biology, Maple Bio and Napigen.

REFERENCES

1. Sonenberg,N. and Hinnebusch,A.G. (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, **136**, 731–745.
2. Jackson,R.J., Hellen,C.U.T. and Pestova,T.V. (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.*, **11**, 113–127.
3. Allen,G.S., Zavialov,A., Gursky,R., Ehrenberg,M. and Frank,J. (2005) The cryo-EM structure of a translation initiation complex from Escherichia coli. *Cell*, **121**, 703–712.
4. Seo,S.W., Yang,J.-S., Kim,I., Yang,J., Min,B.E., Kim,S. and Jung,G.Y. (2013) Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency. *Metab. Eng.*, **15**, 67–74.
5. Salis,H.M. (2011) The ribosome binding site calculator. *Methods Enzymol.*, **498**, 19–42.
6. Bonde,M.T., Pedersen,M., Klausen,M.S., Jensen,S.I., Wulff,T., Harrison,S., Nielsen,A.T., Herrgård,M.J. and Sommer,M.O.A. (2016) Predictable tuning of protein expression in bacteria. *Nat. Methods*, **13**, 233–236.
7. Dvir,S., Velten,L., Sharon,E., Zeevi,D., Carey,L.B., Weinberger,A. and Segal,E. (2013) Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, E2792–E2801.
8. Noderer,W.L., Flockhart,R.J., Bhaduri,A., Diaz de Arce,A.J., Zhang,J., Khavari,P.A. and Wang,C.L. (2014) Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol. Syst. Biol.*, **10**, 748.
9. Cuperus,J.T., Groves,B., Kuchina,A., Rosenberg,A.B., Jojic,N., Fields,S. and Seelig,G. (2017) Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res.*, **27**, 2015–2024.
10. Decoene,T., Peters,G., De Maeseneire,S.L. and De Mey,M. (2018) Toward predictable 5′UTRs in saccharomyces cerevisiae: development of a yUTR calculator. *ACS Synth. Biol.*, **7**, 622–634.
11. Ben-Yehezkel,T., Atar,S., Zur,H., Diament,A., Goz,E., Marx,T., Cohen,R., Dana,A., Feldman,A., Shapiro,E. *et al.* (2015) Rationally designed, heterologous S. cerevisiae transcripts expose novel expression determinants. *RNA Biol.*, **12**, 972–984.
12. Shine,J. and Dalgarno,L. (1974) The 3′-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci. U.S.A.*, **71**, 1342–1346.
13. Ludwig,P., Huber,M., Lehr,M., Wegener,M., Zerulla,K., Lange,C. and Soppa,J. (2018) Non-canonical Escherichia coli transcripts lacking a Shine-Dalgarno motif have very different translational efficiencies and do not form a coherent group. *Microbiology*, **164**, 646–658.
14. Leppek,K., Das,R. and Barna,M. (2017) Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat. Rev. Mol. Cell Biol.*, **19**, 158–174.
15. Grillo,G., Turi,A., Licciulli,F., Mignone,F., Liuni,S., Banfi,S., Gennarino,V.A., Horner,D.S., Pavesi,G., Picardi,E. *et al.* (2010) UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.*, **38**, D75–D80.
16. Tuller,T., Kupiec,M. and Ruppin,E. (2009) Co-evolutionary networks of genes and cellular processes across fungal species. *Genome Biol.*, **10**, R48.
17. Ringnér,M. and Krogh,M. (2005) Folding free energies of 5′-UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLoS Comput. Biol.*, **1**, e72.
18. Hinnebusch,A.G., Dever,T.E. and Asano,K. (2007) 9 Mechanism of translation initiation in the yeast saccharomyces cerevisiae. *Cold Spring Harbor Monogr. Arch.*, **48**, 225–268.
19. Zhang,Z. and Dietrich,F.S. (2005) Mapping of transcription start sites in Saccharomyces cerevisiae using 5′ SAGE. *Nucleic Acids Res.*, **33**, 2838–2851.
20. Kozak,M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, **361**, 13–37.
21. Robbins-Pianka,A., Rice,M.D. and Weir,M.P. (2010) The mRNA landscape at yeast translation initiation sites. *Bioinformatics*, **26**, 2651–2655.
22. Kochetov,A.V. (2005) AUG codons at the beginning of protein coding sequences are frequent in eukaryotic mRNAs with a suboptimal start codon context. *Bioinformatics*, **21**, 837–840.
23. Nakagawa,S., Niimura,Y., Gojobori,T., Tanaka,H. and Miura,K.-I. (2008) Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res.*, **36**, 861–871.
24. Tzani,I., Ivanov,I.P., Andreev,D.E., Dmitriev,R.I., Dean,K.A., Baranov,P.V., Atkins,J.F. and Loughran,G. (2016) Systematic analysis of the PTEN 5′ leader identifies a major AUU initiated proteoform. *Open Biol.*, **6**, 150203.
25. Ben-Yehezkel,T., Zur,H., Marx,T., Shapiro,E. and Tuller,T. (2013) Mapping the translation initiation landscape of an S. cerevisiae gene using fluorescent proteins. *Genomics*, **102**, 419–429.
26. Diaz de Arce,A.J., Noderer,W.L. and Wang,C.L. (2018) Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. *Nucleic Acids Res.*, **46**, 985–994.
27. Chew,G.-L., Pauli,A. and Schier,A.F. (2016) Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nat. Commun.*, **7**, 11663.
28. Tholen,M., Hillebrand,L.E., Tholen,S., Sedelmeier,O., Arnold,S.J. and Reinheckel,T. (2014) Out-of-frame start codons prevent translation of truncated nucleo-cytosolic cathepsin L in vivo. *Nat. Commun.*, **5**, 4931.
29. Zur,H. and Tuller,T. (2013) New universal rules of eukaryotic translation initiation fidelity. *PLoS Comput. Biol.*, **9**, e1003136.
30. Kozak,M. (1986) Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 2850–2854.
31. Kozak,M. (1987) At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J. Mol. Biol.*, **196**, 947–950.
32. Kozak,M. (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.*, **15**, 8125–8148.
33. Kozak,M. (1995) Adherence to the first-AUG rule when a second AUG codon follows closely upon the first. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 2662–2666.
34. Harte,R.A., Farrell,C.M., Loveland,J.E., Suner,M.-M., Wilming,L., Aken,B., Barrell,D., Frankish,A., Wallin,C., Searle,S. *et al.* (2012) Tracking and coordinating an international curation effort for the CCDS Project. *Database*, **2012**, bas008.
35. Pesole,G., Gissi,C., Grillo,G., Licciulli,F., Liuni,S. and Saccone,C. (2000) Analysis of oligonucleotide AUG start codon context in eukariotic mRNAs. *Gene*, **261**, 85–91.
36. Nour-Eldin,H.H., Hansen,B.G., Nørholm,M.H.H., Jensen,J.K. and Halkier,B.A. (2006) Advancing uracil-excision based cloning towards an ideal technique for cloning PCR fragments. *Nucleic Acids Res.*, **34**, e122.
37. Gietz,R.D. and Schiestl,R.H. (2007) Quick and easy yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.*, **2**, 35–37.
38. Jensen,N.B., Strucko,T., Kildegaard,K.R., David,F., Maury,J.J., Mortensen,U.H., Forster,J., Nielsen,J. and Borodina,I. (2014) EasyClone: Method for iterative chromosomal integration of multiple genes in Saccharomyces cerevisiae. *FEMS Yeast Res.*, **14**, 238–248.
39. Cormack,B.P., Bertram,G., Egerton,M., Gow,N.A., Falkow,S. and Brown,A.J. (1997) Yeast-enhanced green fluorescent protein

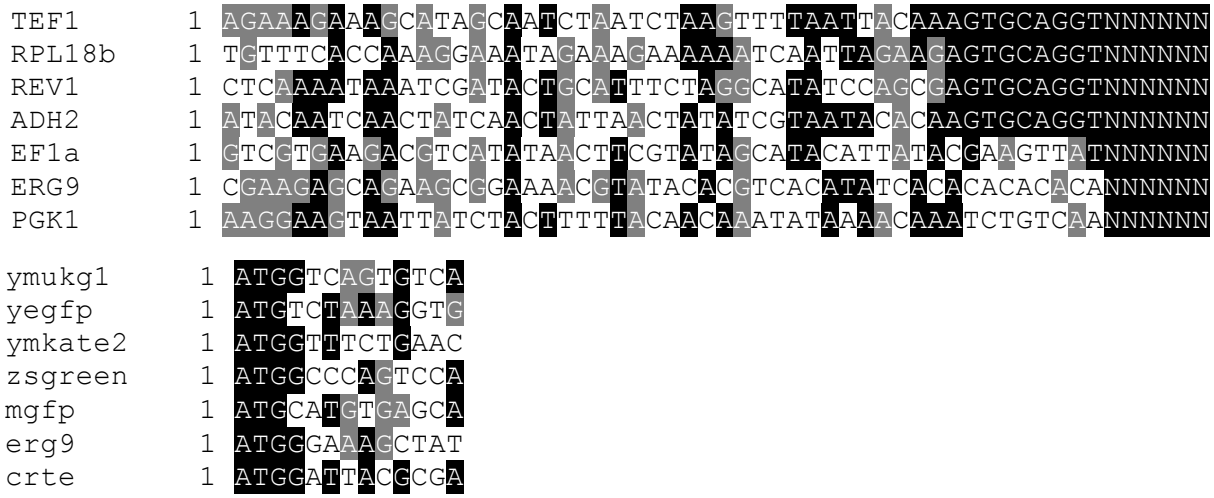
(yEGFP): a reporter of gene expression in Candida albicans. *Microbiology*, **143**, 303–311.
40. Bernard,P., Gabant,P., Bahassi,E.M. and Couturier,M. (1994) Positive-selection vectors using the F plasmid ccdB killer gene. *Gene*, **148**, 71–74.
41. Jessop-Fabre,M.M., Jakočiūnas,T., Stovicek,V., Dai,Z., Jensen,M.K., Keasling,J.D. and Borodina,I. (2016) EasyClone-MarkerFree: a vector toolkit for marker-less integration of genes into Saccharomyces cerevisiae via CRISPR-Cas9. *Biotechnol. J.*, **11**, 1110–1117.
42. Jakočiūnas,T., Rajkumar,A.S., Zhang,J., Arsovska,D., Rodriguez,A., Jendresen,C.B., Skjødt,M.L., Nielsen,A.T., Borodina,I., Jensen,M.K. *et al.* (2015) CasEMBLR: Cas9-Facilitated multiloci genomic integration of in vivo assembled DNA parts in Saccharomyces cerevisiae. *ACS Synth. Biol.*, **4**, 1226–1234.
43. Jakočiūnas,T., Bonde,I., Herrgård,M., Harrison,S.J., Kristensen,M., Pedersen,L.E., Jensen,M.K. and Keasling,J.D. (2015) Multiplex metabolic pathway engineering using CRISPR/Cas9 in Saccharomyces cerevisiae. *Metab. Eng.*, **28**, 213–222.
44. Kaishima,M., Ishii,J., Matsuno,T., Fukuda,N. and Kondo,A. (2016) Expression of varied GFPs in Saccharomyces cerevisiae: codon optimization yields stronger than expected expression and fluorescence intensity. *Sci. Rep.*, **6**, 35932.
45. Lee,S., Lim,W.A. and Thorn,K.S. (2013) Improved blue, green, and red fluorescent protein tagging vectors for S. cerevisiae. *PLoS One*, **8**, e67902.
46. Verwaal,R., Wang,J., Meijnen,J.P., Visser,H., Sandmann,G., Van Den Berg,J.A. and Van Ooyen,A.J.J. (2007) High-level production of beta-carotene in Saccharomyces cerevisiae by successive transformation with carotenogenic genes from Xanthophyllomyces dendrorhous. *Appl. Environ. Microbiol.*, **73**, 4342–4350.
47. Lee,J.S., Kallehauge,T.B., Pedersen,L.E. and Kildegaard,H.F. (2015) Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway. *Sci. Rep.*, **5**, 8572.
48. Grav,L.M., Lee,J.S., Gerling,S., Kallehauge,T.B., Hansen,A.H., Kol,S., Lee,G.M., Pedersen,L.E. and Kildegaard,H.F. (2015) One-step generation of triple knockout CHO cell lines using CRISPR/Cas9 and fluorescent enrichment. *Biotechnol. J.*, **10**, 1446–1456.
49. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, **17**, 10–12.
50. Petzoldt,T. (2017) growthrates: Estimate Growth Rates from Experimental Data. *R package version 0.7.1*. https://CRAN.R-project.org/package=growthrates.
51. Kildegaard,K.R., Adiego-Pérez,B., Doménech Belda,D., Khangura,J.K., Holkenbrink,C. and Borodina,I. (2017) Engineering of Yarrowia lipolytica for production of astaxanthin. *Synth. Syst. Biotechnol.*, **2**, 287–294.
52. Kozak,M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, **44**, 283–292.
53. Lee,M.E., Aswani,A., Han,A.S., Tomlin,C.J. and Dueber,J.E. (2013) Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay. *Nucleic Acids Res.*, **41**, 10668–10678.
54. Tuite,M.F., Dobson,M.J., Roberts,N.A., King,R.M., Burke,D.C., Kingsman,S.M. and Kingsman,A.J. (1982) Regulated high efficiency expression of human interferon-alpha in Saccharomyces cerevisiae. *EMBO J.*, **1**, 603–608.
55. Zhang,L., Patel,H.N., Lappe,J.W. and Wachter,R.M. (2006) Reaction progress of chromophore biogenesis in green fluorescent protein. *J. Am. Chem. Soc.*, **128**, 4766–4772.
56. Salis,H.M., Mirsky,E.A. and Voigt,C.A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.*, **27**, 946–950.
57. Reider Apel,A., d'Espaux,L., Wehrs,M., Sachs,D., Li,R.A., Tong,G.J., Garber,M., Nnadi,O., Zhuang,W., Hillson,N.J. *et al.* (2017) A Cas9-based toolkit to program gene expression in *Saccharomycescerevisiae*. *Nucleic Acids Res.*, **45**, 496–508.
58. Lahtvee,P.-J., Sánchez,B.J., Smialowska,A., Kasvandik,S., Elsemman,I.E., Gatto,F. and Nielsen,J. (2017) Absolute quantification of protein and mRNA abundances demonstrate variability in Gene-Specific translation efficiency in yeast. *Cell Syst.*, **4**, 495–504.
59. Kim,J.Y., Kim,Y.-G. and Lee,G.M. (2012) CHO cells in biotechnology for production of recombinant proteins: current state and further potential. *Appl. Microbiol. Biotechnol.*, **93**, 917–930.
60. Wang,H.H., Isaacs,F.J., Carr,P.A., Sun,Z.Z., Xu,G., Forest,C.R. and Church,G.M. (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, **460**, 894–898.
61. Ro,D.-K., Paradise,E.M., Ouellet,M., Fisher,K.J., Newman,K.L., Ndungu,J.M., Ho,K.A., Eachus,R.A., Ham,T.S., Kirby,J. *et al.* (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, **440**, 940–943.
62. Mitchell,L.A., Chuang,J., Agmon,N., Khunsriraksakul,C., Phillips,N.A., Cai,Y., Truong,D.M., Veerakumar,A., Wang,Y., Mayorga,M. *et al.* (2015) Versatile genetic assembly system (VEGAS) to assemble pathways for expression in S. cerevisiae. *Nucleic Acids Res.*, **43**, 6620–6630.
63. Asadollahi,M.A., Maury,J., Møller,K., Nielsen,K.F., Schalk,M., Clark,A. and Nielsen,J. (2008) Production of plant sesquiterpenes in Saccharomyces cerevisiae: effect of ERG9 repression on sesquiterpene biosynthesis. *Biotechnol. Bioeng.*, **99**, 666–677.
64. Smith,S.J., Crowley,J.H. and Parks,L.W. (1996) Transcriptional regulation by ergosterol in the yeast Saccharomyces cerevisiae. *Mol. Cell. Biol.*, **16**, 5427–5432.
65. Martin,V.J.J., Pitera,D.J., Withers,S.T., Newman,J.D. and Keasling,J.D. (2003) Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat. Biotechnol.*, **21**, 796–802.
66. Ben Yehezkel,T., Rival,A., Raz,O., Cohen,R., Marx,Z., Camara,M., Dubern,J.-F., Koch,B., Heeb,S., Krasnogor,N. *et al.* (2016) Synthesis and cell-free cloning of DNA libraries using programmable microfluidics. *Nucleic Acids Res.*, **44**, e35.
67. Espah Borujeni,A., Channarasappa,A.S. and Salis,H.M. (2014) 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.
68. Arribere,J.A. and Gilbert,W.V. (2013) Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res.*, **23**, 977–987.
69. Huang,M., Bao,J., Hallström,B.M., Petranovic,D. and Nielsen,J. (2017) Efficient protein production by yeast requires global tuning of metabolism. *Nat. Commun.*, **8**, 1131.
70. Canton,B., Labno,A. and Endy,D. (2008) Refinement and standardization of synthetic biological parts and devices. *Nat. Biotechnol.*, **26**, 787–793.
71. Garst,A.D., Bassalo,M.C., Pines,G., Lynch,S.A., Halweg-Edwards,A.L., Liu,R., Liang,L., Wang,Z., Zeitoun,R., Alexander,W.G. *et al.* (2016) Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nat. Biotechnol.*, **35**, 48–55.
72. Barbieri,E.M., Muir,P., Akhuetie-Oni,B.O., Yellman,C.M. and Isaacs,F.J. (2017) Precise editing at DNA replication forks enables multiplex genome engineering in eukaryotes. *Cell*, **171**, 1453–1467.

2.1

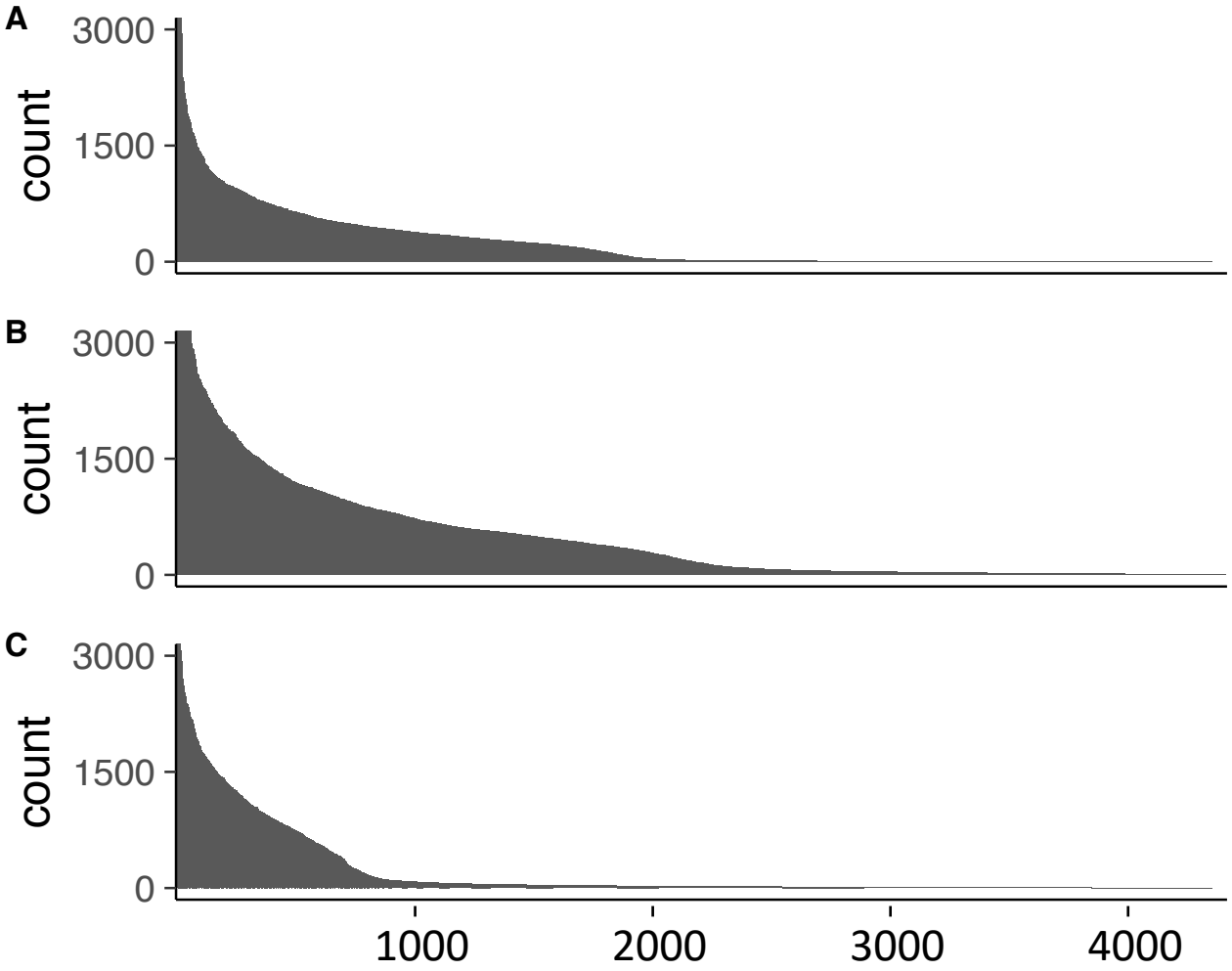
Supplementary material for

Modular 5'UTR hexamers for context-independent tuning of protein expression in eukaryotes

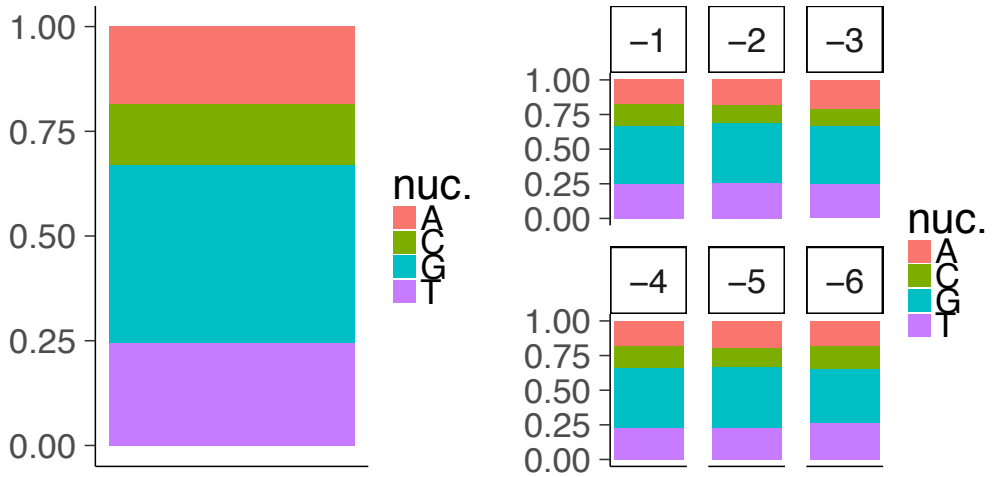
Søren D. Petersen, Jie Zhang, Jae S. Lee, Tadas Jakočiūnas, Lise M. Grav, Helene F. Kildegaard, Jay D. Keasling and Michael K. Jensen



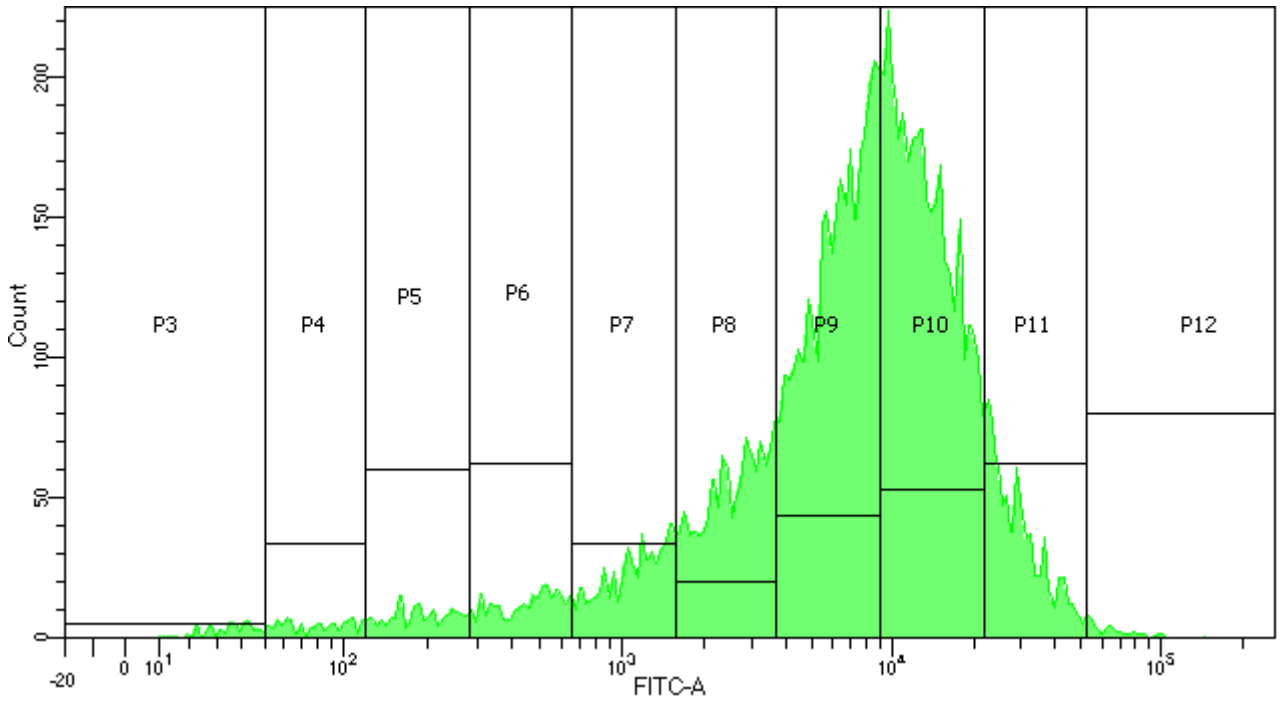
Supplementary Figure S1. Alignment of promoter (from position -56 to -1) and gene (from position +1 to 13) sequences used in study.



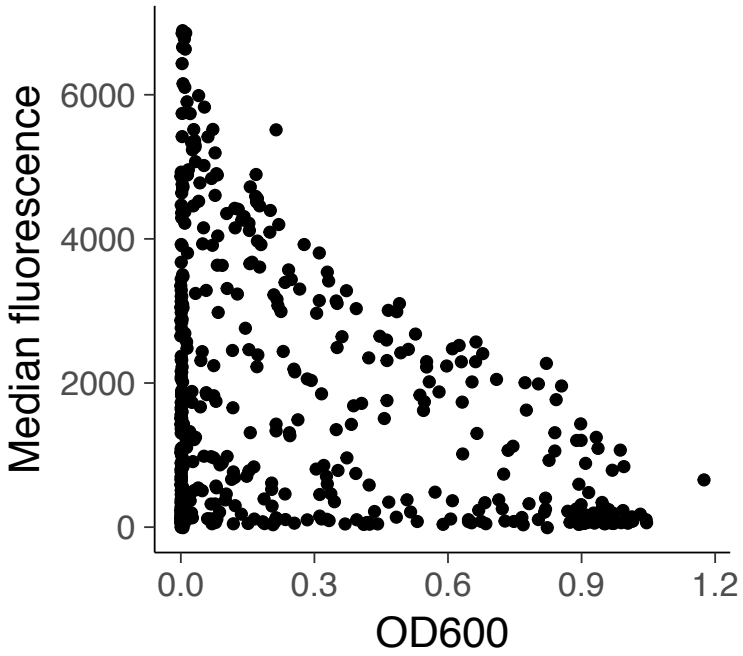
Supplementary Figure S2. Count of TIS sequence reads for promoter libraries A: REV1, B: RPL18b, and C: TEF1.



Supplementary Figure S3. Frequency of nucleotides overall and per position in the TIS sequence of the TEF1 library**.**



Supplementary Figure S4. 10 gates used for selection of single clonal variants from the TIS library in context of the TEF1 promoter



Supplementary Figure S5. Median fluorescence as a function of OD_{600} measured by SynergyMx microplate reader (BioTek) for 480 single cell cultures selected from the TEF1 promoter library.

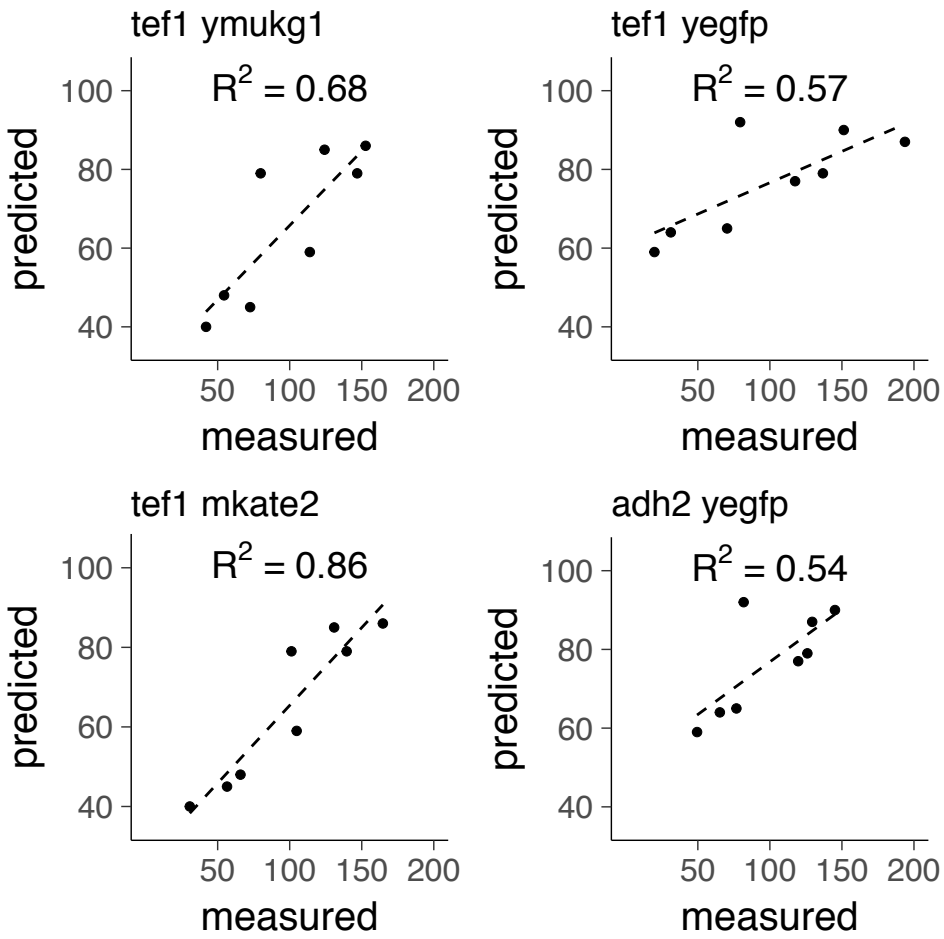


Figure S6. Linear correlations between normalized fluorescence measurements and translation efficiency predicted by the dinuclotide PWM published by Noderer et al. 2014 (1) in Supplementary Table S2 for eight TIS sequences in four context. The four context are TEF1 promoter and reporter ymUkG1, yeGFP, and mKate2 as well as ADH2 promoter and yeGFP reporter cultured with ethanol as carbon source. The fluorescence values are means of normalized median fluorescence values from at least three biological replicates.

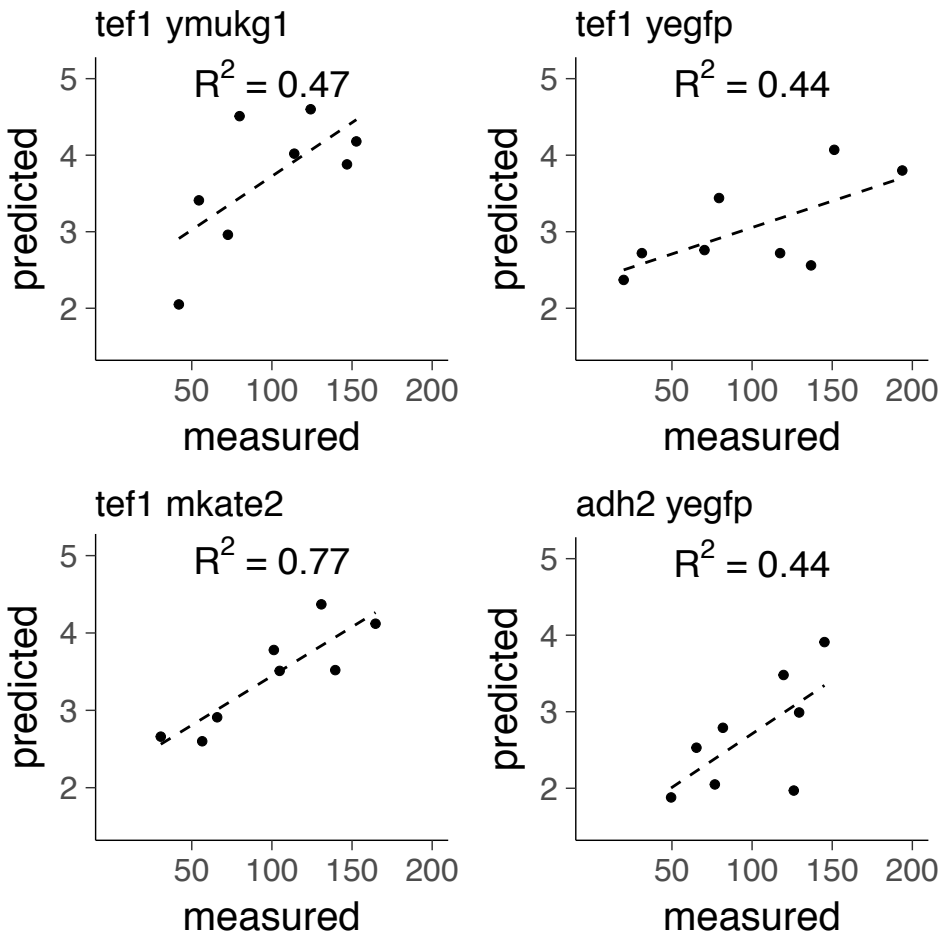


Figure S7. Linear correlations between normalized fluorescence measurements and translation efficiency predicted by the PLS regression model by Decoene et al. 2018 (2) using the yUTR-calculator_reverse_engineering script for eight TIS sequences in four context. The four context are TEF1 promoter and reporter ymUkG1, yeGFP, and mKate2 as well as ADH2 promoter and yeGFP reporter cultured with ethanol as carbon source. The fluorescence values are means of normalized median fluorescence values from at least three biological replicates.

