Title
Tools, strategies, and applications of synthetic biology in Saccharomyces cerevisiae

Author
Lee, Michael Eun-Suk

Publication Date
2015

Peer reviewed|Thesis/dissertation

Tools, strategies, and applications of synthetic biology in *Saccharomyces cerevisiae*

by

Michael Eun-Suk Lee

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Joint Doctor of Philosophy
with University of California, San Francisco

in

Bioengineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John E. Dueber, Chair
Professor Adam P. Arkin
Professor Jamie H. D. Cate
Professor Hana El-Samad

Spring 2015

Tools, strategies, and applications of synthetic biology in *Saccharomyces cerevisiae*

Abstract

Tools, strategies, and applications of synthetic biology in *Saccharomyces cerevisiae*

by

Michael Eun-Suk Lee

Joint Doctor of Philosophy

with University of California, San Francisco

in Bioengineering

University of California, Berkeley

Professor John E. Dueber, Chair

Synthetic biology is founded on the idea that cells are living machines that execute genetically encoded programs, and that we as engineers can reprogram them to perform new functions. Unlike man-made machines that are designed from the ground up, cells have been shaped and molded by evolution, and this makes them much more difficult to engineer. As synthetic biology has grown and matured, the field has shifted from focusing primarily on simpler bacterial chassis to engineering more complex and more powerful eukaryotic hosts, especially the budding yeast, *Saccharomyces cerevisiae*. Here we present technologies and strategies for efficiently engineering yeast, with an emphasis on metabolic engineering. First, we describe a practical framework for designing and constructing DNA for expression in yeast. This framework standardizes—and as a consequence, accelerates—the process of building new strains, enabling more rapid iteration and experimentation. We then develop a strategy for optimizing metabolic pathways by assembling combinatorial libraries that simultaneously titrate the expression of many genes. We identify strains with improved pathway flux for violacein biosynthesis and xylose utilization using mathematical modeling and selection, respectively. Finally, we attempt to alter the specificity of a native hexose transporter to exclusively import xylose without inhibition by glucose. In summary, this work is a combination of developing fundamental tools for engineering yeast and the application of those tools for lignocellulosic biomass fermentation.

Table of Contents

List of figures

List of tables

Acknowledgements

First and foremost, I would like to thank John for his mentorship these last five years. You taught me how to be a scientist. You trained me in the lab, and you showed me the importance of telling a good story. You gave me the freedom to explore, and for that I will always be grateful. I am incredibly lucky to have had you as an advisor.

I would like to thank my fellow members of the Dueber Lab, especially my partner in crime, Will. We've been here since the beginning of the lab, and while we may have lengthened our stay with various deviations from our research, it was well worth the experience.

I would like to thank my friends, especially my roommate, Dawn. From studying at MIT, to finding an apartment, to battling through these late stages of grad school, I couldn't have hoped for a better friend to share this time with.

Finally, I would like to thank my family for their unending love and support. To my parents, you raised me to have a passion for learning and provided help and guidance throughout this journey. To my sister, I can always turn to you for advice and for adventure. And to my brother, you inspire me to be a better person and remind me never to take anything for granted.

Chapter 1. Introduction

1.1 Motivation

Engineering cells is fundamentally difficult. We have the tools at our disposal to edit genes and genomes with relative ease, but with each modification we make, we are competing with millennia of evolution. We hope that we can predict the effects of knocking out or introducing a heterologous gene, and on a qualitative level, we often can. But when these effects are compounded after dozens or hundreds of manipulations, the smallest of details that was initially ignored can rise to the surface and break even the most carefully thought-out plans.

So how does one engineer a system that is so resistant to being engineered? This is the question that drives the field of synthetic biology forward and the challenge that drives the work presented in this dissertation.

In the early 1970's, biologists discovered restriction enzymes[1,2], used them to generate recombinant DNA molecules[3], and cloned a recombinant plasmid that was functional and could propagate in living cells[4]. Soon after, scientists invented techniques for sequencing[5,6], amplifying[7], and synthesizing DNA[8,9]. These technological advances were monumental in shaping modern molecular biology—it became possible to envision redesigning living organisms.

In recent years, we have come even closer to realizing that goal. We learned how to sequence[10-12], synthesize[13], and even replace[14] whole genomes. But while our knowledge about the inner workings of the cell has greatly expanded, we are simultaneously humbled by the realization of how much is yet to be understood. Still, we are persistent, and though we have had to temper our expectations, we have also become more methodical and creative in our approach to engineering biology. As synthetic biologists, we must walk the line between design and discovery, and be willing to learn from biology as much as we want to teach it to do new things.

1.2 Organization

In engineering disciplines, abstraction allows complex problems and systems to be broken down into smaller, well-behaved modules. In many fields, the rules for how these modules behave can be derived from first principles. In biology, the rules are less clearly understood, and are typically inferred from observations. However, the idea of abstraction is still a useful way to deconstruct large systems, even if the rules we follow are imprecise. In Chapter 2, we describe a standardized framework for designing and constructing DNA for heterologous expression in *Saccharomyces cerevisiae*. We also present characterization data of a sample of DNA parts constructed within that framework. These tools serve both as a means to more easily build new systems, but also as a way to rapidly iterate through designs in order to make more observations and improve our understanding of the rules of biology.

Another key component of engineering is modeling. Models are a simplified representation of a real-world system. The accuracy of a model is therefore a function of both our understanding of the system, and also the degree of simplification we desire. Simple models can be useful for making rough estimates about system behavior, and with the knowledge that much of biology is beyond our comprehension, hoping to construct a perfectly predictive model of the cell is not only unrealistic, but also may be unnecessary. In Chapter 3, we discuss using a linear regression model to describe a heterologous metabolic pathway expressed in yeast[15]. Using a set of characterized promoters, we construct a combinatorial library of the five genes in the violacein biosynthesis pathway. We develop a novel genotyping technique for analyzing random clones from this library, and after measuring metabolite production from a small sample of the total combinatorial space, we train our model to predict the production levels for arbitrary combinations from the library. This strategy infers the production landscape of a pathway from a relatively small number of measurements. These broad scale estimates are useful for directing future optimization efforts, which may involve more fine-tuning than any model could predict.

At some point, we must face the reality that biology is not today a traditional engineering discipline. However, despite its challenges, biology also has an advantage over other engineering fields: selection. Evolution is the source of many of the problems we face as biological engineers, but it is also a powerful tool if wielded adeptly. In Chapter 4, we take a similar approach as in Chapter 3 to construct a combinatorial library of the genes involved in xylose catabolism, but rather than use modeling, we harness the power of selection to identify the best expression levels of our genes[16]. Because we are able to link fitness to the function of our pathway (growth on xylose), we can coerce the cell into navigating the high-dimensional space to arrive at its optimal solution. Using this strategy, we interrogate the effects of including or excluding certain steps of the pathway, selecting in different growth conditions, and using mutants of one enzyme in the pathway. The results from this study suggest a number of bottleneck steps in the pathway that would be clear targets for further optimization.

Finally, in continuing to leverage selection as a tool, adaptation is an effective way to identify additional control points that may not have been obvious *a priori*. The combinatorial library strategies in Chapters 3 and 4 assume that the genes we choose have a significant impact on phenotype. However, in Chapter 4 we show that depending on which genes you include in the library, the results can be very different. Thus, in Chapter 5, we examine what happens if we remove our assumptions and allow random mutations to be selected. In this chapter, we attempt to engineer a transporter that specifically imports xylose into the cell and is uninhibited by its native substrate, glucose. We use a combination of rational and random mutagenesis strategies on the transporter gene itself, as well as an extended adaptation under selective pressure. While we do not identify any novel, beneficial mutations in the transporter in our first pass, the strain acquires mutations in its genome that confer a growth advantage. We sequence the genomes of some of these strains to try and locate the causative mutations, and we identify a number that may be worthy of future investigation.

1.3 References

1. Smith, H. O. & Wilcox, K. W. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *J Mol Biol* **51,** 379–391 (1970).
2. Kelly, T. J. & Smith, H. O. A restriction enzyme from Hemophilus influenzae. II. *J Mol Biol* **51,** 393–409 (1970).
3. Jackson, D. A., Symons, R. H. & Berg, P. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proc Natl Acad Sci U S A* **69,** 2904–2909 (1972).
4. Cohen, S. N., Chang, A. C., Boyer, H. W. & Helling, R. B. Construction of biologically functional bacterial plasmids in vitro. *Proc Natl Acad Sci U S A* **70,** 3240–3244 (1973).
5. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74,** 560–564 (1977).
6. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74,** 5463–5467 (1977).
7. Mullis, K. B. & Faloona, F. A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Meth Enzymol* **155,** 335–350 (1987).
8. Matteucci, M. D. & Caruthers, M. H. Synthesis of Deoxyoligonucleotides on a Polymer Support. *J Am Chem Soc* **103,** 3185–3191 (1981).
9. Hunkapiller, M. *et al.* A microchemical facility for the analysis and synthesis of genes and proteins. *Nature* **310,** 105–111 (1984).
10. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269,** 496–512 (1995).
11. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
12. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
13. Gibson, D. G. *et al.* Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science* **319,** 1215–1220 (2008).
14. Gibson, D. G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329,** 52–56 (2010).
15. Lee, M. E., Aswani, A., Han, A. S., Tomlin, C. J. & Dueber, J. E. Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay. *Nucleic Acids Res* **41,** 10668–10678 (2013).
16. Latimer, L. N. *et al.* Employing a combinatorial expression approach to characterize xylose utilization in Saccharomyces cerevisiae. *Metab Eng* **25,** 20–29 (2014).

Chapter 2. A Highly-characterized Yeast Toolkit for Modular, Multi-part Assembly

2.1 Introduction

Synthetic biology is driven by the desire to engineer novel biological functions that push the boundaries of what can be accomplished within living cells. Unfortunately, the potential power of the cell also brings with it a level of complexity that makes engineering biological systems extremely difficult. Synthetic biologists have sought ways to abstract the layers of complexity into components with predictable interactions, making it more feasible to undertake large engineering projects. Despite these efforts, the inner workings of the cell continue to elude understanding, and while certain elements can be highly predictable, the system behavior as a whole is difficult to anticipate. These challenges have led to an additional, and equally important, aspect to synthetic biology: rapid prototyping[1-4]. Because manipulations to the cell often lead to unexpected results, progress is best made by rapidly iterating through highly parallelized experiments to explore a wide parameter space[5,6]. It is the combination of these two principles—predictable parts and rapid prototyping—that give synthetic biologists the ability to approach difficult problems in energy[7,8], agriculture[9], and human health[10-12]. *Saccharomyces cerevisiae* is growing in popularity as a chassis for synthetic biology due to its powerful genetic tools[13-15], extensively studied biology[16-19], and long history of industrial applications[20-22]. In this work, we present a synthetic biology toolkit for engineering yeast that simplifies and accelerates experimentation in this important model organism.

Abstraction is a fundamental principle in any engineering discipline. It allows an engineer to focus on an individual component with the assurance that it will interface correctly with other components, both existing and future. When applied to synthetic biology, abstraction typically refers to the level of complexity of the DNA that is being built or introduced into cells. "Parts" are often thought of as one of the most basic DNA sequence elements that can be assigned a function. For example, a coding sequence, a transcriptional terminator, and an origin of replication could all be described as parts. Although these parts can be broken down further—they contain, among other things, a start and stop codon, a hairpin, and a protein binding site, respectively—the benefit of abstraction is the ability to ignore those lower level details and work with a part based solely on its reported function. Extensive efforts by others in the field have contributed to the Registry of Standard Biological Parts, a catalog of DNA sequences and characterization data that continues to grow each year (http://partsregistry.org)[23]. The Registry, however, is notably biased towards working in bacterial systems, particularly *Escherichia coli*, and with growing interest in yeast as a synthetic biology host, it is becoming apparent that the field needs a more comprehensive set of standard yeast parts. For this toolkit, we collected, constructed, and characterized a starter set of useful parts to lay the foundation for a standardized engineering platform.

Prototyping is a more necessary step in synthetic biology than in other engineering fields, as synthetic biologists lack the ability to accurately predict behavior, even of de-

vices made from parts of known function[24-26]. When working in fast-growing cells such as yeast, cloning is often the bottleneck step in an experimental cycle. The lag between having a DNA design and actually obtaining the physical DNA is far too long to support a robust prototyping workflow. The solution that many groups have developed is standardization of cloning[27-32]. For example, the BioBrick standard (and its relatives) defines a set of restriction enzyme sites that are used to flank each part in a vector[27,30]. When those restriction enzymes are used to join two parts, the junction contains an assembly "scar", and the resulting plasmid reconstitutes the sites external to the newly combined parts (an idempotent operation). This enables an endless number of cycles of pair-wise assembly.

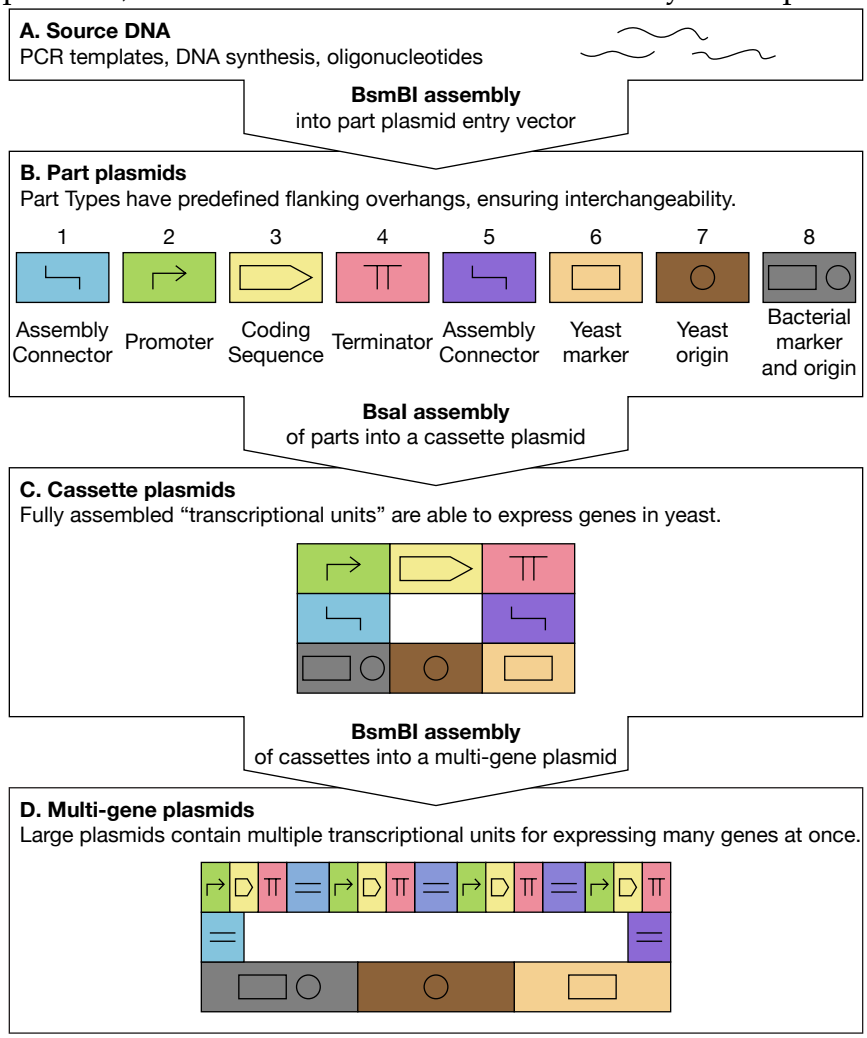


Figure 2-1. Standardized, hierarchical assembly strategy based on MoClo. (A) Source DNA is obtained via PCR, DNA synthesis, or oligonucleotides, then assembled using BsmBI into a part plasmid entry vector. (B) Part plasmids of a particular Type have unique upstream and downstream BsaI-generated overhangs. All part plasmids of the same Type are therefore interchangeable. Plasmids at this stage typically confer chloramphenicol resistance in *E. coli*. One part plasmid of each Type is assembled using BsaI to form a cassette plasmid. (C) Cassette plasmids contain a complete transcriptional unit (TU), and can be transformed directly into yeast. Plasmids at this stage typically confer ampicillin resistance in *E. coli*. Alternatively, cassette plasmids can be further assembled using BsmBI to form a multi-gene plasmid. (D) Multi-gene plasmids contain multiple TUs, the order of which is dictated by the Assembly Connector parts used to flank the individual cassettes. Plasmids at this stage typically confer kanamycin resistance in *E. coli*.

More recently, Golden Gate assembly-based methods have increased in popularity due to the added flexibility provided by the use of Type IIs restriction enzymes, which cut outside their recognition sequence and provide unique cohesive ends to enable directional, multi-insert, one-pot cloning[33]. One example is the MoClo (modular cloning) system, which categorizes parts as "types" based on their function and location in a completed device (*e.g.*, promoter types or coding sequence types) and designates particular overhangs that flank each type, allowing all parts of a particular type to be interchangeable[32]. In this work, we adapted the MoClo strategy specifically to build yeast expression devices. The major advantage of using a standardized system such as MoClo is that once parts are constructed, they are immediately available for incorporation into devices and no longer require synthesis of oligonucleotides, PCR amplification and purification, or verification by sequencing. This allows us to construct from parts, a plasmid carrying multiple gene expression devices in as little as two days.

2.2 Results and Discussion

2.2.1 Definition of an Assembly Standard for Yeast

Our standard for assembling DNA for expression in yeast is a bottom-up hierarchical approach to DNA construction (**Figure 2-1**). A description of the assembly scheme, part types, and overhang sequences are discussed briefly here and in more detail in **Section 2.3**. For brevity, Golden Gate assemblies using either BsaI or BsmBI are referred to as "BsaI assembly" and "BsmBI assembly".

Our workflow for assembling complex plasmids for expressing multiple genes in yeast has multiple steps that correspond to our abstraction layers. First, source DNA is obtained through PCR, synthesis or another user-preferred method. That source DNA is "domesticated" via BsmBI assembly into a universal entry vector, resulting in a "part" plasmid. Part plasmids come in different Types, numbered 1 through 8 (with some optional subtypes). Each part Type is defined by the sequences of the upstream and downstream flanking overhangs generated when digested by BsaI. All parts of a particular Type are interchangeable, which lends the system well to combinatorial experiments. Part plasmids are joined in a BsaI assembly to form a "cassette" plasmid that, in most cases, is used to express a single gene in yeast (a transcriptional unit, TU, comprised of a promoter, coding sequence, and terminator). These cassettes can optionally be joined in a final BsmBI assembly to form "multi-gene" plasmids that, as the name suggests, are used to simultaneously express multiple genes. The multi-gene assembly is enabled by the use of Assembly Connectors (Type 1 and 5) that, in similar fashion to each part plasmid's unique BsaI overhangs, contain unique BsmBI overhangs that flank each cassette. At each round of assembly, the antibiotic selection is changed to minimize background (typically, chloramphenicol → ampicillin → kanamycin). Using this workflow, we can construct a multi-gene plasmid from PCR templates in only three days. This construction time is typically reduced to only two days, since, in most cases, the final multi-gene plasmids are built from existing parts.

There are many benefits to the standard we defined, which should prove useful to synthet-

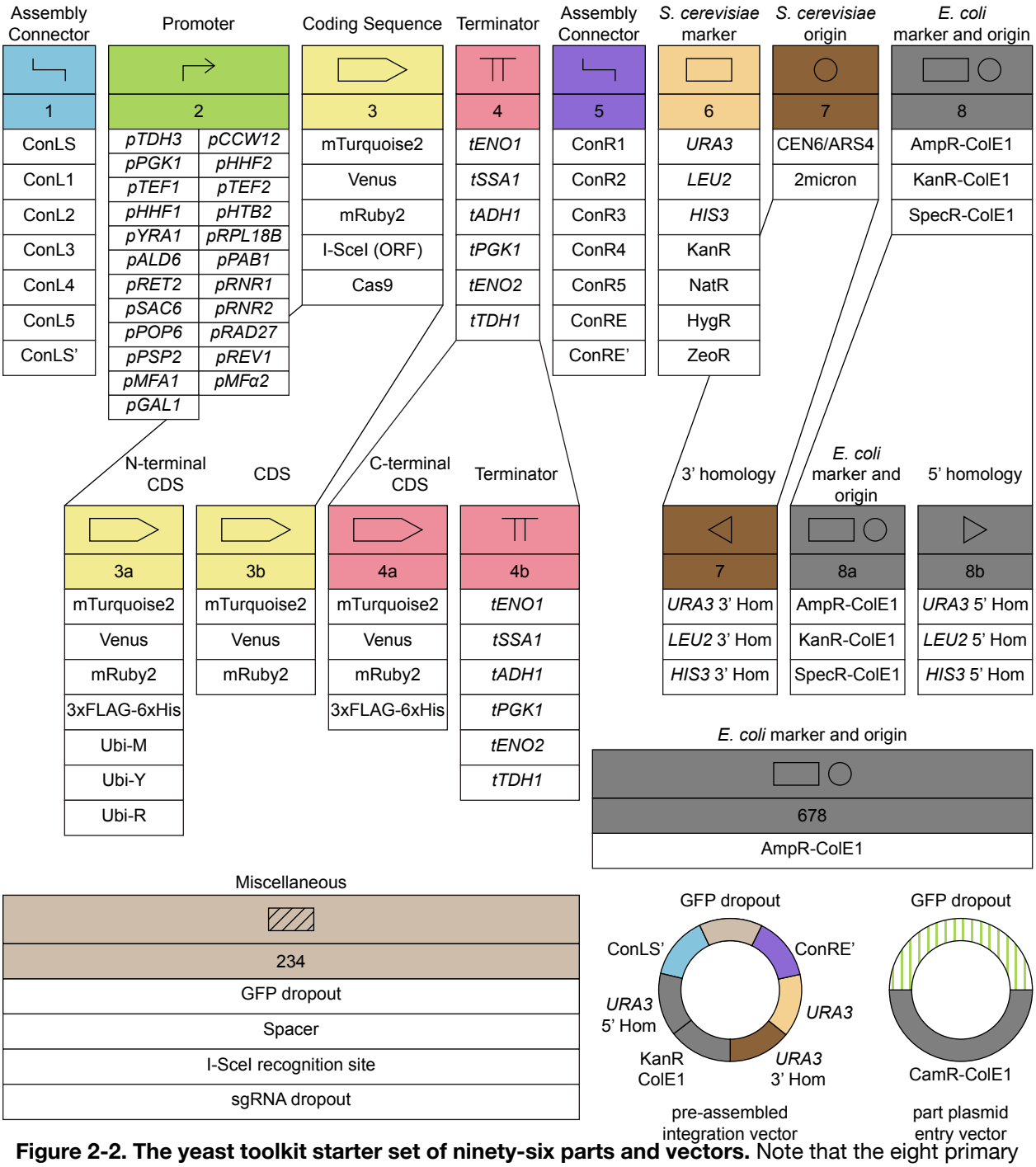


Figure 2-2. The yeast toolkit starter set of ninety-six parts and vectors. Note that the eight primary part Types can be further divided into subtypes (e.g., 3a/3b), or combined to make composite types (e.g. 234). Each Type has a unique upstream and downstream overhang pair, and a complete cassette can be assembled when a complete path can be drawn from left to right (1 to 8). For example, the pre-assembled integration vector is assembled from: ConLS' (1), GFP dropout (234), ConRE' (5), *URA3* (6), *URA3* 3' Homology (7), KanR-ColE1 (8a), and *URA3* 5' Homology (8b). A transcriptional unit (promoter, coding sequence, terminator) can be assembled into this vector, replacing the BsaI-flanked GFP dropout. A set of cassettes can also be assembled into this vector, due to the special Assembly Connectors ConLS' and ConRE' that have the BsmBI recognition sites in the reverse orientation (**Section 2.3.1**). The part plasmid entry vector is used for constructing new parts. A table of plasmid names, parts, and Types is included in **Table 2-1**.

ic biologists with a wide range of needs. First, the cloning protocols are extremely simple, requiring no PCR amplification or purification steps after the initial part creation. Second, the standardized Golden Gate assemblies are highly robust. It was previously shown that for a 10-part assembly with an optimized set of overhangs, 97% of isolated transformants contained a correctly assembled plasmid[34]. We observed comparable efficiencies in this work and screened only one transformant for almost all plasmid assemblies described here. Because PCR- and oligonucleotide-derived point mutations cannot occur after the construction of part plasmids, we do not typically sequence downstream assemblies and instead use simple restriction mapping to verify size. Third, our workflow supports a simple method for chromosomal integration in which plasmids designed for integration can be transformed directly after being linearized via a NotI digestion[35]. Fourth, our design specification includes unique restriction enzyme sites that make cassettes both BioBrick- and BglBrick-compatible, and multi-gene plasmids BioBrick-compatible. While a variety of restriction sites (BamHI, BbsI, BglII, BsaI, BsmBI, EcoRI, NotI, PstI, SpeI, XbaI, and XhoI) have been removed from all parts in the toolkit for this increased flexibility, only BsaI, BsmBI, and NotI must be removed from new parts to conform to the complete assembly scheme described here. Finally, the Assembly Connectors, in addition to harboring BsmBI sites, can also act as homology sequences for recombination-based cloning, such as sequence and ligation-independent cloning (SLIC)[36], Gibson Assembly[37,38], Ligase Cycling Reaction (LCR)[39], or yeast *in vivo* assembly[40], if those methods are preferred.

2.2.2 A Toolkit of Yeast Parts

Although an assembly standard has some inherent value, its utility is determined in large part by the availability of parts. To this end, we have compiled a collection of 96 parts compatible with this standard for efficiently engineering yeast strains (**Figure 2-2** and **Table 2-1**). This starter collection contains an assortment of promoters, terminators, fluorescent proteins, peptide tags, selectable markers, and origins of replication, as well as a part entry vector into which new parts can be cloned. Additionally, we have included sequences targeting chromosomal loci for integration, and genome-editing tools for introducing double-strand breaks to stimulate homologous recombination. Finally, rather than provide a large array of different vectors, the assembly standard enables construction of custom vectors directly from parts in the toolkit, and one such vector is included as an example (see **Section 2.3.2**).

2.2.3 Characterization of Promoters

We have characterized twenty constitutive promoters, two mating-type-specific promoters, and one inducible promoter, all cloned from the yeast genome (although synthetic promoters[41,42] could easily be ported into the system as well). The promoters were selected to span a wide range of transcriptional strengths while minimizing variability between growth conditions[43]. In general, they constitute the 700bp directly upstream of the native start codon, although in some cases where another ORF was less than 700bp away, we cloned only the intergenic non-coding region. To examine the strength of each promoter, we cloned it upstream of a fluorescent reporter (mRuby2, Venus, and mTurquoise2) and measured bulk fluorescence on a plate reader.

Table 2-1. List of toolkit plasmids.

Plasmid	Type	Description	*E. coli* marker
pYTK001	entry vector	Part Plasmid Entry Vector	CamR
pYTK002	1	ConLS	CamR
pYTK003	1	ConL1	CamR
pYTK004	1	ConL2	CamR
pYTK005	1	ConL3	CamR
pYTK006	1	ConL4	CamR
pYTK007	1	ConL5	CamR
pYTK008	1	ConLS'	CamR
pYTK009	2	*pTDH3*	CamR
pYTK010	2	*pCCW12*	CamR
pYTK011	2	*pPGK1*	CamR
pYTK012	2	*pHHF2*	CamR
pYTK013	2	*pTEF1*	CamR
pYTK014	2	*pTEF2*	CamR
pYTK015	2	*pHHF1*	CamR
pYTK016	2	*pHTB2*	CamR
pYTK017	2	*pYRA1*	CamR
pYTK018	2	*pRPL18B*	CamR
pYTK019	2	*pALD6*	CamR
pYTK020	2	*pPAB1*	CamR
pYTK021	2	*pRET2*	CamR
pYTK022	2	*pRNR1*	CamR
pYTK023	2	*pSAC6*	CamR
pYTK024	2	*pRNR2*	CamR
pYTK025	2	*pPOP6*	CamR
pYTK026	2	*pRAD27*	CamR
pYTK027	2	*pPSP2*	CamR
pYTK028	2	*pREV1*	CamR
pYTK029	2	*pMFA1*	CamR
pYTK030	2	*pMFα2*	CamR
pYTK031	2	*pGAL1*	CamR
pYTK032	3	mTurquoise2	CamR
pYTK033	3	Venus	CamR
pYTK034	3	mRuby2	CamR
pYTK035	3	I-SceI (ORF)	CamR
pYTK036	3	Cas9	CamR
pYTK037	3a	mTurquoise2	CamR
pYTK038	3a	Venus	CamR
pYTK039	3a	mRuby2	CamR
pYTK040	3a	3XFLAG-6XHis	CamR
pYTK041	3a	Ubi-M	CamR
pYTK042	3a	Ubi-Y	CamR
pYTK043	3a	Ubi-R	CamR
pYTK044	3b	mTurquoise2	CamR
pYTK045	3b	Venus	CamR
pYTK046	3b	mRuby2	CamR
pYTK047	234r	GFP dropout	CamR
pYTK048	234	Spacer	CamR

pYTK049	234	I-SceI recognition site	CamR
pYTK050	234	sgRNA Dropout	CamR
pYTK051	4	*tENO1*	CamR
pYTK052	4	*tSSA1*	CamR
pYTK053	4	*tADH1*	CamR
pYTK054	4	*tPGK1*	CamR
pYTK055	4	*tENO2*	CamR
pYTK056	4	*tTDH1*	CamR
pYTK057	4a	mTurquoise2	CamR
pYTK058	4a	Venus	CamR
pYTK059	4a	mRuby2	CamR
pYTK060	4a	3XFLAG-6XHis	CamR
pYTK061	4b	*tENO1*	CamR
pYTK062	4b	*tSSA1*	CamR
pYTK063	4b	*tADH1*	CamR
pYTK064	4b	*tPGK1*	CamR
pYTK065	4b	*tENO2*	CamR
pYTK066	4b	*tTDH1*	CamR
pYTK067	5	ConR1	CamR
pYTK068	5	ConR2	CamR
pYTK069	5	ConR3	CamR
pYTK070	5	ConR4	CamR
pYTK071	5	ConR5	CamR
pYTK072	5	ConRE	CamR
pYTK073	5	ConRE’	CamR
pYTK074	6	*URA3*	CamR
pYTK075	6	*LEU2*	CamR
pYTK076	6	*HIS3*	CamR
pYTK077	6	KanamycinR	CamR
pYTK078	6	NourseothricinR	CamR
pYTK079	6	HygromycinR	CamR
pYTK080	6	ZeocinR	CamR
pYTK081	7	CEN6/ARS4	CamR
pYTK082	7	2micron	CamR
pYTK083	8	AmpR-ColE1	AmpR
pYTK084	8	KanR-ColE1	KanR
pYTK085	8	SpecR-ColE1	SpecR
pYTK086	7	*URA3* 3’ Homology	CamR
pYTK087	7	*LEU2* 3’ Homology	CamR
pYTK088	7	*HIS3* 3’ Homology	CamR
pYTK089	8a	AmpR-ColE1	AmpR
pYTK090	8a	KanR-ColE1	KanR
pYTK091	8a	SpecR-ColE1	SpecR
pYTK092	8b	*URA3* 5’ Homology	CamR
pYTK093	8b	*LEU2* 5’ Homology	CamR
pYTK094	8b	*HIS3* 5’ Homology	CamR
pYTK095	678	AmpR-ColE1	AmpR
pYTK096	cassette	Pre-Assembled *URA3* Integration Vector	KanR

It was previously shown that the strength of constitutive promoters cloned from the yeast genome was largely independent of the downstream coding sequence[44], an important distinction between controlling expression in bacteria and yeast. This held true for the twenty constitutive promoters characterized in this work (which include some overlap with Lee et al, 2013) (**Figure 2-3**). The promoters span a range of up to three orders of magnitude, and there are also some promoters that have very similar expression strengths, allowing them to be interchanged so as to reduce the risk of undesired homologous recombination in multi-gene plasmids due to repeated sequences. Although we only tested these promoters in one type of media, the majority of native yeast promoters have been shown to maintain their relative expression strengths in different growth conditions, although the absolute strengths may change[45].