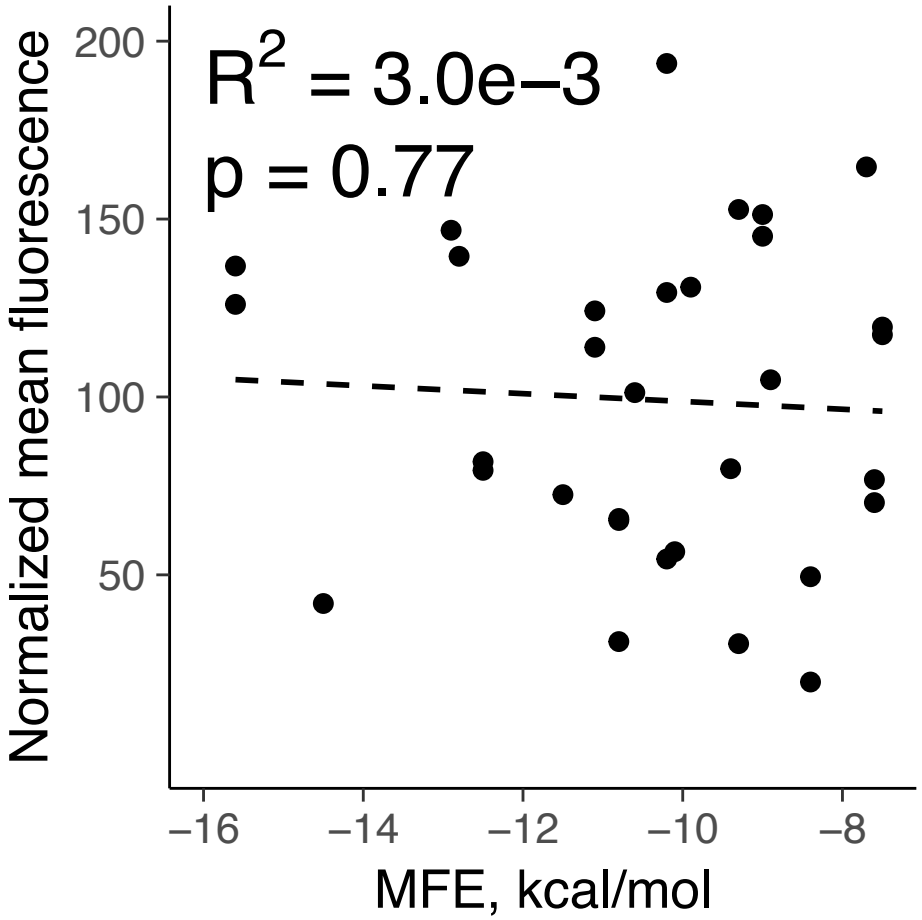
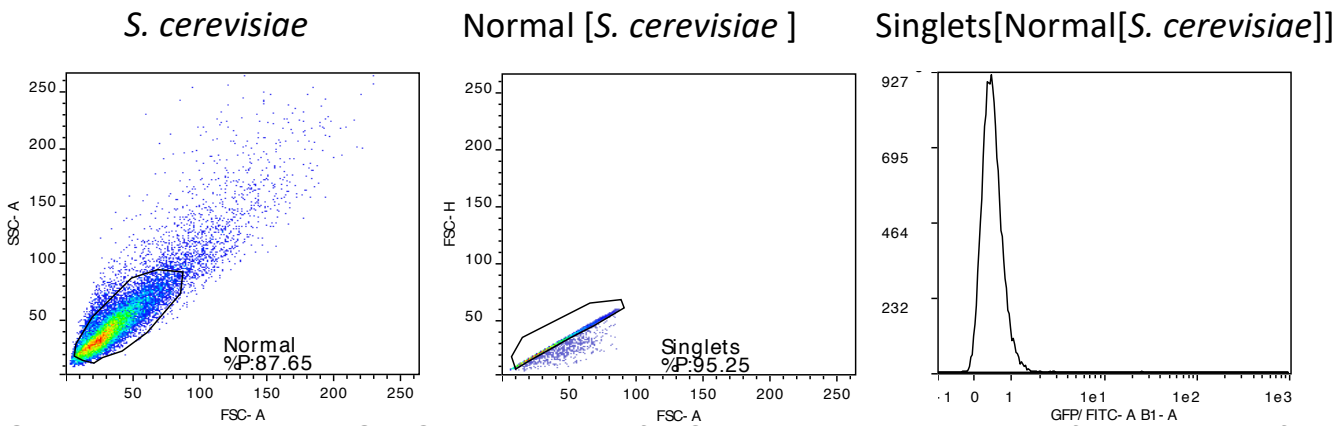
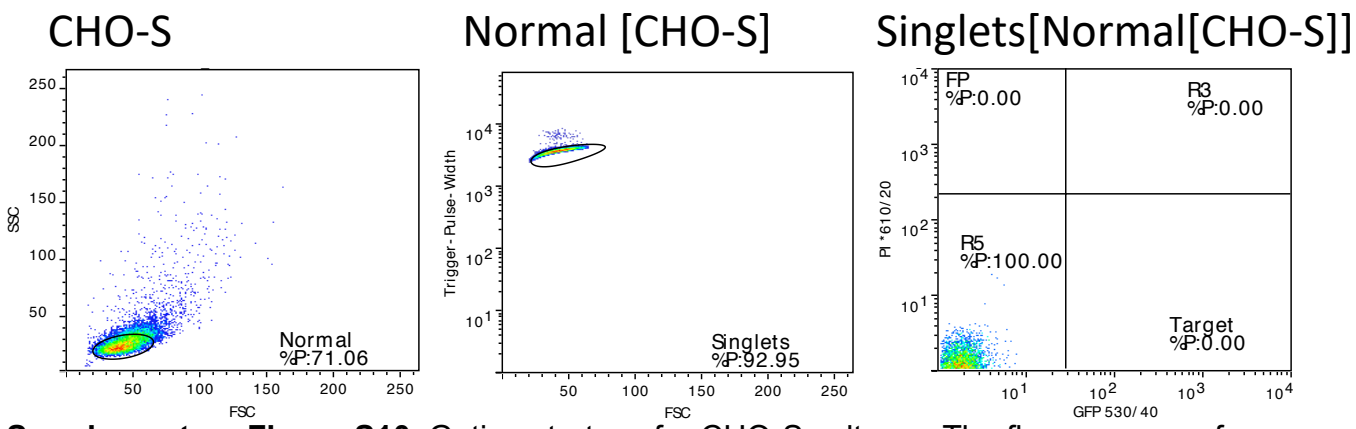


Figure S8. Minimum free energy (kcal/mol) as calculated for bases at positions -15 to +50 by RNAfold as a function of normalised mean fluorescence values for 32 yeast strains (3).



Supplementary Figure S9. Gating strategy for *S. cerevisiae cultures*. The fluorescence of yeast strains was determined as the median of the fluorescence signal in the singlets panel.



Supplementary Figure S10. Gating strategy for CHO-S cultures**.** The fluorescence of CHO cell lines and pools was determined as the median of the fluorescence signal in Target gate of the singlets panel**.**

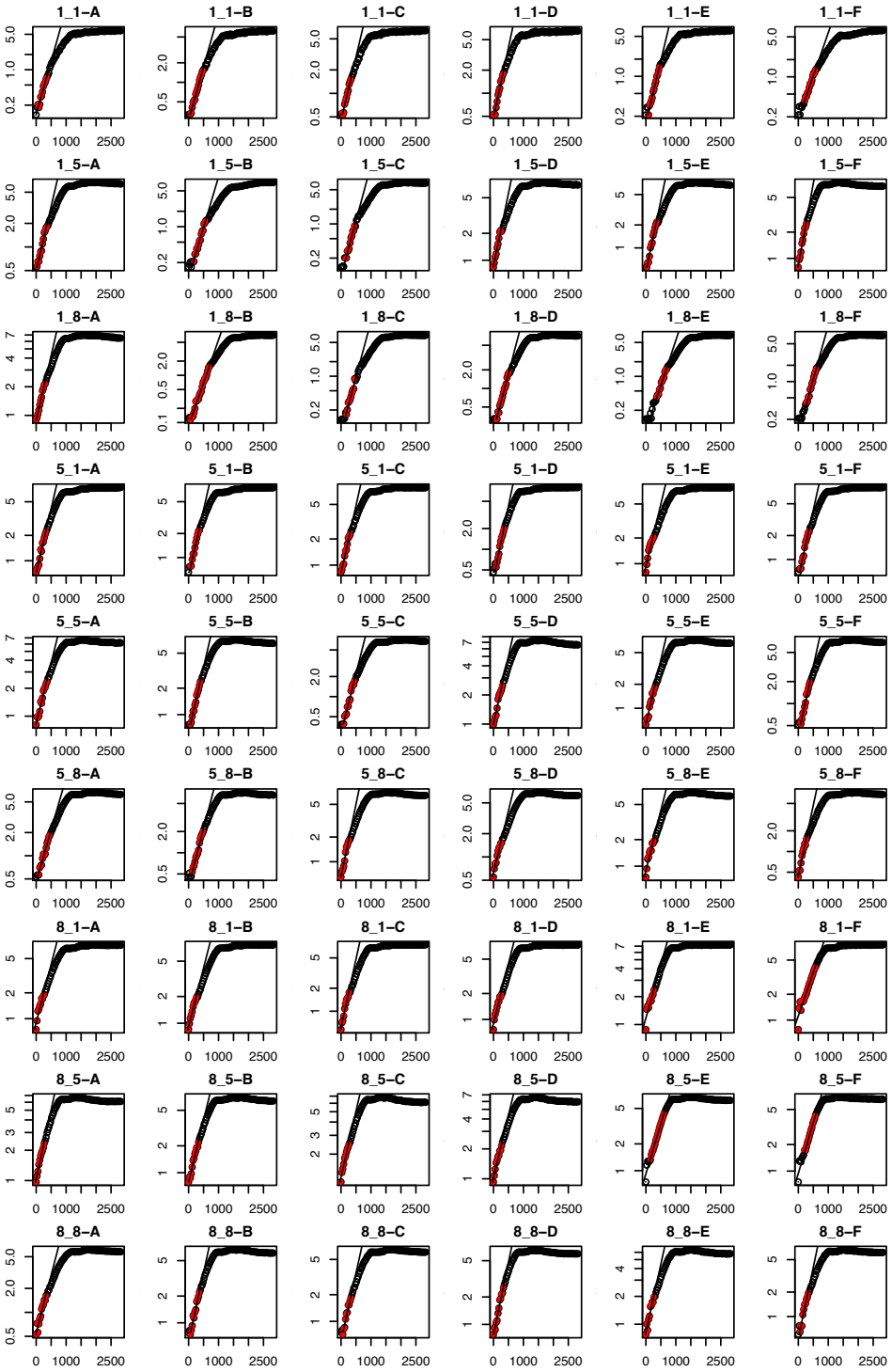


Figure S11. OD600 as a function of time (min) for 9 carotenoid strains in 6 biological replicates. OD600 is shown on a log scale. A linear model is fitted to 10 consecutive points that are highlighted in red. Each plot is described as follows: TIS no. erg9_TIS no. crtE_replicate.

Supplementary Table S1. Primers used in study. Features of interest are separated by a space.

Name	Sequence (5' - 3')	Use
TIS libraries		
pREV1_Fw	CACGCGAU GCTTTGAGTTGGGGTAG	Forward promoter REV1, overhang to CfB2336 backbone
pREV1_Rv	ACCTGCACU CGCTGGATATGCCTAG	Reverse promoter REV1, overhang to *yeGFP*
pRPL18B_Fw	CACGCGAU GGCGTCGTTGTTAATT	Forward promoter RPL18b, overhang to CfB2336 backbone
pRPL18B_Rv	ACCTGCACU CTTCTAATTGATTTTTCTTTCT	Reverse promoter RPL18b, overhang to *yeGFP*
pTEF1_fw	ACCTGCACU TTGTAATTAAAACTTAG	Forward promoter TEF1, overhang to CfB2336 backbone
pTEF1_rv	CACGCGAU GCACACACCATAGCTTC	Reverse promoter TEF1, overhang to *yeGFP*
NNNNNN_GFP_Fw	AGTGCAGGU NNNNNN ATGTCTAAAGGTGAAGAATTATT	Forward yeGFP, randomised hexamer and overhang to promoter part
yeGFP_Rv	CGTGCGAU TTATTTGTACAATTCATCC	Reverse *yeGFP,* overhang to CfB2336 backbone
CfB2336_Fw	ATCGCGTGU ATTCATCCGCTCTAACCGAA	Forward CfB2336, overhang to yeGFP
CfB2336_Rv	ATCGCACGU ATTCCGTTGGTAGATACGTTGTT	Reverse CfB2336, overhang to promoter part
Promoter and reporter strains		
pTEF1-TGATAT-ymUkG1_fw	AATCTAAGTTTTAATTACAAAGT GCAGGT TGATAT ATGGTCAGTGTCATCAAAGAA	Forward ymUkG1, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-CGACTT-ymUkG1_fw	AATCTAAGTTTTAATTACAAAGT GCAGGT CGACTT ATGGTCAGTGTCATCAAAGAA	Forward ymUkG1, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-ACGTTC-ymUkG1_fw	AATCTAAGTTTTAATTACAAAGT GCAGGT ACGTTC ATGGTCAGTGTCATCAAAGAA	Forward ymUkG1, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-GGGGGT-ymUkG1_fw	AATCTAAGTTTTAATTACAAAGT GCAGGT GGGGGT ATGGTCAGTGTCATCAAAGAA	Forward ymUkG1, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-TAGGTT-ymUkG1_fw	AATCTAAGTTTTAATTACAAAGT GCAGGT TAGGTT ATGGTCAGTGTCATCAAAGAA	Forward ymUkG1, overhang containing defined hexamer and overlap to promoter TEF1

pTEF1-AGGACA-ymUkG1_fw	AATCTAAGTTTTAATTACAAAGTGCAGGT AGGACA ATGGTCAGTGTCATCAAAGAA	Forward ymUkG1, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-TGTGAA-ymUkG1_fw	AATCTAAGTTTTAATTACAAAGTGCAGGT TGTGAA ATGGTCAGTGTCATCAAAGAA	Forward ymUkG1, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-TCGGTC-ymUkG1_fw	AATCTAAGTTTTAATTACAAAGTGCAGGT TCGGTC ATGGTCAGTGTCATCAAAGAA	Forward ymUkG1, overhang containing defined hexamer and overlap to promoter TEF1
tADH1-ymUkG1_rv	CAACGTATCTACCAACGGAATACGTGCGAT TTACTTAGAAGCTTGAGATGGC	Reverse ymUkG1, overhang to terminator ADH1
pTEF1-TGATAT-yeGFP_fw	AATCTAAGTTTTAATTACAAAGTGCAGGT TGATAT ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-CGACTT-yeGFP_fw	AATCTAAGTTTTAATTACAAAGTGCAGGT CGACTT ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-ACGTTC-yeGFP_fw	AATCTAAGTTTTAATTACAAAGTGCAGGT ACGTTC ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-GGGGGT-yeGFP_fw	AATCTAAGTTTTAATTACAAAGTGCAGGT GGGGGT ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-TAGGTT-yeGFP_fw	AATCTAAGTTTTAATTACAAAGTGCAGGT TAGGTT ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-AGGACA-yeGFP_fw	AATCTAAGTTTTAATTACAAAGTGCAGGT AGGACA ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-TGTGAA-yeGFP_fw	AATCTAAGTTTTAATTACAAAGTGCAGGT TGTGAA ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-TCGGTC-yeGFP_fw	AATCTAAGTTTTAATTACAAAGTGCAGGT TCGGTC ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter TEF1

tADH1-yeGFP_rv	CAACGTATCTACCAACGGAATACGTGCGAT TTATTTGTACAATTCATCCATACCA	Reverse yeGFP, overhang to terminator ADH1
pTEF1-TGATAT-mKATE2_fw	TAATCTAAGTTTTAATTACAAAGTGCAGGT TGATAT ATGGTTTCTGAACTCATCAAG	Forward mKate2, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-CGACTT-mKATE2_fw	TAATCTAAGTTTTAATTACAAAGTGCAGGT CGACTT ATGGTTTCTGAACTCATCAAG	Forward mKate2, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-ACGTTC-mKATE2_fw	TAATCTAAGTTTTAATTACAAAGTGCAGGT ACGTTC ATGGTTTCTGAACTCATCAAG	Forward mKate2, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-GGGGGT-mKATE2_fw	TAATCTAAGTTTTAATTACAAAGTGCAGGT GGGGGT ATGGTTTCTGAACTCATCAAG	Forward mKate2, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-TAGGTT-mKATE2_fw	TAATCTAAGTTTTAATTACAAAGTGCAGGT TAGGTT ATGGTTTCTGAACTCATCAAG	Forward mKate2, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-AGGACA-mKATE2_fw	TAATCTAAGTTTTAATTACAAAGTGCAGGT AGGACA ATGGTTTCTGAACTCATCAAG	Forward mKate2, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-TGTGAA-mKATE2_fw	TAATCTAAGTTTTAATTACAAAGTGCAGGT TGTGAA ATGGTTTCTGAACTCATCAAG	Forward mKate2, overhang containing defined hexamer and overlap to promoter TEF1
pTEF1-TCGGTC-mKATE2_fw	TAATCTAAGTTTTAATTACAAAGTGCAGGT TCGGTC ATGGTTTCTGAACTCATCAAG	Forward mKate2, overhang containing defined hexamer and overlap to promoter TEF1
tADH1-mKATE2_rv	CAACGTATCTACCAACGGAATACGTGCGAT TTATCTGTGTCCCAACTTAGATG	Reverse mKate2, overhang to terminator ADH1
pADH2-TGATAT-yeGFP_fw	TTAACTATATCGTAATACACAAGTGCAGGT TGATAT ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter ADH2
pADH2-CGACTT-yeGFP_fw	TTAACTATATCGTAATACACAAGTGCAGGT CGACTT ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter ADH2
pADH2-ACGTTC-yeGFP_fw	TTAACTATATCGTAATACACAAGTGCAGGT ACGTTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter ADH2

	ATGTCTAAAGGTGAAGAATTATTC	
pADH2-GGGGGT-yeGFP_fw	TTAACTATATCGTAATACACAAGTGCAGGT GGGGGT ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter ADH2
pADH2-TAGGTT-yeGFP_fw	TTAACTATATCGTAATACACAAGTGCAGGT TAGGTT ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter ADH2
pADH2-AGGACA-yeGFP_fw	TTAACTATATCGTAATACACAAGTGCAGGT AGGACA ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter ADH2
pADH2-TGTGAA-yeGFP_fw	TTAACTATATCGTAATACACAAGTGCAGGT TGTGAA ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter ADH2
pADH2-TCGGTC-yeGFP_fw	TTAACTATATCGTAATACACAAGTGCAGGT TCGGTC ATGTCTAAAGGTGAAGAATTATTC	Forward yeGFP, overhang containing defined hexamer and overlap to promoter ADH2
XII-5_UP_fw	AAAGTATAGGAACTTCTGAAGTGG	Forward XII-5 upstream homology region and terminator ADH1
tADH1_rv	ATCGCACGTATTCCGTTGG	Reverse XII-5 upstream homology region and terminator ADH1
tCYC1_fw	ATCGCGTGTATTCATCCGC	Forward terminator CYC1, HIS5 cassette, and downstream homology region
XII-5-DW_rv	AACTTCACTTCATTTTATTTAAATTTGC	Reverse terminator CYC1, HIS5 cassette, and downstream homology region
tCYC1-pTEF1_fw	TTTCGGTTAGAGCGGATGAATACACGCG ATGCACACACCATAGCTTCAA	Forward promoter TEF1, overlap to terminator CYC1
pTEF1_rv	ACCTGCACTTTGTAATTAAAACTTAGATTAGATTGCTATG	Reverse promoter TEF1
tCYC1-pADH2_fw	TTTCGGTTAGAGCGGATGAATACACGCG ATTCCGGGAAACACAGTACC	Forward promoter ADH2, overlap to terminator CYC1
pADH2_rv	ACCTGCACTTGTGTATTACGATATAGTTAATAGTTGATAGTTG	Reverse promoter ADH2
Carotenoid strains		
TJOS-24F	AGTGCAGGU ATGGGAAAAGAACAAGATCAGG	Forward *crtI*, overhang to promoter part

TJOS-24R	CGTGCGAU TCAGAAAGCAAGAACACCAACG	Reverse *crtI*, overhang to pCfB390
TJOS-23F	ATCTGTCAU ATGACGGCTCTCGCATATTA	Forward *crtYB*, overhang to promoter part
TJOS-23R	CACGCGAU TTACTGCCCTTCCCATCCGC	Reverse *crtYB*, overhang to pCfB390
erg9_repair_ TGATAT	AGAAGCGGAAAACGTATACACGT CACATATCACACACACACA TGATAT ATGGGAAAGCTATTACAATTGGC ATTGCATCCGGTCGAGATG	Duplex oligo containing defined hexamer, overlab to promoter erg9 and erg9
erg9_repair_ TAGGTT	AGAAGCGGAAAACGTATACACGT CACATATCACACACACACA TAGGTT ATGGGAAAGCTATTACAATTGGC ATTGCATCCGGTCGAGATG	Duplex oligo containing defined hexamer, overlab to promoter erg9 and erg9
erg9_repair_ TCGGTC	AGAAGCGGAAAACGTATACACGT CACATATCACACACACACA TCGGTC ATGGGAAAGCTATTACAATTGGC ATTGCATCCGGTCGAGATG	Duplex oligo containing defined hexamer, overlab to promoter erg9 and erg9
crtE_fw	ATGGATTACGCGAACATCC	Forward crtE, used with XII-5-DW_rv
ATATCA_pP GK1_rv	AATTGCTGTGAGGATGTTCGCGT AATCCAT ATATCA TTGACAGATTTGTTTTATATTTG T	Reverse crtE, overhang containing defined hexamer and overlap promoter PGK1, used with XII-5-UP_fw
AACCTA_pP GK1_rv	AATTGCTGTGAGGATGTTCGCGT AATCCAT AACCTA TTGACAGATTTGTTTTATATTTG T	Reverse *crtE*, overhang containing defined hexamer and overlap promoter PGK1, used with XII-5-UP_fw
GACCGA_p PGK1_rv	AATTGCTGTGAGGATGTTCGCGT AATCCAT GACCGA TTGACAGATTTGTTTTATATTTG T	Reverse crtE, overhang containing defined hexamer and overlap promoter PGK1, used with XII-5-UP_fw
CHO reporter cell lines		
LoxP-kozak_mcherr y_LA_fwd	AGTCGGTGU ATAACTTCGTATAGCATACATTATA CGAAGTTATCGCCACCATGGTGAGC A	Forward mCherry, overhang containing linker A and loxP. To construct loxP-mCherry- lox2272-BGHpA template.
Lox2272-mcherry_O4_r ev	AGACTGTGU ATAACTTCGTATAAAGTATCCTATA CGAAGTTAT CTACTTGTACAGCTCGT	Reverse mCherry, overhang containing linker O4 and lox2272. To construct loxP-mCherry- lox2272-BGHpA template.
BGH pA_O4_fwd	ACACAGTCU CTGTGCCTTCTAGTTGCC	Forward BGHpA, overhang containing linker O4. To construct loxP-mCherry- lox2272-BGHpA template.

BGH pA_O5_rev	ACGCAAGU CCATAGAGCCCACCGCAT	Reverse BGHpA, overhang containing linker O5. To construct loxP-mCherry- lox2272-BGHpA template.
T2 5' arm_750bp_LA_fwd	AGTCGGTGU GTTTCTACACCATCCTAAAGA	Forward 5' homology arm, overhang containing linker A. To construct final HDR-donor plasmid.
T2 5' arm_750bp_LB_rev	ACGCTGCTU TGAATTAGAGGCCAGTCTGAT	Reverse 5' homology arm, overhang containing linker B. To construct final HDR-donor plasmid.
T2 3' arm_750bp_LD_fwd	AGGTCTGAGU AGTCCACAGATTCTGAATGAA	Forward 3' homology arm, overhang containing linker D. To construct final HDR-donor plasmid.
T2 3' arm_750bp_O1_rev	AGCGACGU TACAAAGTCAAGCTAACCATAG	Reverse 3' homology arm, overhang containing linker O1. To construct final HDR-donor plasmid.
TGATAT-ZsGreen_fw	AGCATACAU TATACGAAGTTAT TGATAT ATGGCCCAGTCCAAGC	Forward ZsGreen, overhang containing defined hexamer, loxP, and RMSE backbone
TAGGTT-ZsGreen_fw	AGCATACAU TATACGAAGTTAT TAGGTT ATGGCCCAGTCCAAGC	Forward ZsGreen, overhang containing defined hexamer, loxP, and RMSE backbone
TCGGTC-ZsGreen_fw	AGCATACAU TATACGAAGTTAT TCGGTC ATGGCCCAGTCCAAGC	Forward ZsGreen, overhang containing defined hexamer, loxP, and RMSE backbone
ZsGreen_rv	AAGTATCCU ATACGAAGTTATCTAGGGCAAGG	Reverse *ZsGreen*, overhang to RMSE backbone
TGATAT-eGFP_fw	AGCATACAU TATACGAAGTTAT TGATAT ATGCATGTGAGCAAGGG	Forward eGFP, overhang containing defined hexamer, loxP, and RMSE backbone
TAGGTT-eGFP_fw	AGCATACAU TATACGAAGTTAT TAGGTT ATGCATGTGAGCAAGGG	Forward eGFP, overhang containing defined hexamer, loxP, and RMSE backbone
TCGGTC-eGFP_fw	AGCATACAU TATACGAAGTTAT TCGGTC ATGCATGTGAGCAAGGG	Forward eGFP, overhang containing defined hexamer, loxP, and RMSE backbone
GCCACC-eGFP_fw	AGCATACAU TATACGAAGTTAT GCCACC ATGCATGTGAGCAAGGG	Forward eGFP, overhang containing defined hexamer, loxP, and RMSE backbone
eGFP_rv	AAGTATCCU ATACGAAGTTATCTACTTGTACA GCTC	Reverse *eGFP*, overhang to RMSE backbone
RMSE_fw	AGGATACTU TATACGAAGTTATACTTGCGTAG TGA	Forward RMSE backbone, overhang to CDS part
RMSE_rv	ATGTATGCU ATACGAAGTTATACACCGACTGA G	Reverse RMSE backbone, overhang to CDS part
TIS library sequencing		

S_REV1_NNNNNN-yeGFP fw	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGCATAGTCTAAAGACCTGGTTC	Forward hexamer region of REV1 promoter library, overlap containing adapter for illumina sequencing
S_RPL18B-NNNNNN-yeGFP fw	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CGACAACTATGATAAAATTCTGAAG	Forward hexamer region of RPL18b promoter library, overlap containing adapter for illumina sequencing
S_TEF-NNNNNN-yeGFP fw	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CAAGTTTCAGTTTCATTTTTCTTG	Forward hexamer region of TEF1 promoter library, overlap containing adapter for illumina sequencing
S_NNNNNN-yeGFP_rv	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG AAATTGGGACAACACCAGTG	Reverse hexamer region of promoter library, overlap containing adapter for illumina sequencing

Supplementary Table S2. Plasmids used in study.

Name	Description	Reference
TIS libraries		
pCfB261	Integrative plasmid (XII-5), SpHIS5	(6)
pCFB8168	pCfB261-ccdB	This study
pCFB8169	pCFB8168 pTEF1-NNNNNN-yeGFP-tCYC1	This study
pCFB8170	pCFB8168 pRPL18b-NNNNNN-yeGFP-tCYC1	This study
pCFB8171	pCFB8168 pREV1-NNNNNN-yeGFP-tCYC1	This study
pCfB3050	2 µ, pESC-LEU2, pSNR52-gRNA_XII-5-tSUP4	(7)
Carotenoid strains		
pCfB176	pRS414-TRP1-TEF1p-Cas9-CYC1t	(8)
pCfB390	Integrative plasmid (XI-3), KlURA3	(6)
YIplac211-YB/I/E	YIplac211 pTDH3-crtYB-tCYC1 pTDH3-crtI-tCYC1 pTDH3-crtE*-tCYC1	(9)
pTAJAK-11	Integrative plasmid (XI-3), KlURA3, tADH1-crtI-pTDH3_pTEF1-crtYB-tCYC1	This study
pTAJAK-13	Integrative plasmid (XII-5), SpHIS5, pPGK1-*crtE*-tCYC1	This study
pCfB3045	2 µ, pESC-LEU2, pSNR52-gRNA_XI-3-tSUP4	(7)
pCFB8172	2 µ, pESC-LEU2, pSNR52-gRNA_XII-5-tSUP4 pSNR52-gRNA_erg9-tSUP4	this study
CHO reporter cell lines		
pCFB8173	T2-mCherry-HDR-TI-vector	This study
pJ204	RMCE backbone	(10)
OGS591	PSF-CMV-CRE recombinase vector	Sigma-Aldrich
pCFB8174	pEF1α-loxP-TGATAT-ZsGreen-lox2272-BGHpA	This study
pCFB8175	pEF1α-loxP-TAGGTT-ZsGreen-lox2272-BGHpA	This study

pCFB8176	pEF1α-loxP-TCGGTC-ZsGreen-lox2272-BGHpA	This study
pCFB8177	pEF1α-loxP-GCCACC-eGFP-lox2272-BGHpA	This study
pCFB8178	pEF1α-loxP-TGATAT-eGFP-lox2272-BGHpA	This study
pCFB8179	pEF1α-loxP-TAGGTT-eGFP-lox2272-BGHpA	This study

Supplementary Table S3. *S. cerevisiae* strains used in study.

Name	Genotype	Reference
TIS libraries		
CEN.PK2-1C	MATa his3D1 leu2-3_112 ura3-52 trp1-289 MAL2-8c SUC2	EUROSCARF
TC-3	CEN.PK2-1C pCfB176	(11)
y1_A3	TC-3 XII-5::pTEF1-NNNNNN-*yeGFP-tADH1(spHIS5)*	this study
y1_A6	TC-3 XII-5::pRPL18b-NNNNNN-*yeGFP-tADH1*(spHIS5)	this study
y1_A9	TC-3 XII-5::pREV1-NNNNNN-*yeGFP-tADH1*(spHIS5)	this study
Promoter and reporter strains		
yp10_A1	TC-3 XII-5::pTEF1-TGATAT-ymUkG1-tADH1(spHIS5)	this study
yp10_A2	TC-3 XII-5::pTEF1-CGACTT-ymUkG1-tADH1(spHIS5)	this study
yp10_A3	TC-3 XII-5::pTEF1-ACGTTC-ymUkG1-tADH1(spHIS5)	this study
yp10_A4	TC-3 XII-5::pTEF1-GGGGGT-ymUkG1-tADH1(spHIS5)	this study
yp10_A5	TC-3 XII-5::pTEF1-TAGGTT-ymUkG1-tADH1(spHIS5)	this study
yp10_A6	TC-3 XII-5::pTEF1-AGGACA-ymUkG1-tADH1(spHIS5)	this study
yp10_A7	TC-3 XII-5::pTEF1-TGTGAA-ymUkG1-tADH1(spHIS5)	this study
yp10_A8	TC-3 XII-5::pTEF1-TCGGTC-ymUkG1-tADH1(spHIS5)	this study
yp10_D1	TC-3 XII-5::pTEF1-TGATAT-yeGFP-tADH1(spHIS5)	this study
yp10_D2	TC-3 XII-5::pTEF1-CGACTT-yeGFP-tADH1(spHIS5)	this study
yp10_D3	TC-3 XII-5::pTEF1-ACGTTC-yeGFP-tADH1(spHIS5)	this study
yp10_D4	TC-3 XII-5::pTEF1-GGGGGT-yeGFP-tADH1(spHIS5)	this study
yp10_D5	TC-3 XII-5::pTEF1-TAGGTT-yeGFP-tADH1(spHIS5)	this study
yp10_D6	TC-3 XII-5::pTEF1-AGGACA-yeGFP-tADH1(spHIS5)	this study
yp10_D7	TC-3 XII-5::pTEF1-TGTGAA-yeGFP-tADH1(spHIS5)	this study
yp10_D8	TC-3 XII-5::pTEF1-TCGGTC-yeGFP-tADH1(spHIS5)	this study
yp10_G1	TC-3 XII-5::pTEF1-TGATAT-mKate2-tADH1(spHIS5)	this study
yp10_G2	TC-3 XII-5::pTEF1-CGACTT-mKate2-tADH1(spHIS5)	this study
yp10_G3	TC-3 XII-5::pTEF1-ACGTTC-mKate2-tADH1(spHIS5)	this study
yp10_G4	TC-3 XII-5::pTEF1-GGGGGT-mKate2-tADH1(spHIS5)	this study
yp10_G5	TC-3 XII-5::pTEF1-TAGGTT-mKate2-tADH1(spHIS5)	this study
yp10_G6	TC-3 XII-5::pTEF1-AGGACA-mKate2-tADH1(spHIS5)	this study
yp10_G7	TC-3 XII-5::pTEF1-TGTGAA-mKate2-tADH1(spHIS5)	this study
yp10_G8	TC-3 XII-5::pTEF1-TCGGTC-mKate2-tADH1(spHIS5)	this study
yp14_D1	TC-3 XII-5::pADH2-TGATAT-yeGFP-tADH1(spHIS5)	this study
yp14_D2	TC-3 XII-5::pADH2-CGACTT-yeGFP-tADH1(spHIS5)	this study

yp14_D3	TC-3 XII-5::pADH2-ACGTTC-yeGFP-tADH1(spHIS5)	this study
yp14_D4	TC-3 XII-5::pADH2-GGGGGT-yeGFP-tADH1(spHIS5)	this study
yp14_D5	TC-3 XII-5::pADH2-TAGGTT-yeGFP-tADH1(spHIS5)	this study
yp14_D6	TC-3 XII-5::pADH2-AGGACA-yeGFP-tADH1(spHIS5)	this study
yp14_D7	TC-3 XII-5::pADH2-TGTGAA-yeGFP-tADH1(spHIS5)	this study
yp14_D8	TC-3 XII-5::pADH2-TCGGTC-yeGFP-tADH1(spHIS5)	this study
Carotenoid strains		
TC-9	CEN.PK2-1C XI-3::tADH1-*crtI*-pTDH3_pTEF1-*crtYB*-tCYC1(*KlURA3*)	this study
TC-10	TC-9 pCfB176	this study
yp11_B10	TC-10 XII-5::pPGK1-TGATAT-*crtE*(*spHIS5*) pERG9-TGATAT-*ERG9*	this study
yp11_C4	TC-10 XII-5::pPGK1-TGATAT-*crtE*(*spHIS5*) pERG9-TAGGTT-*ERG9*	this study
yp11_C7	TC-10 XII-5::pPGK1-TGATAT-*crtE*(*spHIS5*) pERG9-TCGGTC-*ERG9*	this study
yp11_E4	TC-10 XII-5::pPGK1-TAGGTT-*crtE*(*spHIS5*) pERG9-TGATAT-*ERG9*	this study
yp11_E10	TC-10 XII-5::pPGK1-TAGGTT-*crtE*(*spHIS5*) pERG9-TAGGTT-*ERG9*	this study
yp11_F1	TC-10 XII-5::pPGK1-TAGGTT-*crtE*(*spHIS5*) pERG9-TCGGTC-*ERG9*	this study
yp11_F7	TC-10 XII-5::pPGK1-TCGGTC-*crtE*(*spHIS5*) pERG9-TGATAT-*ERG9*	this study
yp11_G1	TC-10 XII-5::pPGK1-TCGGTC-*crtE*(*spHIS5*) pERG9-TAGGTT-*ERG9*	this study
yp11_G4	TC-10 XII-5::pPGK1-TCGGTC-*crtE*(*spHIS5*) pERG9-TCGGTC-*ERG9*	this study

Supplementary Table S4. CHO strains used in study.