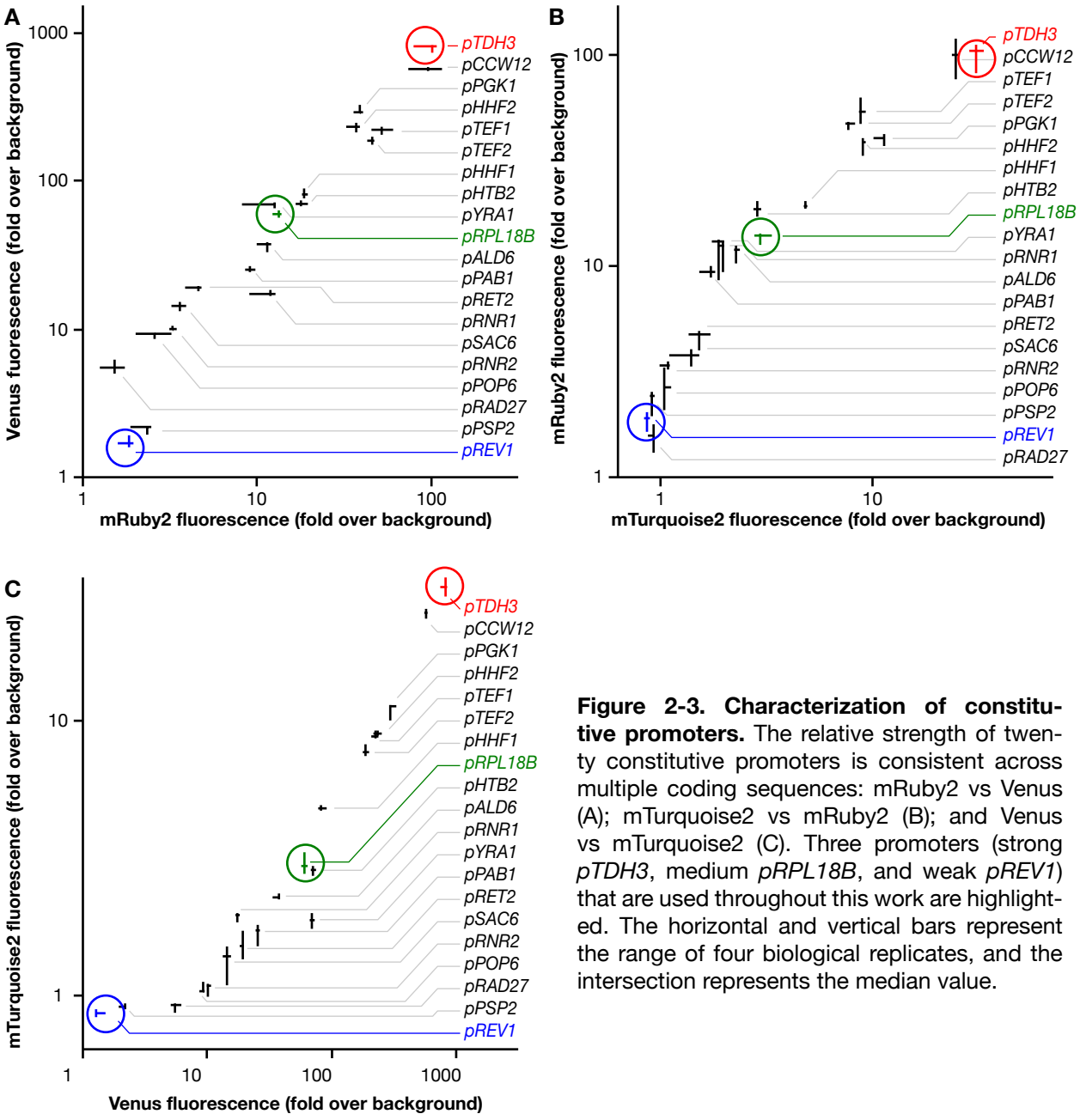


Figure 2-3. Characterization of constitutive promoters. The relative strength of twenty constitutive promoters is consistent across multiple coding sequences: mRuby2 vs Venus (A); mTurquoise2 vs mRuby2 (B); and Venus vs mTurquoise2 (C). Three promoters (strong *pTDH3*, medium *pRPL18B*, and weak *pREV1*) that are used throughout this work are highlighted. The horizontal and vertical bars represent the range of four biological replicates, and the intersection represents the median value.

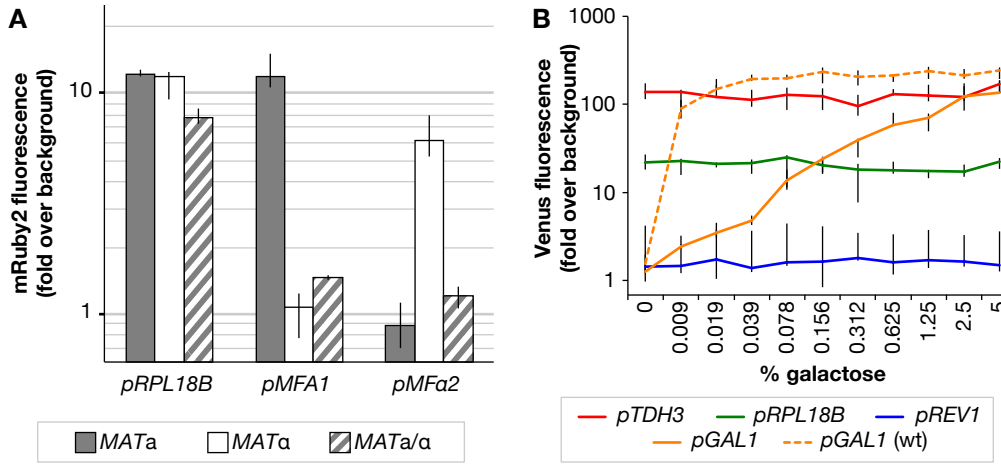


Figure 2-4. Characterization of additional promoters. (A) The mating-type-specific promoter, *pMFA1*, is only active in the *MAT*a haploid; *pMFα2* is only active in *MAT*α haploids; neither promoter is active in the opposite haploid or in the diploid. The expression level of *pRPL18B* in the three strains is shown for reference. The height of the bars represents the median value of four biological replicates, and the error bars show the range. (B) Galactose induction of *pGAL1* increases expression from background levels up to the highest expressing constitutive promoter, *pTDH3*. All solid line data were collected from a *Δgal2* strain. The dashed line shows a much more sensitive response to galactose induction in a wild type strain. Points represent the median value of four biological replicates, and error bars show the range.

It is sometimes useful to have genes under dynamic control, and for this we provide two tools: mating-type-specific and galactose-inducible promoters. We tested *pMFA1* and *pMFα2* and found that they have very close to background levels of fluorescence in both the opposite mating-type haploid and diploid strains and a 6- to 10-fold induction in the appropriate haploid (**Figure 2-4A**). We also tested *pGAL1* in varying concentrations of galactose and observed a 100-fold induction (**Figure 2-4B**). Although the promoter can be used in wild-type strains, the response is very sensitive to low concentrations of galactose; a strain with the *GAL2* transporter knocked out should be used for more graded control over expression[46].

2.2.4 Characterization of Terminators

The impact of different transcriptional terminators on gene expression can vary considerably, and could provide a secondary mode of control to complement the promoters. However, for simplicity, we opted in this toolkit to provide terminators that yielded approximately the same expression output. Using expression data from the whole yeast genome[43], we selected six of the most highly expressed genes and cloned the 225bp immediately downstream of the stop codon. We assembled these terminators with each of our three fluorescent reporters and each using three promoters. The largest difference in expression we observed between terminators for a given promoter and fluorescent protein was 3.6-fold (**Figure 2-5**). In general, the fold-changes produced by different promoters were greater than those effected by the terminators, but this was not always the case. If applications are sensitive to small fold-changes of expression, we advise characterizing individual promoter-terminator pairs to ensure that the desired levels of expression are obtained.

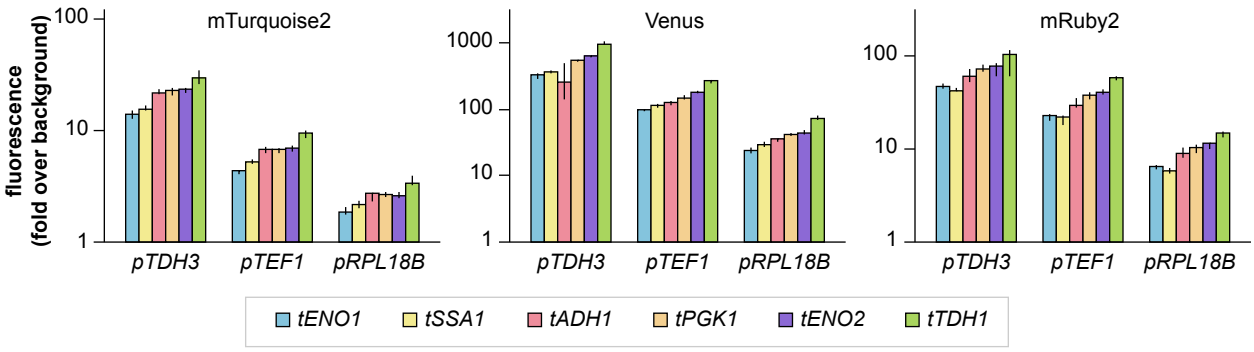


Figure 2-5. Characterization of terminators. Six terminators were cloned behind three fluorescent proteins, each driven by three promoters. The relative expression levels for this set of terminators are largely independent of the coding sequence and the promoter. The height of the bars represents the median value of four biological replicates, and the error bars show the range.

2.2.5 Protein Degradation Tags

In addition to controlling transcript levels, protein levels can be tuned by fusing degradation tags to the N-terminus. We have included three such tags of varying strengths—Ubi-M (weak), Ubi-Y (medium), and Ubi-R (strong)—which can be used to adjust the rate of protein turnover[47]. We fused these tags to the N-terminus of mRuby2, and expressed them using a strong, moderate, and weak promoter—*pTDH3*, *pRPL18B*, and *pREV1*, respectively. The strong degradation tag (Ubi-R) resulted in no detectable fluorescence at any expression level, while the medium strength degradation tag (Ubi-Y) resulted in detectable levels of fluorescence at only the highest expression level (**Figure 2-6**).

2.2.6 Copy Number, Gene Expression, and Single-Cell Variability

When engineering yeast strains expressing multiple heterologous proteins, it is important to consider the relative expression of those proteins. As described above, protein levels can be controlled by changing promoters, terminators, or degradation rates. However, another important consideration is the copy number of the gene(s). Typically, one of three systems is used to express genes in yeast: single-copy integrations into the chromosome, low-copy centromeric plasmids, and high-copy 2micron plasmids. One could easily as-

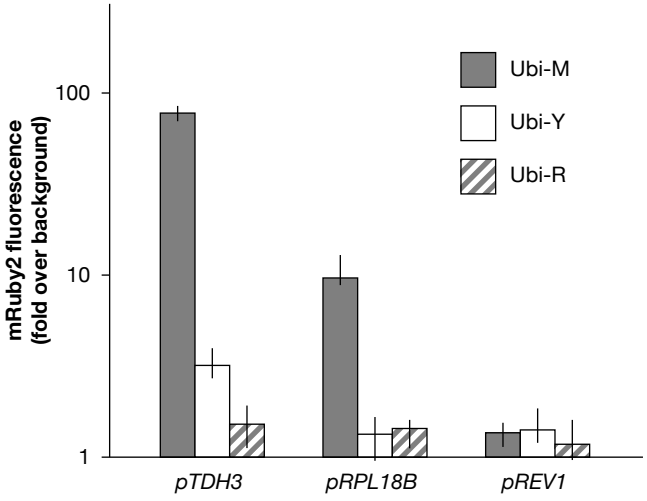


Figure 2-6. Protein degradation tags. Three N-terminal degradation tags were fused to mRuby2 and expressed using three different promoters. Steady-state fluorescence levels are dependent on the difference between the strength of the promoter and the strength of the degradation tag. The height of the bars represents the median value of six biological replicates, and the error bars show the range.

sume that the differences in copy number simply titrate gene expression accordingly, but we observed that there are subtle, but important, effects that could influence the decision to use one system over another.

We cloned cassettes expressing either mRuby2 or Venus under strong, moderate, and weak promoters (*pTDH3*, *pRPL18B*, and *pREV1*, respectively). Versions of these cassettes were made for each of the three copy numbers. Finally, each of the nine possible combinations of the three promoters and two genes were either assembled in tandem onto a single chromosomal locus/plasmid or kept separate in two loci/plasmids. We measured bulk fluorescence of both fluorescent proteins to compare protein expression levels of the cell populations at the three copy-numbers (**Figure 2-7**).

In the chromosomally integrated strains, the different promoter combinations fill out the points of a regular grid, as expected. In the low-copy CEN6/ARS4 plasmid system, the absolute fluorescence is generally higher compared to the chromosome, again, as expected. Interestingly, the range between the highest and lowest expression is actually slightly greater in the CEN6/ARS4 plasmid system. Compared to low-copy plasmids, the high-copy 2micron plasmids showed considerably more irregular expression patterns. In the two-plasmid, 2micron system, the grid is preserved, but compressed at higher expression levels, suggesting that some expression machinery in the cell is limiting and that having more copies of the DNA has little effect on increasing expression—*i.e.*, the average

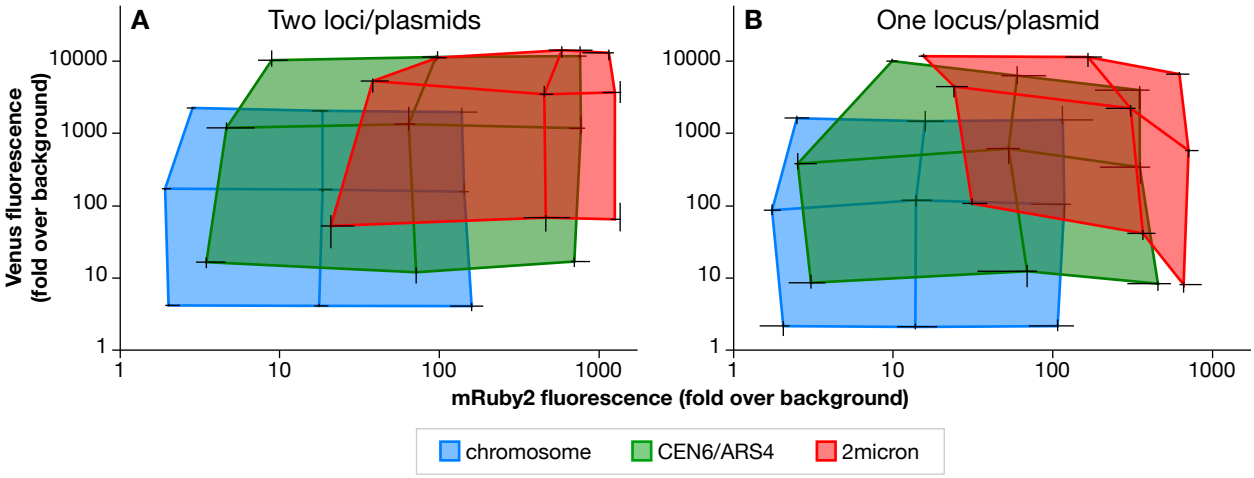


Figure 2-7. The effect of copy number on gene expression. Three promoters (*pTDH3*, *pRPL18B*, and *pREV1*) drive two fluorescent proteins (mRuby2 and Venus) in all nine possible combinations. These nine combinations are integrated into the chromosome (blue), expressed from a low-copy plasmid (green), and expressed from a high-copy plasmid (red). The translucent, shaded boxes show the range of expression spanned by each respective copy number. For lower strength promoters, increasing copy number gives higher fluorescence; but for the strongest promoter, there is a much smaller difference between the low- and high-copy plasmids. Each gene is integrated in a separate locus or expressed from a separate plasmid (A); both genes are integrated in tandem at the same locus or expressed from a single plasmid (B). When the two genes are on the same high copy plasmid, higher expression of one gene appears to reduce the average expression of the second. Single-cell fluorescence data in **Figure 2-8** shows two subpopulations, one with the expected level of fluorescence and another with reduced fluorescence, which explains the lowered average of bulk measurements. The horizontal and vertical bars represent the range of four biological replicates, and the intersection represents the median value.

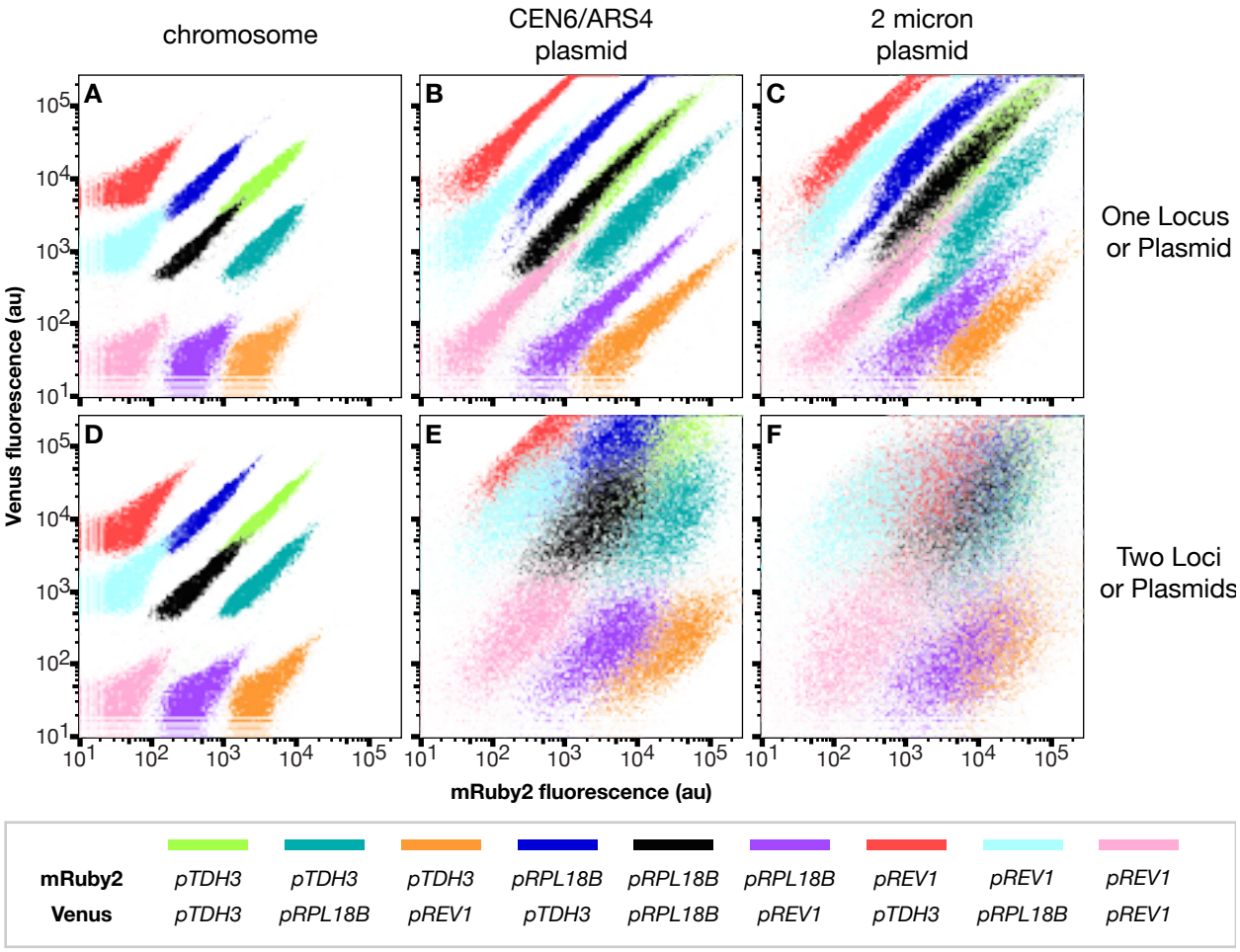


Figure 2-8. The effect of copy number on single-cell gene expression. The same strains expressing mRuby2 and Venus that were measured for bulk fluorescence in **Figure 2-7** were run on a flow cytometer: chromosomally integrated in a single locus (A) or two loci (D); on a single (B) or two (E) low-copy plasmids; on a single (C) or two (F) high-copy plasmids. As copy number increased, the variability of expression also increased. For all single-locus strains, the expression of the two fluorescent proteins was well correlated, suggesting that copy number is the main contributor to variation in expression. When expressed from two plasmids, correlation between fluorescent proteins is lost, suggesting that the copy number of each plasmid is independent of the other. Dot plots for each sample represent 10,000 events.

fluorescence of cells with the strongest promoter is similar between low- and high-copy plasmids. In the single-plasmid, 2micron system, not only is the grid compressed, but also it appears that high expression of one gene seems to reduce the expression of the second gene (**Figure 2-7B**). Based on flow cytometry, there appears to be a bimodal distribution for some of these populations (**Figure 2-8C**, *e.g.*, *pTDH3*-mRuby2/*pRPL18B*-Venus), which is consistent with previous studies comparing the distribution of expression in 2micron and chromosomally integrated systems[48]. Interestingly, this effect is not nearly as pronounced in the chromosome or on the low-copy plasmid. It is unclear why this would be specific to the high-copy plasmid. Based on these data, we believe that use of high-copy 2micron plasmids should generally be avoided, since the highest expression levels accessible by them are very nearly accessible by low-copy CEN6/ARS4 plasmids, and low-copy plasmids give greater access to lower expression, and in general have less erratic expression patterns.

Another parameter we examined was cell-to-cell variability in the relative expression of two genes. While it has been shown that strains expressing fluorescent proteins from chromosomally integrated genes display much tighter distributions compared to those expressing from 2micron plasmids[48], we were curious about any additional effects of propagating one versus multiple plasmids. We took the same cultures used to measure bulk fluorescence and ran them on a flow cytometer to measure single-cell fluorescence of the two fluorescent proteins (**Figure 2-8**). As expected, the single-cell measurements revealed that the variability in fluorescence increased considerably when moving from the chromosome to a low-copy to a high-copy plasmid, indicating that the precise copy number of these plasmids is not tightly regulated. When expressed from a single locus/plasmid, the expression of the two fluorescent proteins was well correlated, as evidenced by the distribution of each strain along the diagonal. This result suggests that DNA copy number is the primary source of added variation in plasmid-based expression systems, a model which is further supported by the data from two loci/plasmids. Strains expressing mRuby2 and Venus from two separate loci in the chromosome showed distributions that are nearly identical to the single locus, chromosomal strains. In contrast, the low-copy and high-copy plasmids lose their tight correlation between the two fluorescent proteins when the two genes are expressed from separate plasmids. Thus, not only is the copy number of a plasmid highly variable, the relative copy numbers of two plasmids in the same cell are not well correlated. Therefore, we would recommend that genes be integrated into the chromosome whenever possible. If, however, higher expression than can be attained from the chromosome is required, use of a low-copy plasmid is preferred, and all genes should be expressed from the same plasmid rather than split onto multiple plasmids. Accordingly, the assembly standard we provide here accommodates the facile assembly of up to six genes on a single plasmid or in a single chromosomal locus, with more possible if additional Assembly Connectors are designed.

2.2.7 High-efficiency Integrations into the Chromosome

Yeast is a very powerful genetic organism due to its efficient homologous recombination machinery. This allows for site-specific integration of DNA into the chromosome by simply transforming linear DNA flanked by sequences homologous to the target locus. However, compared to plasmid transformation, integrations usually result in almost an order of magnitude reduction in colony counts, which is one reason why the use of plasmids is often preferred. Given the desire for chromosomal integrations described earlier, it is evident that a higher efficiency method for integrating into the chromosome is necessary, particularly when working with large libraries. Fortunately, it was previously shown that transformation efficiency could be dramatically improved by using a homing endonuclease to generate a double-strand break in the chromosome and stimulate recombination[49]. More recently, the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR-associated (Cas) system has been used for similar purposes[50-52]. We tested both systems to directly compare their effects on chromosomal library construction.

First, we prepared an experimental strain by integrating a "landing pad" using conventional homologous recombination of linear DNA (**Figure 2-9A**). This landing pad contained an I-SceI recognition site; the I-SceI recognition site conveniently contains an NGG

protospacer adjacent motif (PAM) close to the I-SceI cutting site, and we added an extra 10bp upstream of the site to create a 20bp targeting sequence for a single guide RNA (sgRNA); we also included a partial *URA3* coding sequence and terminator that by itself is non-functional.

Next, we designed the repair DNA we were integrating to contain a Venus-expressing cassette and a *HIS3*-expressing cassette, flanked by homology to the sequence upstream

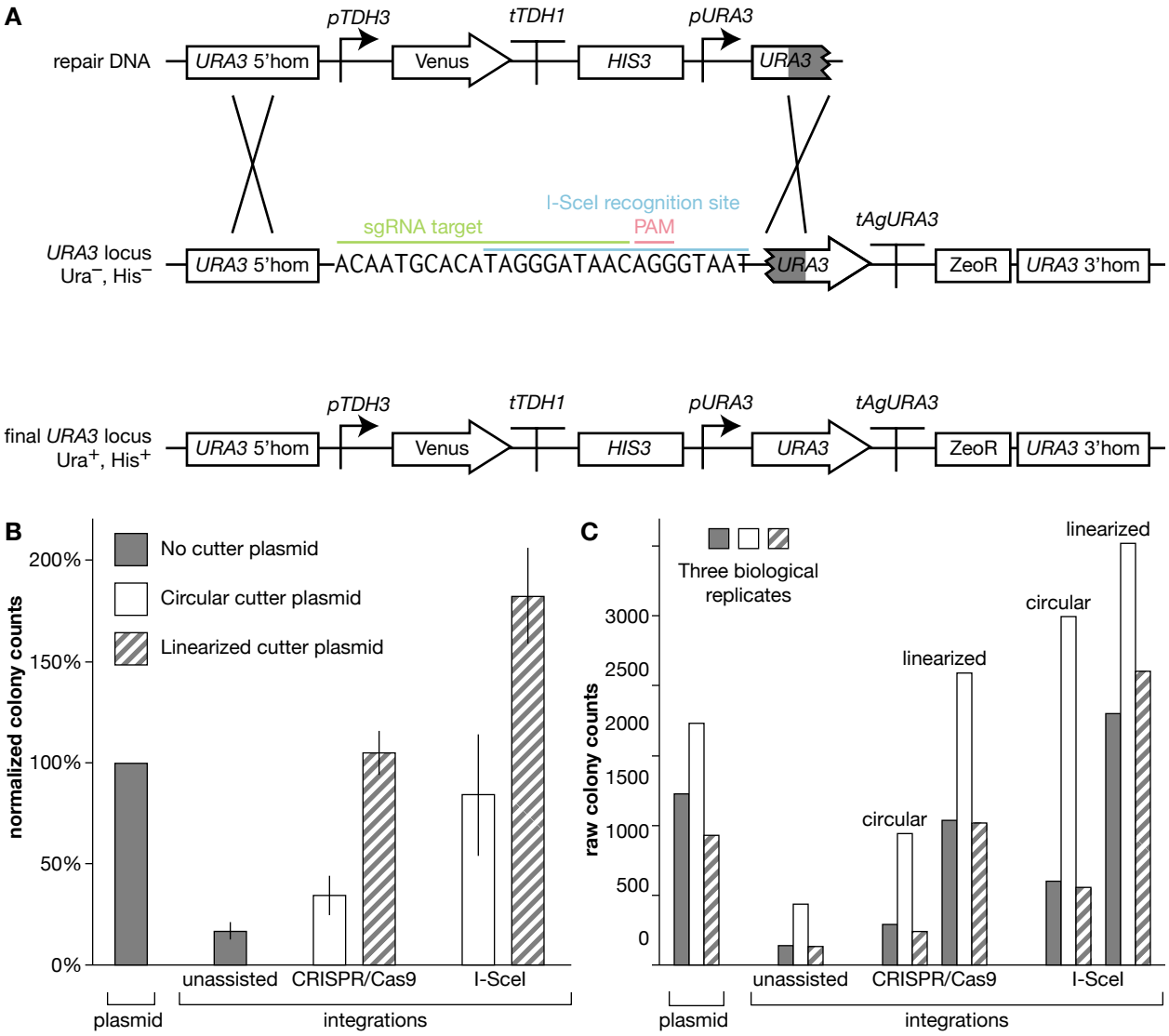


Figure 2-9. High-efficiency integration into the chromosome. (A) Schematic of the integration "landing pad" at the *URA3* locus, the repair DNA that targets the landing pad, and the final *URA3* locus after successful integration. (B) Integration of linear DNA into the chromosome by homologous recombination yields 6-fold fewer colonies (compare shaded bars). Adding in a Cas9 or I-SceI improves transformation efficiency by 2-fold and 5-fold, respectively (compare white bars to unassisted integration). Linearizing the Cas9 or I-SceI expression plasmid prior to transformation further improves transformation efficiency to match plasmid transformation efficiency or exceed it by 1.8-fold, respectively (compare striped bars to plasmid transformation). Colony counts are normalized to the number of colonies from the plasmid transformation for each replicate. The height of the bars represents the mean value of three biological replicates, and the error bars show the standard error of the mean. (C) Individual replicate data from (B).

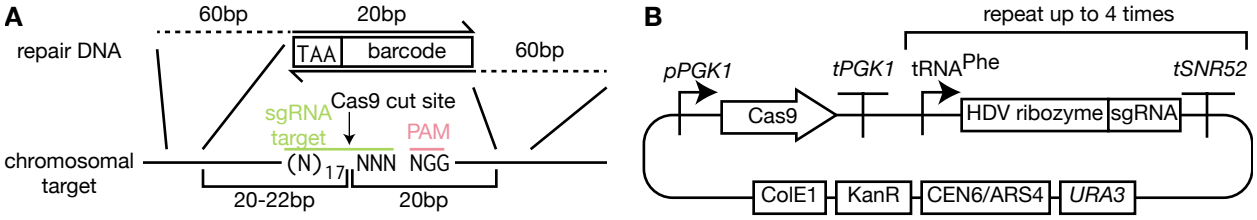


Figure 2-10. Design of CRISPR/Cas9 knockout constructs. (A) Repair DNA was designed with 60bp of flanking homology to the chromosomal target. The upstream homology region was chosen such that the repair DNA would introduce an in-frame stop codon and delete the protospacer sequence. (B) Map of the Cas9/sgRNA expression plasmid; up to four sgRNAs were assembled onto a single plasmid.

of the landing pad and to the partial *URA3* marker. Thus, when the DNA integrated successfully, the cells would be prototrophic for histidine and uracil, and they would be fluorescent. If the DNA integrated off-target, the cells would be prototrophic for histidine and fluorescent, but they would remain auxotrophic for uracil, allowing us to measure the rate of off-target integration by selecting on 5-fluoroorotic acid (5-FOA).

Finally, we compared the efficiency of integration when the repair DNA was transformed unassisted, with a transient "cutter" plasmid (we did not select for it) expressing Cas9 and an sgRNA, or with a transient cutter plasmid expressing I-SceI. As a control for cell competency, we transformed a circular version of the repair DNA that also contained an origin of replication (**Figure 2-9B and C**). Compared to the plasmid transformation, unassisted integration gave approximately 6-fold fewer colonies. When a CRISPR/Cas9 cutter plasmid was co-transformed, there was a 2-fold increase in colony count over unassisted integration; when an I-SceI cutter plasmid was co-transformed, there was an additional 2.5-fold increase (5-fold compared to unassisted). We were able to further improve the efficiency for both the Cas9 and I-SceI systems by linearizing (with a restriction digest) the cutter plasmid prior to transformation. Doing so brought the efficiency of Cas9-assisted integration to match that of plasmid transformation. Incredibly, the linearized I-SceI-expressing DNA actually increased integration efficiency to 1.8-fold over the rate of plasmid transformation. We measured the rate of off-target integration for this most efficient method (linearized I-SceI) and found that only 0.02% of transformants were 5-FOA resistant, and therefore had integrated the repair DNA improperly. By measuring Venus fluorescence, we found that 0.14% of transformants contained multiple integrations. It is unclear why linearizing the cutter plasmid increases the transformation

Table 2-2. List of protospacer sequences for CRISPR/Cas9 knockouts.

Target	Protospacer Sequence
LEU2 1	AAGAAGATCGTCGTTTTGCC
LEU2 2	GGTGACCACGTTGGTCAAGA
HIS3 1	AGTAAAGCGTATTACAAATG
HIS3 2	ATTGCGATCTCTTTAAAGGG
MET15 1	GATACTGTTCAACTACACGC
MET15 2	GCCAAGAGAACCCTGGTGAC
TRP1 1	ATTAATTTCACAGGTAGTTC
TRP1 2	GGTCCATTGGTGAAAGTTTG

efficiency, but one possibility is that linear DNA enters the cell and/or nucleus more efficiently than circular plasmids. Regardless, the ease with which sequences can now be integrated into the chromosome should further encourage the use of integrations over plasmids, even with the high transformation efficiency requirements when working with large libraries.