Name	Genotype	Reference
CHO-S		ThermoFisher
c1_A1	CHO-S pEF1α-loxP-TGATAT-ZsGreen-lox2272-BGHpA	this study
c1_A2	CHO-S pEF1α-loxP-TAGGTT-ZsGreen-lox2272-BGHpA	this study
c1_A3	CHO-S pEF1α-loxP-TCGGTC-ZsGreen-lox2272-BGHpA	this study
c1_A4	CHO-S pEF1α-loxP-GCCACC-eGFP-lox2272-BGHpA	this study
c1_A5	CHO-S pEF1α-loxP-TGATAT-eGFP-lox2272-BGHpA	this study
c1_A6	CHO-S pEF1α-loxP-TAGGTT-eGFP-lox2272-BGHpA	this study
c1_A7	CHO-S pEF1α-loxP-TCGGTC-eGFP-lox2272-BGHpA	this study

Supplementary Table S5. Nucleic acid sequence of promoters used in study. TIS sequence is in bold.

>TEF1
GCACACACCATAGCTTCAAAATGTTTCTACTCCTTTTTTACTCTTCCAGATTTTCTCGGACTCCGC
GCATCGCCGTACCACTTCAAAACACCCAAGCACAGCATACTAAATTTCCCCTCTTTCTTCCTCTAG
GGTGTCGTTAATTACCCGTACTAAAGGTTTGGAAAAGAAAAAAGAGACCGCCTCGTTTCTTTTTCT
TCGTCGAAAAAGGCAATAAAAATTTTTATCACGTTTCTTTTTCTTGAAAATTTTTTTTTTTGATTT
TTTTCTCTTTCGATGACCTCCCATTGATATTTAAGTTAATAAACGGTCTTCAATTTCTCAAGTTTC
AGTTTCATTTTTCTTGTTCTATTACAACTTTTTTTACTTCTTGCTCATTAGAAAGAAAGCATAGCA
ATCTAATCTAAGTTTTAATTACAAAGTGCAGGT**NNNNNN**
>RPL18b
GGCGTCGTTGTTAATTTCGAAGAGGATGTCCAATATTTTTTTTAAGGAATAAGGATACTTCAAGAC
TAGATTCCCCCCTGCATTCCCATCAGAACCGTAAACCTTGGCGCTTTCCTTGGGAAGTATTCAAGA
AGTGCCTTGTCCGGTTTCTGTGGCTCACAAACCAGCGCGCCCGATATGGCTTTCTTTTCACTTATG
AATGTACCAGTACGGGACAATTAGAACGCTCCTGTAACAATCTCTTTGCAAATGTGGGGTTACATT
CTAACCATGTCACACTGCTGACGAAATTCAAAGTAAAAAAAAAATGGGACCACGTCTTGAGAACGAT
AGATTTTCTTTATTTTACATTGAACAGTCGTTGTCTCAGCGCGCTTTATGTTTTCATTCATACTTC
ATATTATAAAATAACAAAAGAAGAATTTCATATTCACGCCCAAGAAATCAGGCTGCTTTCCAAATG
CAATTGACACTTCATTAGCCATCACACAAAACTCTTTCTTGCTGGAGCTTCTTTTAAAAAAGACCT
CAGTACACCAAACACGTTACCCGACCTCGTTATTTTACGACAACTATGATAAAATTCTGAAGAAAA
AATAAAAAAATTTTCATACTTCTTGCTTTTATTTAAACCATTGAATGATTTCTTTTGAACAAAACT
ACCTGTTTCACCAAAGGAAATAGAAAGAAAAAATCAATTAGAAGAGTGCAGGT**NNNNNN**
>REV1
GCTTTGAGTTGGGGTAGATTATCGCAAATTACTCATCACATTTATTGACTACGAACTTGCTGATGT
CCTTTTTTTATTTATATTTTTCTTCAGTGAAGCGATTTTTTTTTTACACAGACCAAGACGGAAAAA
AGTAGCTAAGGAAGAAAACAAAATCATGAAAAAAATGTGAAGTGATCATGCACATCGCATCAACTT
AAACATTGGCTTAGAGATATATAGAGTTAGAGTTTACGGCAACCTTTAAGCACCAATACCTTTTGG
CATAGTCTAAAGACCTGGTTCTTAATTTTAAACAAATTTAACTAAAGATTTCCCTATCAAAGAAGT
AACGAGTTGACAGATTTTCTCAAAATAAATCGATACTGCATTTCTAGGCATATCCAGCGAGTGCAG
GT**NNNNNN**
>ADH2
TCCGGGAAACACAGTACCGATACTTCCCAATTCGTCTTCAGAGCTCATTGTTTGTTTGAAGAGACT
AATCAAAGAATCGTTTTCTCAAAAAAATTAATATCTTAACTGATAGTTTGATCAAAGGGGCAAAAC
GTAGGGGCAAACAAACGGAAAAATCGTTTCTCAAATTTTCTGATGCCAAGAACTCTAACCAGTCTT
ATCTAAAAATTGCCTTATGATCCGTCTCTCCGGTTACAGCCTGTGTAACTGATTAATCCTGCCTTT
CTAATCACCATTCTAATGTTTTAATTAAGGGATTTTGTCTTCATTAACGGCTTTCGCTCATAAAAA
TGTTATGACGTTTTGCCCGCAGGCGGGAAACCATCCACTTCACGAGACTGATCTCCTCTGCCGGAA
CACCGGGCATCTCCAACTTATAAGTTGGAGAAATAAGAGAATTTCAGATTGAGAGAATGAAAAAAA
AAAAAAAAAAAAAGGCAGAGGAGAGCATAGAAATGGGGTTCACTTTTTGGTAAAGCTATAGCATGC
CTATCACATATAAATAGAGTGCCAGTAGCGACTTTTTTCACACTCGAAATACTCTTACTACTGCTC
TCTTGTTGTTTTTTATCACTTCTTGTTTCTTCTTGGTAAATAGAATATCAAGCTACAAAAAGCATAC
AATCAACTATCAACTATTAACTATATCGTAATACACAAGTGCAGGT**NNNNNN**
>EF1a
GTGAGGCTCCGGTGCCCGTCAGTGGGCAGAGCGCACATCGCCCACAGTCCCCGAGAAGTTGGGGGG
AGGGGTCGGCAATTGAACCGGTGCCTAGAGAAGGTGGCGCGGGGTAAACTGGGAAAGTGATGTCGT
GTACTGGCTCCGCCTTTTTCCCGAGGGTGGGGGAGAACCGTATATAAGTGCAGTAGTCGCCGTGAA
CGTTCTTTTTCGCAACGGGTTTGCCGCCAGAACACAGGTAAGTGCCGTGTGTGGTTCCCGCGGGCC
TGGCCTCTTTACGGGTTATGGCCCTTGCGTGCCTTGAATTACTTCCACCTGGCTCCAGTACGTGAT
TCTTGATCCCGAGCTGGAGCCAGGGGCGGGCCTTGCGCTTTAGGAGCCCCTTCGCCTCGTGCTTGA
GTTGAGGCCTGGCCTGGGCGCTGGGGCCGCCGCGTGCGAATCTGGTGGCACCTTCGCGCCTGTCTC
GCTGCTTTCGATAAGTCTCTAGCCATTTAAAATTTTTGATGACCTGCTGCGACGCTTTTTTTCTGG

CAAGATAGTCTTGTAAATGCGGGCCAGGATCTGCACACTGGTATTTCGGTTTTTGGGGCCGCGGGC
GGCGACGGGGCCCGTGCGTCCCAGCGCACATGTTCGGCGAGGCGGGGCCTGCGAGCGCGGCCACCG
AGAATCGGACGGGGGTAGTCTCAAGCTGGCCGGCCTGCTCTGGTGCCTGGCCTCGCGCCGCCGTGT
ATCGCCCCGCCCTGGGCGGCAAGGCTGGCCCGGTCGGCACCAGTTGCGTGAGCGGAAAGATGGCCG
CTTCCCGGCCCTGCTCCAGGGGGCTCAAAATGGAGGACGCGGCGCTCGGGAGAGCGGGCGGGTGAG
TCACCCACACAAAGGAAAGGGGCCTTTCCGTCCTCAGCCGTCGCTTCATGTGACTCCACGGAGTAC
CGGGCGCCGTCCAGGCACCTCGATTAGTTCTGGAGCTTTTGGAGTACGTCGTCTTTAGGTTGGGGG
GAGGGGTTTTATGCGATGGAGTTTCCCCACACTGAGTGGGTGGAGACTGAAGTTAGGCCAGCTTGG
CACTTGATGTAATTCTCCTTGGAATTTGCCCTTTTTGAGTTTGGATCTTGGTTCATTCTCAAGCCT
CAGACAGTGGTTCAAAGTTTTTTTCTTCCATTTCAGGTGTCGTGAAGACGTCATATAACTTCGTAT
AGCATACATTATACGAAGTTAT**NNNNNN**

>ERG9

AAAAGTGCAGCTCAGAGCCCCCAGCACCAGTATTAGAGGTCATAATGGGCTGCGAAGCCTGCTAAA
ATGCAGTGGAGGCCGTGTACCCTTTGCCAAATTGGCTATTGGAATCGGCAGAGAACCTGGGTCCCG
TTCTAGAGACCCTGCGAGCGTGTCCCGGTGGGTTCTGGGAGCTCTAACTCCGCAGGAACTACAAAC
CTTGCTTACACAGAGTGAACCTGCTGCCTGGCGTGCTCTGACTCAGTACATTTCATAGCCCATCTT
CAACAACAATACCGACTTACCATCCTATTTGCTTTGCCCTTTTTCTTTTCCACTGCACTTTGCATC
GGAAGGCGTTATCGGTTTTGGGTTTAGTGCCTAAACGAGCAGCGAGAACACGACCACGGGCTATAT
AAATGGAAAGTTAGGACAGGGGCAAAGAATAAGAGCACAGAAGAAGAGAAAAGACGAAGAGCAGAA
GCGGAAAACGTATACACGTCACATATCACACACACACA**NNNNNN**

>PGK1

GGAAGTACCTTCAAAGAATGGGGTCTTATCTTGTTTTGCAAGTACCACTGAGCAGGATAATAATAG
AAATGATAATATACTATAGTAGAGATAACGTCGATGACTTCCCATACTGTAATTGCTTTTAGTTGT
GTATTTTTAGTGTGCAAGTTTCTGTAAATCGATTAATTTTTTTTTCTTTCCTCTTTTTATTAACCT
TAATTTTTATTTTAGATTCCTGACTTCAACTCAAGACGCACAGATATTATAACATCTGCATAATAG
GCATTTGCAAGAATTACTCGTGAGTAAGGAAAGAGTGAGGAACTATCGCATACCTGCATTTAAAGA
TGCCGATTTGGGCGCGAATCCTTTATTTTGGCTTCACCCTCATACTATTATCAGGGCCAGAAAAAG
GAAGTGTTTCCCTCCTTCTTGAATTGATGTTACCCTCATAAAGCACGTGGCCTCTTATCGAGAAAG
AAATTACCGTCGCTCGTGATTTGTTTGCAAAAAGAACAAAACTGAAAAAACCCAGACACGCTCGAC
TTCCTGTCTTCCTATTGATTGCAGCTTCCAATTTCGTCACACAACAAGGTCCTAGCGACGGCTCAC
AGGTTTTGTAACAAGCAATCGAAGGTTCTGGAATGGCGGGAAAGGGTTTAGTACCACATGCTATGA
TGCCCACTGTGATCTCCAGAGCAAAGTTCGTTCGATCGTACTGTTACTCTCTCTCTTTCAAACAGA
ATTGTCCGAATCGTGTGACAACAACAGCCTGTTCTCACACACTCTTTTCTTCTAACCAAGGGGGTG
GTTTAGTTTAGTAGAACCTCGTGAAACTTACATTTACATATATATAAACTTGCATAAATTGGTCAA
TGCAAGAAATACATATTTGGTCTTTTCTAATTCGTAGTTTTTCAAGTTCTTAGATGCTTTCTTTTT
CTCTTTTTTACAGATCATCAAGGAAGTAATTATCTACTTTTTACAACAAATATAAAACAAATCTGT
CAA**NNNNNN**

Supplementary Table S6. Nucleic acid sequence of genes used in study

>ymUkG1

ATGGTCAGTGTCATCAAAGAAGAAATGAAGATCAAGTTGCACATGGAAGGTAACGTTAATGGTCAT
GCCTTTGTTATTGAAGGTGATGGTAAAGGTAAACCATACGATGGTACTCAAACTTTGAACTTGACT
GTCAAAGAAGGTGCTCCATTGCCATTCTCTTACGATATTTTGACTAACGCCTTCCAATACGGTAAT
AGAGCTTTTACTAAGTACCCAGCCGATATCCCAGATTACTTTAAGCAAACTTTTCCAGAAGGTTAC
TCCTGGGAAAGAACTATGTCTTACGAAGATAACGCTATCTGCAACGTCAGATCCGAAATTTCTATG
GAAGGTGATTGCTTCATCTACAAGATCAGATTCGATGGTAAGAACTTTCCACCAAATGGTCCAGTC
ATGCAAAAAAAGACTTTGAAGTGGGAACCATCCACCGAAATGATGTATGTTAGAGATGGTTTCTTG
ATGGGTGATGTCAATATGGCTTTGTTGTTGGAAGGTGGTGGTCATCATAGATGTGATTTCAAGACT
TCTTACAAGGCCAAGAAGGTTGTTCAATTGCCAGATGCTCATAAGATCGATCACAGAATCGAAATC

TTGTCCCACGATAGAGATTACTCCAAGGTTAAGTTGTACGAAAACGCTGTTGCTAGAAACTCTTTG
TTGCCATCTCAAGCTTCTAAGTAA

>yeGFP

ATGTCTAAAGGTGAAGAATTATTCACTGGTGTTGTCCCAATTTTGGTTGAATTAGATGGTGATGTT
AATGGTCACAAATTTTCTGTCTCCGGTGAAGGTGAAGGTGATGCTACTTACGGTAAATTGACCTTA
AAATTTATTTGTACTACTGGTAAATTGCCAGTTCCATGGCCAACCTTAGTCACTACTTTCGGTTAT
GGTGTTCAATGTTTTGCTAGATACCCAGATCATATGAAACAACATGACTTTTTCAAGTCTGCCATG
CCAGAAGGTTATGTTCAAGAAAGAACTATTTTTTTCAAAGATGACGGTAACTACAAGACCAGAGCT
GAAGTCAAGTTTGAAGGTGATACCTTAGTTAATAGAATCGAATTAAAAGGTATTGATTTTAAAGAA
GATGGTAACATTTTAGGTCACAAATTGGAATACAACTATAACTCTCACAATGTTTACATCATGGCT
GACAAACAAAAGAATGGTATCAAAGTTAACTTCAAAATTAGACACAACATTGAAGATGGTTCTGTT
CAATTAGCTGACCATTATCAACAAAATACTCCAATTGGTGATGGTCCAGTCTTGTTACCAGACAAC
CATTACTTATCCACTCAATCTGCCTTATCCAAAGATCCAAACGAAAAGAGAGACCACATGGTCTTG
TTAGAATTTGTTACTGCTGCTGGTATTACCCATGGTATGGATGAATTGTACAAATAA

>ymKate2

ATGGTTTCTGAACTCATCAAGGAAAACATGCACATGAAACTTTACATGGAAGGTACTGTGAACAAT
CATCATTTTAAGTGTACATCCGAGGGTGAAGGCAAACCTTACGAAGGAACTCAAACTATGAGAATT
AAAGCTGTAGAAGGTGGACCATTACCTTTTGCATTTGATATCTTGGCAACATCATTCATGTATGGG
AGCAAGACATTCATAAACCATACTCAAGGTATACCAGACTTTTTCAAACAGAGTTTTCCAGAGGGT
TTTACATGGGAAAGAGTAACAACGTACGAGGATGGAGGTGTATTGACAGCCACTCAAGACACATCA
CTTCAAGATGGGTGTTTAATCTACAATGTCAAGATTAGAGGCGTCAATTTCCCTTCTAATGGTCCA
GTTATGCAGAAAAAGACATTAGGCTGGGAAGCGTCAACCGAAACCCTGTACCCTGCTGATGGTGGC
CTAGAAGGCAGAGCTGACATGGCCCTTAAACTGGTTGGTGGAGGGCATCTAATCTGCAATTTGAAA
ACCACTTATCGTTCTAAAAAGCCAGCCAAAAACCTAAAGATGCCAGGTGTTTACTACGTCGACCGA
AGATTAGAAAGGATTAAAGAGGCTGATAAAGAGACTTATGTTGAACAACACGAAGTGGCAGTGGCT
AGATACTGTGATTTGCCATCTAAGTTGGGACACAGATAA

>ZsGreen

ATGGCCCAGTCCAAGCACGGCCTGACCAAGGAGATGACCATGAAGTACCGCATGGAGGGCTGCGTG
GACGGCCACAAGTTCGTGATCACCGGCGAGGGCATCGGCTACCCCTTCAAGGGCAAGCAGGCCATC
AACCTGTGCGTGGTGGAGGGCGGCCCCTTGCCCTTCGCCGAGGACATCTTGTCCGCCGCCTTCATG
TACGGCAACCGCGTGTTCACCGAGTACCCCCAGGACATCGTCGACTACTTCAAGAACTCCTGCCCC
GCCGGCTACACCTGGGACCGCTCCTTCCTGTTCGAGGACGGCGCCGTGTGCATCTGCAACGCCGAC
ATCACCGTGAGCGTGGAGGAGAACTGCATGTACCACGAGTCCAAGTTCTACGGCGTGAACTTCCCC
GCCGACGGCCCCGTGATGAAGAAGATGACCGACAACTGGGAGCCCTCCTGCGAGAAGATCATCCCC
GTGCCCAAGCAGGGCATCTTGAAGGGCGACGTGAGCATGTACCTGCTGCTGAAGGACGGTGGCCGC
TTGCGCTGCCAGTTCGACACCGTGTACAAGGCCAAGTCCGTGCCCCGCAAGATGCCCGACTGGCAC
TTCATCCAGCACAAGCTGACCCGCGAGGACCGCAGCGACGCCAAGAACCAGAAGTGGCACCTGACC
GAGCACGCCATCGCCTCCGGCTCCGCCTTGCCCTAG

>mEGFP

ATGCATGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGC
GACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTG
ACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTG
ACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCC
GCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACC
CGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTC
AAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATC
ATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGC
AGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCC

GACAACCACTACCTGAGCACCCAGTCCAAGCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATG
GTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAG

>erg9

ATGGGAAAGCTATTACAATTGGCATTGCATCCGGTCGAGATGAAGGCAGCTTTGAAGCTGAAGTTT
TGCAGAACACCGCTATTCTCCATCTATGATCAGTCCACGTCTCCATATCTCTTGCACTGTTTCGAA
CTGTTGAACTTGACCTCCAGATCGTTTGCTGCTGTGATCAGAGAGCTGCATCCAGAATTGAGAAAC
TGTGTTACTCTCTTTTATTTGATTTTAAGGGCTTTGGATACCATCGAAGACGATATGTCCATCGAA
CACGATTTGAAAATTGACTTGTTGCGTCACTTCCACGAGAAATTGTTGTTAACTAAATGGAGTTTC
GACGGAAATGCCCCCGATGTGAAGGACAGAGCCGTTTTGACAGATTTCGAATCGATTCTTATTGAA
TTCCACAAATTGAAACCAGAATATCAAGAAGTCATCAAGGAGATCACCGAGAAAATGGGTAATGGT
ATGGCCGACTACATCTTAGATGAAAATTACAACTTGAATGGGTTGCAAACCGTCCACGACTACGAC
GTGTACTGTCACTACGTAGCTGGTTTGGTCGGTGATGGTTTGACCCGTTTGATTGTCATTGCCAAG
TTTGCCAACGAATCTTTGTATTCTAATGAGCAATTGTATGAAAGCATGGGTCTTTTCCTACAAAAA
ACCAACATCATCAGAGATTACAATGAAGATTTGGTCGATGGTAGATCCTTCTGGCCCAAGGAAATC
TGGTCACAATACGCTCCTCAGTTGAAGGACTTCATGAAACCTGAAAACGAACAACTGGGGTTGGAC
TGTATAAACCACCTCGTCTTAAACGCATTGAGTCATGTTATCGATGTGTTGACTTATTTGGCCGGT
ATCCACGAGCAATCCACTTTCCAATTTTGTGCCATTCCCCAAGTTATGGCCATTGCAACCTTGGCT
TTGGTATTCAACAACCGTGAAGTGCTACATGGCAATGTAAAGATTCGTAAGGGTACTACCTGCTAT
TTAATTTTGAAATCAAGGACTTTGCGTGGCTGTGTCGAGATTTTTGACTATTACTTACGTGATATC
AAATCTAAATTGGCTGTGCAAGATCCAAATTTCTTAAAATTGAACATTCAAATCTCCAAGATCGAA
CAGTTTATGGAAGAAATGTACCAGGATAAATTACCTCCTAACGTGAAGCCAAATGAAACTCCAATT
TTCTTGAAAGTTAAAGAAAGATCCAGATACGATGATGAATTGGTTCCAACCCAACAAGAAGAAGAG
TACAAGTTCAATATGGTTTTATCTATCATCTTGTCCGTTCTTCTTGGGTTTTATTATATATACACT
TTACACAGAGCGTGA

>crtE

ATGGATTACGCGAACATCCTCACAGCAATTCCACTCGAGTTTACTCCTCAGGATGATATCGTGCTC
CTTGAACCGTATCACTACCTAGGAAAGAACCCTGGAAAAGAAATTCGATCACAACTCATCGAGGCT
TTCAACTATTGGTTGGATGTCAAGAAGGAGGATCTCGAGGTCATCCAGAACGTTGTTGGCATGCTA
CATACCGCTAGCTTATTAATGGACGATGTGGAGGATTCATCGGTCCTCAGGCGTGGGTCGCCTGTG
GCCCATCTAATTTACGGGATTCCGCAGACAATAAACACTGCAAACTACGTCTACTTTCTGGCTTAT
CAAGAGATCTTCAAGCTTCGCCCAACACCGATACCCATGCCTGTAATTCCTCCTTCATCTGCTTCG
CTTCAATCATCCGTCTCCTCTGCATCCTCCTCCTCCTCGGCCTCGTCTGAAAACGGGGGCACGTCA
ACTCCTAATTCGCAGATTCCGTTCTCGAAAGATACGTATCTTGATAAAGTGATCACAGACGAGATG
CTTTCCCTCCATAGAGGGCAAGGCCTGGAGCTATTCTGGAGAGATAGTCTGACGTGTCCTAGCGAA
GAGGAATATGTGAAAATGGTTCTTGGAAAGACGGGAGGTTTGTTCCGTATAGCGGTCAGATTGATG
ATGGCAAAGTCAGAATGTGACATAGACTTTGTCCAGCTTGTCAACTTGATCTCAATATACTTCCAG
ATCAGGGATGACTATATGAACCTTCAGTCTTCTGAGTATGCCCATAATAAGAATTTTGCAGAGGAC
CTCACAGAAGGAAAATTCAGTTTTCCCACTATCCACTCGATTCATGCCAACCCCTCATCGAGACTC
GTCATCAATACGTTGCAGAAGAAATCGACCTCTCCTGAGATCCTTCACCACTGTGTAAACTACATG
CGCACAGAAACCCACTCATTCGAATATACTCAGGAAGTCCTCAACACCTTGTCAGGTGCACTCGAG
AGAGAACTAGGAAGGCTTCAAGGAGAGTTCGCAGAAGCTAACTCAAAGATTGATCTTGGAGACGTA
GAGTCGGAAGGAAGAACGGGGAAGAACGTCAAATTGGAAGCGATCCTGAAAAAGCTAGCCGATATC
CCTCTGTGA

Supplementary Table S6. Translation efficiencies of yeast native genes with TISs characterized in this study. Translation efficiencies were obtained from Lahtvee *et al*. (2017)(12).

TIS no.	Genes*	Lathvee *et al.* (2017) (SI_Table S6)	Protein-mRNA ratio	Scaled from 1 to 100
1				
2	YCL048W			
3	YJL194W			
4				
5	YDL149W, YGR070W, YLL003W, YPL137C	YLL003W	73.27	0.00192
6	YER180C-A, YJL029C, YML102W, YMR214W	YJL029C, YMR214W	215.94, 766.85	0.00576, 0.0206
7	YPR175W	YPR175W	0	-0.00005
8				

*According to CEN.PK113-7d genome (Genbank ID: AEHG01000000; Saccharomyces cerevisiae CEN.PK113-7D, whole genome shotgun sequencing project)

REFERENCES

1. Noderer,W.L., Flockhart,R.J., Bhaduri,A., Diaz de Arce,A.J., Zhang,J., Khavari,P.A. and Wang,C.L. (2014) Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol. Syst. Biol.*, **10**, 748.
2. Decoene,T., Peters,G., De Maeseneire,S. and De Mey,M. (2018) Toward predictable 5'UTRs in Saccharomyces cerevisiae: Development of a yUTR calculator. 10.1021/acssynbio.7b00366.
3. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker T A N D,M. and Schuster,P. (1994) Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte ftir Chemie*, **125**, 167–188.
4. Lahtvee,P.J., Sánchez,B.J., Smialowska,A., Kasvandik,S., Elsemman,I.E., Gatto,F. and Nielsen,J. (2017) Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. *Cell Syst.*, **4**, 495–504.e5.
5. Tyner,C., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Eisenhart,C., Fischer,C.M., Gibson,D., Gonzalez,J.N., Guruvadoo,L., *et al.* (2017) The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.*, **45**, D626–D634.
6. Jensen,N.B., Strucko,T., Kildegaard,K.R., David,F., Maury,J.J., Mortensen,U.H., Forster,J., Nielsen,J. and Borodina,I. (2014) EasyClone: Method for iterative chromosomal integration of multiple genes in Saccharomyces cerevisiae. *FEMS Yeast Res.*, **14**, 238–248.
7. Jessop-Fabre,M.M., Jakočiūnas,T., Stovicek,V., Dai,Z., Jensen,M.K., Keasling,J.D. and Borodina,I. (2016) EasyClone-MarkerFree: A vector toolkit for marker-less integration of genes into Saccharomyces cerevisiae via CRISPR-Cas9. *Biotechnol. J.*, **11**, 1110–1117.
8. DiCarlo,J.E., Norville,J.E., Mali,P., Rios,X., Aach,J. and Church,G.M. (2013) Genome

engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res.*, **41**, 4336–4343.
9. Verwaal,R., Wang,J., Meijnen,J.P., Visser,H., Sandmann,G., Van Den Berg,J.A. and Van Ooyen,A.J.J. (2007) High-level production of beta-carotene in Saccharomyces cerevisiae by successive transformation with carotenogenic genes from Xanthophyllomyces dendrorhous. *Appl. Environ. Microbiol.*, **73**, 4342–4350.
10. Lund,A.M., Kildegaard,H.F., Petersen,M.B.K., Rank,J., Hansen,B.G., Andersen,M.R. and Mortensen,U.H. (2014) A Versatile System for USER Cloning-Based Assembly of Expression Vectors for Mammalian Cell Engineering. *PLoS One*, **9**, e96693.
11. Jakočinas,T., Bonde,I., Herrgård,M., Harrison,S.J., Kristensen,M., Pedersen,L.E., Jensen,M.K. and Keasling,J.D. (2015) Multiplex metabolic pathway engineering using CRISPR/Cas9 in Saccharomyces cerevisiae. *Metab. Eng.*, **28**, 213–222.
12. Lahtvee,P.-J., Sánchez,B.J., Smialowska,A., Kasvandik,S., Elsemman,I.E., Gatto,F. and Nielsen,J. (2017) Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. Cell Syst, 4, 495–504.e5.

3. Predictive engineering and optimization of tryptophan metabolism in yeast through a combination of mechanistic and machine learning models

Jie Zhang[1#], Søren D. Petersen[1#], Tijana Radivojevic[2,5,8], Andrés Ramirez[3], Andrés Pérez[3], Eduardo Abeliuk[4], Benjamín J. Sánchez[1], Zachary Costello[2,5,8], Yu Chen[9,10], Mike Fero[4], Hector Garcia Martin[2,5,8,11], Jens Nielsen[1,9,12] , Jay D. Keasling[1-2,5-7], & Michael K. Jensen[1*]

[1] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kgs. Lyngby, Denmark

[2] Joint BioEnergy Institute, Emeryville, CA, USA

[3] TeselaGen SpA, Santiago, Chile

[4] TeselaGen Biotechnology, San Francisco, CA 94107, USA

[5] Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[6] Department of Chemical and Biomolecular Engineering & Department of Bioengineering, University of California, Berkeley, CA, USA

[7] Center for Synthetic Biochemistry, Institute for Synthetic Biology, Shenzhen Institutes of Advanced Technologies, Shenzhen, China

[8] DOE Agile BioFoundry, Emeryville, CA, USA

[9] Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden

[10] Novo Nordisk Foundation Center for Biosustainability, Chalmers University of Technology, Gothenburg, Sweden

[11] BCAM, Basque Center for Applied Mathematics, Bilbao, Spain

[12] BioInnovation Institute, Ole Maaløes Vej 3, DK-2200 Copenhagen N, Denmark

* To whom correspondence should be addressed. Michael K. Jensen: Email: mije@biosustain.dtu.dk, Tel: +45 6128 4850

[#] These authors contributed equally to this study

SUMMARY

In combination with advanced mechanistic modeling and the generation of high-quality multi-dimensional data sets, machine learning is becoming an integral part of understanding and engineering living systems. Here we show that mechanistic and machine learning models can complement each other and be used in a combined approach to enable accurate genotype-to-phenotype predictions. We use a genome-scale model to pinpoint engineering targets and produce a large combinatorial library of metabolic pathway designs with different promoters which, once phenotyped, provide the basis for machine learning algorithms to be trained and used for new design recommendations. The approach enables successful forward engineering of aromatic amino acid metabolism in yeast, with the new recommended designs improving tryptophan production by up to 17% compared to the best designs used for algorithm training, and ultimately producing a total increase of 106% in tryptophan accumulation compared to optimized reference designs. Based on a single high-throughput data-generation iteration, this study highlights the power of combining mechanistic and machine learning models to enhance their predictive power and effectively direct metabolic engineering efforts.

KEYWORDS

Machine learning, genome-scale metabolic modeling, yeast, biosensor, tryptophan

INTRODUCTION

Metabolic engineering is the directed improvement of cell properties through the modification of specific biochemical reactions (Stephanopoulos, 1999). Beyond offering an improved understanding of basic cellular metabolism, the field of metabolic engineering also envisions sustainable production of biomolecules for health, food, and manufacturing industries, by fermenting feedstocks into value-added biomolecules using engineered cells (Keasling, 2010). These promises leverage tools and technologies developed over recent decades which include mechanistic metabolic modeling, targeted genome engineering, and robust bioprocess optimization; ultimately aiming for accurate and scalable predictions of cellular phenotypes from deduced genotypes (Nielsen and Keasling, 2016; Choi et al., 2019; Liu and Nielsen, 2019).

Among the different types of mechanistic models for simulating metabolism, genome-scale models (GSMs) are one of the most popular approaches, as they are

genome-complete, covering thousands of metabolic reactions. These computational models not only provide qualitative mapping of cellular metabolism (Hefzi et al., 2016; Monk et al., 2017; Lu et al., 2019), but have also been successfully applied for the discovery of novel metabolic functions (Guzmán et al., 2015), and to guide engineering designs towards desired phenotypes (Yang et al., 2018).As GSMs are built based only on the stoichiometry of metabolic reactions, several methods have been developed to account for additional layers of information regarding the chemical intermediates and the catalyzing enzymes participating in the metabolic pathways of interest (Lewis et al., 2012). However, the predictive power of these enhanced models is often hampered by the limited knowledge and data available for any of such parameters affecting metabolic regulation (Gardner, 2013; Khodayari et al., 2015; Long and Antoniewicz, 2019).

Machine learning provides a complementary approach to guide metabolic engineering by learning patterns on systems behavior from large experimental data sets (Camacho et al., 2018). As such, machine learning models differ from mechanistic models by being purely data-driven. Indeed, machine learning methods for the generation of predictive models on living systems are becoming ubiquitous, including applications within genome annotation, *de novo* pathway discovery, product maximization in engineered microbial cells, pathway dynamics, and transcriptional drivers of disease states (Alonso-Gutierrez et al., 2015; Carro et al., 2010; Costello and Martin, 2018; Jervis et al., 2019; Mellor et al., 2016; Schläpfer et al., 2017). While being able to provide predictive power based on complex multivariate relationships (Presnell and Alper, 2019), the training of machine learning algorithms requires large datasets of high quality, and thereby imposes certain standards for the experimental workflows. For instance, for genotype-to-phenotype predictions, it is desirable that datasets contain a high variation between both genotypes and phenotypes (Carbonell et al., 2019). Also, measurements on the individual experimental unit, e.g. a strain, should be accurate and obtainable in a high-throughput manner, in order to limit the number of iterative design-build-test cycles needed in order to reach the desired output.

While mechanistic models require *a priori* knowledge of the living system of interest, and machine learning-guided predictions require ample multivariate experimental data for training, the combination of mechanistic and machine learning models holds promise for improved performance of predictive engineering of cells by uniting the advantages of the causal understanding of mechanism from mechanistic models with the predictive power of machine learning (Zampieri et al., 2019; Presnell and Alper, 2019). Metabolic pathways are known to be

regulated at multiple levels, including transcriptional, translational, and allosteric levels (Chubukov et al., 2014). To cost-effectively move through the design and build steps of complex metabolic pathways regulated at multiple levels, combinatorial optimization of metabolic pathways, in contrast to sequential genotype edits, has been demonstrated to effectively facilitate identification of global optima for outputs of interest (i.e. production; Jeschek et al., 2017). Searching global optima using combinatorial approaches involves facing an exponentially growing number of designs (known as the combinatorial explosion), and requires efficient building of multi-parameterized combinatorial libraries. However, this challenge can be mitigated by the use of intelligently designed condensed libraries which allow uniform discretisation of multidimensional spaces: e.g. by using well-characterized sets of DNA elements controlling the expression of candidate genes at defined levels (Jeschek et al., 2016; Lee et al., 2013). As cellular metabolism is regulated at multiple levels (Feng et al., 2014; Lahtvee et al., 2017), an efficient search strategy for global optima using combinatorial approaches should also take this into consideration, e.g. by using mechanistic models, 'omics data repositories, and *a priori* biological understanding.

Here we combine mechanistic and machine learning models to enable robust genotype-to-phenotype predictions as a tool for metabolic engineering. The approach is exemplified for predictive engineering and optimization of the complexly regulated aromatic amino acid pathway that produces tryptophan in baker's yeast *Saccharomyces cerevisiae*. We defined a 7,776-membered combinatorial library design space, based on 5 genes selected from GSM simulations and *a priori* biological understanding, each controlled at the level of gene expression by 6 different promoters from a total set of 30 promoters selected from transcriptomics data mining. In order to train predictive models for high-tryptophan biosynthesis rate in yeast, we collected >144,000 experimental data points using a tryptophan biosensor, exploring this way approximately 4% of the genetic designs of the library design space. Based on a single Design-Build-Test-Learn cycle focused on sequencing data, growth profiles, and biosensor output, we trained various machine learning algorithms. Predictive models based on these algorithms enabled construction of designs exhibiting tryptophan biosynthesis rates 106% higher than a state-of-the-art high-tryptophan reference strain (Hartmann et al., 2003; Rodriguez et al., 2015), and up to 17% higher rate than best designs used for training the models.

RESULTS

Model-guided design of high tryptophan production

One prime example of the multi-tiered complexity regulating metabolic fluxes, is the shikimate pathway, driving the central metabolic route leading to aromatic amino acid biosynthesis in microorganisms (Lingens et al., 1967; Braus, 1991; Averesch and Krömer, 2018). This pathway has enormous industrial relevance, since it has been used to produce bio-based replacements of a wealth of fossil fuel-derived aromatics, polymers, and potent human therapeutics (Curran et al., 2013; Suástegui and Shao, 2016).

To search for gene targets predicted to perturb tryptophan production, we initially performed constraint-based modeling for predicting single gene targets, with a simulated objective of combining growth and tryptophan production (Orth et al., 2010; Ferreira et al., 2019). From this analysis, we retrieved 192 genes, covering 259 biochemical reactions, that showed considerable changes as production shifted from growth towards tryptophan production (Figure 1A-B, Table S4). By performing an analysis for statistical over-representation of genome-scale modelled metabolic pathways, we observed that both the pentose phosphate pathway and glycolysis were among the top pathways with a significantly higher number of gene targets compared to the representation of all metabolic genes (Figure 1C, Table S5). Among the predicted gene targets in those pathways, *CDC19*, *TKL1*, *TAL1* and *PCK1* were initially selected as targets for combinatorial library construction (Figure 1B), as these genes have all been experimentally validated to be directly linked or to have an indirect impact on the shikimate pathway precursors erythrose 4-phosphate (E4P) and phosphoenolpyruvate (PEP). Specifically, *CDC19* encodes the major isoform of pyruvate kinase converting PEP into pyruvate to fuel the tricarboxylic acid (TCA) cycle, while *TKL1* and *TAL1* that encode the major isoform of transketolase and transaldolase, respectively, in the reversible non-oxidative pentose phosphate pathway (PPP), have been reported to impact the supply of E4P (Patnaik and Liao, 1994; Curran et al., 2013). Additionally, focusing on the E4P and PEP linkage, *PCK1* encoding PEP carboxykinase, was also selected due to its regeneration capacity of PEP from oxaloacetate (Yin, 1996). Lastly, while not being predicted as a target by the constraint-based modeling approach, the *PFK1* gene, encoding the alpha subunit of heterooctameric phosphofructokinase, catalyzing the irreversible conversion of fructose 6-phosphate (F6P) to fructose 1,6-bisphosphate (FBP), was selected, as insufficient activity of this enzyme is known to cause divergence of carbon flux towards the pentose phosphate pathway in different

organisms across different kingdoms (Wang et al., 2013; Zhang et al., 2016).

Next, we mined transcriptomics data sets for the selection of promoters to control the expression of the five selected candidate genes. Here we focused on well-characterized and sequence-diverse promoters, to ensure rational designs spanning large absolute levels of promoter activities and limit the risk of recombination within strain designs and loss of any genetic elements, respectively (Figure S1; Rajkumar et al., 2019; Reider Apel et al., 2017). Together, this mining resulted in the selection of 25 sequence-diverse promoters, which together with the five promoters natively regulating the selected candidate genes, constitutes the parts catalog for combinatorial library design (Figure 1D; Figure S1, Table S6).

Creation of a platform strain for a combinatorial library

To construct a combinatorial library targeting equal representation of thirty promoters expressing five candidate genes, we harnessed high-fidelity homologous recombination in yeast together with the targetability of CRISPR/Cas9 genome engineering for a one-pot assembly of a maximum of 7,776 (6^5) different combinatorial designs. Due to the dramatic decrease in transformation efficiency when simultaneously targeting multiple loci in the genome (Jakočiūnas et al., 2015), we targeted the sequential deletion of all five selected target genes from their original genomic loci, and next assemble a cluster of five expression cassettes into a single genomic landing as recently successfully reported for the "single-locus glycolysis" in yeast (Kuijpers et al., 2016)(Figure 2A). However, as *CDC19* is an essential gene, and deletion of *PFK1* causes growth retardation (Breslow et al., 2008; Cherry et al., 2012), this genetic background would be unsuitable for efficient one-pot transformation. For this reason our platform strain for library construction had a galactose-curable plasmid introduced expressing *PFK1*, *CDC19*, *TKL1* and *TAL1* under their native promoters (see METHODS DETAILS), before performing two sequential rounds of genome engineering to delete *PCK1*, *TKL1* and *TAL1*, and knock-down *CDC19* and *PFK1* using the weak promoters *RNR2* and *REV1*, respectively (Figure 2A). Furthermore, prior to one-pot assembly of the combinatorial library, we integrated the two feedback-inhibited shikimate pathway enzymes 3-deoxy-D-arabinose-heptulosonate-7-phosphate (DAHP) synthase ($ARO4^{K229L}$) and anthranilate synthase ($TRP2^{S65R,\ S76L}$) into our platform strain (Hartmann et al., 2003; Graf et al., 1993), thereby aiming to maximise the impact from transcriptional regulation of candidate genes on the overall tryptophan output, as removal of allosteric feedback inhibition is known to increase amino acid accumulation in microbial cells (Park et al., 2014; Vogt et al., 2014).

One-pot construction of the combinatorial library

For library construction, we first tested the transformation by constructing five control strains, including a strain with native promoters in front of each of the five selected genes (herein labeled the reference strain; Table S7). Next, we transformed in one-pot the platform strain with equimolar amounts (1 pmol/part) of double-stranded DNA encoding each of the thirty promoters, the five open reading frames encoding the candidate genes with native terminators, a *HIS3* expression cassette for selection, and two 500-bps homology-regions for targeted repair of the genomic integration site. In total, this design combination included 38 different parts for 7,776 unique 20 kb 13-parts assemblies at the targeted genomic locus (Chr. XII, EasyClone site V; Figure 2A). Following transformation, we randomly sampled 480 colonies from the library, together with 27 colonies from the five control strains (507 in total), and successfully cured 423 out of 461 (92%) sufficiently growing strains of the complementation plasmid by means of galactose-induced expression of the dosage-sensitive gene *ACT1* (Figures 2B & S6; Liu et al., 1992; Makanae et al., 2013). Next, genotyping all promoter-gene junctions by sequencing (Figure S2), identified 380 out of 461 (82%) of the sufficiently growing strains to be correctly assembled with only 9 out of 245 (3.7%) of the fully filtered library genotypes observed in duplicates (245 = 250 library and control genotypes - 5 control genotypes)(Figure 2B). Based on a Monte Carlo simulation with 10,000 repeated samplings of 10,000 library colonies, and assuming percent correct assemblies and promoter distribution as determined for the library sample (Figure 2), the expected no. of unique genotypes among all library colonies was calculated to be 3,759. This equals an estimated library coverage of 48% (3,759/7,776). Importantly, all thirty promoters from the one-pot transformation mix were represented in the genotyped designs, with promoters *PGK1* (no. 14) and *MLS1* (no. 15), represented the least (1%) and most (35%), respectively (Figure 2C).

Taken together, these results demonstrate high transformation efficiency of the platform strain, high fidelity of parts assembly, and expected high coverage of the genetically diverse combinatorial library design.

Engineering a tryptophan biosensor for high-throughput library characterization

In order to support high-throughput analysis of tryptophan accumulation in library strains, we harnessed the power of modular engineering allosterically regulated transcription factors as small-molecule *in vivo* biosensors (Mahr and Frunzke, 2016; Rogers et al., 2016). Here, a yeast

tryptophan biosensor was developed based on the *trpR* repressor of the *trp* operon from *E. coli* (Roesser and Yanofsky, 1991; Gunsalus and Yanofsky, 1980). In order to engineer *trpR* as a tryptophan biosensor in yeast, we first tested trpR-mediated transcriptional repression by expressing *trpR* together with a GFP reporter gene under the control of the strong *TEF1* promoter containing a palindromic consensus *trpO* sequence (5'-GTACTAGTT-AACTAGTAC-3'; Yang et al., 1996) downstream of the TATA-like element (TATTTAAG; Figure 3A; Rhee and Pugh, 2012). From this, we observed that *trpR* was able to repress GFP expression by 2.4-fold (Figure S3A). Next, to turn the native *trpR* repressor into an activator with a positively correlated biosensor-tryptophan readout we fused the Gal4 activation domain to the N-terminus of codon-optimized *trpR* (*GAL4*$_{AD}$*-trpR*) expressed under the control of the weak *REV1* promoter (Figure S3B). For the reporter promoter, we placed *trpO* 97 bp upstream of the TATA-like element of the *TEF1* promoter (Figure S3B), and observed that *trpR* was able to activate GFP expression by a maximum of 1.75-fold upon supplementing tryptophan to the cultivation medium (Figure S3B). To further optimize the dynamic range of the reporter output, the GFP reporter was expressed under a hybrid promoter consisting of tandem repeats of triple *trpO* sequences (i.e., in total 6x *trpO* sequences) located 88 bp upstream of the TATA box in an engineered *GAL1* core promoter without Gal4 binding sites, ultimately enabling *GAL4*$_{AD}$*-trpR*-mediated biosensing with a dynamic output range of 5-fold, and an operational input range spanning supplemented tryptophan concentrations from ~2-200 mg/L (Figure 3B).

To further validate the designed biosensor we measured fluorescence output in strains engineered for expression of feedback-resistant versions of ARO4 and TRP2 (ARO4^{K229L} and TRP2$^{S65R,\ S76L}$; (Hartmann et al., 2003; Graf et al., 1993), and observed high biosensor outputs from these strains in line with previously demonstrated high enzyme activities in strains expressing ARO4^{K229L} and TRP2$^{S65R,\ S76L}$ (Hartmann et al., 2003; Graf et al., 1993), and thus corroborating the ability of the tryptophan biosensor to monitor changes in endogenously produced tryptophan pools (Figure 3C). Most importantly, we confirmed the biosensor readout as a valid proxy for tryptophan levels, by comparing external tryptophan titers measured by HPLC with a change in GFP intensities for 6 library strains spanning 2.5-fold changes in GFP intensities (R^2 = 0.75; Figure 3D).

Having established a biosensor for high-throughput screening of the combinatorial library, we next sought to explore the maximal resolution of the biosensor readout at the single-design level of growing isoclonal strains, with the intention to define optimal data sampling time point. To do so, we measured time-series data of OD and GFP in triplicates for all

507 colonies, covering a total of >144,000 data points (Figure S4). Here, as we observed that the fluorescence per cell generally stabilized at an OD value of 0.075 and started to decrease beyond an OD value of 0.15 (Figure 3E, Figure S4, see METHODS DETAILS), and the between strains variation in fluorescence at the single-cell level was relatively high within this OD-interval, we chose this interval for determining the GFP synthesis rate as a proxy for tryptophan flux. By sampling all variant designs, average GFP synthesis rate was observed to vary between 43.7 and 255.7 MFI/h (approx. 6-fold; Figure 3F), with an average standard error of the mean of 6.6 MFI/h corresponding to an average coefficient of variation for the mean values of 4.3%. By comparison, the GFP synthesis rate of the platform strain, expressing $ARO4^{K229L}$ and $TRP2^{S65R, S76L}$ together with all five candidate genes under native promoters, was 144.8 MFI/h (Figure 3F).

Using machine learning to predict metabolic pathway designs

Having successfully established a combinatorial genetic library and a large phenotypic data set thereof, we next assessed the potential of using machine learning to predict promoter combinations expected to improve tryptophan productivity. Since there is no algorithm which is optimal for all learning tasks (Wolpert, 1996), we used two different machine learning approaches: the Automated Recommendation Tool (ART) and EVOLVE algorithm (Radivojević et al., 2019; TeselaGen, 2019). The input for both algorithms was the promoter combination and tryptophan productivity (measured through the GFP proxy, Figure S4). Briefly, ART uses a Bayesian ensemble approach where eight regressors from the scikit-learn library (Pedregosa et al., 2011) are allowed to “vote” on a prediction with a weight proportional to their accuracy; the EVOLVE algorithm is inspired by Bayesian Optimization and uses an ensemble of estimators as a surrogate model that predicts the outcome of the process to be optimized (see METHODS DETAILS). As the quality of the data is of paramount importance for machine learning predictions, we initially filtered our data to avoid genotypes with insufficient growth, no sequencing data, incorrect assembly, no plasmid curation, or which exhibited more than one genotype (see METHOD DETAILS; Figure S5). Following this, approximately 58% (266/461) of the growing strains remained after filtering, while another 3% of the remaining data was removed because of lack of reproducibility (high error in triplicate measurements)(Figure S5).

Both modeling approaches, ART and EVOLVE, were able to recapitulate the data they were trained on. The average (obtained from 10 independent runs) training mean absolute error (MAE) of the predicted tryptophan production compared to the measured values was 13.8 and 11.9 MFI/h for the ART and EVOLVE model approaches, respectively, when calculated for the

whole data set (Figure 4A-B). These MAEs represent ~7% and 6% of the full range of measurements (50 to 200 MFI/h). The train MAE uncertainty (represented by the shaded area in Figure 4A-B and quantified as the 95% confidence interval from 10 runs) decreased slightly with increasing size of the training data set for ART, whereas the overall uncertainty was smaller for the EVOLVE model approach (Figure 4A-B). The ability to predict the production for new promoter combinations the algorithms had not been trained on was tested by cross-validation, i.e. by training the model on 90% of the data, and then testing the predictions of this model against measurements for the remaining 10% (10-fold cross-validation). Here, the average cross-validated MAE (test MAE) was 21.4 and 22.4 MFI/h for ART and EVOLVE model approaches, respectively (Figure 4A-B), which represent ~11% of the full range of measurements. The test MAE decreased systematically with the size of the data set, yet the decrease rate declined markedly as more data was added. However, while the two approaches had similar average cross-validated MAEs, the uncertainty of the MAEs was slightly smaller for ART than for EVOLVE algorithm (Figure 4A-B).

Machine learning-guided engineering of designs with high tryptophan productivity

Next, beyond enabling prediction of tryptophan production, we used an exploitative approach implemented in the ART model and an explorative one adopting the EVOLVE algorithm to recommend two sets of 30 prioritized designs aiming for high tryptophan production (Tables S8 and S9). The exploitative model focuses on exploiting the predictive power to recommend promoter combinations that improve production, whereas the exploratory model combines predictive power with the estimated uncertainty of each prediction, to recommend promoter combinations (Radivojević et al., 2019; TeselaGen, 2019).

Among the recommendations from each of the two machine learning approaches, two overlapped (SP588 and SP627, Table S8-S9). Interestingly, while use of *PGK1* promoter to control *TKL1* expression was underrepresented in the original library sample (Figure 2C), the explorative set of recommendations included eight (even top-three) designs with *PGK1* promoter for expression control of *TKL1*, and the exploitative approach included none (Table S5; Figure 4C-D). From construction of these recommendations, we used the same genome engineering approach as for library construction (Figure 2A) to successfully construct 19 individual assemblies of the explorative recommendations and 24 individual assemblies of the exploitative recommendations. Interestingly, we were not able to construct any of the eight designs with *PGK1* promoter, partially explaining the lower number of viable strains found with

the explorative approach.

Of the 41 recommendations constructed, the predictions from both sets generally fitted well with the measurements, and both approaches successfully enabled predictive strain engineering for high-performing GFP synthesis rates, with the best recommendation having a measured GFP synthesis rate 106% higher than the already improved platform design, and 17% higher than the best one in the library sample (Figure 4E-F). Moreover, eight recommendations were found in the top-ten of productivity, of which four were from the exploitative set, three were from the explorative set, and one overlapping between the two sets. Comparing the output of the ART and EVOLVE approaches, the variation in measurements was higher for strains recommended with the explorative EVOLVE approach than for strains recommended with the exploitative ART approach (Figure 4E-F), and the explorative approach included recommendations based on a more diverse set of promoters than the exploitative approach (Figure 4C-D). Still, taken together, both approaches successfully enabled predictive engineering of a strain with tryptophan productivity beyond those previously observed (Figure 4E-F).

DISCUSSION

We have demonstrated that mechanistic and machine learning approaches can complement and enhance each other, enabling a more effective predictive engineering of living systems. Using a single design-build-test-learn cycle, this study i) leveraged mechanistic genome-scale models to select and rank reactions/genes most likely to affect production, ii) included the efficient one-pot construction of a library with different promoter combinations for these reactions, and iii) used machine learning algorithms trained on the ensuing phenotyping data to choose novel promoter combinations that further enhance tryptophan productivity. In total, we managed to increase the tryptophan synthesis rate by 106% compared to an already improved reference strain ($ARO4^{K229L}$ and $TRP2^{S65R, S76L}$).

To gather the large data sets required to enable machine learning approaches, we developed a biosensor which enabled the sampling of >144,000 GFP intensity measurements as a proxy for tryptophan flux for 1,728 isoclonal designs in a high-throughput fashion (Figures 3E, S5A). Indeed, while requiring a few design iterations (Figures 3A, S3), the tryptophan biosensor ultimately allowed us to i) phenotypically characterize an order of magnitude higher number of strains than in previous machine learning-guided metabolic engineering studies (Alonso-Gutierrez et al., 2015; Lee et al., 2013a; Redding-Johanson et al., 2011; Zhou et al.,

2018a), and ii) identify optimal sampling points that displayed the largest differences between genotypes (Figures 3C, S4). Likewise, one-pot CRISPR/Cas9-mediated genome editing was a vital enabling technology for this project, since it allowed us to efficiently create a diverse 20-kb clustered combinatorial library with representation of all 30 specified sequence- and expression-diverse promoters to control five expression units, including very few duplicate designs (Figure 2B-C).

Enabled by this high-quality data set, we used two different machine learning models for predicting productivity (ART and EVOLVE algorithm), and two different approaches to recommend new strains (exploitative and explorative). Cross-validation showed that both models could be trained to show good correlations (MAE approximately 11% of the measurement range) between predictions and measurements for data they had not seen previously (test data). The test MAE was basically the same for the two models, and plateaued quickly as a function of the number of genotypes in the training data set (Figure 4A-B). Whereas the uncertainty in predictive accuracy decreased considerably with the number of genotypes in the data set, this decrease was similar for both models. With this in mind, a relevant guideline for choosing a recommendation approach should focus on the desired outcome: the explorative approach providing a more diverse set of recommendations (Figure 4C-D), whereas the exploitative approach provides less varied recommendations. We observed the largest improvement in productivity when using the exploitative approach (Figure 4E-F). However, if subsequent design-build-test-learn cycles are performed, the diversity of recommendations of the explorative approach could help avoid local optima of tryptophan production(Figure 4E-F).

Notably, while the recommendations were able to improve production, the predictions from both machine learning models were noticeably worse than for the library, reflecting the general challenge of extrapolating outside of the previous range of measurements. As such, we envision that future machine learning approaches will need to focus on models able to extrapolate more efficiently.

With respect to advancing biological understanding of tryptophan metabolism, the results provided examples of anticipated results as well as non-intuitive predictions. The best performing strain (SP606, Table S8) predicted by machine-learning, displayed knock-downs of both *CDC19* and *PFK1*, corroborating our intuitive strategies for increasing precursor availability: i.e. lower pyruvate kinase activity would lead to higher PEP pools, while limiting glycolysis redirects carbon flux into PPP and subsequently increases E4P. However, this strain also had low expression of *TKL1* and high expression of *TAL1*, despite the report that

overexpression of *TKL1*, rather than *TAL1*, leads to higher aromatic amino acid production in both *E. coli* and yeast (Curran et al., 2013). This finding remarks the importance of carefully considering the systems-level context of these "metabolic rules of thumb" (e.g. overexpress TKL1 instead of TAL1 for higher amino acid production) to ensure their validity. Consistently, both the second (SP616) and third (SP624) best performing strains, also predicted by machine learning, had low expression of *TKL1* and high expression of *TAL1*, together with very low expression (*TPK2* promoter) for *PFK1* and high expression of *CDC19*. One possible explanation is that, although normally expressed, the pyruvate kinase activity could be limited by low level of its allosteric activator FBP due to limited PFK expression. Another plausible explanation is that medium-high expression of *PCK1* (conversion of oxaloacetate to PEP) by *ACT1* or *TDH3* promoters in these two strains can replenish PEP pools consumed by pyruvate kinase. The fact that 8 out of 10 top-performing strains had high expression of *PCK1*, which was not predicted to be impactful on glucose by the GSM approach, indicates that this indeed has a positive effect on tryptophan biosynthesis rate, and stresses the importance of combining mechanistic and machine learning approaches.

Ultimately, in our case study, machine learning models have demonstrated significant predictive power. However, this predictive power is heavily dependent on the availability of high quality experimental data, which is not a prerequisite for mechanistic GSMs. Without any experimental input, GSMs are able to guide metabolic engineering using various constraint-based algorithms, which, however, predict a large number of potential targets and may also miss some effective ones, e.g. *PFK1* in our study. This could be due to the lack of other information beyond metabolism e.g. regulation in GSMs. To address this problem, manual efforts are currently needed to filter out less relevant targets, and add intuitively promising ones based on existing knowledge and literature mining. Additionally, future GSMs that include more biological aspects and suitable predicting algorithms are envisioned to further improve gene target selection. Irrespective of the ongoing efforts for model-guided engineering of living cells, this study highlights the enhanced predictive power obtained by combining GSMs for selecting genetic targets with machine learning algorithms for leveraging experimental data. Finally, as even more efficient methods for combining data-driven machine learning algorithms and GSMs are developed, we envision dramatic improvements in our ability to engineer virtually any cell system effectively.

ACKNOWLEDGMENTS

This work was supported by the Novo Nordisk Foundation and the European Commission Horizon 2020 programme (grant agreement No. 722287 and No. 686070). This work was also part of the DOE Agile BioFoundry (http://agilebiofoundry.org), supported by the U.S. Department of Energy, Energy Efficiency and Renewable Energy, Bioenergy Technologies Office, and the DOE Joint BioEnergy Institute (http://www.jbei.org), supported by the Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). H.G.M. was also supported by the Basque Government through the BERC 2014-2017 program and by Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa excellence accreditation SEV-2013-0323. This work was also supported by the Chilean economic development agency, Corfo, through grant 17IEAT-73382.

AUTHOR CONTRIBUTIONS

JZ, SDP, JDK, JN and MKJ conceived the study. JZ and SDP conducted all experimental work, YC and BJS all mechanistic modelling, and TR, ZC, and HGM developed and applied statistical modelling and recommendations based on ART, while EA, AR, and MF developed and applied statistical modelling and recommendations based on TeselaGen EVOLVE model. SDP, JZ, and MKJ wrote the manuscript.

DECLARATION OF INTERESTS

JDK has a financial interest in Amyris, Lygos, Demetrix, Maple Bio, and Napigen. EA and MF have a financial interest in TeselaGen Biotechnology.

FIGURE LEGENDS

Figure 1. Selection of gene targets and promoters for combinatorial engineering of tryptophan metabolism in *S. cerevisiae*. (A) Gene-gene interaction network built with Cytoscape (Shannon et al., 2003), showing that pentose phosphate pathway and glycolysis are both in the core of metabolism in close proximity to many genes. Nodes are all 909 genes in yeast metabolism (Aung et al., 2013), sharing connections based on the number of shared metabolites by the corresponding reactions that the genes are related to: the thicker the edge, the higher the number of shared metabolites. Currency metabolites such as water, protons, ATP, etc. are removed from the analysis. The prefuse force directed layout is used for displaying the network. Genes are highlighted with a yellow border if they are selected targets by the mechanistic modeling approach, and in orange and dark blue if they belong to the pentose phosphate pathway or glycolysis, respectively. (B) Simplified map of metabolism showing the selected gene targets from glycolysis (dark blue) and pentose phosphate pathway (orange) based on a combination of mechanistic genome-scale modeling and literature studies for optimizing tryptophan production. Black dashed lines indicate multi-step reactions. Dashed green line indicates allosteric activation. G6P, glucose 6-phosphate; F6P, fructose 6-phosphate; FBP, fructose 1,6-bisphosphate; GAP, glyceraldehyde 3-phosphate; DHAP, dihydroxyacetone phosphate; PEP, phosphoenolpyruvate; OAA, oxaloacetate; 6PG, 6-phosphogluconate; E4P, erythrose 4-phosphate; S7P, sedoheptulose 7-phosphate; DAHP, 3-deoxy-7-phosphoheptulonate; Tyr, tyrosine; Phe, phenylalanine; Trp, tryptophan. (C) Percentage of genes in glycolysis (dark blue) and pentose phosphate pathway (orange) that were predicted by the mechanistic modelling to increase tryptophan production compared to the percentage of genes predicted as targets from the whole metabolism. *** = P-value < 0.05, Fisher's exact testing. (D) Relative mRNA abundance, calculated for each gene as the proportion of mRNA reads obtained for any given promoter relative to the total sum of mRNA reads from each bin of six promoters. Absolute abundances for the 30 promoters were measured in S. cerevisiae CEN.PK 113-7D in the mid-log phase (Rajkumar et al., 2019). The promoters are grouped according to intended combinatorial gene associations.

Figure 2. Construction and validation of the 13-parts assembled 20 kb combinatorial promoter:gene library. (A) Strategy for library construction including a 13-part *in vivo* assembly for the reintegration of target genes into a single genomic locus. The platform strain used for

one-pot transformation includes a total of 9 genome edits for knowck-out, knock-down and heterologous expression of candidate genes (see METHODS DETAILS). (B) Key descriptive statistics for the library construction and genotyping. (C) Promoter distribution (name, % representation) by gene. Color intensity correlates with promoter strength (see Figure 1D).

Figure 3. Phenotypic library characterization using an engineered tryptophan biosensor. (A) Schematic illustration of the design of the tryptophan (Trp) biosensor ($trpR_{AD}$) engineered in this study. The $trpR_{AD}$ indicates the engineering tryptophan biosensor comprised of the *E. coli* TrpR fused to the GAL4 activation domain. The biosensor regulates and engineered reporter (yeGFP) *GAL1*-promoter including 6x copies of TrpR binding sites (*trpO*), placed upstream the TATA box of *GAL1* promoter (*pGAL1_6x_trpO*). (B) Fluorescence normalized by optical density (OD600) for two strains related to concentration of tryptophan supplemented media (Mean Fluorescence Intensity/OD, MFI/OD with standard errors, n = 3). Both strains contain the yeGFP reporter under the control of the *pGAL1_6x_trpO* reporter promoter, and only one strain expresses the Gal4 activation domain fused to trpR (in green). (C) Fluorescence normalized by OD600 for a wild-type strain and strains with expression of feedback-resistant versions of ARO4 and TRP2, $ARO4^{K229L}$ and $TRP2^{S65R,S76L}$, respectively (mean fluorescence intensity, MFI/h with standard errors, n = 3). (D) Extracellular tryptophan normalized by OD600 related to fluorescence normalized by OD600 (mean values with standard errors, n = 3). (E) Fluorescence divided by OD600 related to OD600 for library and control strains. Dashed lines are shown at OD600 equals 0.075 and 0.15. (F) Measured mean green fluorescent protein synthesis rate. MFI/h with standard errors, n = 3. The data is ranked according to increasing mean rate. The strain with five native promoters expressing the five candidate genes is highlighted in green. MFI = Mean Fluorescence Intensity. OD600 = Optical density (600 nm). a.u. = arbitrary units.

Figure 4. Machine learning-guided predictive engineering of tryptophan metabolism. (A-B) Learning curves for ART and EVOLVE algorithms, respectively. Mean absolute error (MAE) from model training and testing as a function of the number of genotypes in the dataset. Shaded areas represent 95% confidence intervals. Blue curves indicate MAE when calculated for the whole data set (Train), while red curves indicate the cross-validation, i.e. by training the models on 80% of the data and then testing the predictions of this model against measurements for the remaining 20% (Test). (C-D) Promoter distributions for the 30 recommendations of the exploitative (ART) and explorative (EVOLVE) approach, respectively. The orders and colors of

promoters correspond to those in Figure 1C. (E-F) Cross-validated predictions vs average of measured GFP synthesis rate for the exploitative (ART) and explorative (EVOLVE) approach, respectively. Data is shown for library and controls strains (grey markers; green markers show the platform strain expressing ARO4^{K229L} and TRP2^{S65R,S76L}), as well as for recommended strains (blue markers; orange markers show recommendations that overlap between the two approaches).

STAR*METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
    - Mechanistic modeling of high tryptophan flux
    - Promoter selection
    - General strain construction
    - Platform strain construction
    - Construction of combinatorial library
    - Development of tryptophan biosensor
    - Validation of biosensor by HPLC
    - Genomic DNA sequencing
    - Measuring fluorescence and growth
- QUANTIFICATION AND STATISTICAL ANALYSIS
    - Modelling
- DATA AND SOFTWARE AVAILABILITY

STAR*METHODS

Detailed methods are provided in the online version of this paper and include the following:

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
yeast synthetic drop-out media	Sigma	P#:Y2001
LB medium	Sigma	P#:L3522
Ampicillin	Sigma	P#:A0166
L-Leucine	Sigma	P#:L8912
Uracil	Sigma	P#:U1128
L-Tryptophan	Sigma	P#: T0254
PEG	Sigma	Cat#P3640-1KG
LiAc	Sigma	Cat#517992-100G
Salmon sperm	Sigma	Cat#D9156
Critical Commercial Assays		
PlateSeq PCR Kits	Eurofins	PID:3094-000PPP
Deposited Data		
RNAseq data (Arun)	(Rajkumar et al., 2019)	N/A
Genotypes	The Joint BioEnergy Institute's Inventory of Composable Elements (ICE; https://public-registry.jbei.org)	Zhang and Petersen et al. 2019
Time series	The Joint BioEnergy Institute's Experiment Data Depot (EDD; https://public-edd.jbei.org)	Zhang and Petersen et al. 2019
Experimental Models: Organisms/Strains		
*MAT***a** *his3*Δ1, *LEU2*, *ura3*-52, *TRP1 MAL2*-8c *SUC2*	EUROSCARF	CEN.PK113-11C
*MAT***a** *his3*Δ1, *leu2*-3_112, *ura3*-52, *trp1*-289, *MAL2*-8c *SUC2*	EUROSCARF	CEN.PK2-1C
*MAT***a** $P_{GAL1core\_6xtrpO}$-*yEGFP*-T_{ADH1}, P_{TEF1_trpO}-*mKate2*-T_{CYC1}, pCfB176	This study	TrpA-1
*MAT***a** $P_{GAL1core\_6xtrpO}$-*yEGFP*-T_{ADH1}, P_{TEF1_trpO}-mKate2-T_{CYC1}, $ARO4^{wt}$::$ARO4^{K229L}$, pCfB176	This study	TrpA-2
*MAT***a** $P_{GAL1core\_6xtrpO}$-*yEGFP*-T_{ADH1}, P_{TEF1_trpO}-mKate2-T_{CYC1}, $TRP2^{wt}$::$TRP2^{S65R, S76L}$, pCfB176	This study	TrpA-3
*MAT***a** $P_{GAL1core\_6xtrpO}$-*yEGFP*-T_{ADH1}, P_{TEF1_trpO}-mKate2-T_{CYC1}, $ARO4^{wt}$::$ARO4^{K229L}$, $TRP2^{wt}$::$TRP2^{S65R, S76L}$, pCfB176	This study	TrpA-4

*MATa tkl1*Δ *tal1*Δ *pck1*Δ, P_{PFK1}::P_{REV1}-*PFK1*, P_{CDC19}::P_{RNR2}-*CDC19*, P_{PFK1}-*GAL4*$_{ad}$-*trpR*-T_{ADH1}, $P_{GAL1core\_3xtrpO}$-*yEGFP*-T_{ADH1}, P_{TEF1_trpO}-*mKate2*-T_{CYC1}, P_{PGK1}-*ARO4*^{K229L}-T_{ADH1}, P_{TEF1}-*TRP2*$^{S65R, S76L}$-T_{CYC1}, pCfB176, pCfB9307	This study	TrpNA-W
Recombinant DNA		
Plasmids used in the study, see Table S2	This study	N/A
Oligonucleotides		
Primers for strain construction, plasmid construction and sequencing, see Table S1	This study	N/A
Software and Algorithms		
Chromeleon™ Chromatography Data System Software v7.1.3	Thermo fisher (https://www.thermofisher.com/)	Chromeleon™ CDS 7.1.3
Python and standard packages for data analysis	Python (https://www.python.org)	N/A
S. cerevisiae v7 consensus genome scale model	Sourceforge (https://sourceforge.net/projects/yeast/)	Yeast 7.0
COBRA Toolbox	Github (https://github.com)	opencobra/cobratoolbox
GSM analysis	Github (https://github.com)	biosustain/trp-scores
ART	Github (https://github.com)	JBEI/AutomatedRecommendationTool
Teselagen EVOLVE model	TeselaGen's platform (https://teselagen.com)	EVOLVE module
Code for preprocessing and ART modelling approach	Github (https://github.com)	Zhang and Petersen et al. 2019 (sorpet/Zhang_and_Petersen_et_al_2019)

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Michael Krogh Jensen (mije@biosustain.dtu.dk).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Saccharomyces cerevisiae strains were derived from CEN.PK2-1C (EUROSCARF, Germany). These were cultivated in yeast synthetic drop-out media (Sigma-Aldrich) at 30 °C.