2.2.8 Multiplex, Markerless Genome Editing Using CRISPR/Cas9

Although the high-efficiency integration method described above is very powerful for integrating sequences into the chromosome, it requires a selectable marker be present in

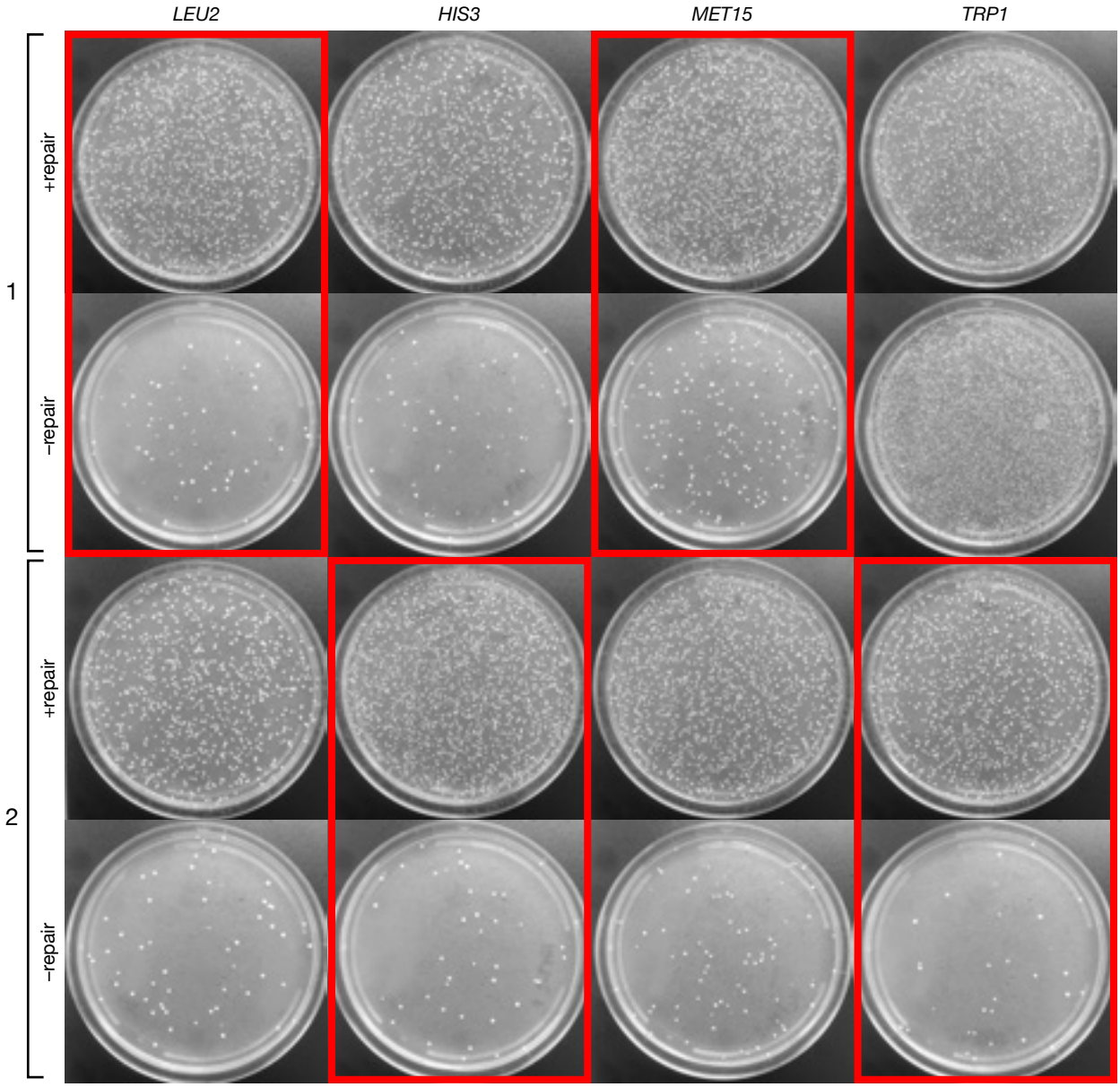
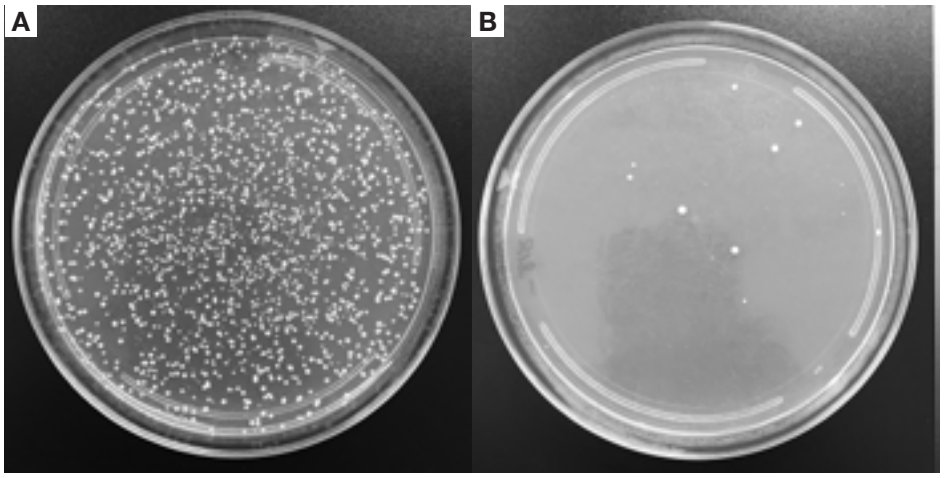


Figure 2-11. CRISPR/Cas9 knockouts. Transformation plates of single target Cas9/sgRNA plasmids with and without repair DNA. Two guides were tested for each of four loci, and the specific guides chosen to use in the multiplex experiments are highlighted in red. *TRP1* guide 1 was ineffective at targeting, as shown by the high colony yields in the no repair control.

the integrated DNA. There are some cases where this requirement is undesirable, such as knocking out multiple genes in a single strain. In this case, a unique marker is needed for each locus; markers must be introduced and then removed for each sequential knockout; or complex mating and screening strategies have to be used to collect all the mutants in a single strain. To avoid these tedious procedures, we adapted the recently described CRISPRm method for making multiple genome edits simultaneously[51].

First, we designed two sgRNAs to target each of four genes—*LEU2*, *HIS3*, *MET15*, and *TRP1*—and repair DNAs (PCR products with 60bp of homology flanking a 20bp barcode, see **Figure 2-10A** and **Table 2-2**) that target those loci and introduce a premature stop codon. We assembled each sgRNA onto a CEN6/ARS4 plasmid containing a Cas9 expression cassette with a *URA3* marker (**Figure 2-10B**). We then transformed each plasmid with or without its cognate repair DNA and selected for transformants on synthetic media lacking uracil. For seven out of eight guides, the transformations with repair DNA had over 100-fold more colonies than those without repair DNA (**Figure 2-11**). This large difference in colony yields suggests that when Cas9 successfully targets a locus in the



C

loci targeted		*LEU2*		*LEU2, HIS3*		*LEU2, HIS3, MET15*		*LEU2, HIS3, MET15, TRP1*	
fraction correct phenotype		39/40	38/40	35/40	26/40	23/40	30/40	11/40	6/40
number of transformants	+repair	110,000	90,000	8,000	3,000	1,200	575	212	70
	–repair	799	779	167	108	250	110	61	53
fraction of correct phenotype with correct genotype		3/3	3/3	3/3	3/3	3/3	3/3	2/3	2/3

Figure 2-12. Multiplex, markerless knockouts. A plasmid expressing Cas9 and an sgRNA that targets the *LEU2* locus was transformed with (A) or without (B) a repair DNA that introduces a stop codon and destroys the sgRNA target. Shown are transformations plated on synthetic media lacking uracil, which selects only for the Cas9/sgRNA plasmid. Thus, selecting for the Cas9/sgRNA plasmid indirectly selects for cells that repaired the locus. (C) Multiple loci were targeted simultaneously, and forty colonies each from two independent experiments were screened for the appropriate phenotype (auxotrophy). The raw number of colonies with and without repair is also shown to demonstrate that both the number of transformants and the fraction of correct clones decrease with an increasing number of targets. Finally, three colonies with the correct phenotype were then screened by colony PCR to verify proper integration of the repair DNA.

genome, the double-strand break it introduces is toxic to the cell; when a repair DNA is present, it removes the target and abolishes the toxicity caused by Cas9. Thus, selecting for the Cas9 plasmid indirectly selects for repairs to the genome at the targeted locus. It should be noted that both here and elsewhere[51,52], it has been shown that the effectiveness of sgRNAs at targeting is variable, and so we recommend that multiple guides be tested until more robust design rules have been determined.

Next, to test multiplexed knockouts, we assembled guides in tandem on a single plasmid with Cas9, targeting one, two, three, or all four loci at once. The standardization and modularity of our assembly scheme made the construction of these multi-targeting plasmids straightforward. We transformed these plasmids, again, with or without their cognate repair DNAs, and selected on synthetic media lacking uracil. We picked forty colonies into different dropout media to determine the fraction of transformants that had the correct phenotype (auxotrophy). Consistent with results from Ryan et al, as the number of simultaneous targets increased, the fraction of correct transformants decreased, but it was still possible to disrupt all four targets at once with ~20% efficiency (**Figure 2-12**). One possible cause of this decrease in efficiency could be recombination within the Cas9 plasmid that excises one or more guides, which could be minimized by using different promoters for each guide, but this has not been tested. Although the intention here is loss-of-function disruptions, only assaying for phenotype does not demonstrate the rate of correct repair DNA incorporation, as a non-homologous end joining could also result in a disruption. To assay this, we screened six (three from each replicate) phenotypically correct colonies for each transformation by colony PCR. For the single, double, and triple knockouts, all six colonies were correct, and for the quadruple knockout, four out of six were correct.

In addition to disrupting genes, this same strategy could be used to integrate large constructs in a markerless fashion or to introduce single nucleotide polymorphisms (SNPs). In cases where SNPs do not completely disrupt sgRNA targeting, two rounds of editing can be performed. The first round sgRNA should target the endogenous sequence, and the repair DNA should destroy the target and/or PAM by introducing a temporary, orthogonal sequence. The second round sgRNA should target the newly introduced sequence, and the repair DNA should reintroduce the endogenous sequence with the desired SNP. Although the same could be accomplished using a counterselectable marker such as *URA3*, using CRISPRm allows for multiple modifications to be made at once.

2.3 Detailed Description of Assembly Standard

2.3.1 Definition of Part Types

There are eight primary part Types in our assembly standard and three of those have options to split into Subtypes. It should be noted that Types are technically defined only by their flanking overhangs, and the contents need not necessarily match the biologically-defined functions we describe (**Figure 2-13**). However, in order to ensure compatibility between all parts designed by all researchers, we recommend that new parts be designed to match the Types defined here.

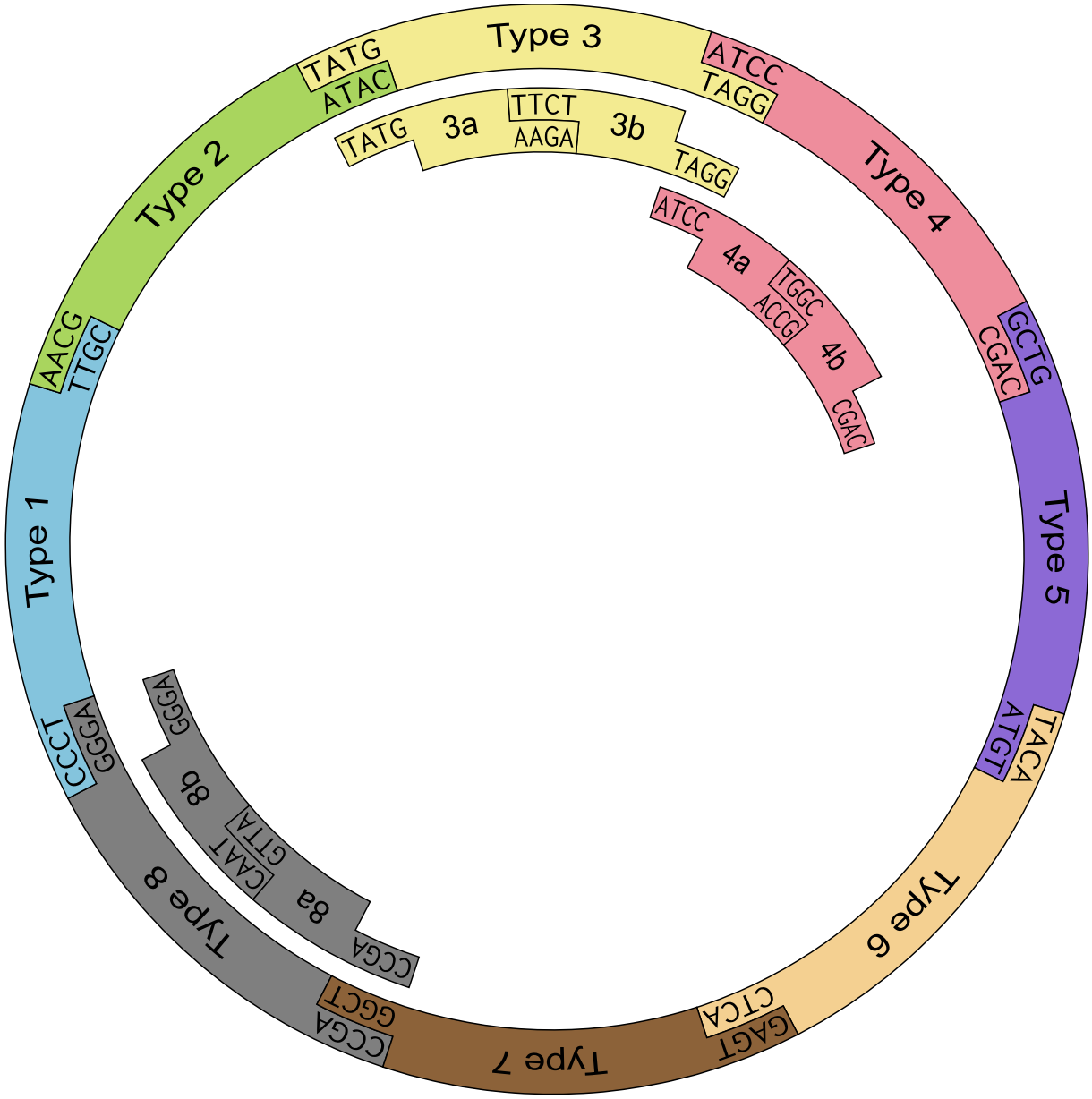


Figure 2-13. Part types and overhangs.

Type 1: 5′ Assembly Connector

Type 1 parts are flanked by CCCT and AACG (**Figure 2-14A**). Typically, this Type contains non-coding, non-regulatory sequences that are used to direct assembly of multigene plasmids.

The Type 1 part plasmids included in the toolkit contain a 143bp concatenation of barcode sequences used in the systematic deletion of yeast genes[53], a BsmBI recognition site and unique overhang, and a 21bp barcode scar (again, from the systematic deletion collection) (**Figure 2-14B**). We designate these sequences as Assembly Connectors, and the nomenclature used is "ConLX" where X = 1, 2, 3, etc. The BsmBI site is oriented such that the restriction enzyme digests the sequence downstream of the recognition sequence.

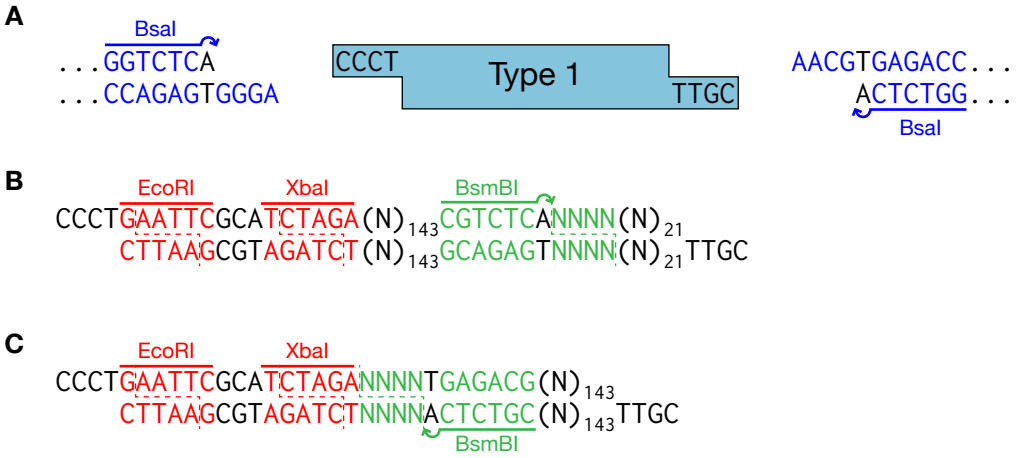


Figure 2-14. Type 1: 5' Assembly Connector.

There are also ConLX' parts, where the structure instead contains first a reversed BsmBI site (digests upstream) followed by the 143bp sequence (and no barcode scar) (**Figure 2-14C**). The purpose of this alternate Assembly Connector is for generation of multi-gene backbone plasmids (see the detailed description of assembly scheme below). To simplify the toolkit, we include a single reversed Assembly Connector, which we designate ConLS' and its cognate forward version, ConLS (S for Start), although any numbered Assembly Connector can have a reversed version if desired.

Finally, the Type 1 parts in this toolkit also include an EcoRI and XbaI site for BioBrick compatibility of the assembled cassettes and multi-gene plasmids.