Escherichia coli DH5α were cultivated in LB medium containing 100 mg/l ampicillin (Sigma-Aldrich) at 37 °C.

METHOD DETAILS

Mechanistic modeling of high tryptophan flux

In order to select targets for increased tryptophan accumulation, we followed a constraint-based strategy implemented in a recent study (Ferreira et al., 2019), similar to the FSEOF approach (Choi et al., 2010). Briefly, flux balance analysis (FBA; Orth et al., 2010) was used to simulate growth of *S. cerevisiae* at 11 different sub-optimal growth conditions ranging from 30% to 80% of the maximum specific growth rate, with all remaining flux oriented towards tryptophan accumulation. Based on these simulations, a score was calculated for each reaction in metabolism as the average simulated flux fold-change compared to maximum growth rate conditions. These reaction scores were in turn used to compute gene scores, by averaging the associated reaction scores. A gene score higher than one means that the gene is associated to reactions that increase in flux as tryptophan production increases, and could point to a target for overexpression. On the other hand, a gene score lower than one signifies that the gene is connected to reactions that decrease their flux as tryptophan production increases, and therefore could be a target for downregulation. The analysis was performed with either glucose or ethanol as carbon sources, so to find candidates under a mixed-fermentation regime, a purely respiratory regime and the overlap between both regimes. The 7th version of the consensus genome-scale model of *S. cerevisiae* (Aung et al., 2013), a parsimonious FBA (pFBA) approach (Lewis et al., 2010), and the COBRA toolbox (Heirendt et al., 2019) were used for all simulations.

Promoter selection

Each of the five gene targets was expressed under six unique promoters. The six promoters included the promoter native to the gene as well as 5 promoters chosen to span a wide expression range All promoters were chosen based on absolute mRNA abundances measured for S. cerevisiae CEN.PK 113-7D in the mid-log phase (Rajkumar et al., 2019), and unless otherwise stated were 1 kb in length by default. To minimize homologous recombination during one-pot transformation for library construction and potential loop-out of promoters and genes following genomic integration, all scanned promoter sequences were aligned to ensure

there were no extensive homologous sequence stretches.

General strain construction

Strains were edited using the CasEMBLR method (Jakočiūnas et al., 2015). All integration were directed towards EasyClone sites (Jensen et al., 2014). Homology regions between DNA parts were by default 30 bp, and homology regions, framing the repair assembly, were about 0.5 kb. Yeast transformations were performed by LiAc/SS carrier DNA/PEG method (Gietz and Schiestl, 2007). DNA parts and plasmids were purified using kits from Macherey-Nagel. PCR products for USER assembly were amplified using Phusion U Hot Start PCR Master Mix (ThermoFisher), bricks for transformation by Phusion High-Fidelity PCR Master Mix with HF Buffer (ThermoFisher), whereas colony PCRs were performed using 2xOneTaq Quick-Load Master Mix with Standard Buffer (New England Biolabs). Genomic DNA was extracted from overnight cultures using Yeast DNA Extraction Kit (Thermo Scientific). Oligos were purchased from IDT. Sequencing was performed by Eurofins. All primers, plasmids, and yeast strains, are listed in Tables S1, S2, and S3, respectively.

Platform strain construction

Several enzymes within the aromatic amino acid (AAA) biosynthesis are subject to allosteric regulations. Specifically, 3-deoxy-D-arabino-heptulosonate-7-phosphate (DAHP) synthase (encoded by *ARO4*), which controls the entry of the shikimate pathway, is feedback inhibited by all three aromatic amino acids, although to different extents. Anthranilate synthase (encoded by *TRP2*), which catalyzes the first committed step towards the tryptophan branch, is also inhibited by its end product tryptophan (Braus, 1991). To maximise the transcriptional regulatory effect on the tryptophan flux, and benchmark with current state-of-the-art in shikimate pathway optimization, feedback resistant variants of these two enzymes, $ARO4^{K229L}$ (Hartmann et al., 2003) and $TRP2^{S65R,\ S76L}$ (Graf et al., 1993), were overexpressed under the *TEF1* and *TDH3* promoters, respectively at EasyClone site XI-3 (Jessop-Fabre et al., 2016; Table S2). Secondly, a tryptophan biosensor system (see Library phenotypic characterization) was introduced by integrating corresponding sensor and reporter sequences into EasyClone sites at Chr. XI-2 and XI-5, respectively (Jensen et al., 2014).

Construction of combinatorial library

Due to the dramatic decrease in transformation efficiency targeting multiple loci in the

genome (Jakočiūnas et al., 2015), we opted for removing all five target genes from their original loci and assemble the five expression units into a single cluster for targeted integration into EasyClone site XII-5 (Jensen et al., 2014), and thereby ensuring comparable genomic accessibility of all genes. While *PCK1*, *TKL1* and *TAL1* were successfully knocked out; deleting *PFK1* and/or *CDC19* was unsuccessful. Alternatively, we replaced *PFK1* and *CDC19* promoters with weak *REV1* and *RNR2* promoters, respectively. Due to an expected loss of activity in phosphofructokinase (PFK1) and pyruvate kinase (CDC19), and consequently slow ATP generation, the resulting strain (TrpNA-W) grew extremely poorly and was barely transformable using linear DNA fragments for assembly. To overcome this limitation, the TrpNA-W strain was complemented with plasmid pCfB9307 (Table S2) harboring *PFK1*, *CDC19*, *TKL1* and *TAL1* genes, which restored the growth to the wild type level. The plasmid backbone carries yeast *ACT1* gene under the control of *GAL1* promoter, which can be used as counter-selection of the plasmid due to the growth arrest caused by *ACT1* overexpression on galactose as the sole carbon source (Makanae et al., 2013, Figure S6).

For combinatorial library construction we adopted CasEMBLR (Jakočiūnas et al., 2015). Briefly, five target genes together with a *HIS3* expression cassette (in the order of *PCK1-TAL1-TKL1-CDC19-PFK1-HIS3*) were assembled in the same orientation and integrated at EasyClone site XII-5 (Jensen et al., 2014). All five target genes (the complete ORFs) together with their terminators (500 bp downstream of the stop codon) were amplified from the genomic DNA of yeast strain CEN.PK113-7D using primers listed in Table S1. All 30 promoters (defined as the 1000 bp upstream the ORF) were amplified using primers with a 30 bp overlap to adjacent DNA parts (i.e. the terminator upstream and the target gene). All promoters can be found in Tables S4. The *HIS3* cassette was amplified from plasmid pRS413-HIS3 (Sikorski and Hieter, 1989) with primers 30 bp overlapping with the *PFK1* terminator and fragment homologous to the downstream of XII-5. The *HIS3* cassette was included as one part of the assembly. The one-pot transformation of all 38 parts (30 promoters, 5 candidate genes, *HIS3* cassette, and up- and down-homology regions for EasyClone site XII-5) was performed with 50 mL the base strain grown to an optical density of 1.0 (equivalent to 6.5 mg of cell dry weight), 5.0 ug of plasmid expressing the guide RNA targeting XII-5, and 1.0 picomole of each of 13 DNA fragments. A total of 480 colonies were picked from 10 transformation plates by dividing the area of each individual plate into 4 subareas of equal size and picking 12 colonies of varying size from each subarea.

Finally, the complementation plasmid introduced was cured by culturing strains to

stationary phase twice in media with galactose instead of glucose as carbon source (Figure S6). The success of curing were then gauged by a growth assay where LEU auxotrophs were considered as cured and prototrophs as not cured. Control strains and recommended strains were constructed similarly to the library strains except that instead of transforming pools of promoter parts for each gene only specific promoters were transformed per gene.

Development of tryptophan biosensor

The yeast tryptophan biosensor was developed based on the *trpR* repressor of the *trp* operon from *E. coli* (Gunsalus and Yanofsky, 1980). The *trpR* gene was amplified from E. coli M1665 genome. All yeast promoters as well as the activator domain of *GAL4* were amplified from *S. cerevisiae* strain CEN.PK113-7D genome. All designs of trpR biosensor and GFP reporter were first cloned into the pRS416 (*URA3*) and pRS413 (*HIS3*) vectors, respectively, by USER cloning (Bitinaite et al., 2007). The activator domain of *GAL4* ($GAL4_{AD}$) was fused to trpR with a GSGSGS linker by USER cloning. The *trpO* sequence was inserted into the *TEF1* promoter 8 bp downstream of the TATA-like element (TATTTAAG) by inverse PCR from a plasmid containing the P_{TEF1}-*yEGFP*-T_{ADH1} cassette, with both primers containing the overhang AACTAGTAC (ie., half of the *trpO* sequence). The linear PCR product was treated with DpnI enzyme to fragmente the template plasmid and self-ligated to generate circular plasmid (Quick Ligation™ Kit, NEB). Promoters containing multiple *trpO* sequences were constructed by USER cloning from a synthetic DNA fragment (Integrated DNA Technologies) of a minimal *GAL1* promoter (-329 to -5 relative to the *GAL1* open reading frame, thus without the *GAL4* binding sequence which is located at -435 to -418) with 3x tandem repeats of *trpO* (separated by 2 nucleotides) inserted at 88 bp upstream of the TATA box (TATATAAA). Plasmids containing the sensor and reporter cassettes were transformed into yeast strain CEN.PK113-11C. To test the biosensor performance, yeast transformants were grown in selection media overnight and regrown in Delft medium supplemented with various tryptophan concentrations (2-1000 mg/L) for 6 hrs (typically reaching early exponential phase). GFP and mKate2 outputs were measured on SynergyMX microtiter plate reader (BioTek) with excitation/emission at 485/515 nm and 588/633 nm, respectively, and always normalized by absorbance at 600 nm (OD600nm). To construct the base strain for library assembly, the tryptophan sensor (P_{REV1}-$GAL4_{AD}$-*trpR*-T_{ADH1}) and the reporter cassette ($P_{GAL1core\_3xtrpO}$-*yEGFP*-T_{ADH1}, P_{TEF1_trpO}-*mKate2*-T_{CYC1}) were integrated into strain TC-3 (Jakočiūnas et al., 2015) at the EasyClone sites XI-2 and XI-5 (Jessop-Fabre et al., 2016), respectively.

Validation of biosensor by HPLC

To validate the correlation between biosensor reporter gene output and tryptophan production, we quantified extracellular tryptophan levels by HPLC using a method described by Luo et al. (2019). Supernatants of cultivated strains were separated from the culture broth following 24 hrs of cultivation in synthetic dropout medium without tryptophan and histidine. From this 200 µl was used for HPLC and the data were processed using Chromeleon™ Chromatography Data System Software v7.1.3.

Genomic DNA sequencing

Genomic DNA was extracted from overnight cultures using method described by Lõoke et al. (2011). Each extract was used as template in 5 PCR reactions spanning the 5 integrated promoters and amplifying from 1,200 - 1,700 bp. The PCR products were validated using a LabChip GX II (Perkin Elmer) and sequenced using PlateSeq PCR Kits (Eurofins) according to the manufacturer's instructions. From the LabChip results, a PCR reaction was considered as trusted if it showed a strong band of the correct size, not trusted if it showed a strong band of the wrong size, and as no information gained if it showed a weak or no band. From the sequencing results, a sequencing reaction was considered as trusted if it showed an unambiguous sequence of the expected length (i.e. only limited by length of PCR fragment, stretches of the same nucleotide in the promoter or of about 1,000 bp limit of sanger sequencing reactions), not trusted if it showed an unambiguous sequence of the expected length with an assembly error, and no information gained if there were no or bad sequence results. If one or more sequencing results from the same strain showed double peaks in the promoter region the strain was considered as a double population. Finally, the promoter was noted as failed assembly (FA) if either LabChip and or sequencing results were considered not trusted, as no information (NI) if the sequencing result was no information and else as the promoter predicted by pairwise alignment between sequencing results and promoter sequence.

Measuring fluorescence and growth

Yeast cells were cultured ON to saturation, diluted to OD_{600} 0.025 (measured by reading the absorbance at 600 nm on Synergy Mx Microplate Reader, BioTek) and then cultured again in a Synergy Mx Microplate Reader. While culturing, the reader measured OD_{600} and fluorescence with excitation and emission wavelengths of 485 and 515 nm, respectively every

15 min for 20 hrs. All wells were sealed with VIEWseal membrane (Greiner Bio-One).

QUANTIFICATION AND STATISTICAL ANALYSIS

Modelling

All genotype and time series data as well as scripts for preprocessing are publicly available (see section DATA AND SOFTWARE AVAILABILITY). Briefly, all OD and GFP measurements were subtracted background signal (i.e. mean value of OD and GFP measurements in wells containing pure media). Background signals were calculated for each 96-well plate. Strains were quality-controlled based on 5 criteria. The criteria were: 1. Optical densities must cover the whole range up to 0.15 OD units to exclude uninoculated wells and wells with insufficient growth, 2. Sequencing results must exist for all five promoter gene junctions, 3. The integrated sequence must be exactly as designed, 4. The complementation plasmid must be cured, and 5. The sequencing results must not indicate the presence of multiple genotypes (Figure S5A). GFP synthesis rates were calculated in the OD_{600} interval from 0.075 to 0.150, as measured by a Synergy Mx Microplate Reader from BioTek.

In the ART approach, outliers were identified and removed based on replicate differences in GFP synthesis rate relative to the mean value for the strain. Replicates with the one percent most extreme differences were identified and the corresponding strains were removed. GFP synthesis rate was modelled as a function of promoter combination, represented through one-hot encoding, using the Automated Recommendation Tool (ART; Radivojević et al., 2019). Briefly, ART uses a probabilistic ensemble model consisting of eight individual models. The weight of each ensemble model is considered a random variable with a probability distribution characterized by the available training data, and determined through Bayesian inference and Markov Chain Monte Carlo (Brooks et al., 2011). ART uses the trained ensemble model in combination with a Parallel Tempering approach (Earl and Deem, 2005) to recommend 30 new promoter combinations (unseen designs), which are predicted to improve production. The recommended designs were chosen as the 30 strains with the highest expected GFP synthesis rate predicted by the model. This recommendation approach was labelled exploitative since predictions with high uncertainty were not prioritized, although ART can provide both exploitative and explorative recommendations

For the TeselaGen EVOLVE algorithm used in this study, outliers were identified and removed based on a method described by Rousseeuw and Hubert (2011). The decision was

made on a per strain basis taking into account replicate to mean value differences. In cases where just a single replicate was left after filtering, this replicate were excluded as well. Of the remaining strains, GFP synthesis rate were modelled as a function of promoter combination coded as categorical variables using a TeselaGen-developed machine learning algorithm based on Bayesian Optimization (Mockus, 1994). The algorithm was set-up to recommend 30 new promoter combinations (unseen designs), and designs were chosen by highest selection score. The selection score was the expected improvement (Bergstra et al., 2011), calculated based on predicted high GFP synthesis rate and the uncertainty of prediction. The approach was labelled explorative since high uncertainty weighed positively in the selection score calculation. While using EVOLVE for explorative recommendations, thereby complementing the ART approach, it should be mentioned that EVOLVE can be set up to provide both explorative and exploitative recommendations.

DATA AND SOFTWARE AVAILABILITY

The complete flux balance analysis, with additional simulation details and filtering criteria, is publicly available at https://github.com/biosustain/trp-scores. The genotype and time series datasets generated during this study are available at The Joint BioEnergy Institute's Inventory of Composable Elements (ICE; https://public-registry.jbei.org) and Experiment Data Depot (EDD; https://public-edd.jbei.org), respectively under the study 'Zhang and Petersen, et al 2019' (Ham et al., 2012; Morrell et al., 2017). The complete preprocessing and all statistical calculations are documented in a jupyter notebook, available at https://github.com/sorpet/Zhang_and_Petersen_et_al_2019. The notebook also contains the ART approach for modeling and strain recommendations. The Teselagen software is available through commercial and non-commercial licenses (https://teselagen.com).

REFERENCES

Alonso-Gutierrez, J., Kim, E.-M., Batth, T.S., Cho, N., Hu, Q., Chan, L.J.G., Petzold, C.J., Hillson, N.J., Adams, P.D., Keasling, J.D., et al. (2015). Principal component analysis of proteomics (PCAP) as a tool to direct metabolic engineering. Metab. Eng. *28*, 123–133.

Aung, H.W., Henry, S.A., and Walker, L.P. (2013). Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism. Ind. Biotechnol. *9*, 215–228.

Averesch, N.J.H., and Krömer, J.O. (2018). Metabolic Engineering of the Shikimate Pathway for

Production of Aromatics and Derived Compounds—Present and Future Strain Construction Strategies. Front. Bioeng. Biotechnol. *6*.
Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for Hyper-parameter Optimization. In Proceedings of the 24th International Conference on Neural Information Processing Systems, (USA: Curran Associates Inc.), pp. 2546–2554.
Bitinaite, J., Rubino, M., Varma, K.H., Schildkraut, I., Vaisvila, R., and Vaiskunaite, R. (2007). USER™ friendly DNA engineering and cloning method by uracil excision. Nucleic Acids Res. *35*, 1992–2002.
Braus, G.H. (1991). Aromatic amino acid biosynthesis in the yeast Saccharomyces cerevisiae: a model system for the regulation of a eukaryotic biosynthetic pathway. Microbiol. Rev. *55*, 349–370.
Breslow, D.K., Cameron, D.M., Collins, S.R., Schuldiner, M., Stewart-Ornstein, J., Newman, H.W., Braun, S., Madhani, H.D., Krogan, N.J., and Weissman, J.S. (2008). A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. Nat. Methods *5*, 711–718.
Brooks, S., Gelman, A., Jones, G.L., and Meng, X.-L. (2011). Handbook of Markov Chain Monte Carlo (CRC Press).
Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., and Collins, J.J. (2018). Next-Generation Machine Learning for Biological Networks. Cell *173*, 1581–1592.
Carbonell, P., Radivojevic, T., and García Martín, H. (2019). Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation. ACS Synth. Biol. *8*, 1474–1477.
Carro, M.S., Lim, W.K., Alvarez, M.J., Bollo, R.J., Zhao, X., Snyder, E.Y., Sulman, E.P., Anne, S.L., Doetsch, F., Colman, H., et al. (2010). The transcriptional network for mesenchymal transformation of brain tumours. Nature *463*, 318–325.
Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., et al. (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. *40*, D700–D705.
Choi, K.R., Jang, W.D., Yang, D., Cho, J.S., Park, D., and Lee, S.Y. (2019). Systems Metabolic Engineering Strategies: Integrating Systems and Synthetic Biology with Metabolic Engineering. Trends Biotechnol. *37*, 817–837.
Costello, Z., and Martin, H.G. (2018). A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. Npj Syst. Biol. Appl. *4*.
Curran, K.A., Leavitt, J.M., Karim, A.S., and Alper, H.S. (2013). Metabolic engineering of muconic acid production in Saccharomyces cerevisiae. Metab. Eng. *15*, 55–66.
Earl, D.J., and Deem, M.W. (2005). Parallel tempering: Theory, applications, and new perspectives. Phys. Chem. Chem. Phys. *7*, 3910–3916.
Feng, Y., De Franceschi, G., Kahraman, A., Soste, M., Melnik, A., Boersema, P.J., de Laureto, P.P., Nikolaev, Y., Oliveira, A.P., and Picotti, P. (2014). Global analysis of protein structural changes in complex proteomes. Nat. Biotechnol. *32*, 1036–1044.
Ferreira, R., Skrekas, C., Hedin, A., Sánchez, B.J., Siewers, V., Nielsen, J., and David, F. (2019). Model-Assisted Fine-Tuning of Central Carbon Metabolism in Yeast through dCas9-Based Regulation. ACS Synth. Biol.
Gardner, T.S. (2013). Synthetic biology: from hype to impact. Trends Biotechnol. *31*, 123–125.
Gietz, R.D., and Schiestl, R.H. (2007). Quick and easy yeast transformation using the LiAc/SS carrier DNA/PEG method. Nat. Protoc. *2*, 35–37.
Graf, R., Mehmann, B., and Braus, G.H. (1993). Analysis of feedback-resistant anthranilate synthases from Saccharomyces cerevisiae. J. Bacteriol. *175*, 1061–1068.
Gunsalus, R.P., and Yanofsky, C. (1980). Nucleotide sequence and expression of Escherichia

coli trpR, the structural gene for the trp aporepressor. Proc. Natl. Acad. Sci. U. S. A. *77*, 7117–7121.
Guzmán, G.I., Utrilla, J., Nurk, S., Brunk, E., Monk, J.M., Ebrahim, A., Palsson, B.O., and Feist, A.M. (2015). Model-driven discovery of underground metabolic functions in *Escherichia coli*. Proc. Natl. Acad. Sci. *112*, 929–934.
Ham, T.S., Dmytriv, Z., Plahar, H., Chen, J., Hillson, N.J., and Keasling, J.D. (2012). Design, implementation and practice of JBEI-ICE: an open source biological part registry platform and tools. Nucleic Acids Res. *40*, e141–e141.
Hartmann, M., Schneider, T.R., Pfeil, A., Heinrich, G., Lipscomb, W.N., and Braus, G.H. (2003). Evolution of feedback-inhibited / barrel isoenzymes by gene duplication and a single mutation. Proc. Natl. Acad. Sci. *100*, 862–867.
Hefzi, H., Ang, K.S., Hanscho, M., Bordbar, A., Ruckerbauer, D., Lakshmanan, M., Orellana, C.A., Baycin-Hizal, D., Huang, Y., Ley, D., et al. (2016). A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. Cell Syst. *3*, 434-443.e8.
Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S.N., Richelle, A., Heinken, A., Haraldsdóttir, H.S., Wachowiak, J., Keating, S.M., Vlasov, V., et al. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat. Protoc. *14*, 639–702.
Jakočiūnas, T., Rajkumar, A.S., Zhang, J., Arsovska, D., Rodriguez, A., Jendresen, C.B., Skjødt, M.L., Nielsen, A.T., Borodina, I., Jensen, M.K., et al. (2015). CasEMBLR: Cas9-Facilitated Multiloci Genomic Integration of in Vivo Assembled DNA Parts in Saccharomyces cerevisiae. ACS Synth. Biol. *4*, 1226–1234.
Jakočiūnas, T., Bonde, I., Herrgård, M., Harrison, S.J., Kristensen, M., Pedersen, L.E., Jensen, M.K., and Keasling, J.D. (2015). Multiplex metabolic pathway engineering using CRISPR/Cas9 in Saccharomyces cerevisiae. Metab. Eng. *28*, 213–222.
Jensen, N.B., Strucko, T., Kildegaard, K.R., David, F., Maury, J., Mortensen, U.H., Forster, J., Nielsen, J., and Borodina, I. (2014). EasyClone: method for iterative chromosomal integration of multiple genes in Saccharomyces cerevisiae. FEMS Yeast Res. *14*, 238–248.
Jervis, A.J., Carbonell, P., Vinaixa, M., Dunstan, M.S., Hollywood, K.A., Robinson, C.J., Rattray, N.J.W., Yan, C., Swainston, N., Currin, A., et al. (2019). Machine Learning of Designed Translational Control Allows Predictive Pathway Optimization in Escherichia coli. ACS Synth. Biol. *8*, 127–136.
Jeschek, M., Gerngross, D., and Panke, S. (2016). Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort. Nat. Commun. *7*, 11163.
Jeschek, M., Gerngross, D., and Panke, S. (2017). Combinatorial pathway optimization for streamlined metabolic engineering. Curr. Opin. Biotechnol. *47*, 142–151.
Jessop-Fabre, M.M., Jakočiūnas, T., Stovicek, V., Dai, Z., Jensen, M.K., Keasling, J.D., and Borodina, I. (2016). EasyClone-MarkerFree: A vector toolkit for marker-less integration of genes into Saccharomyces cerevisiae via CRISPR-Cas9. Biotechnol. J. *11*, 1110–1117.
Keasling, J.D. (2010). Manufacturing Molecules through Metabolic Engineering. Science *330*, 1355–1358.
Khodayari, A., Chowdhury, A., and Maranas, C.D. (2015). Succinate Overproduction: A Case Study of Computational Strain Design Using a Comprehensive Escherichia coli Kinetic Model. Front. Bioeng. Biotechnol. *2*.
Kuijpers, N.G.A., Solis-Escalante, D., Luttik, M.A.H., Bisschops, M.M.M., Boonekamp, F.J., van den Broek, M., Pronk, J.T., Daran, J.-M., and Daran-Lapujade, P. (2016). Pathway swapping: Toward modular engineering of essential cellular processes. Proc. Natl. Acad. Sci. *113*, 15060–15065.
Lahtvee, P.J., Sánchez, B.J., Smialowska, A., Kasvandik, S., Elsemman, I.E., Gatto, F., and

Nielsen, J. (2017). Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. Cell Syst. *4*, 495-504.e5.
Lee, S., Lim, W.A., and Thorn, K.S. (2013). Improved Blue, Green, and Red Fluorescent Protein Tagging Vectors for S. cerevisiae. PLoS ONE *8*, e67902.
Lewis, N.E., Hixson, K.K., Conrad, T.M., Lerman, J.A., Charusanti, P., Polpitiya, A.D., Adkins, J.N., Schramm, G., Purvine, S.O., Lopez-Ferrer, D., et al. (2010). Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol. Syst. Biol. *6*.
Lewis, N.E., Nagarajan, H., and Palsson, B.O. (2012). Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. Nat. Rev. Microbiol. *10*, 291–305.
Lingens, F., Goebel, W., and Uesseler, H. (1967). Regulation der Biosynthese der aromatischen Aminosäuren in Saccharomyces cerevisiae. Eur. J. Biochem. *1*, 363–374.
Liu, Y., and Nielsen, J. (2019). Recent trends in metabolic engineering of microbial chemical factories. Curr. Opin. Biotechnol. *60*, 188–197.
Liu, H., Krizek, J., and Bretscher, A. (1992). Construction of a GAL1-regulated yeast cDNA expression library and its application to the identification of genes whose overexpression causes lethality in yeast. Genetics *132*, 665–673.
Long, C.P., and Antoniewicz, M.R. (2019). Metabolic flux responses to deletion of 20 core enzymes reveal flexibility and limits of E. coli metabolism. Metab. Eng.
Lõoke, M., Kristjuhan, K., and Kristjuhan, A. (2011). Extraction of genomic DNA from yeasts for PCR-based applications. BioTechniques *50*, 325–328.
Lu, H., Li, F., Sánchez, B.J., Zhu, Z., Li, G., Domenzain, I., Marcišauskas, S., Anton, P.M., Lappa, D., Lieven, C., et al. (2019). A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nat. Commun. *10*.
Luo, H., Hansen, A.S.L., Yang, L., Schneider, K., Kristensen, M., Christensen, U., Christensen, H.B., Du, B., Özdemir, E., Feist, A.M., et al. (2019). Coupling S-adenosylmethionine–dependent methylation to growth: Design and uses. PLOS Biol. *17*, e2007050.
Mahr, R., and Frunzke, J. (2016). Transcription factor-based biosensors in biotechnology: current state and future prospects. Appl. Microbiol. Biotechnol. *100*, 79–90.
Makanae, K., Kintaka, R., Makino, T., Kitano, H., and Moriya, H. (2013). Identification of dosage-sensitive genes in Saccharomyces cerevisiae using the genetic tug-of-war method. Genome Res. *23*, 300–311.
Mellor, J., Grigoras, I., Carbonell, P., and Faulon, J.-L. (2016). Semisupervised Gaussian Process for Automated Enzyme Search. ACS Synth. Biol. *5*, 518–528.
Mockus, J. (1994). Application of Bayesian approach to numerical methods of global and stochastic optimization. J. Glob. Optim. *4*, 347–365.
Monk, J.M., Lloyd, C.J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., et al. (2017). iML1515, a knowledgebase that computes Escherichia coli traits. Nat. Biotechnol. *35*, 904–908.
Morrell, W.C., Birkel, G.W., Forrer, M., Lopez, T., Backman, T.W.H., Dussault, M., Petzold, C.J., Baidoo, E.E.K., Costello, Z., Ando, D., et al. (2017). The Experiment Data Depot: A Web-Based Software Tool for Biological Experimental Data Storage, Sharing, and Visualization. ACS Synth. Biol. *6*, 2248–2259.
Nielsen, J., and Keasling, J.D. (2016). Engineering Cellular Metabolism. Cell *164*, 1185–1197.
Orth, J.D., Thiele, I., and Palsson, B.Ø. (2010). What is flux balance analysis? Nat. Biotechnol. *28*, 245–248.

Park, S.H., Kim, H.U., Kim, T.Y., Park, J.S., Kim, S.-S., and Lee, S.Y. (2014). Metabolic engineering of Corynebacterium glutamicum for L-arginine production. Nat. Commun. *5*.
Patnaik, R., and Liao, J.C. (1994). Engineering of Escherichia coli central metabolism for aromatic metabolite production with near theoretical yield. Appl. Environ. Microbiol. *60*, 3903–3908.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 6.
Presnell, K.V., and Alper, H.S. (2019). Systems Metabolic Engineering Meets Machine Learning: A New Era for Data-Driven Metabolic Engineering. Biotechnol. J. *0*, 1800416.
Radivojević, T., Costello, Z., and Martin, H.G. (2019). ART: A machine learning Automated Recommendation Tool for synthetic biology. ArXiv191111091 Q-Bio Stat.
Rajkumar, A.S., Özdemir, E., Lis, A.V., Schneider, K., Qin, J., Jensen, M.K., and Keasling, J.D. (2019). Engineered Reversal of Function in Glycolytic Yeast Promoters. ACS Synth. Biol. *8*, 1462–1468.
Reider Apel, A., d'Espaux, L., Wehrs, M., Sachs, D., Li, R.A., Tong, G.J., Garber, M., Nnadi, O., Zhuang, W., Hillson, N.J., et al. (2017). A Cas9-based toolkit to program gene expression in Saccharomyces cerevisiae. Nucleic Acids Res. *45*, 496–508.
Rhee, H.S., and Pugh, B.F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature *483*, 295–301.
Rodriguez, A., Kildegaard, K.R., Li, M., Borodina, I., and Nielsen, J. (2015). Establishment of a yeast platform strain for production of p-coumaric acid through metabolic engineering of aromatic amino acid biosynthesis. Metab. Eng. *31*, 181–188.
Roesser, J.R., and Yanofsky, C. (1991). The effects of leader peptide sequence and length on attenuation control of the trp operon of E.coli. Nucleic Acids Res. *19*, 795–800.
Rogers, J.K., Taylor, N.D., and Church, G.M. (2016). Biosensor-based engineering of biosynthetic pathways. Curr. Opin. Biotechnol. *42*, 84–91.
Rousseeuw, P.J., and Hubert, M. (2011). Robust statistics for outlier detection: Robust statistics for outlier detection. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. *1*, 73–79.
Schläpfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., Dreher, K., Chavali, A.K., Nilo-Poyanco, R., Bernard, T., et al. (2017). Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. Plant Physiol. *173*, 2041–2059.
Sikorski, R.S., and Hieter, P. (1989). A System of Shuttle Vectors and Yeast Host Strains Designed for Efficient Manipulation of DNA in Saccharomyces Cerevisiae. Genetics *122*, 19–27.
Stephanopoulos, G. (1999). Metabolic Fluxes and Metabolic Engineering. Metab. Eng. *1*, 1–11.
Suástegui, M., and Shao, Z. (2016). Yeast factories for the production of aromatic compounds: from building blocks to plant secondary metabolites. J. Ind. Microbiol. Biotechnol. *43*, 1611–1624.
TeselaGen (2019). TeselaGen Technology including EVOLVE module.
Vogt, M., Haas, S., Klaffl, S., Polen, T., Eggeling, L., van Ooyen, J., and Bott, M. (2014). Pushing product formation to its limit: Metabolic engineering of Corynebacterium glutamicum for l-leucine overproduction. Metab. Eng. *22*, 40–52.
Wolpert, D.H. (1996). The Lack of A Priori Distinctions Between Learning Algorithms. Neural Comput. *8*, 1341–1390.
Yang, J., Gunasekera, A., Lavoie, T.A., Jin, L., Lewis, D.E.A., and Carey, J. (1996). In vivo and in vitro Studies of TrpR-DNA Interactions. J. Mol. Biol. *258*, 37–52.
Yang, J.E., Park, S.J., Kim, W.J., Kim, H.J., Kim, B.J., Lee, H., Shin, J., and Lee, S.Y. (2018).