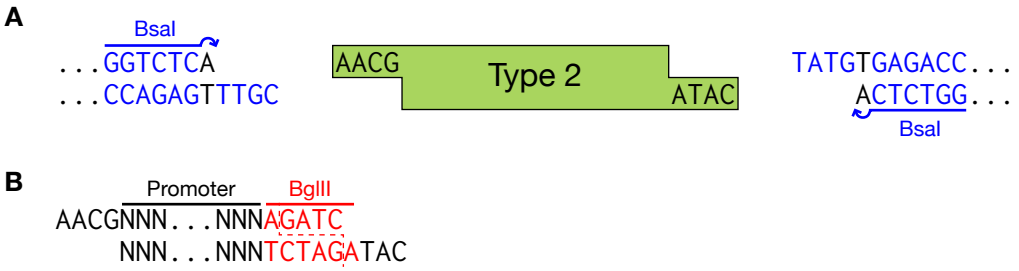


Figure 2-15. Type 2: Promoter.

Type 2: Promoter

Type 2 parts are flanked by AACG and TATG (**Figure 2-15A**). Typically, this Type contains a promoter. The downstream overhang doubles as the start codon for the subsequent Type 3 or 3a coding sequence. Additionally, all the promoters in this toolkit have a BglII site immediately preceding the start codon (overlapping the downstream overhang) for BglBrick compatibility (**Figure 2-15B**).

Type 3: Coding Sequence

Type 3 parts are flanked by TATG and ATCC (**Figure 2-16A**). Typically, this Type contains a coding sequence. As discussed above, the TATG overhang includes a start codon so coding sequences should begin with the second codon. The ATCC overhang was designed to

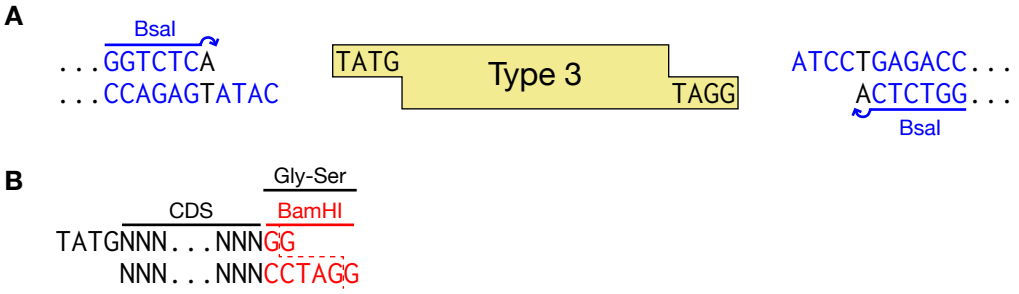


Figure 2-16. Type 3: Coding Sequence.

enable read-through for protein fusions. If a stop codon is omitted from the part, and two bases are added before the overhang, the resulting NNATCC can be used as a two amino acid linker to a Type 4 or 4a C-terminal fusion. The Type 3 parts in this toolkit all omit the stop codon and add a GG, resulting in GGATCC, which serves a dual purpose. First, the resulting Gly-Ser linker is relatively innocuous; and second, the sequence is a BamHI recognition site, which enables BglBrick compatibility (**Figure 2-16B**). We highly recommend following this convention unless the protein in question is sensitive to C-terminal modifications.

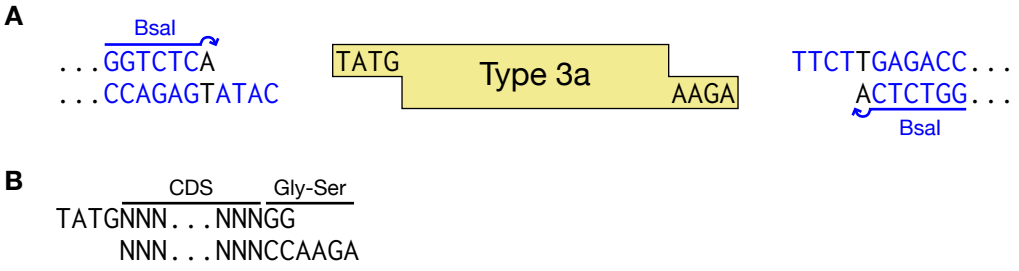


Figure 2-17. Type 3a: N-terminal Coding Sequence.

Type 3a: N-terminal Coding Sequence

Type 3 parts can be split into 3a and 3b parts for greater flexibility for making protein fusions. Type 3a parts are flanked by TATG and TTCT (**Figure 2-17A**). As with Type 3 parts, these typically contain coding sequences, and can be used for fusing N-terminal tags (such as the degradation tags described in the main text). Again, as with Type 3 parts, the stop codon should be omitted and two bases should be added before the TTCT overhang if protein fusions are desired. The Type 3a parts in this toolkit add a GG, resulting in GGTTCT, another Gly-Ser linker (**Figure 2-17B**).

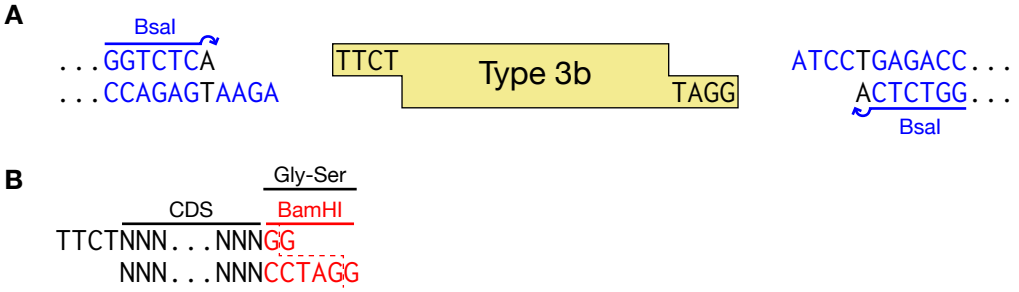


Figure 2-18. Type 3b: Coding Sequence.

Type 3b: Coding Sequence

Type 3b parts are flanked by TTCT and ATCC (**Figure 2-18A**). As with Type 3 and 3a parts, these typically contain coding sequences. Again, the start codon should be removed for direct fusions to the Type 3a preceding it, and two bases should be added before the ATCC overhang if C-terminal fusions are desired. As with the Type 3 parts, all Type 3b parts in this toolkit add a GG, resulting in GGATCC (Gly-Ser/BamHI site) (**Figure 2-18B**).

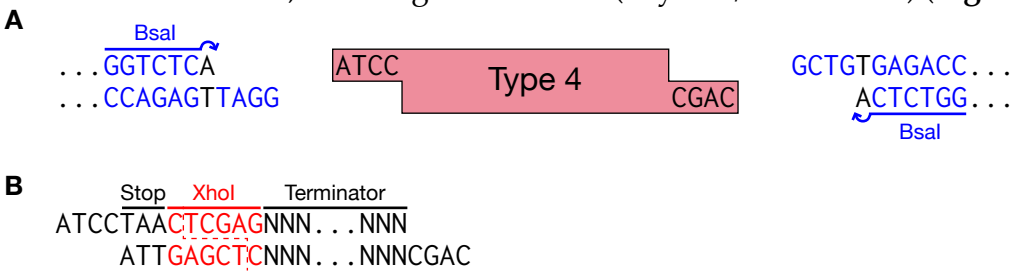


Figure 2-19. Type 4: Terminator.

Type 4: Terminator

Type 4 parts are flanked by ATCC and GCTG (**Figure 2-19A**). Typically, this Type contains a transcriptional terminator. As described above, the convention for a Type 3 or 3b is to omit the stop codon and allow read-through of a GGATCC (Gly-Ser) linker. Therefore, the Type 4 should encode an in-frame stop codon before the transcriptional terminator. The Type 4 parts in this toolkit begin with a TAA stop codon, followed by a XhoI site (CTCGAG, for BglBrick compatibility), then the terminator sequence (**Figure 2-19B**). Commonly used C-terminal fusions, such as purification or epitope tags, may be in included before the stop codon, but we recommend using the 4a/4b subtypes to maintain their modularity.

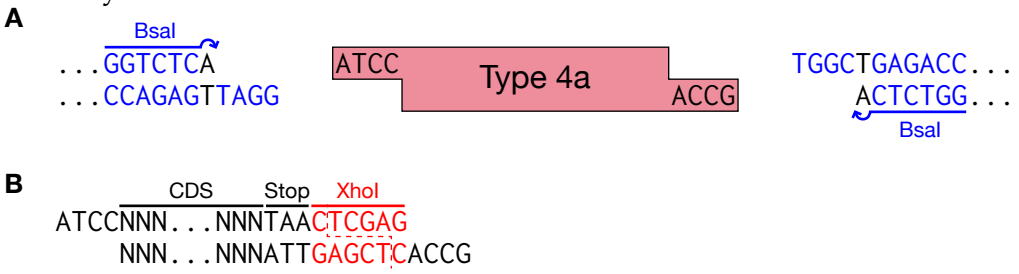


Figure 2-20. Type 4a: C-terminal Coding Sequence.

Type 4a: C-terminal Coding Sequence

Like Type 3 parts, Type 4 parts can be split into 4a and 4b parts for additional modularity. Type 4a parts are flanked by ATCC and TGGC (**Figure 2-20A**). Typically, this Type contains a coding sequence for fusing to the C-terminus of a protein (such as a localization tag, fluorescent protein, or purification tag). However, in contrast to the Type 3 and 3b parts, the convention for 4a parts is to include the stop codon rather than enable read-through of the TGGC overhang (although this is possible if desired). As such, the Type 4a parts in this toolkit end with a TAA stop codon, followed by a XhoI site (CTCGAG, for BglBrick compatibility) (**Figure 2-20B**).

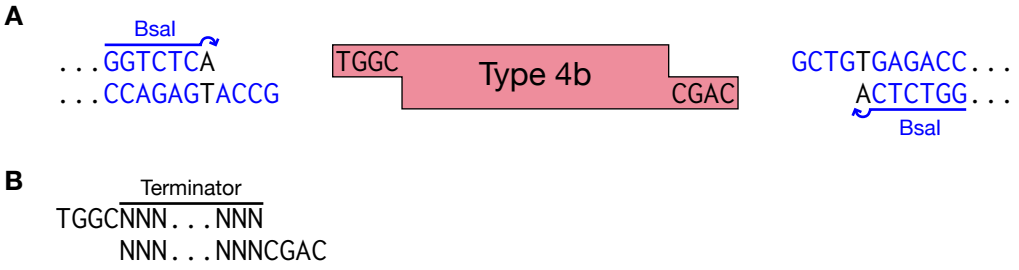


Figure 2-21. Type 4b: Terminator.

Type 4b: Terminator

Type 4b parts are flanked by TGGC and GCTG (**Figure 2-21A**). As with Type 4 parts, these typically contain transcriptional terminators (**Figure 2-21B**). Because the convention of a Type 4a part is to encode the stop codon, one is not necessary in a 4b and so only the terminator sequence is needed.

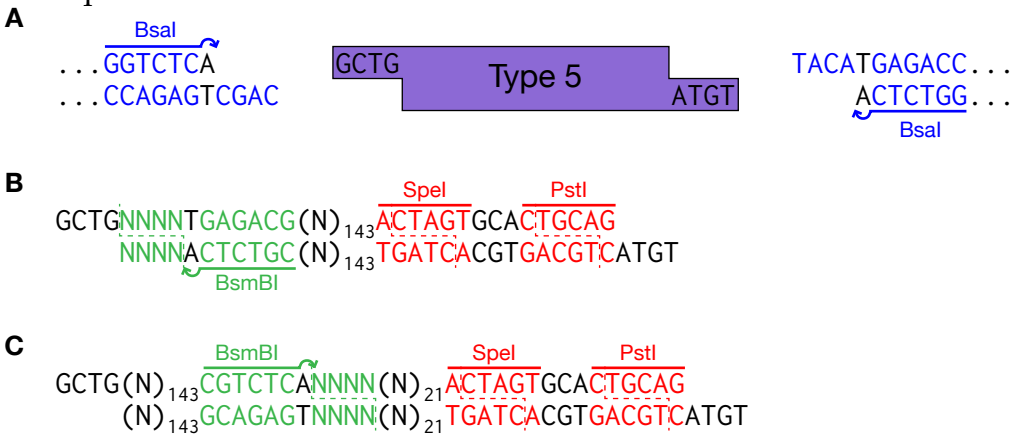


Figure 2-22. Type 5: 3' Assembly Connector.

Type 5: 3′ Assembly Connector

Type 5 parts are flanked by GCTG and TACA (**Figure 2-22A**). As with the Type 1 parts, these parts typically contain Assembly Connectors.

The structure of Type 5 parts is very similar to that of Type 1 parts. First is an upstream-cutting BsmBI site (with a unique overhang), followed by a 143bp concatenated barcode sequence (this structure is identical to that of the Type 1 ConLX') (**Figure 2-22B**). Here, the nomenclature used is "ConRX". Again, there is a special structure for ConRX' parts: the 143bp sequence, a downstream-cutting BsmBI site, and a 20bp barcode (this structure is identical to that of the Type 1 ConLX) (**Figure 2-22C**). We included in this toolkit, a single ConRE' (E for end) part and its cognate forward version, ConRE.

The key to the Type 1 and 5 Assembly Connectors is that the unique overhangs generated by BsmBI digestion should match for parts with the same value of X. For example, the BsmBI overhang generated by ConL1 and by ConR1 is CCAA. This is critical for enabling assembly of multi-gene plasmids, which is described in detail below.

Finally, the Type 5 parts in this toolkit also include a SpeI (ACTAGT) and PstI (CTGCAG) site for BioBrick compatibility of the assembled cassettes and multi-gene plasmids.

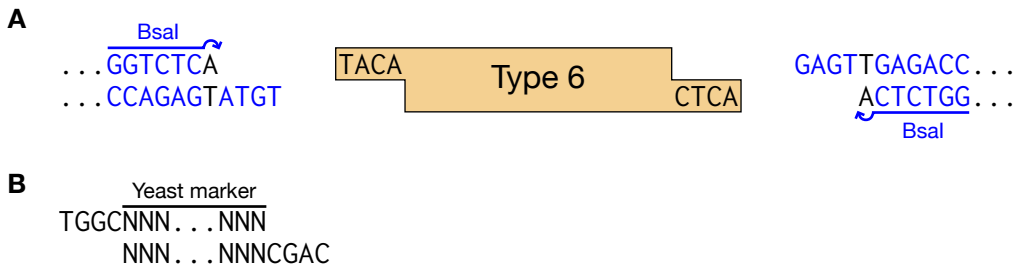


Figure 2-23. Type 6: Yeast Marker.

Type 6: Yeast Marker

Type 6 parts are flanked by TACA and GAGT (**Figure 2-23A**). Typically, this Type contains a selectable marker for *S. cerevisiae*. These parts should include the full expression cassette (promoter, ORF, and terminator) for conferring the selectable phenotype (usually amino acid prototrophy or drug-resistance) (**Figure 2-23B**).

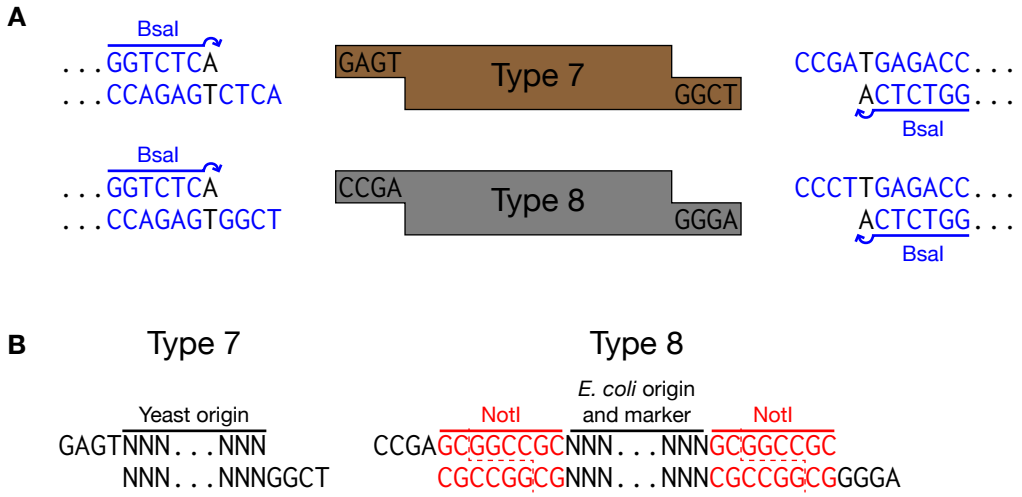


Figure 2-24. Type 7+8: Yeast Plasmid Propagation.

Type 7+8: Yeast Plasmid Propagation

Type 7 and 8 parts can be used in two ways, depending on the application. For plasmid expression in yeast, Type 7 and 8 parts should be used as described in this section; for integration into the yeast chromosome, Type 7, 8a, and 8b parts should be used as described in the next section.

Type 7 parts are flanked by GAGT and CCGA (**Figure 2-24A**). For propagation of a stable plasmid in yeast, this Type contains a yeast origin of replication (**Figure 2-24B**). Type 8 parts are flanked by CCGA and CCCT as well as NotI sites that are useful for restriction mapping to verify new assemblies (**Figure 2-24A**). This Type contains a bacterial origin of replication and antibiotic resistance marker (**Figure 2-24B**).

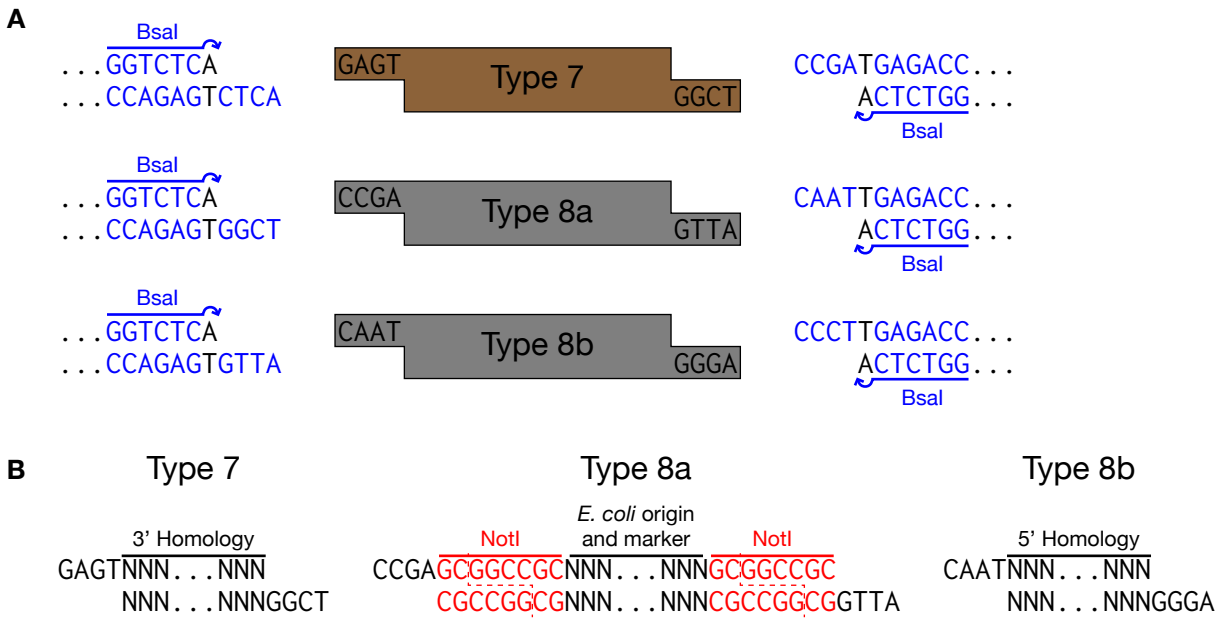


Figure 2-25. Type 7+8a+8b: Yeast Chromosomal Integration.

Type 7+8a+8b: Yeast Chromosomal Integration

For integration into the yeast chromosome, the Type 7 parts (which retain the GAGT and CCGA overhangs) contain sequences that have homology that is downstream (3′) of the target locus (**Figure 2-25A**). Longer homology sequences are more efficient at recombining into the chromosome; therefore, the parts in this toolkit contain 500bp of homology. Additionally, a 20bp barcode sequence is included upstream of the homology region to serve as a forward primer binding site for colony PCR verification of integration into the correct locus.

Type 8a parts are flanked by CCGA and CAAT (**Figure 2-25A**). As with the Type 8 parts, these typically contain a bacterial origin of replication and antibiotic resistance marker (**Figure 2-25B**). These parts are also flanked by NotI sites that can be used to linearize the integration plasmid prior to transformation into yeast (as well as for restriction mapping).

Type 8b parts are flanked by CAAT and CCCT (**Figure 2-25A**). Similar to Type 7 homology parts, these parts contain long sequences of homology to the genome that is upstream (5′) of the target locus (**Figure 2-25B**). Additionally, a 20bp barcode sequence is included downstream of the homology region to serve as a reverse primer binding site for colony PCR verification of integration into the correct locus.

Miscellaneous

In addition to these standard part Types, non-standard Types that span two or more positions can be constructed and are conventionally named as a concatenation of the Type numbers spanned. For example, some cassette plasmids are constructed only as intermediates toward a multi-gene plasmid. These cassettes no longer require any of the yeast maintenance machinery (origin and marker) and so a Type 678 part that only contains a bacterial origin and marker may be appropriate to use.

2.3.2 Detailed Description of Hierarchical Assembly System

The construction of plasmids encoding multiple transcriptional units is done in three steps. The first is construction of part plasmids; second is assembly of cassette plasmids; and third is assembly of multi-gene plasmids (**Figure 2-1**).

Construction of Part Plasmids

The general structure of a part plasmid is as follows: 1) a downstream-facing BsaI site that generates the upstream flanking overhang of the part Type; 2) the part sequence; 3) an upstream-facing BsaI site that generates the downstream flanking overhang of the part Type; and 4) a ColE1 origin of replication and chloramphenicol resistance marker. Detailed descriptions for each part Type can be found in **Section 2.3.1**.

Part plasmids are assembled via a BsmBI Golden Gate reaction into the part entry vector (**Figure 2-26A**). The entry vector contains a ColE1 origin of replication and chloramphenicol resistance marker, as well as a GFP expression dropout for green/white screening. BsaI, BsmBI, and NotI sites should be removed from all parts except in special cases. Additional restriction sites such as BbsI or the BioBrick/BglBrick enzymes may also be removed, but it is not necessary unless future use of those enzymes is anticipated.

Primers for amplifying preexisting templates should be designed as illustrated in **Figure 2-26A** to enable BsmBI assembly into the entry vector, and subsequent BsaI cassette assemblies. The four N's flanking the part should correspond to the flanking overhangs for the specific part Type (e.g., AACG and TATG for a Type 2). Modifications to the part sequence (e.g. restriction site removal) can be easily introduced by dividing the part into multiple DNA inserts in the BsmBI Golden Gate reaction. Internal overhangs in this reaction can be user-selected, but should avoid similarity to the entry vector overhangs TCGG and GACC. Parts made from de novo synthesis should mimic the same structure or be ordered in the entry vector. Finally, small parts can be assembled from overlapping

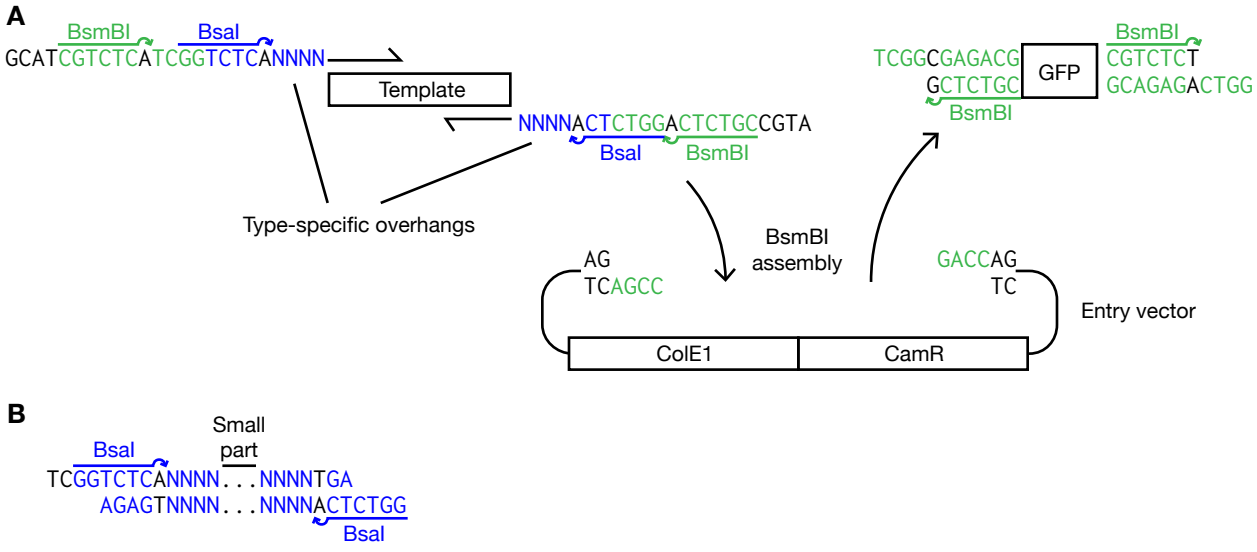


Figure 2-26. Construction of Part Plasmids.

oligonucleotides that drop directly into the entry vector (**Figure 2-26B**). We routinely add annealed oligonucleotides and/or synthesized gene fragments (e.g. gBlocks®) directly to the BsmBI Golden Gate reaction.

A special exception must be made for constructing new Assembly Connectors (Type 1 and 5). In these cases, the Assembly Connectors contain internal BsmBI sites used in multi-gene assemblies. There are two options for constructing these part plasmids. First, primers can be designed as normal, but the Golden Gate assembly protocol should be modified to exclude the final digestion and heat inactivation steps, thereby ending on a ligation. Second, an existing Type 1 or Type 5 plasmid can be digested and gel purified using BsaI, and the new part can then be assembled in using BsaI rather than BsmBI.

Assembly of Cassette Plasmids

The simplest way to assemble a cassette is to include one part of each Type in a BsaI assembly. The Type 8 and 8a parts included in this toolkit serve as the canonical "vectors", and accordingly have an mRFP1 expression dropout for red/white screening.

An alternative approach is to pre-assemble commonly used parts with a GFP dropout that spans the variable region. For example, a Type 234 GFP dropout part is included in

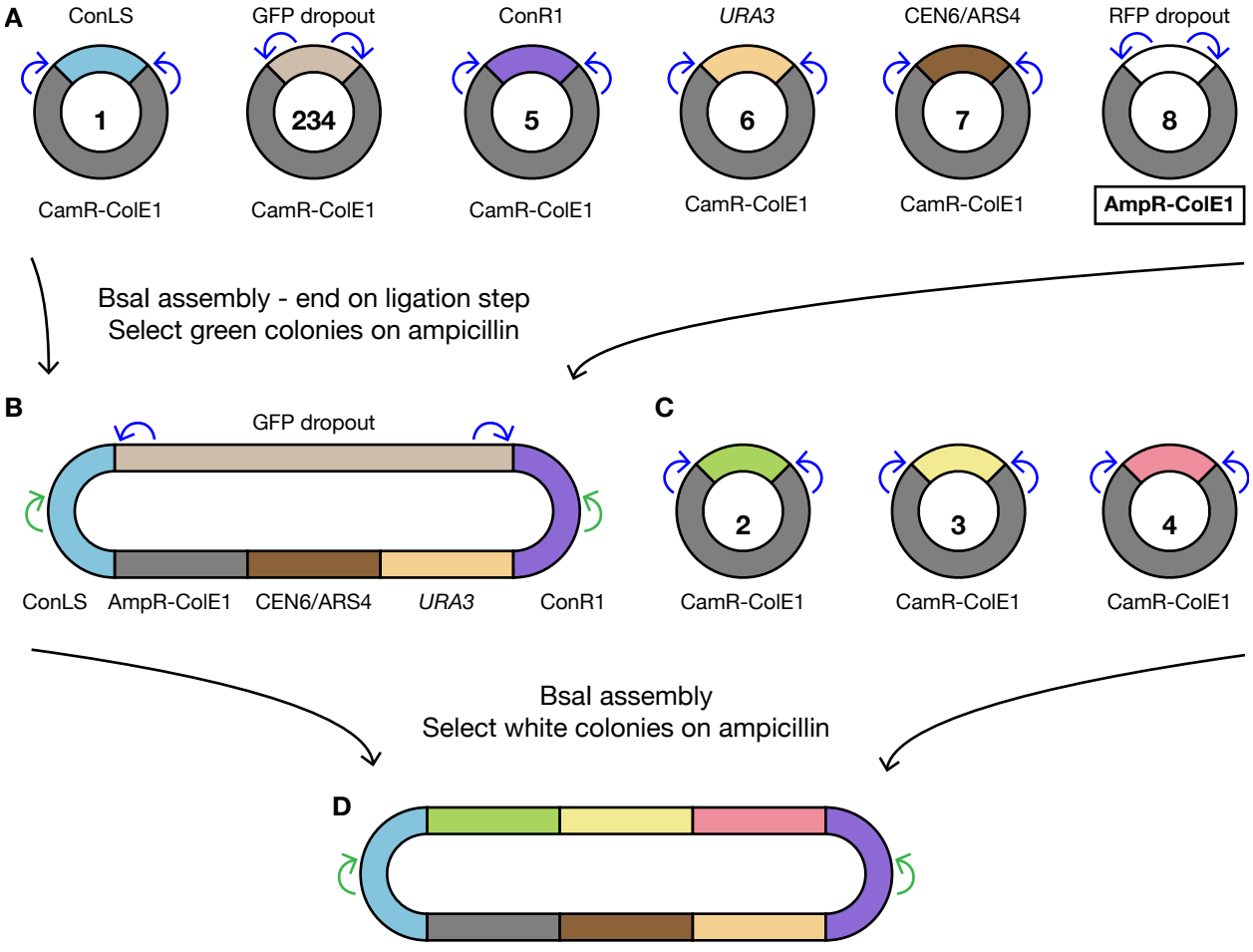


Figure 2-27. Assembly of Cassette Plasmids.

this toolkit. By assembling ConLS, the GFP dropout, ConR1, *URA3*, CEN6/ARS4, and AmpR-ColE1, a cassette "vector" can be made (**Figure 2-27A and B**). By storing this cassette, any future assemblies of transcriptional units (promoter, coding sequence, terminator) with these vector components will require fewer parts, something that is particularly useful for the generation of combinatorial libraries (**Figure 2-27C and D**). It is important to note that the Type 234 GFP dropout part has BsaI sites in the reverse orientation that normal parts do and will remain in the finished vector. Therefore, the Golden Gate assembly protocol should be modified to exclude the final digestion and heat inactivation steps, thereby ending on a ligation. We have observed a significantly higher rate of misassembly for this procedure (~50%). Incorrect products are typically concatenations of part plasmids and contain multiple origins of replication and antibiotics markers. Wrong products can be easily identified because they will confer growth in media with either chloramphenicol or the desired antibiotic, whereas the correct product will not confer growth in media with chloramphenicol.

Assembly of Multi-Gene Plasmids

The construction of a multi-gene plasmid from cassettes requires that the cassettes are flanked by unique pairs of Assembly Connectors, which dictate the order of assembly. The first cassette must contain the ConLS part, and the last cassette must contain the ConRE part. The order of internal Assembly Connectors can be arbitrary, although going in increasing numerical order is recommended to avoid confusion. Thus, before the individual cassettes are made, the structure of the final multi-gene plasmid should be designed because it will determine which Assembly Connectors should be used during cassette assembly.

For example, if three transcriptional units, TU1, TU2, and TU3 are to be assembled into a multi-gene plasmid in that order, one possible design would be to flank TU1 with ConLS and ConR1, TU2 with ConL1 and ConR2, and TU3 with ConL2 and ConRE (**Figure 2-28A**). The "vector" into which these cassettes are assembled is itself another cassette,

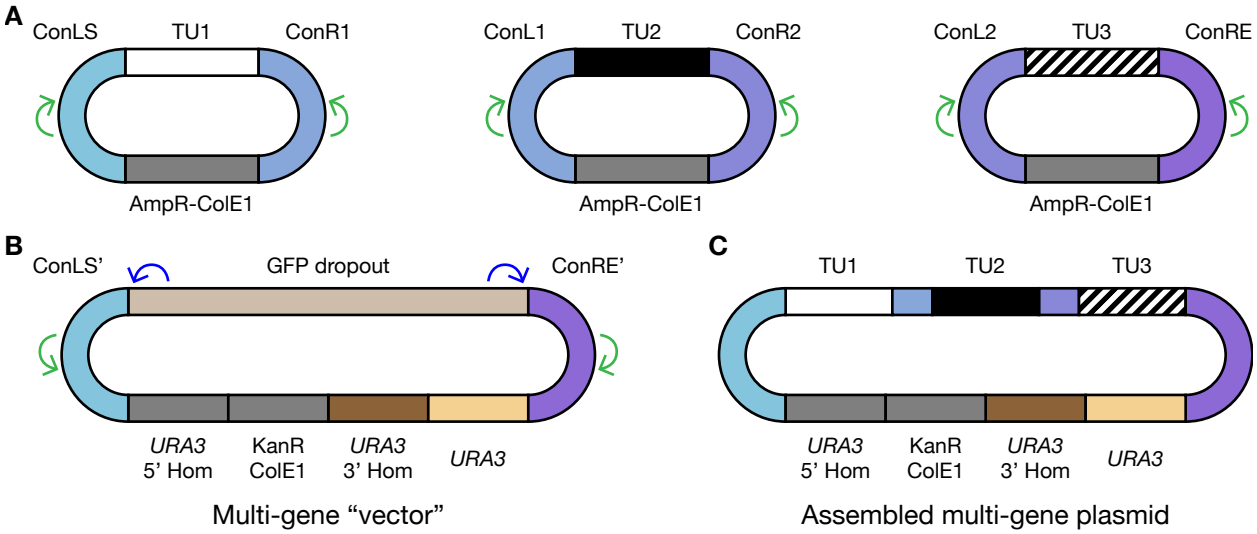


Figure 2-28. Assembly of Multi-Gene Plasmids.

which uses the special ConLS' and ConRE' parts (**Figure 2-28B**). One example of this special cassette is included in this toolkit, designed to target the *URA3* locus for integration. Again, in assembling this "vector" cassette, the Type 234 GFP dropout can be used to enable green/white screening. When the final multi-gene plasmid is assembled, the Assembly Connector junctions leave behind 20bp barcode "scars", which can be used to verify proper assembly by colony PCR or sequencing (**Figure 2-28C**).