One-step fermentative production of aromatic polyesters from glucose by metabolically engineered Escherichia coli strains. Nat. Commun. *9*.
Yin, Z. (1996). Multiple signalling pathways trigger the exquisite sensitivity of yeast gluconeogenic mRNAs to glucose. Mol. Microbiol. *20*, 751–764.
Zampieri, G., Vijayakumar, S., Yaneske, E., and Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling. PLOS Comput. Biol. *15*, e1007084.
Zhang, J., Sonnenschein, N., Pihl, T.P.B., Pedersen, K.R., Jensen, M.K., and Keasling, J.D. (2016). Engineering an NADPH/NADP $^{+}$ Redox Biosensor in Yeast. ACS Synth. Biol. *5*, 1546–1556

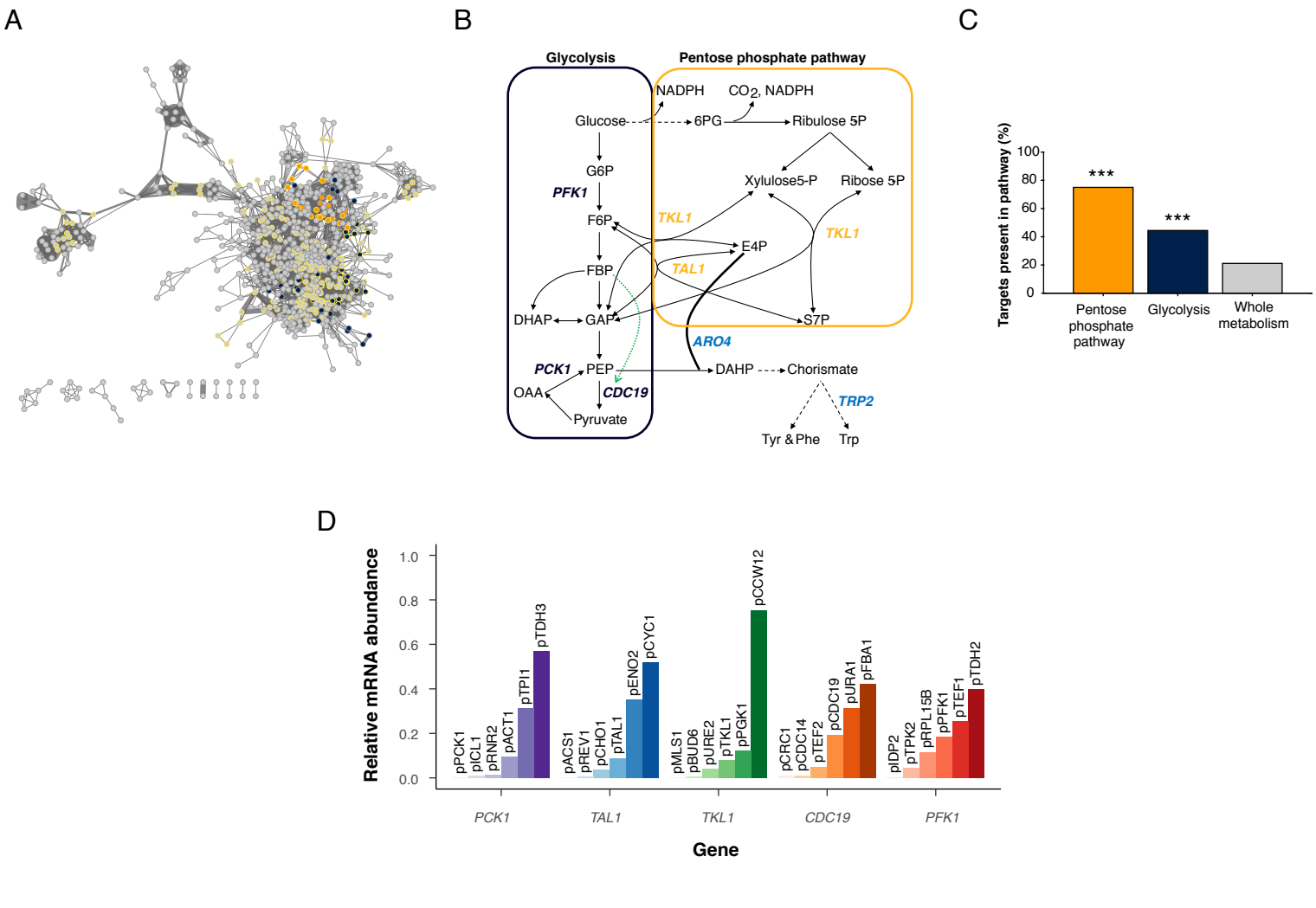
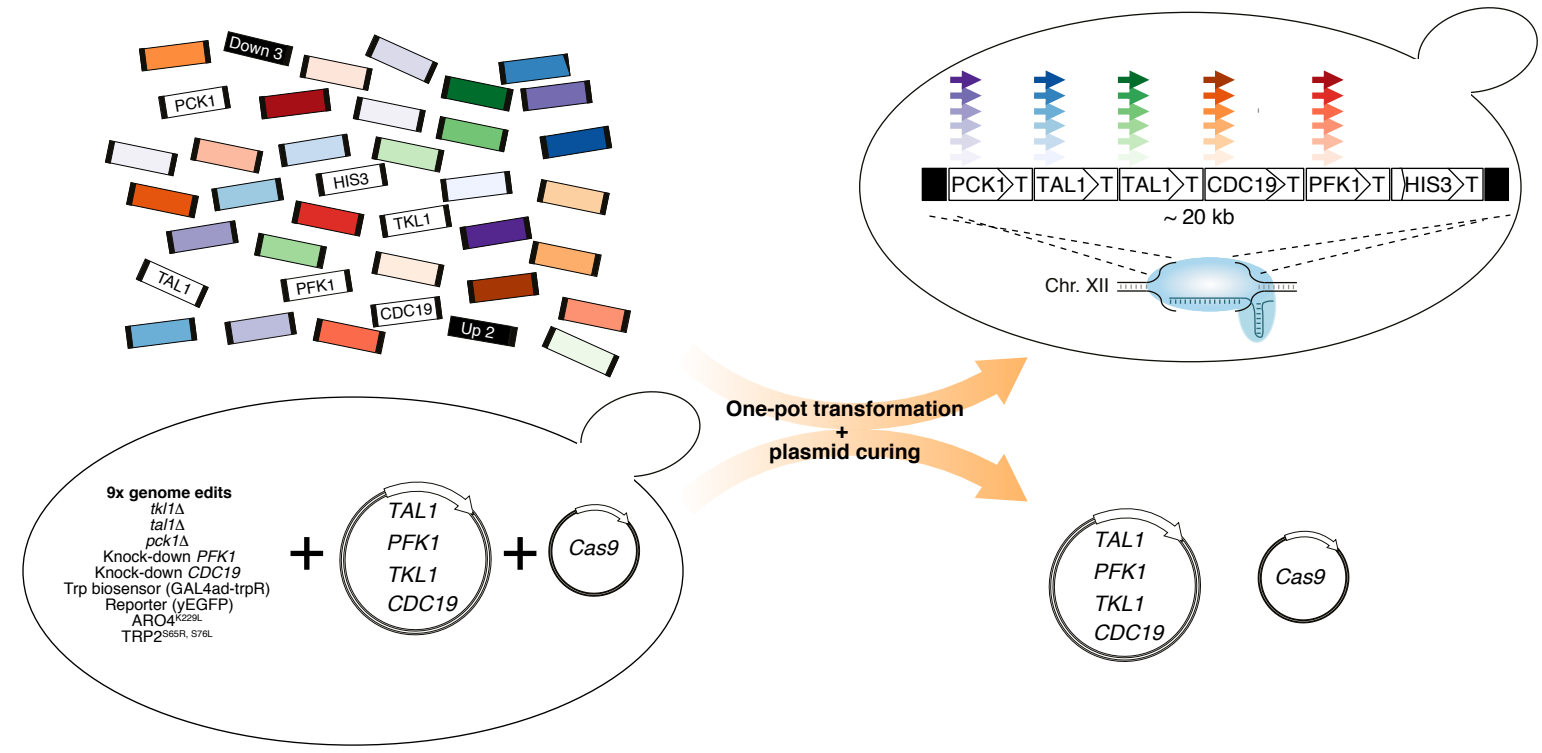

A
B
Glycolysis
Pentose phosphate pathway
NADPH
CO2, NADPH
Glucose
6PG
Ribulose 5P
G6P
PFK1
Xylulose5-P
Ribose 5P
F6P
TKL1
TKL1
E4P
FBP
TAL1
DHAP
GAP
S7P
ARO4
PCK1
PEP
DAHP
Chorismate
OAA
CDC19
TRP2
Pyruvate
Tyr & Phe
Trp
C
Targets present in pathway (%)
100
80
60
40
20
0
***
***
Pentose phosphate pathway
Glycolysis
Whole metabolism
D
Relative mRNA abundance
1.0
0.8
0.6
0.4
0.2
0.0
pPCK1
pICL1
pRNR2
pACT1
pTPI1
pTDH3
pACS1
pREV1
pCHO1
pTAL1
pENO2
pCYC1
pMLS1
pBUD6
pURE2
pTKL1
pPGK1
pCCW12
pCRC1
pCDC14
pTEF2
pCDC19
pURA1
pFBA1
pIDP2
pTPK2
pRPL15B
pPFK1
pTEF1
pTDH2
PCK1
TAL1
TKL1
CDC19
PFK1
Gene

Figure 1

A



B

Potential unique genotypes	7,776
Library colonies	~ 10,000
Library sample	480
Plasmid cured strains	92%
Correct assembly	82%
Repeated genotypes	3.7%

C

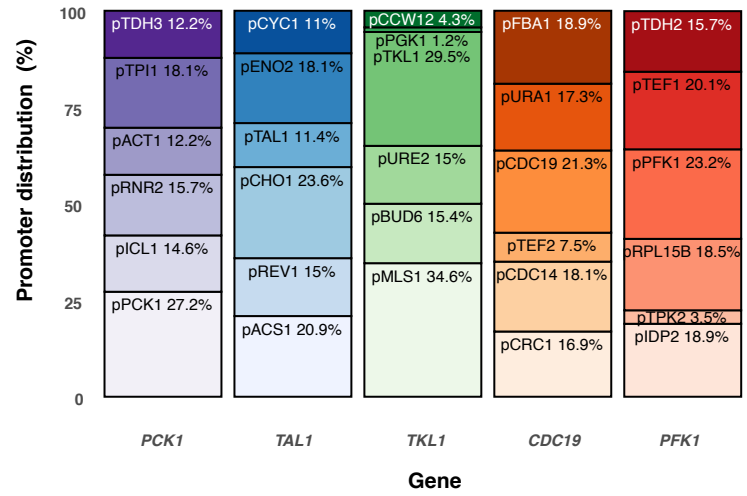


Figure 2

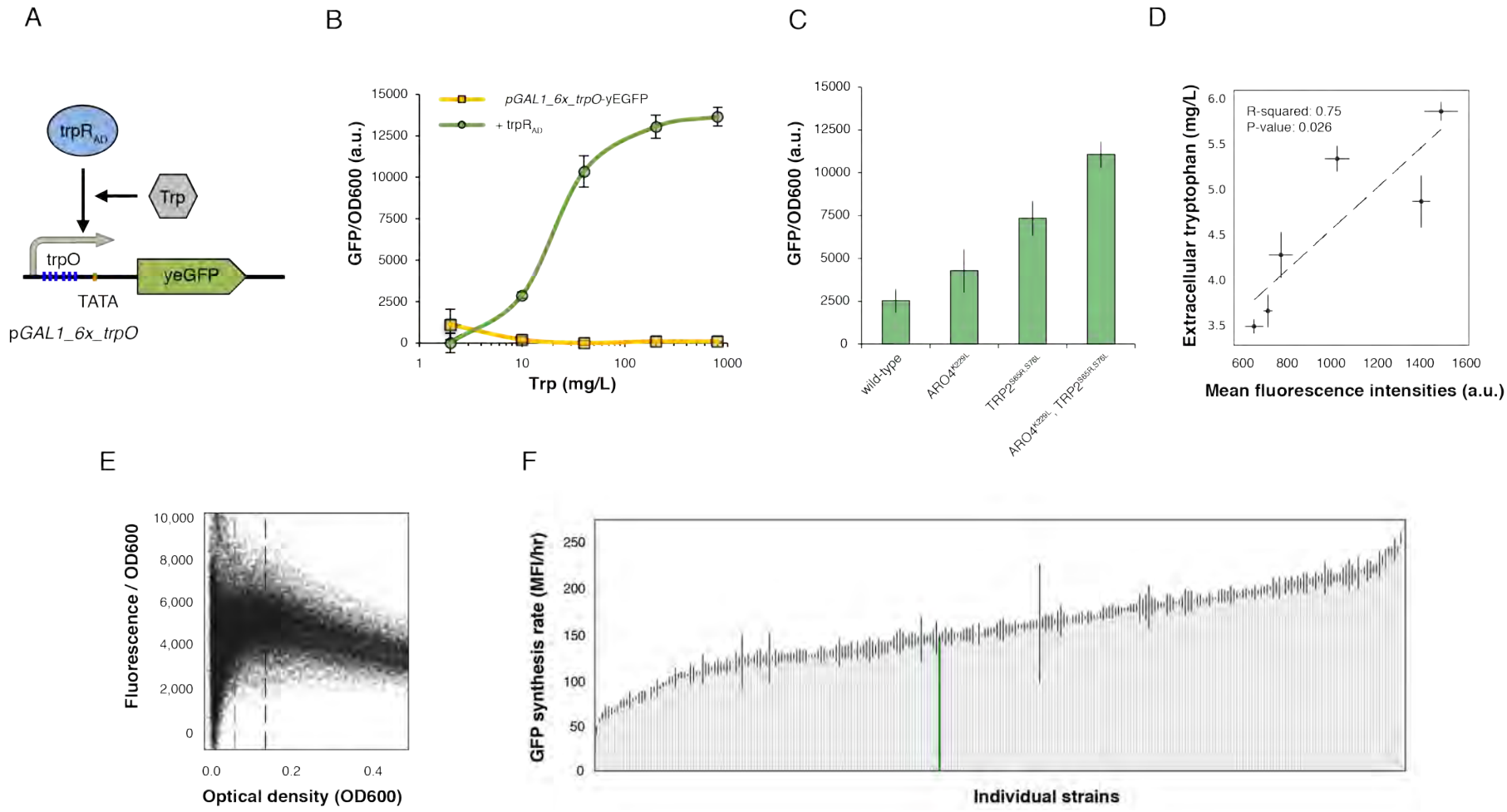

A
trpR_AD
Trp
trpO
yeGFP
TATA
pGAL1_6x_trpO
B
pGAL1_6x_trpO-yEGFP
+ trpR_AD
GFP/OD600 (a.u.)
Trp (mg/L)
C
GFP/OD600 (a.u.)
wild-type
ARO4K229L
TRP2S86R,S76L
ARO4K229L, TRP2S86R,S76L
D
R-squared: 0.75
P-value: 0.026
Extracellular tryptophan (mg/L)
Mean fluorescence intensities (a.u.)
E
Fluorescence / OD600
Optical density (OD600)
F
GFP synthesis rate (MFI/hr)
Individual strains

Figure 3

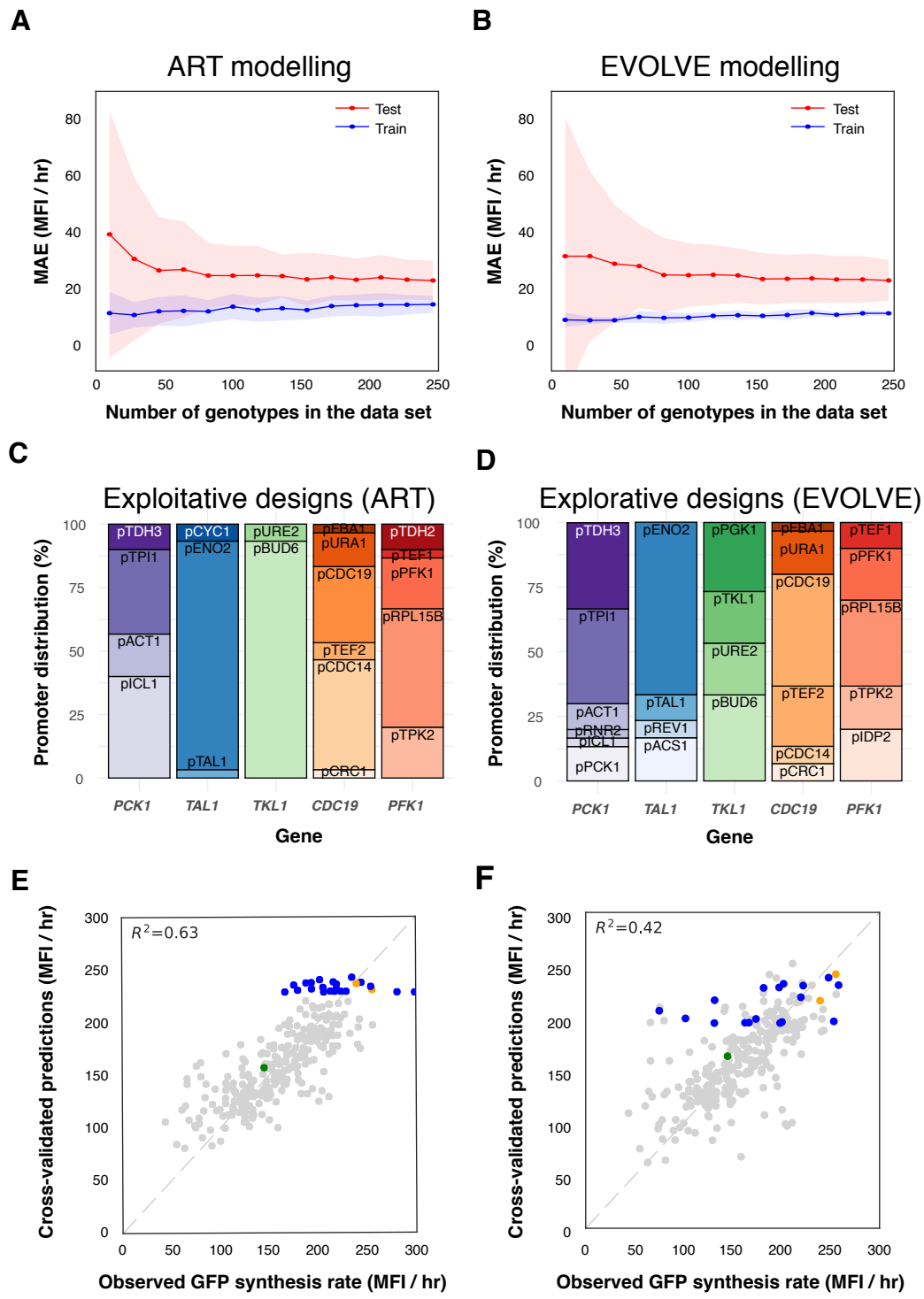

A
ART modelling
Test
Train
MAE (MFI / hr)
Number of genotypes in the data set
B
EVOLVE modelling
Test
Train
MAE (MFI / hr)
Number of genotypes in the data set
C
Exploitative designs (ART)
Promoter distribution (%)
pTDH3
pTPI1
pACT1
pICL1
pCYC1
pENO2
pTAL1
pURE2
pBUD6
pFBA1
pURA1
pCDC19
pTEF2
pCDC14
pCRC1
pTDH2
pTEF1
pPFK1
pRPL15B
pTPK2
PCK1
TAL1
TKL1
CDC19
PFK1
Gene
D
Explorative designs (EVOLVE)
Promoter distribution (%)
pTDH3
pTPI1
pACT1
pRNR2
pICL1
pPCK1
pENO2
pTAL1
pREV1
pACS1
pPGK1
pTKL1
pURE2
pBUD6
pFBA1
pURA1
pCDC19
pTEF2
pCDC14
pCRC1
pTEF1
pPFK1
pRPL15B
pTPK2
pIDP2
PCK1
TAL1
TKL1
CDC19
PFK1
Gene
E
R^2=0.63
Cross-validated predictions (MFI / hr)
Observed GFP synthesis rate (MFI / hr)
F
R^2=0.42
Cross-validated predictions (MFI / hr)
Observed GFP synthesis rate (MFI / hr)

Figure 4

3.1 Supplementary material and extended methods

Predictive engineering and optimization of tryptophan metabolism in yeast through a combination of mechanistic and machine learning models

Jie Zhang[1#], Søren D. Petersen[1#], Tijana Radivojevic[2,5,8], Andrés Ramirez[3], Andrés Pérez[3], Eduardo Abeliuk[4], Benjamín J. Sánchez[1], Zachary Costello[2,5,8], Yu Chen[9,10], Mike Fero[4], Hector Garcia Martin[2,5,8,10], Jens Nielsen[1,9,12] , Jay D. Keasling[1-2,5-7], & Michael K. Jensen[1*]

[1] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kgs. Lyngby, Denmark

[2] Joint BioEnergy Institute, Emeryville, CA, USA

[3] TeselaGen SpA, Santiago, Chile

[4] TeselaGen Biotechnology, San Francisco, CA 94107, USA

[5] Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[6] Department of Chemical and Biomolecular Engineering & Department of Bioengineering, University of California, Berkeley, CA, USA

[7] Center for Synthetic Biochemistry, Institute for Synthetic Biology, Shenzhen Institutes of Advanced Technologies, Shenzhen, China

[8] DOE Agile BioFoundry, Emeryville, CA, USA

[9] Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden

[10] Novo Nordisk Foundation Center for Biosustainability, Chalmers University of Technology, Gothenburg, Sweden

[11] BCAM, Basque Center for Applied Mathematics, Bilbao, Spain

[12] BioInnovation Institute, Ole Maaløes Vej 3, DK-2200 Copenhagen N, Denmark

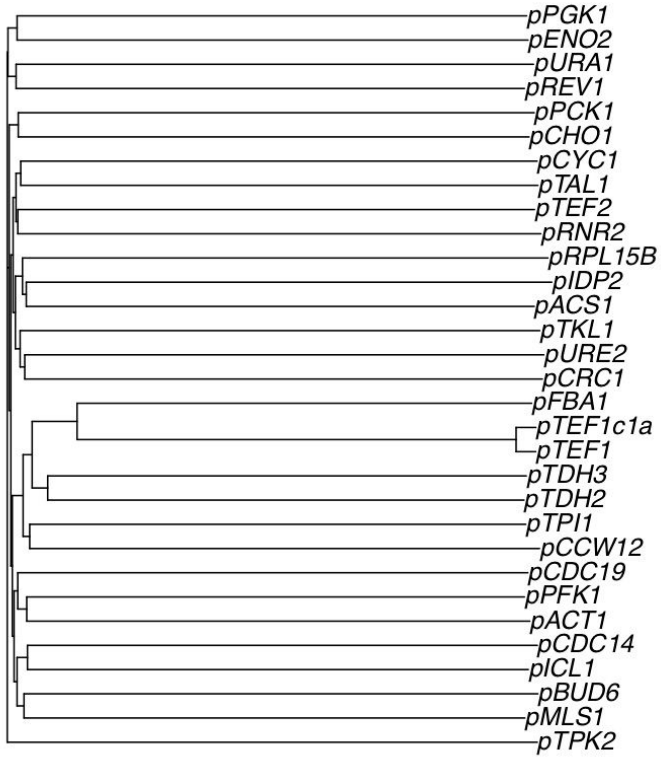


Figure S1. Related to Figure 1. Dendrogram of the sequence diversity of 30 selected native yeast promoters. Sequence pTEF1c1a with a single nucleotide change from pTEF1 has been added as a reference. The dendrogram was constructed using the neighbor-joining method (Saitou and Nei, 1987; Studier and Keppler, 1988).

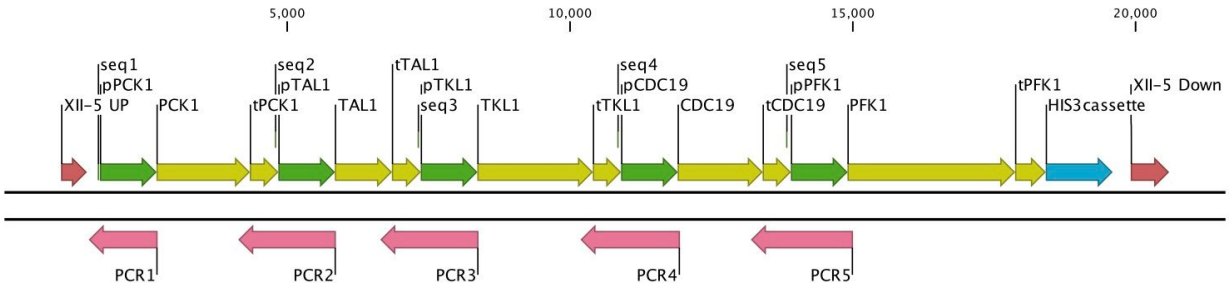


Figure S2. Related to Figure 1. Genotyping strategy. Schematic outline of the genotyping strategy to assess correct *in vivo* junction-junction assemblies of 11 parts, and the integration at EasyClone site XII-5 (Jensen et al., 2014). Marked in red are chromosomal regions of EasyClone site XII-5, whereas green marks the promoters, and yellow the coding sequences and terminators. Marked in blue is the selectable *HIS3* expression cassette, while genotyping PCRs are marked in light red. Primers used for sequencing of the 5 PCR reactions are marked seq1-seq5.

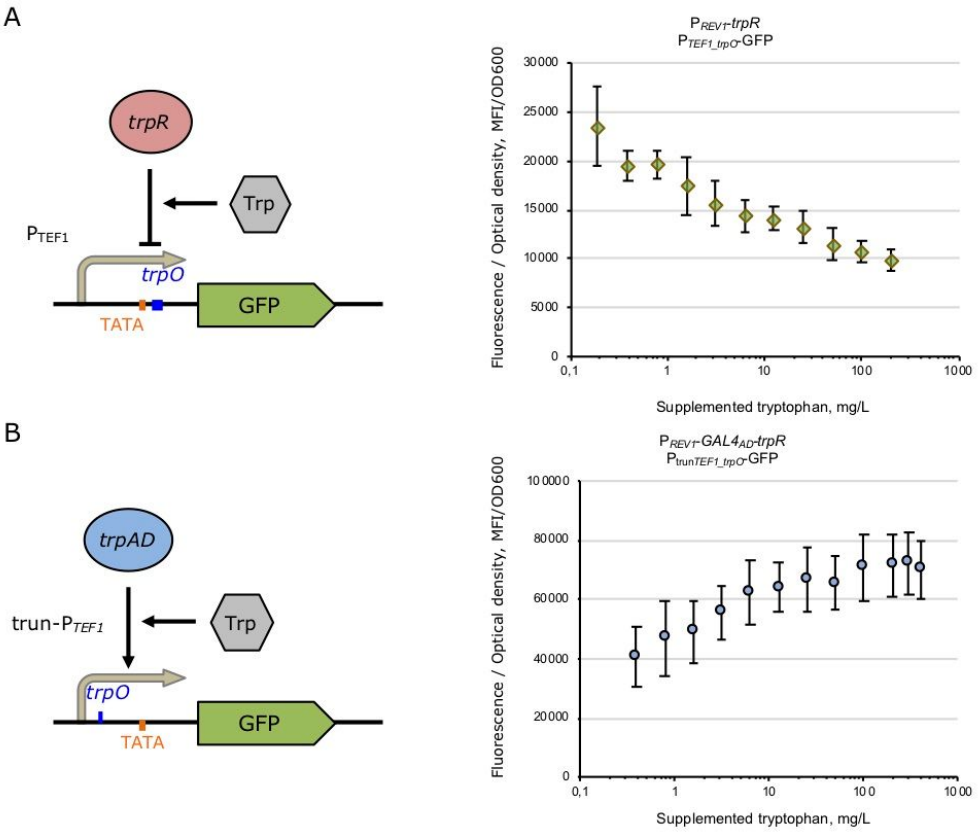


Figure S3. Related to Figure 3. Biosensor development and characterization. Overnight cultures of the strain containing sensor and reporter was used to inoculate fresh media supplemented with various concentrations of tryptophan and grown for 6 hours (early-mid exponential phase). Optical density (measured as absorbance at 600 nm) was used to normalize the green fluorescence (excitation/emission at 485/515 nm). (A) *E. coli trpR* was directly expressed in a yeast strain harboring the yEGFP reporter under the control of *TEF1* promoter containing *trpO* sequence inserted downstream of the TATA-like element. (B) The *trpR* gene was fused to the C-terminus of the activator domain of GAL4 ($GAL4_{ad}$) with a GSGSGS linker, turning this transcriptional repressor into an activator (trpAD). Accordingly, the *trpO* sequence was placed upstream of a truncated *TEF1* promoter (lacking region with multiple Rap1-binding sites).

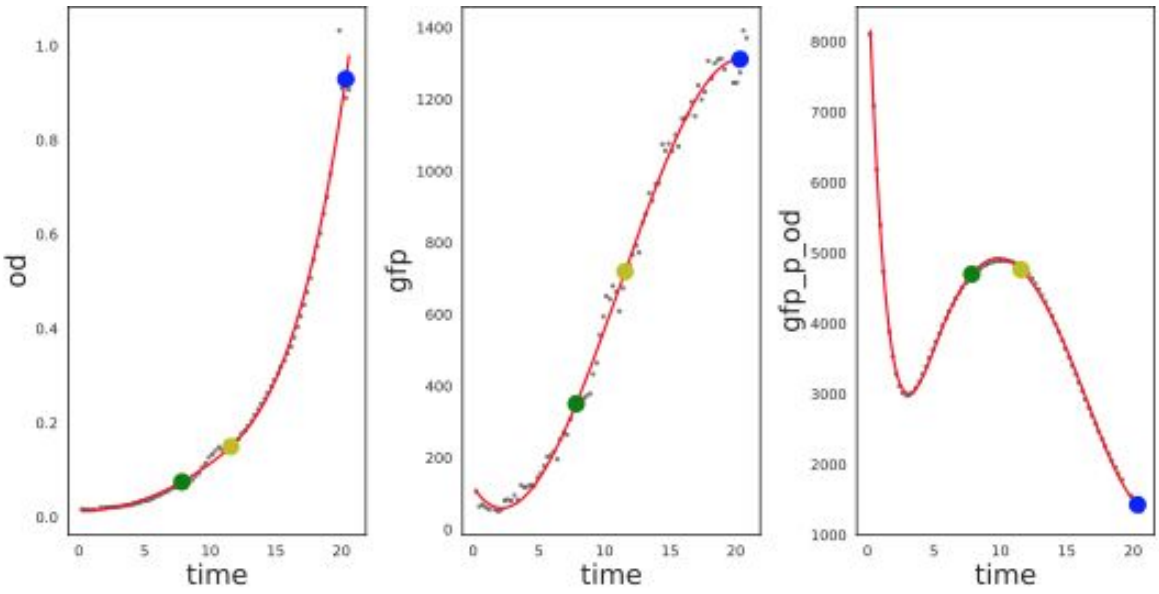


Figure S4. Related to Figure 3E-F. Parameter estimation from time series data. (A) Representative growth curve of *S. cerevisiae* in microtiter plates. *S. cerevisiae* was grown in yeast synthetic drop-out media in 96-well microtiter plates, and cell density measured at 600 nm (OD_{600}) over 20 hrs. (B) Representative tryptophan biosensor output measured as fluorescence (GFP) in *S. cerevisiae* cells (n = 1). *S. cerevisiae* was grown in yeast synthetic drop-out media in 96-well microtiter plates, and GFP measured at 485 nm (OD_{485}) over 20 hrs. (C) Tryptophan biosensor output normalized by absorbance at 600 nm (OD_{600}) over 20 hrs. For (A-C) the red line shows model fitting using a univariate spline. All plots represent a single replicate measurement (n = 1). The green, yellow and blue markers indicate OD_{600} = 0.075, OD_{600} = 0.15, and maximum rate of OD_{600} increase, respectively.

When calculating GFP synthesis rates (increase in GFP/time) we normalized our measurements with the number of cells (GFP/OD_{600}/time), because it is not possible to inoculate the medium with exactly the same number of cells. In order to calculate normalized rates (GFP/OD_{600}) we measured both OD_{600} and GFP over time for all >500 strains. We only calculated rates in the period when GFP/OD_{600} was fairly constant and high (Figure 3E, Figure S4). Here, we observed that this was the case in the early part of the exponential phase, i.e. not in the entire exponential growth phase. From this, we observed that increase in GFP/time declined before OD_{600}/time. This is considered to be due GFP maturation being more sensitive to oxygen than to cell growth. Picking the correct period for calculating rates was necessary to make sure that we got the actual strain characteristics, and not biases due to the specific laboratory setup (e.g. that the

cells begin to shade one another at high OD_{600} and thereby limit detection of GFP, or due to feedback degradation of GFP). By ensuring this we achieved high reproducibility, and thus a higher signal to noise ratio (Figure 3F).

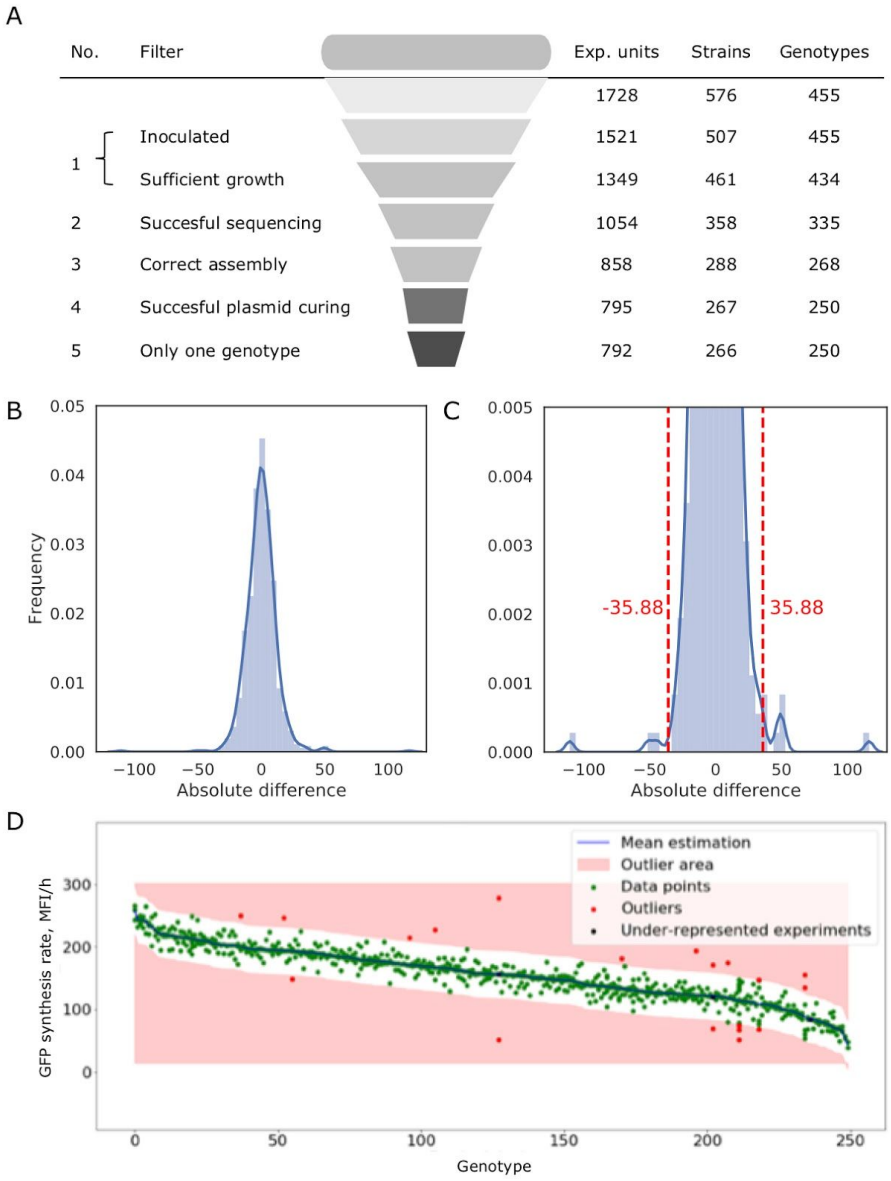


Figure S5. Related to Figures 3-4. Data filtering and outlier removal. (A) Schematic illustration of the various filtering steps applied for data quality control. The six steps used for filtering are indicated by number to the left, and listed to the right are the numbers of unique genotypes as inferred from sequencing, the number of strains, and the number of experimental units (Exp. units, n = 3). (B) The distribution of absolute differences between replicate measurements (n = 3) of strain GFP synthesis rate. (C) Same as in (B), but with y-axis expanded by a factor 10. For (B-C) the dashed red lines delimits the 1% most extreme differences between replicates which were removed in the ART modelling approach. (D) GFP

synthesis rate compared to strain genotype (n = 3). The data is ordered according to decreasing mean GFP synthesis rate. Data points included in the TeselaGen EVOLVE modeling approach are shown in green, whereas data points in red or black were excluded. Red markers indicate outliers whereas black markers indicates strains for which only one replicate is left after outlier removal.

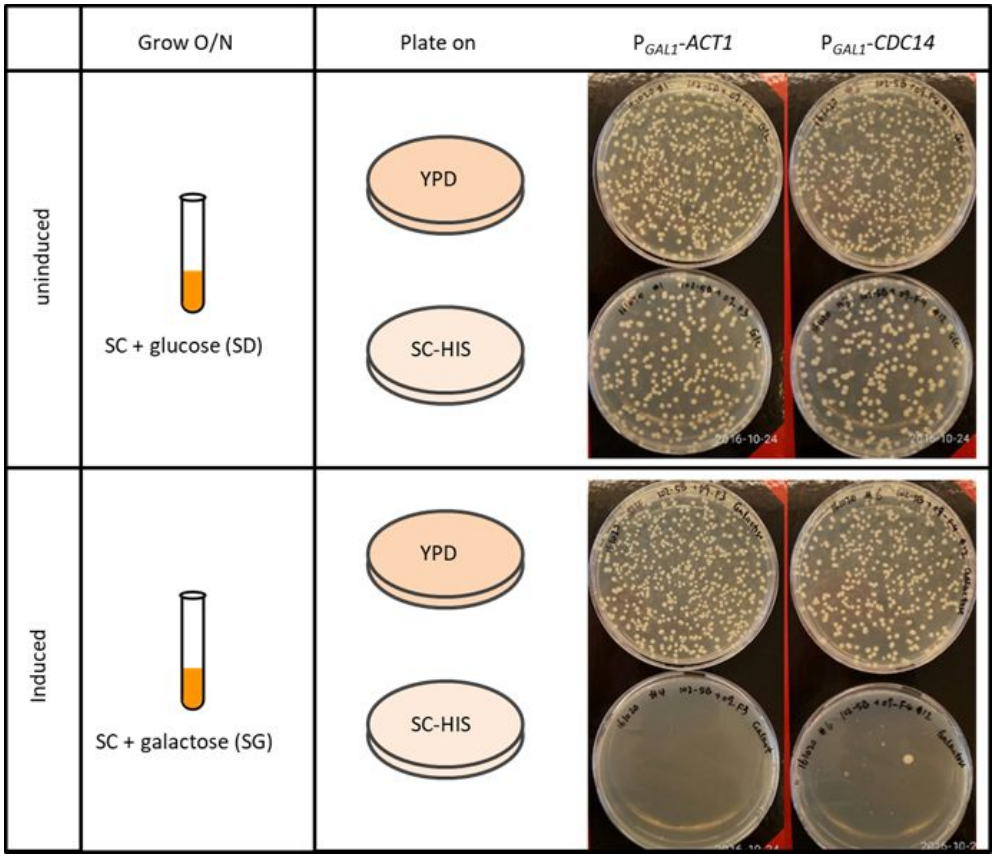


Figure S6. Construction of an easy-curable plasmid using counter selection. Two dosage sensitive genes (*ACT1* & *CDC14*) were expressed under the control of the galactose-inducible *GAL1* promoter and cloned into USER vector pRS413-mKate2 (pCfB2866, Zhang et al., ACS Synth Biol). To test the efficiency of counter selection, yeast strain with a plasmid containing one of the counter selection cassettes (pRS413-HIS3 P_{GAL1}-*ACT1*-T_{IDP1} or P_{GAL1}-*CDC14*-T_{ADH1}) was grown in both non-induction (synthetic complete + glucose) and induction (synthetic complete + galactose) media for 18 hrs. A diluted aliquot of culture was spread onto both YPD (without selection for the *HIS3* selectable marker) and SC-HIS (with selection for the *HIS3* selectable marker) drop out agar plates. Only cultures without growth on SC-HIS selective media were used for further studies.

Table S1. Primers used in study. Sequence features of interest are separated by a space.

Name	Sequence (5' - 3')	Use
Library construction		
tADH1-pPCK1_Fw	CTTATTTAGAAGTGTCAACAACGTATCTAC CACATGTCGACGAGTTT	Forward pPCK1, overhang to EasyClone site XII-5 UP
PCK1_N-pPCK1_Rv	TACTGTAGCATTCATTTTAGAAGGGGACAT GTTGTTATTTTATTATGGAATAATTAGT	Reverse pPCK1, overhang to PCK1 N-terminal
tADH1-pTPI1_Fw	CTTATTTAGAAGTGTCAACAACGTATCTAC AAGGATGAGCCAAGAATAA	Forward pTPI1, overhang to EasyClone site XII-5 UP
PCK1_N-pTPI1_Rv	TACTGTAGCATTCATTTTAGAAGGGGACAT TTTTAGTTTATGTATGTGTTTTTTG	Reverse pTPI1, overhang to PCK1 N-terminal
tADH1-pICL1_Fw	CTTATTTAGAAGTGTCAACAACGTATCTAC TTGGAAATGTAAAGGATAAT G	Forward pICL1, overhang to EasyClone site XII-5 UP
PCK1_N-pICL1_Rv	TACTGTAGCATTCATTTTAGAAGGGGACAT TTTTCGTTGACTTTTTGTTAT	Reverse pICL1, overhang to PCK1 N-terminal
tADH1-pRNR2_Fw	CTTATTTAGAAGTGTCAACAACGTATCTAC TTCTTTATCTTTTTTTCCCTT	Forward pRNR2, overhang to EasyClone site XII-5 UP
PCK1_N-pRNR2_Rv	TACTGTAGCATTCATTTTAGAAGGGGACAT GGTAATTGGACAAATAAATACG	Reverse pRNR2, overhang to PCK1 N-terminal
tADH1-pACT1_Fw	CTTATTTAGAAGTGTCAACAACGTATCTAC ACCCATATAATATAATAACTAAATAAGTAA	Forward pACT1, overhang to EasyClone site XII-5 UP
PCK1_N-pACT1_Rv	TACTGTAGCATTCATTTTAGAAGGGGACAT ACCAGAACCGTTATCAATAAC	Reverse pACT1, overhang to PCK1 N-terminal
tADH1-pTHD3_Fw	CTTATTTAGAAGTGTCAACAACGTATCTAC CTATTTTCGAGGACCTTGT	Forward pTDH3, overhang to EasyClone site XII-5 UP
PCK1_N-pTHD3_Rv	TACTGTAGCATTCATTTTAGAAGGGGACAT TTTGTTTGTTTATGTGTGTTTAT	Reverse pTDH3, overhang to PCK1 N-terminal
tPCK1-pTAL1_Fw	ACTTGTATCGAATCACTTTACCGTTCTTTA TCTATACCCATTGATCGG	Forward pTAL1, overhang to tPCK1
TAL1_N-pTAL1_Rv	CTTTTGTTTCTTTTGAGCTGGTTCAGACAT TATGTACACGTATATGTGACGA	Reverse pTAL1, overhang to TAL1 N-terminal
tPCK1-pENO2_Fw	ACTTGTATCGAATCACTTTACCGTTCTTTA AATCCTACTCTTGCCGTT	Forward pENO2, overhang to tPCK1
TAL1_N-pENO2_Rv	CTTTTGTTTCTTTTGAGCTGGTTCAGACAT TATTATTGTATGTTATAGTATTAGTTGCTT	Reverse pENO2, overhang to TAL1 N-terminal
tPCK1-pACS1_Fw	ACTTGTATCGAATCACTTTACCGTTCTTTA TGGTCTGCAAATGCTTT	Forward pACS1, overhang to tPCK1
TAL1_N-pACS1_Rv	CTTTTGTTTCTTTTGAGCTGGTTCAGACAT AACACAGTGGGGCAAT	Reverse pACS1, overhang to TAL1 N-terminal

tPCK1-pREV1_Fw	ACTTGTATCGAATCACTTTACCGTTCTTTA TTCTTAGGCACAACAATATTTATA	Forward pREV1, overhang to tPCK1
TAL1_N-pREV1_Rv	CTTTTGTTTCTTTTGAGCTGGTTCAGACAT CGCTGGATATGCCTAGA	Reverse pREV1, overhang to TAL1 N-terminal
tPCK1-pCHO1_Fw	ACTTGTATCGAATCACTTTACCGTTCTTTA AAAAAGCGAATCATGTAGAT	Forward pCHO1, overhang to tPCK1
TAL1_N-pCHO1_Rv	CTTTTGTTTCTTTTGAGCTGGTTCAGACAT AACGTCTGTGTCCGTG	Reverse pCHO1, overhang to TAL1 N-terminal
tPCK1-pCYC1_Fw	ACTTGTATCGAATCACTTTACCGTTCTTTA CAGCATTTTCAAAGGTGT	Forward pCYC1, overhang to tPCK1
TAL1_N-pCYC1_Rv	CTTTTGTTTCTTTTGAGCTGGTTCAGACAT TATTAATTTAGTGTGTGTATTTGTGT	Reverse pCYC1, overhang to TAL1 N-terminal
tTAL1-pTKL1_Fw	CCATTTTGACCTGAATCAAACGAATGAATG ATACTGGACTTGGAAATTCC	Forward pTKL1, overhang to tTAL1
TKL1_N-pTKL1_Rv	TAGCTTATCAATGTCAGTGAATTGAGTCAT TTTGTTTGCTAAAGAGGTAACT	Reverse pTKL1, overhang to TKL1 N-terminal
tTAL1-pPGK1_Fw	CCATTTTGACCTGAATCAAACGAATGAATG AGACGCGAATTTTTCGA	Forward pPGK1, overhang to tTAL1
TKL1_N-pPGK1_Rv	TAGCTTATCAATGTCAGTGAATTGAGTCAT TGTTTTATATTTGTTGTAAAAAGTAGATAA	Reverse pPGK1, overhang to TKL1 N-terminal
tTAL1-pMLS1_Fw	CCATTTTGACCTGAATCAAACGAATGAATG TTTAATCTTTAGGGAGGGTAA	Forward pMLS1, overhang to tTAL1
TKL1_N-pMLS1_Rv	TAGCTTATCAATGTCAGTGAATTGAGTCAT TTTCTTAATTCTTTTATGTGCTT	Reverse pMLS1, overhang to TKL1 N-terminal
tTAL1-pBUD6_Fw	CCATTTTGACCTGAATCAAACGAATGAATG CTTTTGAAGACTGCTGCT	Forward pBUD6, overhang to tTAL1
TKL1_N-pBUD6_Rv	TAGCTTATCAATGTCAGTGAATTGAGTCAT CTAATTTTAAAATAATACGAGGATTAC	Reverse pBUD6, overhang to TKL1 N-terminal
tTAL1-pURE2_Fw	CCATTTTGACCTGAATCAAACGAATGAATG CAAGCTGAACTCGCTGA	Forward pURE2, overhang to tTAL1
TKL1_N-pURE2_Rv	TAGCTTATCAATGTCAGTGAATTGAGTCAT TTGGTGTAGAACTTAATTTGC	Reverse pURE2, overhang to TKL1 N-terminal
tTAL1-pCCW12_Fw	CCATTTTGACCTGAATCAAACGAATGAATG AAAGAAACTTAATACGTTATGCC	Forward pCCW12, overhang to tTAL1
TKL1_N-pCCW12_Rv	TAGCTTATCAATGTCAGTGAATTGAGTCAT TATTGATATAGTGTTTAAGCGAA	Reverse pCCW12, overhang to TKL1 N-terminal
tTKL1-pCDC19_Fw	AGTCGAAAAGGCTAATCTAGAAAATCGATT ACTTGAGATGTGTGTCAATG	Forward pCDC19, overhang to tTKL1
CDC19_N-pCDC19_Rv	TAATGAGGTCAATCTTTCTAATCTAGACAT TGTGATGATGTTTTATTTGTTT	Reverse pCDC19, overhang to CDC19 N-terminal

tTKL1-pURA1_Fw	AGTCGAAAAGGCTAATCTAGAAAATCGATT GTTGTATTAATTTTCTCGAAGG	Forward pURA1, overhang to tTKL1
CDC19_N-pURA1_Rv	TAATGAGGTCAATCTTTCTAATCTAGACAT GTTTGGTACGGAAGTTCA	Reverse pURA1, overhang to CDC19 N-terminal
tTKL1-pCRC1_Fw	AGTCGAAAAGGCTAATCTAGAAAATCGATT TAGTTGATTTATTTCCCTGC	Forward pCRC1, overhang to tTKL1
CDC19_N-pCRC1_Rv	TAATGAGGTCAATCTTTCTAATCTAGACAT TACTGACACGATGACGTTT	Reverse pCRC1, overhang to CDC19 N-terminal
tTKL1-pCDC14_Fw	AGTCGAAAAGGCTAATCTAGAAAATCGATT GTTGTGTATTTCGTACCTATGTAT	Forward pCDC14, overhang to tTKL1
CDC19_N-pCDC14_Rv	TAATGAGGTCAATCTTTCTAATCTAGACAT TTATAAGCGTACTTTGTAGTCC	Reverse pCDC14, overhang to CDC19 N-terminal
tTKL1-pTEF2_Fw	AGTCGAAAAGGCTAATCTAGAAAATCGATT GTAGGTGTTCCTTGAGCTAC	Forward pTEF2, overhang to tTKL1
CDC19_N-pTEF2_Rv	TAATGAGGTCAATCTTTCTAATCTAGACAT GTTTAGTTAATTATAGTTCGTTGACC	Reverse pTEF2, overhang to CDC19 N-terminal
tTKL1-pFBA1_Fw	AGTCGAAAAGGCTAATCTAGAAAATCGATT ACTGGTAGAGAGCGACTTT	Forward pFBA1, overhang to tTKL1
CDC19_N-pFBA1_Rv	TAATGAGGTCAATCTTTCTAATCTAGACAT TTTGAATATGTATTACTTGGTTATG	Reverse pFBA1, overhang to CDC19 N-terminal
tCDC19-pPFK1_Fw	ACGCGGGCAGATTCAATTAGTGTCCTAAAT ACCTCATCTATAATTTTTACCCT	Forward pPFK1, overhang to tCDC19
PFK1_N-pPFK1_Rv	AACACCGTAGCATGAATCTTGAGATTGCAT CTTTGATATGATTTTGTTTCAG	Reverse pPFK1, overhang to PFK1 N-terminal
tCDC19-pTDH2_Fw	ACGCGGGCAGATTCAATTAGTGTCCTAAAT CTAGATCAGAGGGTGGTAAAT	Forward pTDH2, overhang to tCDC19
PFK1_N-pTDH2_Rv	AACACCGTAGCATGAATCTTGAGATTGCAT TTTGTTTTGTTTGTTTGTGT	Reverse pTDH2, overhang to PFK1 N-terminal
tCDC19-pIDP2_Fw	ACGCGGGCAGATTCAATTAGTGTCCTAAAT AATAGTCTTACACCAATGAGC	Forward pIDP2, overhang to tCDC19
PFK1_N-pIDP2_Rv	AACACCGTAGCATGAATCTTGAGATTGCAT TACGATTTTATATATATACGTACGTTAC	Reverse pIDP2, overhang to PFK1 N-terminal
tCDC19-pTPK2_Fw	ACGCGGGCAGATTCAATTAGTGTCCTAAAT CAACAAGTCTGAAACTTTCA	Forward pTPK2, overhang to tCDC19
PFK1_N-pTPK2_Rv	AACACCGTAGCATGAATCTTGAGATTGCAT ACCGACAATTTTCAACAG	Reverse pTPK2, overhang to PFK1 N-terminal
tCDC19-pRPL15B_Fw	ACGCGGGCAGATTCAATTAGTGTCCTAAAT GTACTGCTGGCCATTTTTAT	Forward pRPL15B, overhang to tCDC19
PFK1_N-pRPL15B_Rv	AACACCGTAGCATGAATCTTGAGATTGCAT TGCTTGTGTGGTAGGTAATT	Reverse pRPL15B, overhang to PFK1 N-terminal

tCDC19-pTEF1_Fw	ACGCGGGCAGATTCAATTAGTGTCCTAAAT CTTCATCGGTATCTTCGC	Forward pTEF1, overhang to tCDC19
PFK1_N-pTEF1_Rv	AACACCGTAGCATGAATCTTGAGATTGCAT TTTGTAATTAAAACTTAGATTAGATTG	Reverse pTEF1, overhang to PFK1 N-terminal
PCK1_N_Fw	ATGTCCCCTTCTAAAATGA	Forward PCK1
tPCK1_Rv-1	TAAAGAACGGTAAAGTGATTC	Reverse PCK1
TAL1_N_Fw	ATGTCTGAACCAGCTCAA	Forward TAL1
tTAL1_Rv-1	CATTCATTCGTTTGATTCA	Reverse TAL1
TKL1_N_Fw	ATGACTCAATTCACTGACATT	Forward TKL1
tTKL1_Rv-1	AATCGATTTTCTAGATTAGCC	Reverse TKL1
CDC19_N_Fw	ATGTCTAGATTAGAAAGATTGACC	Forward CDC19
tCDC19_Rv-1	ATTTAGGACACTAATTGAATCTG	Reverse CDC19
PFK1_N_Fw	ATGCAATCTCAAGATTCATG	Forward PFK1
tPFK1_Rv-1	CACTAGTTTCCATTTTTCCA	Reverse PFK1
EC_UP_Fw	AAAGTATAGGAACTTCTGAAGTGG	Forward XII-5 UP
tADH1_Rv	GTAGATACGTTGTTGACACTTCTAAATA	Reverse tADH1
tCYC1_Fw	ATCCGCTCTAACCGAAAAG	Forward tCYC1
EC_DW_Rv	AACTTCACTTCATTTTATTTAAATTTGC	Reverse XII-5 DW
tPFK1_pHIS3_Fw	GTTTCTTTTTATCTTTCCGCTGGAAAAATGGAAACTAG TG CGTTTTAAGAGCTTGGTGAG	Forward pHIS3, overhang to tPFK1
tCYC1_tHIS3_Rv	CTAACTCCTTCCTTTTCGGTTAGAGCGGAT NNNNNNNNN ATAGATCCGTCGAGTTCAAGA	Reverse tHIS3, overhang to tCYC1
Sequencing for construction validation		
XII-5-up-out-sq	CCACCGAAGTTGATTTGCTT	Forward upstream sequence integrated at EasyClone site XII-5
TADH1_towards out	GTTGACACTTCTAAATAAGCGAATTTC	Reverse beginning of integrated sequencing at EasyClone site XII-5
DW_towards out	CCTGCAGGACTAGTGCTGAG	Forward end of integrated sequence at EasyClone site XII-5
XII-5-down-out-sq	GTGGGAGTAAGGGATCCTGT	Reverse downstream sequence integrated at EasyClone site XII-5
USER_XhoI_Fw	ACTCTCGAG AGCGACCTCATGCTATACC	Forward tADH1, designed to test junctions around promoter at position 1
PCK1_N_Rv	ATTCATTTTAGAAGGGGACAT	Reverse N-terminal of PCK1, designed to test junctions around promoter at position 1
PCK1_C_Fw	CGATTTTCAATCTTCAAGTAC	Forward C-terminal PCK1, designed to test junctions around promoter at position 2
TAL1_N_Rv	TTTGAGCTGGTTCAGACAT	Reverse N-terminal of TAL1, designed to test junctions around promoter at position 2

TAL1_C_Fw	CTTTCCCAAGAGTTTTGG	Forward C-terminal TAL1, designed to test junctions around promoter at position 3
TKL1_N_Rv	CAATGTCAGTGAATTGAGTCAT	Reverse N-terminal of TKL1, designed to test junctions around promoter at position 3
TKL1_C_Fw-2	TCCAATCATGTCTGTTGAA	Forward C-terminal TKL1, designed to test junctions around promoter at position 4
CDC19_N_Rv-2	CAGCAACAACGTTTAATGA	Reverse N-terminal of CDC19, designed to test junctions around promoter at position 4
CDC19_C_Fw	ACTTGTACAGAGGTGTCTTCC	Forward C-terminal CDC19, designed to test junctions around promoter at position 5
PFK1_N_Rv	CTAGAGTGTGATAAAAGTGAATG	Reverse N-terminal of PFK1, designed to test junctions around promoter at position 5
ADH1_test_fw	GAAATTCGCTTATTTAGAAGTGTC	Forward tADH1, designed to identify promoter at position 1
seq_junction_2	GGTTACCGGAATGATTCACCG	Forward tPCK1, designed to identify promoter at position 2
seq_junction_3	CGATGCTGTAAACGTCCCTG	Forward tTAL1, designed to identify promoter at position 3
seq_junction_4	GATCACCAATGGCGGAAGC	Forward tTKL1, designed to identify promoter at position 4
seq_junction_5	GTTCAGCTTCTGGCCTTCG	Forward tCDC19, designed to identify promoter at position 5

Table S2. Plasmids constructed and used in study.

Name	Description	Reference
Tryptophan biosensor development		
pCfB4107	CEN6/ARS4 pRS413U-*HIS3*, P_{TEF2_trpO}-*yEGFP*-T_{ADH1}	This study
pCfB4108	CEN6/ARS4 pRS416U-*HIS3*, P_{REV1}-*trpR*-T_{ADH1}	This study
pCfB4743	CEN6/ARS4 pRS416U-*URA3*, P_{REV1}-$GAL4_{ad}$_*trpR*-T_{ADH1}	This study
pCfB4747	CEN6/ARS4 pRS416U-*URA3*, P_{REV1}-$GAL4_{ad}$-T_{ADH1}	This study
pCfB4750	CEN6/ARS4 pRS413U-*HIS3*, P_{TEF2}-*mKate2*-T_{IDP1}, $P_{trunTEF1\_trpO}$-yEGFP-T_{ADH1}	This study
pCfB5397	CEN6/ARS4, pRS413U-*HIS3*, $P_{GAL1core\_3xtrpO}$-*yEGFP*-T_{ADH1}, P_{TEF1_trpO}-*mKate2*-T_{CYC1}	This study
pCfB5399	CEN6/ARS4, pRS413U-*HIS3*, $P_{GAL1core\_6xtrpO}$-*yEGFP*-T_{ADH1}, P_{TEF1_trpO}-*mKate2*-T_{CYC1}	This study
Platform and library strain construction		
pCfB176	CEN6/ARS4, pRS414-*TRP1*, P_{TEF1}-SpCas9-T_{CYC1}	DiCarlo et al., 2013
pCfB4672	CEN6/ARS4, pRS413U-*HIS3*, P_{TEF2}-*mKate2*-T_{IDP1}, P_{GAL1}-*ACT1*-T_{ADH1}	This study
pCfB4673	CEN6/ARS4, pRS413U-*HIS3*, P_{TEF2}-*mKate2*-T_{IDP1}, P_{GAL1}-*CDC14*-T_{ADH1}	This study
pCfB9303	CEN6/ARS4, pRS415U-*LEU2*, P_{GAL1}-*ACT1*-T_{IDP1}	This study
pCfB9307	CEN6/ARS4, pRS415U-*LEU2*, P_{GAL1}-*ACT1*-T_{IDP1}, *TKL1-TAL1-PFK1-CDC19* (native expression cassettes)	This study
pCfB6842	2 *μ*, pESC-*LEU2,* P_{SNR52}-*ARO4*_gRNA-T_{SUP4}	This study

pCfB6843	2 *μ*, pESC-*LEU2,* P_{SNR52}-*ARO4*_gRNA-T_{SUP4}, P_{SNR52}-*TRP2*_gRNA_1-T_{SUP4}, P_{SNR52}-*TRP2*_gRNA_2-T_{SUP4}	This study
pCfB6844	2 *μ*, pESC-*LEU2,* P_{SNR52}-*TRP2*_gRNA_1-T_{SUP4}, P_{SNR52}-*TRP2*_gRNA_2-T_{SUP4}	This study
pCfB6903	2 *μ*, pESC-*LEU2,* P_{SNR52}-XI-2_gRNA-T_{SUP4}	This study
pCfB6904	2 *μ*, pESC-*LEU2,* P_{SNR52}-XI-3_gRNA-T_{SUP4}	This study
pCfB6909	2 *μ*, pESC-*LEU2*-P_{SNR52}-XII-5_gRNA-T_{SUP4}	This study
pCfB6916	2 *μ*, pESC-*URA3,* P_{SNR52}-XI-5_gRNA-T_{SUP4}	This study
pCfB6895	2 μ, pESC-*URA3*, P_{SNR52}-*PCK1*_gRNA-$T_{SUP4,}$ P_{SNR52}-*TAL1*_gRNA_1-$T_{SUP4,}$ P_{SNR52}-*TAL1*_gRNA_2-$T_{SUP4,}$ P_{SNR52}-*TKL1*_gRNA-T_{SUP4}	This study
pCfB9306	2 μ, pESC-*URA3*, P_{SNR52}-*pPFK1*_gRNA_1-$T_{SUP4,}$ P_{SNR52}-*pPFK1*_gRNA_2-$T_{SUP4,}$ P_{SNR52}-p*CDC19*_gRNA_1-$T_{SUP4,}$ P_{SNR52}-p*CDC19*_gRNA_2-T_{SUP4}	This study

Table S3. Yeast strains engineered and used in study.

Name	**Genotype**	**Reference**
CEN.PK113-11C	*MAT***a** *his3*Δ1, *LEU2*, *ura3*-52, *TRP1 MAL2*-8c *SUC2*	EUROSCARF
CEN.PK2-1C	*MAT***a** *his3*Δ1, *leu2*-3_112, *ura3*-52, *trp1*-289, *MAL2*-8c *SUC2*	EUROSCARF
TrpA-1	*MAT***a** $P_{GAL1core\_6xtrpO}$-*yEGFP*-T_{ADH1}, P_{TEF1_trpO}-*mKate2*-T_{CYC1}, pCfB176	this study
TrpA-2	*MAT***a** $P_{GAL1core\_6xtrpO}$-*yEGFP*-T_{ADH1}, P_{TEF1_trpO}-mKate2-T_{CYC1}, *ARO4*^{wt}::*ARO4*^{K229L}, pCfB176	this study
TrpA-3	*MAT***a** $P_{GAL1core\_6xtrpO}$-*yEGFP*-T_{ADH1}, P_{TEF1_trpO}-mKate2-T_{CYC1}, *TRP2*^{wt}::*TRP2*$^{S65R, S76L}$, pCfB176	this study
TrpA-4	*MAT***a** $P_{GAL1core\_6xtrpO}$-*yEGFP*-T_{ADH1}, P_{TEF1_trpO}-mKate2-T_{CYC1}, *ARO4*^{wt}::*ARO4*^{K229L}, *TRP2*^{wt}::*TRP2*$^{S65R, S76L}$, pCfB176	this study
TrpNA-W	*MAT***a** *tkl1*Δ *tal1*Δ *pck1*Δ, P_{PFK1}::P_{REV1}-*PFK1*, P_{CDC19}::P_{RNR2}-*CDC19*, P_{PFK1}-*GAL4*$_{ad}$-*trpR*-T_{ADH1}, $P_{GAL1core\_3xtrpO}$-*yEGFP*-T_{ADH1}, P_{TEF1_trpO}-*mKate2*-T_{CYC1}, P_{PGK1}-*ARO4*^{K229L}-T_{ADH1}, P_{TEF1}-*TRP2*$^{S65R, S76L}$-T_{CYC1}, pCfB176, pCfB9307	this study

Table S4. Related to Figure 1. Gene scores of all 192 genome-scale modelled (FBA) genes with significant changes in flux towards tryptophan production under glucose and ethanol conditions. A score higher than one means the gene is an up-regulation candidate, a score between zero and one means the gene is a down-regulation candidate, a score equal to zero means the gene is a knockout candidate, and a blank score means the gene is associated to reactions that do not change significantly in flux as tryptophan production increases under that particular condition. The four out of five gene targets identified by FBA and selected for this study are marked in bold.

Gene name	**Glucose**	**Ethanol**		**Gene name**	**Glucose**	**Ethanol**		**Gene name**	**Glucose**	**Ethanol**
YLR438W	1000	0		YER070W	0,51899	0,433024		YML008C		17,67071
YBR249C	636,7273			YGR180C	0,51899	0,433024		YGL055W		6,313433

YKL120W	636,7273			YIL066C	0,51899	0,433024		YKL182W		4,098839
YHR208W	500,275	0,129706		YJL026W	0,51899	0,433024		YPL231W		4,098839
YBR068C	448,3277	353,2292		YBR218C	0,5093	636,7273		YDR353W		2,608208
YBR069C	448,3277	353,2292		YGL062W	0,5093	636,7273		YLR058C		2,443903
YDR046C	448,3277	353,2292		YEL039C	0,49376			**YKR097W (PCK1)**		**1,576565**
YKR039W	448,3277	353,2292		YJR048W	0,49376			YMR170C		1,29168
YOL020W	448,3277	353,2292		YOR222W	0,44886	0,303233		YGR254W		1,228317
YDR035W	321,6998			YPL134C	0,44886	0,303233		YHR174W		1,228317
YHR137W	318,4523	46,12524		**YAL038W (CDC19)**	**0,39342**			YKL152C		1,228317
YBR117C	303,4223	2,944678		YOR347C	0,39342			YIR031C		1,066515
YPR074C (TKL1)	**303,4223**	**2,944678**		YOL126C	0,25278	1,357649		YKL060C		1,031041
YJR148W	250,275	200,22		YBR252W	0	15,60595		YLR377C		1,031041
YOR311C	167,5			YDL174C	0			YDR050C		1,029272
YER019W	143,5891	0		YDR272W	0			YER065C		1,028886
YJL121C	136,8566	5,113079		YDR300C	0	1000		YCR012W		1,009732
YPR113W	122,016	363,9394		YEL071W	0			YGR192C		1,009732
YBR029C	122,0159			YKL029C	0	45,94508		YJL052W		1,009732
YDR367W	109,9168	250,25		YML004C	0			YJR009C		1,009732
YKL004W	109,9168	250,25		YOR323C	0	1000		YDR226W		0,546538
YDR072C	91,72753	0		YDL078C		1000		YER091C		0,543448
YDR454C	61,21212	30,96364		YER170W		1000		YGL125W		0,543448
YDR007W	56,78997	42,47148		YGL080W		1000		YPL023C		0,543448
YDR354W	56,78997	42,47148		YGL205W		1000		YDR502C		0,526292
YER090W	56,78997	42,47148		YGR243W		1000		YER043C		0,526292
YGL026C	56,78997	42,47148		YHR002W		1000		YJR105W		0,526292
YKL211C	56,78997	42,47148		YHR162W		1000		YLR180W		0,526292
YGR209C	45,93409	222,1165		YIL160C		1000		YBL039C		0,512229
YDR127W	7,356177	5,589245		YKL188C		1000		YJR103W		0,512229
YGL148W	7,356177	5,589245		YKR009C		1000		YDL022W		0,504723
YBR291C	6,672427			YKR080W		1000		YOL059W		0,504723
YMR241W	6,672427	0		YLR056W		1000		YLR153C		0,488052
YBL068W	6,473537	4,935219		YLR109W		1000		YJL153C		0,458956
YER099C	6,473537	4,935219		YLR284C		1000		YPL087W		0,420455
YHL011C	6,473537	4,935219		YLR348C		1000		YDR287W		0,410628
YKL181W	6,473537	4,935219		YMR015C		1000		YHR046C		0,410628
YOL061W	6,473537	4,935219		YMR272C		1000		YPL061W		0,364362
YER081W	3,494777	5,449761		YNL009W		1000		YER069W		0,27162
YGR208W	3,494777	5,449761		YNL202W		1000		YMR062C		0,27162
YIL074C	3,494777	5,449761		YOR180C		1000		YOL140W		0,27162
YOR184W	3,494777	5,449761		YPL147W		1000		YOR130C		0,27162
YPR021C	3,10273	0,138041		YBL015W		636,3939		YFL030W		0,059875
YPR035W	2,765313	2,189179		YNL117W		500,5333		YNL037C		0,030022
YOR095C	2,221708	5,143796		YDR297W		500,5		YOR136W		0,030022
YDR384C	1,202505	1,032684		YMR165C		500,5		YAL044C		0
YGR121C	1,202505	1,032684		YMR208W		500,2608		YAR035W		0

YNL142W	1,202505	1,032684		YAL054C		500,244		YBR036C		0
YPR138C	1,202505	1,032684		YLR027C		500,221		YBR084W		0
YCR024CA	1,174844			YLR304C		500,022		YBR161W		0
YEL017CA	1,174844			YGR204W		500,0102		YDR019C		0
YGL008C	1,174844			YJR139C		500		YDR148C		0
YPL036W	1,174844			YML126C		500		YER024W		0
YBR196C	1,138695			YNL104C		500		YFL018C		0
YLL052C	1,108704	1,100811		YPL028W		500		YHR144C		0
YPR192W	1,108704	1,100811		YBR183W		364,2727		YIL125W		0
YLR354C (TAL1)	**1,071089**			YLR043C		334,3861		YLR089C		0
YDL085W	1,070445	1,17891		YMR246W		333,6667		YLR174W		0
YMR145C	1,070445	1,17891		YOR317W		333,6667		YML042W		0
YGR248W	1,064769			YGR170W		318,8636		YMR189W		0
YGR256W	1,064769			YOR245C		309,5818		YNL169C		0
YHR163W	1,064769			YCR048W		258,0765		YOR100C		0
YHR183W	1,064769			YNR019W		258,0765		YOR108W		0
YNL241C	1,064769			YMR202W		38,00156		YPL057C		0

Table S5. FBA results for all pathways in metabolism, including the number of gene targets predicted in each pathway, the total size of each pathway, the fraction of genes in each pathway that are gene targets, and the significance of that representation in each pathway compared to the rest of metabolism ("Whole metabolism"), indicated by a P-value computed with a Fisher's exact test. General pathways such as "carbon metabolism" and "biosynthesis of amino acids" were filtered out of the analysis.

KEGG pathway	Gene count	Pathway size	Coverage (%)	P-value
Pyruvate metabolism	29	47	61,70	5.64e-10
Pentose phosphate pathway	18	24	75,00	1.44e-08
Peroxisome	15	21	71,43	7.54e-07
Valine, leucine and isoleucine degradation	12	19	63,16	7.19e-05
Glyoxylate and dicarboxylate metabolism	16	30	53,33	7.73e-05
Glycolysis	24	54	44,44	8.85e-05
Gluconeogenesis	24	54	44,44	8.85e-05
One carbon pool by folate	12	20	60,00	1.47e-04

Aminoacyl-tRNA biosynthesis	0	37	0,00	2.56e-04
Citrate cycle (TCA cycle)	16	33	48,48	3.35e-04
Phenylalanine, tyrosine and tryptophan biosynthesis	11	19	57,89	4.39e-04
Glycine, serine and threonine metabolism	15	33	45,45	1.57e-03
Lysine degradation	8	13	61,54	1.67e-03
Starch and sucrose metabolism	1	35	2,86	4.79e-03
Oxidative phosphorylation	6	70	8,57	5.71e-03
Nicotinate and nicotinamide metabolism	0	23	0,00	7.45e-03
Sulfur relay system	3	3	100,00	9.31e-03
2-Oxocarboxylic acid metabolism	14	35	40,00	9.76e-03
Meiosis - yeast	0	21	0,00	1.20e-02
Galactose metabolism	0	19	0,00	1.94e-02
Purine metabolism	19	58	32,76	3.05e-02
Phagosome	0	17	0,00	3.14e-02
Propanoate metabolism	9	22	40,91	3.17e-02
Carbapenem biosynthesis	2	2	100,00	4.44e-02
ABC transporters	2	2	100,00	4.44e-02
Synthesis and degradation of ketone bodies	2	2	100,00	4.44e-02
Pyrimidine metabolism	12	33	36,36	4.71e-02
Whole metabolism	192	909	21,12	-

Table S6. Related to Figure 1. The 30 selected native yeast promoters, and their position in the combinatorial cluster.

Number	Name	Systematic name	Position in cluster
01	pPCK1	YKR097W	01

02	pTPI1	YDR050C	01
03	pICL1	YER065C	01
04	pRNR2	YJL026W	01
05	pACT1	YFL039C	01
06	pTDH3	YGR192C	01
07	pTAL1	YLR354C	02
08	pENO2	YHR174W	02
09	pACS1	YAL054C	02
10	pREV1	YOR346W	02
11	pCHO1	YER026C	02
12	pCYC1	YJR048W	02
13	pTKL1	YPR074C	03
14	pPGK1	YCR012W	03
15	pMLS1	YNL117W	03
16	pBUD6	YLR319C	03
17	pURE2	YNL229C	03
18	pCCW12	YLR110C	03
19	pCDC19	YAL038W	04
20	pURA1	YKL216W	04
21	pCRC1	YOR100C	04
22	pCDC14	YFR028C	04
23	pTEF2	YBR118W	04
24	pFBA1	YKL060C	04
25	pPFK1	YGR240C	05
26	pTDH2	YJR009C	05
27	pIDP2	YLR174W	05
28	pTPK2	YPL203W	05
29	pRPL15B	YMR121C	05
30	pTEF1	YPR080W	05

Table S7. Related to Figure 1D and 3D. Promoter combinations of library control strains. The numbers in each row refer to promoter numbers as shown in Table S5. Design no. 1 contains the promoters that are native to the genes at the five positions.

Design	Position 1	Position 2	Position 3	Position 4	Position 5
1	1	7	13	19	25
2	6	12	18	23	28
3	4	11	17	24	30
4	2	8	14	23	28

5	3	9	15	20	26

Table S8. Related to Figure 1 and 4C. Top-30 promoter combinations as recommended by ART. Size of color bars indicate promoter expression strength (see Figure 1), and column "dgfp/dt" shows predicted GFP synthesis rate.

Priority	Promoter *PCK1*	Promoter *TAL1*	Promoter *TKL1*	Promoter *CDC19*	Promoter *PFK1*	Constructed	Line Name	dgfp_dt
1	pICL1	pENO2	pBUD6	pCDC14	pTPK2	Yes	SP609	234.72
2	pICL1	pENO2	pBUD6	pCDC19	pTPK2	No		
3	pICL1	pENO2	pBUD6	pCDC14	pRPL15B	Yes	SP610	201.91
4	pICL1	pENO2	pBUD6	pCDC19	pRPL15B	Yes	SP605	215.80
5	pICL1	pENO2	pBUD6	pCDC14	pPFK1	Yes	SP607	244.19
6	pICL1	pENO2	pBUD6	pCDC19	pTDH2	Yes	SP603	193.35
7	pICL1	pENO2	pBUD6	pCDC14	pTDH2	Yes	SP608	187.92
8	pTDH3	pENO2	pBUD6	pCDC14	pRPL15B	Yes	SP627	239.33
9	pICL1	pENO2	pBUD6	pCDC19	pPFK1	Yes	SP602	219.01
10	pTPI1	pENO2	pBUD6	pCDC19	pPFK1	Yes	SP586	175.43
11	pTPI1	pENO2	pBUD6	pCDC14	pTPK2	No		
12	pACT1	pENO2	pBUD6	pCDC14	pRPL15B	Yes	SP620	253.94
13	pTPI1	pENO2	pBUD6	pCDC19	pTPK2	No		
14	pTPI1	pENO2	pBUD6	pCDC14	pPFK1	Yes	SP591	205.20
15	pTPI1	pENO2	pBUD6	pTEF2	pRPL15B	No		
16	pTPI1	pENO2	pBUD6	pCDC14	pRPL15B	Yes	SP593	193.57
17	pICL1	pENO2	pURE2	pCDC14	pRPL15B	Yes	SP612	218.93
18	pTPI1	pENO2	pBUD6	pCDC19	pRPL15B	Yes	SP588	255.34
19	pTPI1	pENO2	pBUD6	pURA1	pPFK1	Yes	SP589	179.41
20	pACT1	pENO2	pBUD6	pCDC19	pRPL15B	Yes	SP617	218.75
21	pACT1	pENO2	pBUD6	pCDC14	pPFK1	Yes	SP618	212.79
22	pTPI1	pENO2	pBUD6	pTEF2	pTEF1	No		
23	pICL1	pENO2	pBUD6	pFBA1	pTDH2	Yes	SP611	223.48
24	pACT1	pENO2	pBUD6	pCDC14	pTPK2	Yes	SP619	228.85
25	pTDH3	pCYC1	pBUD6	pURA1	pRPL15B	Yes	SP633	217.05
26	pTDH3	pCYC1	pURE2	pURA1	pRPL15B	No		
27	pICL1	pTAL1	pBUD6	pCDC14	pRPL15B	Yes	SP601	166.32
28	pTPI1	pENO2	pBUD6	pURA1	pRPL15B	Yes	SP590	205.96
29	pACT1	pENO2	pBUD6	pCDC19	pTPK2	Yes	SP616	280.88
30	pICL1	pENO2	pBUD6	pCRC1	pRPL15B	Yes	SP606	298.80

Table S9. Related to Figure 1 and 4C. Top-30 promoter combinations as recommended by TeselaGen EVOLVE. Size of color bars indicate promoter expression strength (see Figure 1), and column “dgfp/dt” shows predicted GFP synthesis rate.

Priority	Promoter *PCK1*	Promoter *TAL1*	Promoter *TKL1*	Promoter *CDC19*	Promoter *PFK1*	Constructed	*Line Name*	dgfp_dt
1	pTDH3	pENO2	pPGK1	pCDC19	pRPL15B	No		
2	pACT1	pENO2	pPGK1	pTEF2	pIDP2	No		
3	pPCK1	pENO2	pPGK1	pCDC19	pRPL15B	No		
4	pTDH3	pENO2	pURE2	pCDC19	pIDP2	Yes	SP628	219.66
5	pTDH3	pENO2	pBUD6	pCDC14	pRPL15B	Yes	SP627	239.33
6	pTPI1	pENO2	pPGK1	pURA1	pPFK1	No		
7	pPCK1	pENO2	pPGK1	pTEF2	pTPK2	No		
8	pTDH3	pACS1	pTKL1	pCDC19	pTPK2	Yes	SP630	162.63
9	pRNR2	pENO2	pBUD6	pCDC19	pRPL15B	Yes	SP614	201.95
10	pTDH3	pENO2	pBUD6	pCDC19	pTPK2	Yes	SP624	258.17
11	pTPI1	pENO2	pPGK1	pCDC14	pTPK2	No		
12	pTDH3	pENO2	pBUD6	pCDC19	pTEF1	Yes	SP625	222.04
13	pICL1	pACS1	pTKL1	pTEF2	pPFK1	Yes	SP613	198.45
14	pTPI1	pENO2	pURE2	pCDC19	pRPL15B	No		
15	pTPI1	pACS1	pTKL1	pCDC19	pPFK1	Yes	SP600	131.34
16	pTPI1	pTAL1	pBUD6	pCRC1	pIDP2	Yes	SP582	101.84
17	pTPI1	pENO2	pURE2	pTEF2	pPFK1	No		
18	pTPI1	pENO2	pURE2	pCDC19	pIDP2	Yes	SP597	131.58
19	pTDH3	pENO2	pURE2	pCDC19	pRPL15B	Yes	SP629	247.94
20	pTDH3	pACS1	pPGK1	pFBA1	pRPL15B	No		
21	pTPI1	pENO2	pPGK1	pTEF2	pIDP2	No		
22	pTPI1	pENO2	pBUD6	pCDC19	pRPL15B	Yes	SP588	255.34
23	pACT1	pREV1	pTKL1	pCRC1	pTEF1	Yes	SP621	166.70
24	pPCK1	pENO2	pURE2	pCDC19	pPFK1	Yes	SP580	181.63
25	pTPI1	pTAL1	pBUD6	pURA1	pIDP2	Yes	SP581	173.88
26	pTDH3	pENO2	pBUD6	pURA1	pRPL15B	Yes	SP626	197.43
27	pTPI1	pENO2	pBUD6	pTEF2	pTPK2	No		
28	pACT1	pREV1	pTKL1	pTEF2	pPFK1	Yes	SP622	253.30
29	pPCK1	pTAL1	pBUD6	pURA1	pRPL15B	Yes	SP577	75.05
30	pTDH3	pACS1	pTKL1	pURA1	pTEF1	Yes	SP631	200.35

REFERENCES

Jensen, N.B., Strucko, T., Kildegaard, K.R., David, F., Maury, J., Mortensen, U.H., Forster, J., Nielsen, J., and Borodina, I. (2014). EasyClone: method for iterative chromosomal integration of multiple genes in Saccharomyces cerevisiae. FEMS Yeast Res. *14*, 238–248.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*, 406–425.

Studier, J.A., and Keppler, K.J. (1988). A note on the neighbor-joining algorithm of Saitou and Nei. Mol. Biol. Evol. *5*, 729–731.

4. Main conclusions and perspectives

This work consists of contributions to the development of methods for speeding up different steps of the Design-Build-Test-Learn (DBTL) cycle. The study gives some answers, but it also raises some new questions.

Paper no. 1 contributes mainly to the design phase. Here, it is shown that TIS sequences can be used for predictive tuning of protein expression context-independently from yeast to CHO cells. And it is shown that TISs can be used to control the flux of metabolic pathways in cell factories. Methods for context-independent tuning of protein activity using hexamers is believed to have broad interest. Pristovšek et al. (2019) have already used the technique in a study on recombinant gene expression for programmable mammalian cell engineering. The method's transferability to Aspergillus fungi is currently being investigated at DTU Bioengineering in the Uffe Mortensen group (Mortensen and Dorota Jarczynska, 2018, personal communication). It could be really interesting to investigate to what extent the method can be transferred to other organisms, for example, plants and higher-order animals. For instance, do all hexamers contribute with the same strength in such organisms and are the effects fully context-independent. In *E. coli*, it has been shown that the effect of the Translation Initiation Region (TIR equivalent to TIS) depends on the initial bases within the open reading frame (Mirzadeh et al., 2015, 2016). The availability of well-characterized regulatory genetic parts could be an advantage in the study of an organism by enabling controlled system perturbations. It would be relevant to test this tool against other tools for predictable tuning of protein expression such as promoter replacement, mainly in terms of accuracy. Accurate tools are critical to facilitate robust outputs from ML algorithms used as a tool for metabolic engineering efforts (Opgenorth et al., 2019). In the study we observed that the activity of fluorescent proteins changed with changes in the hexameric region of the TIS sequence. We believe this is due to a change in translation efficiency. However, it may also be due to other factors such as transcription efficiency. Therefore it would be relevant to measure the translation efficiency more directly with techniques such as ribosome profiling (Brar and Weissman, 2015) such as it has been done by (Sample et al., 2019). It would also be relevant to control for other

factors e.g by quantification of mRNA. One interesting method of mRNA quantification is the use of RNA aptamer-fluorophore complexes (Paige et al., 2011).

Paper no. 2 is broader in the sense that it contributes to several steps in the DBTL cycle. However, my main and most innovative contribution is within the learn step in the form of using a combination of mechanistic and machine learning modeling to provide the predictive power needed to fruitfully lead metabolic engineering efforts. Machine learning has also in other studies been very effective in finding patterns in large and complex data sets (Bonde et al., 2016; Costello and Martin, 2018; Zhou et al., 2018; Jervis et al., 2019). In the near future, I expect that machine learning approaches based on ensembles of ML models, as in the present project, will be used in a variety of other projects. This applies not at least for projects that include product maximization by directed evolution.

The tryptophan synthesis rate in yeast can be increased significantly in one optimization cycle when using a GSM to assist the selection of genetic targets and ML for prediction. What would be the next step? More iterations with recombination of the same genetic edits could be done to further optimize the fluxes. It is argued that recombining some of the best edits, as opposed to focusing only on the single best edit, is important in order not to lose useful genetic diversity (Zhang et al., 2002; Figure 9). The decision of which edits to recombine can be supported by ML algorithms. It could also be relevant to expand the space of edits with edits from other genomic loci than the five investigated. Furthermore, the edit types could be extended from only promoter replacement to include also e.g. gene replacement. Thus, other genetic edits that also have an effect on tryptophan production could be identified e.g. by using a biosensor, and the overall best ones from the total set of edits could then be combined based on ML. The identification of new edits could be based on GSM modeling. Alternatively, relevant edits can be sought out more unbound by existing mechanistic understanding, even by random mutagenesis (Aharonowitz and Cohen, 1981). Such work on large genetic diversity and directed evolution go well in hand with the ongoing strong expansion of high throughput characterization methods (Chao et al., 2017; Hillson et al., 2019).

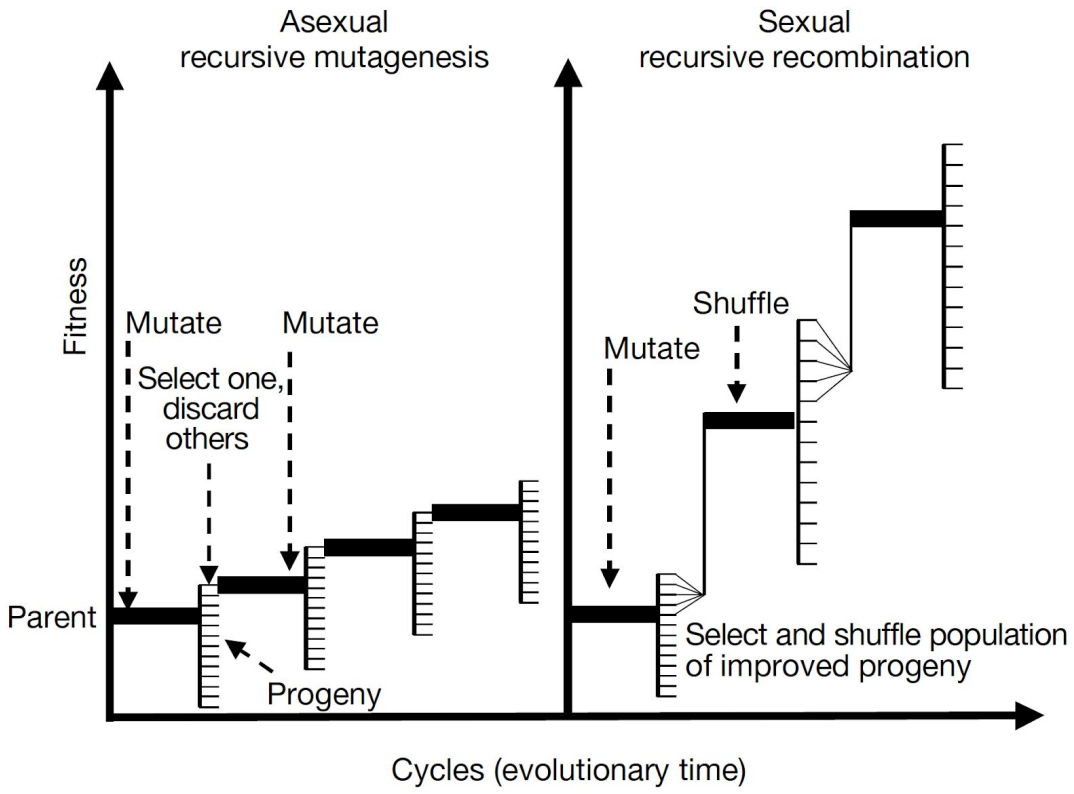


Figure 9. Evolution of an asexual versus a sexual organism. The asexual organism evolves by recursive mutagenesis i.e. where individuals evolve alone by accumulating mutations. The sexual organism evolves by recursive recombination i.e. where individuals evolve by sharing genetic information. The sharing of genetic information provides a mechanism for the combination of useful mutations and the loss of harmful ones. Figure from Zhang et al., (2002).

Flow-seq is one such high throughput method that can be used in the process to associate genotype and phenotype of individual cells in a heterologous population (Kosuri et al., 2013; Bonde et al., 2016; Kotopka and Smolke, 2019). In Flow-seq, FACS is first used to separate the library cells into groups of increasing fluorescence. Then high-throughput sequencing is used to count the number of times each genotype appears in the different groups. By using this distribution of read counts, an average fluorescence expression can finally be associated to each genotype. Flow-seq could be used to characterize libraries such as the one we constructed in paper no. 2 if the previously mentioned challenges with target enrichment for long-read sequencing is solved. Such a setup should also provide us with the capacity to characterize our library under different environmental conditions such as different carbon sources.

The advances in method development and biological understanding derived from this study have addressed limitations in the predictive engineering of living cells. However, they also point at new potentials for improved workflows in metabolic engineering efforts that ultimately may support the routine development of cell factories.

5 References

Aharonowitz, Y., and Cohen, G. (1981). The Microbiological Production of Pharmaceuticals. Sci. Am. *245*, 140–153.

Arsenic, R., Treue, D., Lehmann, A., Hummel, M., Dietel, M., Denkert, C., and Budczies, J. (2015). Comparison of targeted next-generation sequencing and Sanger sequencing for the detection of PIK3CA mutations in breast cancer. BMC Clin. Pathol. *15*.

Bailey, J.E. (1991). Toward a Science of Metabolic Engineering. Science *252*, 1668–1675.

Balleza, E., Kim, J.M., and Cluzel, P. (2017). A systematic characterization of maturation kinetics of fluorescent proteins in live cells. *6*, 1–10.

Becker, J., Zelder, O., Häfner, S., Schröder, H., and Wittmann, C. (2011). From zero to hero—Design-based systems metabolic engineering of Corynebacterium glutamicum for l-lysine production. Metab. Eng. *13*, 159–168.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Bialecka-Fornal, M., Makushok, T., and Rafelski, S.M. (2016). A review of fluorescent proteins for use in yeast. In Methods in Molecular Biology, pp. 309–346.

Bonde, M.T., Pedersen, M., Klausen, M.S., Jensen, S.I., Wulff, T., Harrison, S., Nielsen, A.T., Herrgård, M.J., and Sommer, M.O.A. (2016). Predictable tuning of protein expression in bacteria. Nat. Methods *13*, 233–236.

Brar, G.A., and Weissman, J.S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. Nat. Rev. Mol. Cell Biol. *16*, 651–664.

Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., and Collins, J.J. (2018). Next-Generation Machine Learning for Biological Networks. Cell *173*, 1581–1592.

Carbonell, P., Jervis, A.J., Robinson, C.J., Yan, C., Dunstan, M., Swainston, N., Vinaixa, M., Hollywood, K.A., Currin, A., Rattray, N.J.W., et al. (2018). An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals. Commun. Biol. *1*, 1–10.

Carbonell, P., Radivojevic, T., and García Martín, H. (2019). Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation. ACS Synth. Biol. *8*, 1474–1477.

Chao, R., Mishra, S., Si, T., and Zhao, H. (2017). Engineering biological systems using automated biofoundries. Metab. Eng. *42*, 98–108.

Chomvong, K., Benjamin, D.I., Nomura, D.K., and Cate, J.H.D. (2017). Cellobiose Consumption Uncouples Extracellular Glucose Sensing and Glucose Metabolism in Saccharomyces cerevisiae. MBio *8*.

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

Cossarizza, A., Chang, H.-D., Radbruch, A., Akdis, M., Andrä, I., Annunziato, F., Bacher, P., Barnaba, V., Battistini, L., Bauer, W.M., et al. (2017). Guidelines for the use of flow cytometry and cell sorting in immunological studies. Eur. J. Immunol. *47*, 1584–1797.

Costello, Z., and Martin, H.G. (2018). A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. Npj Syst. Biol. Appl. *4*.

Cranfill, P.J., Sell, B.R., Baird, M.A., Allen, J.R., Lavagnino, Z., De Gruiter, H.M., Kremers, G.J., Davidson, M.W., Ustione, A., and Piston, D.W. (2016). Quantitative assessment of fluorescent

proteins. Nat. Methods *13*, 557–562.
Culler, S. (2016). A Bioengineering Platform to Industrialize Biotechnology. Chem. Eng. Prog. 10.
Davy, A.M., Kildegaard, H.F., and Andersen, M.R. (2017). Cell Factory Engineering. Cell Syst. *4*, 262–275.
Dey, A. (2016). Machine Learning Algorithms : A Review. p.
Dietterich, T.G. (2000). Ensemble Methods in Machine Learning. In Multiple Classifier Systems, (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 1–15.
van Dijk, E.L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The Third Revolution in Sequencing Technology. Trends Genet. *34*, 666–681.
Dodd, M.S., Papineau, D., Grenne, T., Slack, J.F., Rittner, M., Pirajno, F., O'Neil, J., and Little, C.T.S. (2017). Evidence for early life in Earth's oldest hydrothermal vent precipitates. Nature *543*, 60–64.
Eelderink-Chen, Z., Mazzotta, G., Sturre, M., Bosman, J., Roenneberg, T., and Merrow, M. (2010). A circadian clock in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. *107*, 2043–2047.
Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *323*, 7.
Endy, D. (2005). Foundations for engineering biology. Nature *438*, 449–453.
Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J., and Turner, S.W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat. Methods *7*, 461–465.
Gardner, T.S. (2013). Synthetic biology: from hype to impact. Trends Biotechnol. *31*, 123–125.
Herzenberg, L.A., Sweet, R.G., and Herzenberg, L.A. (1976). Fluorescence-activated cell sorting. Sci. Am. *234*, 108–117.
Hillson, N., Caddick, M., Cai, Y., Carrasco, J.A., Chang, M.W., Curach, N.C., Bell, D.J., Le Feuvre, R., Friedman, D.C., Fu, X., et al. (2019). Building a global alliance of biofoundries. Nat. Commun. *10*, 2040.
Jakočiūnas, T., Rajkumar, A.S., Zhang, J., Arsovska, D., Rodriguez, A., Jendresen, C.B., Skjødt, M.L., Nielsen, A.T., Borodina, I., Jensen, M.K., et al. (2015). CasEMBLR: Cas9-Facilitated Multiloci Genomic Integration of in Vivo Assembled DNA Parts in Saccharomyces cerevisiae. ACS Synth. Biol. *4*, 1226–1234.
Jakočiūnas, T., Jensen, M.K., and Keasling, J.D. (2016). CRISPR/Cas9 advances engineering of microbial cell factories. Metab. Eng. *34*, 44–59.
Jervis, A.J., Carbonell, P., Vinaixa, M., Dunstan, M.S., Hollywood, K.A., Robinson, C.J., Rattray, N.J.W., Yan, C., Swainston, N., Currin, A., et al. (2019). Machine Learning of Designed Translational Control Allows Predictive Pathway Optimization in *Escherichia coli*. ACS Synth. Biol. *8*, 127–136.
Jeschek, M., Gerngross, D., and Panke, S. (2016). Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort. Nat. Commun. *7*, 11163.
Kalinovski, J. (2019).
Keasling, J.D. (2010). Manufacturing Molecules through Metabolic Engineering. Science *330*, 1355–1358.
Kosuri, S., and Church, G.M. (2014). Large-scale de novo DNA synthesis: technologies and applications. Nat. Methods *11*, 499–507.
Kosuri, S., Goodman, D.B., Cambray, G., Mutalik, V.K., Gao, Y., Arkin, A.P., Endy, D., and Church, G.M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proc. Natl. Acad. Sci. *110*, 14024–14029.
Kotopka, B.J., and Smolke, C.D. (2019). Model-driven generation of artificial yeast promoters.

BioRxiv 748616.
Liu, Y., and Nielsen, J. (2019). Recent trends in metabolic engineering of microbial chemical factories. Curr. Opin. Biotechnol. *60*, 188–197.
Long, C.P., and Antoniewicz, M.R. (2019). Metabolic flux responses to deletion of 20 core enzymes reveal flexibility and limits of E. coli metabolism. Metab. Eng. *55*, 249–257.
Mahr, R., and Frunzke, J. (2016). Transcription factor-based biosensors in biotechnology: current state and future prospects. Appl. Microbiol. Biotechnol. *100*, 79–90.
Mali, P., Aach, J., Stranges, P.B., Esvelt, K.M., Moosburner, M., Kosuri, S., Yang, L., and Church, G.M. (2013). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat. Biotechnol. *31*, 833–838.
Martin, V.J.J., Pitera, D.J., Withers, S.T., Newman, J.D., and Keasling, J.D. (2003). Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. Nat. Biotechnol. *21*, 796–802.
Mirzadeh, K., Martínez, V., Toddo, S., Guntur, S., Herrgård, M.J., Elofsson, A., Nørholm, M.H.H., and Daley, D.O. (2015). Enhanced Protein Production in Escherichia coli by Optimization of Cloning Scars at the Vector–Coding Sequence Junction. ACS Synth. Biol. *4*, 959–965.
Mirzadeh, K., Toddo, S., Nørholm, M.H.H., and Daley, D.O. (2016). Codon Optimizing for Increased Membrane Protein Production: A Minimalist Approach. In Heterologous Expression of Membrane Proteins: Methods and Protocols, I. Mus-Veteau, ed. (New York, NY: Springer New York), pp. 53–61.
Mortensen, U.H., and Dorota Jarczynska, Z. (2018). Personal communication.
Nakamura, C.E., and Whited, G.M. (2003). Metabolic engineering for the microbial production of 1,3-propanediol. Curr. Opin. Biotechnol. *14*, 454–459.
Nielsen, J., and Keasling, J.D. (2016). Engineering Cellular Metabolism. Cell *164*, 1185–1197.
Ohler, U., Liao, G., Niemann, H., and Rubin, G.M. (2002). Computational analysis of core promoters in the Drosophila genome. Genome Biol. *3*, RESEARCH0087.
Opgenorth, P., Costello, Z., Okada, T., Goyal, G., Chen, Y., Gin, J., Benites, V., de Raad, M., Northen, T.R., Deng, K., et al. (2019). Lessons from Two Design–Build–Test–Learn Cycles of Dodecanol Production in *Escherichia coli* Aided by Machine Learning. ACS Synth. Biol. *8*, 1337–1351.
Paige, J.S., Wu, K.Y., and Jaffrey, S.R. (2011). RNA Mimics of Green Fluorescent Protein. Science *333*, 642–646.
Picot, J., Guerin, C.L., Le Van Kim, C., and Boulanger, C.M. (2012). Flow cytometry: retrospective, fundamentals and recent instrumentation. Cytotechnology *64*, 109–130.
Presnell, K.V., and Alper, H.S. (2019). Systems Metabolic Engineering Meets Machine Learning: A New Era for Data-Driven Metabolic Engineering. Biotechnol. J. *0*, 1800416.
Pristovšek, N., Nallapareddy, S., Grav, L.M., Hefzi, H., Lewis, N.E., Rugbjerg, P., Hansen, H.G., Lee, G.M., Andersen, M.R., and Kildegaard, H.F. (2019). Systematic Evaluation of Site-Specific Recombinant Gene Expression for Programmable Mammalian Cell Engineering. ACS Synth. Biol. *8*, 758–774.
Radivojević, T., Costello, Z., and Martin, H.G. (2019). ART: A machine learning Automated Recommendation Tool for synthetic biology. ArXiv191111091 Q-Bio Stat.
Ro, D.-K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A., Ham, T.S., Kirby, J., et al. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature *440*, 940–943.
Rokach, L. (2010). Ensemble-based classifiers. Artif. Intell. Rev. *33*, 1–39.
Rossell, S., Solem, C., Jensen, P.R., and Heijnen, J.J. (2011). Towards a quantitative prediction of the fluxome from the proteome. Metab. Eng. *13*, 253–262.

Sample, P.J., Wang, B., Reid, D.W., Presnyak, V., McFadyen, I.J., Morris, D.R., and Seelig, G. (2019). Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. Nat. Biotechnol. *37*, 803–809.
Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.-P.Z., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778.
Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., and Waterston, R.H. (2017). DNA sequencing at 40: past, present and future. Nature *550*, 345–353.
Snapp, E. (2005). Design and Use of Fluorescent Fusion Proteins in Cell Biology. Curr. Protoc. Cell Biol. Editor. Board Juan Bonifacino Al *CHAPTER*, Unit-21.4.
Song, C.-X., Yi, C., and He, C. (2012). Mapping recently identified nucleotide variants in the genome and transcriptome. Nat. Biotechnol. *30*, 1107–1116.
Stucky, B.J. (2012). SeqTrace: A Graphical Tool for Rapidly Processing DNA Sequencing Chromatograms. J. Biomol. Tech. JBT *23*, 90–93.
Wolpert, D.H. (1996). The Lack of A Priori Distinctions Between Learning Algorithms. Neural Comput. *8*, 1341–1390.
Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J.D., Osterhout, R.E., Stephen, R., et al. (2011). Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. Nat. Chem. Biol. *7*, 445–452.
Zhang, J., Jensen, M.K., and Keasling, J.D. (2015). Development of biosensors and their application in metabolic engineering. Curr. Opin. Chem. Biol. *28*, 1–8.
Zhang, Y., Nicholatos, J., Dreier, J.R., Ricoult, S.J.H., Widenmaier, S.B., Hotamisligil, G.S., Kwiatkowski, D.J., and Manning, B.D. (2014). Coordinated regulation of protein synthesis and degradation by mTORC1. Nature *513*, 440–443.
Zhou, Y., Li, G., Dong, J., Xing, X., Dai, J., and Zhang, C. (2018). MiYA, an efficient machine-learning workflow in conjunction with the YeastFab assembly strategy for combinatorial optimization of heterologous metabolic pathways in Saccharomyces cerevisiae. Metab. Eng. *47*, 294–302.