One issue with this style of assembly is that cassettes are locked into their position based on the Assembly Connectors flanking them. For example, in the same three-TU multi-gene plasmid described above, if TU2 were to be omitted, there would be a gap that would require reassembling either TU1 or TU3 to replace the right or left Assembly Connectors, respectively. One solution to this we provide in the toolkit is the Type 234 "Spacer" part. This part can be used to assemble filler cassettes—in this case, a ConL1-Spacer-ConR2 cassette. The advantage of assembling a filler cassette rather than a reassembled TU cassette is that the filler can be used again in future assemblies when that gap needs to be filled.

Construction of CRISPR/Cas9 sgRNAs

To construct sgRNAs for targeting Cas9 to a site in the genome, we have provided a Type 234 sgRNA dropout, which is effectively an entry vector for these parts. This vector is based on the CRISPRm sgRNA architecture: a phenylalanine tRNA, a HDV ribozyme, a 20bp targeting sequence, the sgRNA, and an *SNR52* terminator. For the dropout, the targeting sequence is replaced by a BsmBI-flanked GFP expression cassette. Unfortunately, the 4bp immediately upstream and downstream of the targeting sequence are CTTT and GTTT, which could incorrectly ligate, so the upstream overhang was moved two bases upstream to GACT. Consequently, two additional T's should be added before the targeting sequence when ordering oligonucleotides to anneal and ligate into the dropout (**Figure 2-29**). Once constructed, the sgRNA 234 part can be assembled into a cassette with appropriate connectors. This cassette should then be assembled into a multi-gene plasmid that also includes Cas9 expression and additional sgRNAs (optional).

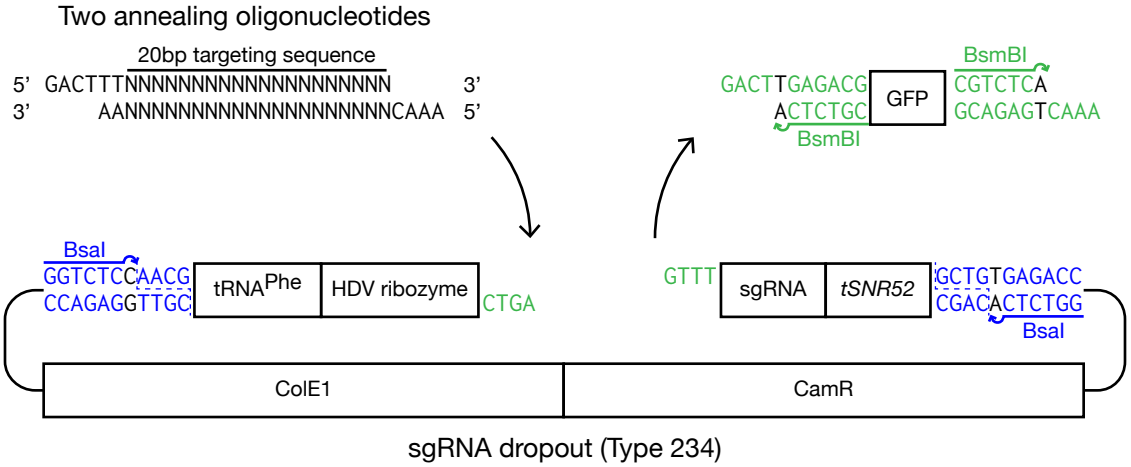


Figure 2-29. Construction of CRISPR/Cas9 sgRNAs.

2.3.3 Differences from MoClo

The hierarchical assembly system described in this work borrows heavily from the MoClo[32] system with some modifications.

First, the specific overhang sequences that flank the parts and the cassettes are different from those used by the MoClo creators. This change was made to support a greater number of part Types as well as the in-frame protein fusions enabled by several of those Types. Additionally, in an attempt to minimize assembly errors due to misligation of incompatible overhangs, we tried to find a more optimal set. It has been previously reported that when three out of four contiguous nucleotides match between two overhangs, a misligation event can occur (e.g., ATCG and ATCA)[34]. Furthermore, we had observed that even three non-contiguous matches could result in a misligation event (e.g., ATCG and ATAG, or ATCG and ATGA). Therefore, we tried to find a set of overhangs where the fewest such matches were present.

Second, the MoClo system requires an extensive series of vectors to support the various possibilities of multi-gene assemblies. Rather than creating the exhaustive set of possible vectors up front, our system utilizes the Assembly Connector parts to enable on-the-fly construction of vectors. One advantage of this approach is that rather than defining transcriptional units as being in Position 1, 2, 3…, they are defined as being between Assembly Connectors X and Y. Thus, a transcriptional unit could be cloned between ConL1 and ConR4, or ConL3 and ConR2, as long as the final sequence begins with ConLS and ends with ConRE and has no repeated Assembly Connectors.

Finally, MoClo utilizes a third Type IIs restriction enzyme to enable indefinite assembly of multiple transcriptional units. As described, our system is limited to one round of multigene assembly, although it could easily be extended to include this added functionality if desired. We have removed a third Type IIs restriction site (BbsI) from all parts in this toolkit for such purposes.

2.3.4 Alternative Assembly Methods

Although Golden Gate is the preferred assembly method to be used in this system, there are a number of alternative methods that can be used at some steps of the process.

The initial part plasmid construction can be performed using any method, as long as the resulting plasmid has the appropriate BsaI overhangs flanking it.

The cassette plasmid assembly must be performed using Golden Gate. Other methods such as Gibson or SLIC can be used, but they will require unique primers for every new junction. Only Golden Gate assembly will preserve modularity at this step.

Once a cassette plasmid has been assembled, there is much more flexibility in terms of downstream assembly steps. If all the conventions described above are followed, the cassettes will be flanked by BioBrick restriction enzyme sites, enabling BioBrick cloning of cassettes with each other, or with existing BioBrick plasmids that have not been converted to this new system. Second, cassettes also contain BglBrick restriction enzyme sites

that flank the coding sequence, enabling BglBrick cloning for fusing coding sequences with existing BglBrick parts. Third, the purpose of the 143bp sequences in the Assembly Connector parts is to facilitate modular recombination-based assembly methods, such as Gibson, SLIC, or *in vivo* yeast assembly. As with the BsmBI overhangs in the Assembly Connectors, the 143bp sequences of ConLX and ConRX parts with the same value of X will be exactly the same, so cassettes can be designed in the same way for both Golden Gate and recombination-based assembly. Note that the final sequence of the multi-gene plasmid will be different depending on which method (Golden Gate, BioBrick, or recombination) is used for the assembly.

2.4 Summary

We have described a methodology and an accompanying toolkit of essential parts for engineering yeast. This MoClo-derived assembly standard supports the rapid cloning of multi-gene expression devices. We characterized a set of promoters and terminators, which are by no means exhaustive or perfect, but nonetheless diverse, in order to support the construction of multi-gene plasmids with minimal risk of unwanted homologous recombination. As a distinct method of controlling protein concentration, degradation tags are also characterized. Additionally, we have illustrated an important difference between using plasmids and chromosomal integrations and encourage expression from the chromosome whenever possible. To facilitate this, our system is designed to make integrations as straightforward as plasmid transformations. We also present two options, using I-SceI or CRISPR/Cas9, for generating double-stranded breaks in the chromosome that increase integration efficiencies to match or even exceed that of plasmid transformations. Finally, we adapted the CRISPRm method to our standardized assembly scheme to enable multiplexed knockouts of endogenous genes. In summary, we believe this work will be a useful resource for both novice and experienced yeast biologists and engineers, and lays the foundation for a community that shares novel parts, as well as leads to greater consistency and reproducibility.

2.5 Materials and Methods

Strains and growth media

The *S. cerevisiae* strain used for measuring most promoters, terminators, degradation tags, copy number, and chromosomal integrations was BY4741 (*MAT*a *his3*Δ*1 leu2*Δ*0 met15*Δ*0 ura3*Δ*0*). The mating-type-specific promoters were also tested in BY4742 (*MAT*α *his3*Δ*1 leu2*Δ*0 lys2*Δ*0 ura3*Δ*0*) and BY4743 (diploid cross of BY4741 and 4742). The galactose-induction experiments were conducted in a *GAL2* knockout of BY4741. The multiplex CRISPR/Cas9 knockout experim19ents were conducted in an S288C *MAT*a haploid with a complete *URA3* coding sequence deletion.

Constitutive promoter and terminator characterization experiments were conducted in synthetic media with 2% (w/v) Dextrose (Fisher Scientific), 0.67% (w/v) Yeast Nitrogen Base without amino acids (VWR International), 0.2% (w/v) Drop-out Mix Complete w/o

Yeast Nitrogen Base (US Biological), 0.85% (w/v) MOPS Free Acid (Sigma), 0.1M Dipotassium phosphate (Sigma), 100μg/L Zeocin (Life Technologies), buffered to pH 7.

Galactose inductions were performed in synthetic media with 2% (w/v) Raffinose (Fisher Scientific), 0.67% (w/v) Yeast Nitrogen Base without amino acids (VWR International), 0.2% (w/v) Drop-out Mix Synthetic Minus Uracil w/o Yeast Nitrogen Base (US Biological), plus 0%-5% (w/v) Galactose (Fisher Scientific).

All other expeirments were conducted in synthetic media with 2% (w/v) Dextrose (Fisher Scientific), 0.67% (w/v) Yeast Nitrogen Base without amino acids (VWR International), 0.2% (w/v) Drop-out Mix Synthetic Minus appropriate amino acids w/o Yeast Nitrogen Base (US Biological).

YPD was used for preparing cells for transformation and recovery after heat shock: 1% (w/v) Bacto Yeast Extract, 2% (w/v) Bacto Peptone, 2% (w/v) Dextrose.

TG1 chemically competent *E. coli* was used for all cloning experiments. Transformed cells were selected on Lysogeny Broth (LB) with the appropriate antibiotics (ampicillin, chloramphenicol, or kanamycin).

Yeast transformations

Yeast colonies were grown to saturation overnight in YPD, then diluted 1:100 in 50mL of fresh media and grown for 4-6hrs to OD600~0.8. Cells were pelleted and washed once with water and twice with 100mM Lithium Acetate (Sigma). Cells were then mixed by vortexing with 2.4mL of 50% PEG-3350 (Fisher Scientific), 360μL of 1M Lithium Acetate, 250μL of salmon sperm DNA (Sigma), and 500μL of water. DNA was added to 100-350μL of transformation mixture and incubated at 42C for 25min. When selecting for prototrophy, the transformation mixture was pelleted, resuspended in water, and plated directly onto solid agar plates. When selecting for drug resistance, the transformation mixture was pelleted, resuspended in YPD, incubated at 30C for 2hrs with shaking, pelleted and washed with water, then plated onto solid agar plates.

Plasmids designed for chromosomal integration (*i.e.*, containing 5' and 3' genome homology regions without a yeast origin of replication) were digested with NotI for 10 minutes prior to transformation to stimulate homologous recombination. The entire digestion reaction (without DNA cleanup) was included in the transformation in place of plasmid DNA.

Golden Gate Assembly protocol

A Golden Gate reaction mixture was prepared as follows: 0.5μL of each DNA insert or plasmid, 1μL T4 DNA Ligase buffer (NEB), 0.5μL T7 DNA Ligase (NEB), 0.5μL restriction enzyme, and water to bring the final volume to 10μL. The restriction enzymes used were either BsaI or BsmBI (both 10,000 U/mL from NEB). The amount of DNA inserts can optionally be normalized to equimolar concentrations (~20 fmol each) to improve assembly efficiencies.

Reaction mixtures were incubated in a thermocycler according to the following program: 25 cycles of digestion and ligation (42C for 2min, 16C for 5min) followed by a final digestion step (60C for 10min), and a heat inactivation step (80C for 10min). In some cases, where noted in the text, the final digestion and heat inactivation steps were omitted.

Cloning of parts

See Supporting Information for details on construction of new parts.

Promoter, terminator, and degradation tag characterization

Promoter, terminator, and degradation tag testing constructs were integrated into the *URA3* locus of the yeast chromosome. Constitutive promoter, terminator, and degradation tag testing constructs were selected using a Zeocin resistance cassette; mating-type and galactose promoter testing constructs were selected for uracil prototrophy.

Colonies were picked and grown in 500μL of media in 96-deep-well blocks at 30C in an ATR shaker, shaking at 750RPM until saturated. Cultures were diluted 1:100 in fresh media, grown for 12-16hrs, then diluted 1:3 in fresh media, and fluorescence was measured on a TECAN Safire2. For the galactose inductions, the media was switched during the dilution step from 2% dextrose to 2% raffinose with different concentrations of galactose.

Excitation and emission wavelengths used to measure fluorescent proteins were: mTurquoise2 at 435nm/478nm, Venus at 516nm/530nm, and mRuby2 at 559nm/600nm. Raw fluorescence values were first normalized to the OD600 of the cultures, and then normalized to the background fluorescence of cells not expressing any fluorescent protein. The median log value of biological replicates was calculated and plotted with the range.

Copy number characterization

mRuby2 expression cassettes were assembled onto *URA3* plasmids or integrated into the *URA3* locus; Venus expression cassettes were assembled onto *LEU2* plasmids or integrated into the *LEU2* locus. For constructs where the two fluorescent proteins were expressed in tandem from the same locus/plasmid, they were assembled onto *URA3* plasmids or integrated into the *URA3* locus; the strain used for these constructs was prototrophic for leucine.

Four colonies of each strain were picked into 400μL of synthetic media lacking uracil and leucine, and grown in 96-deep-well blocks at 30C in an ATR shaker, shaking at 750RPM until saturated. The saturated cultures were measured for bulk fluorescence in a TECAN Safire2. The cultures were then diluted 1:100 into fresh media, grown for 4hrs, and measured on a Fortessa X-20 flow cytometer.

Excitation and emission wavelengths used to measure bulk fluorescence on the TECAN were: Venus at 516nm/530nm and mRuby2 at 560nm/590nm. Fluorescence values were normalized and reported in the same manner as the promoter characterization experiments.

The lasers and filters used on the flow cytometer were: a 488nm laser and a FITC filter (505LP 530/30) for Venus; a 561nm laser and PE-Texas Red filter (595LP 610/20) for mRuby2. Voltages for each channel were kept constant for all samples at all copy numbers. Cytometry data were analyzed using FlowJo (http://www.flowjo.com).

Note: the selectable auxotrophic markers for uracil and leucine used in these experiments were different from those included in the toolkit. At the time these experiments were conducted, we had designed markers that encoded for the native Ura3p and Leu2p proteins, but used alternate codons for almost every position. We also used the respective terminator sequence from *Ashbya gossypii*, although we used the native *S. cerevisiae* promoter. The reason for these changes was an attempt to construct selectable markers with orthogonal sequences that would minimize undesired recombination with the chromosome, particularly for strains that did not have clean deletions of those genes as BY4741 does. Unfortunately, the changes resulted in a reduced growth rate on selective media, and were abandoned in favor of the native sequences. The only other experiment to use the alternative markers was the high-efficiency integration experiment (which also used *HIS3*).

High-efficiency integrations

The experimental strain used for the integration efficiency experiments was prepared by integrating the landing pad into BY4741 as depicted in **Figure 2-9**. The repair DNA was constructed in two ways, with and without a CEN6/ARS4 origin. The plasmid with an origin was transformed and used to normalize the colony counts of all other transformations. The plasmid without an origin was linearized using NotI prior to transformation. The cutting plasmids expressing either I-SceI or Cas9/sgRNA were constructed onto CEN6/ARS4 plasmids with a *HIS3* selection marker, but were never selected for and were presumably present only transiently in cells. The cutting plasmids either were or were not also linearized with NotI prior to transformation. 100fmol of each DNA (cutter and/or repair) was added to 350μL of transformation mix. After heat shock, 1/10th of the transformation was plated onto synthetic media lacking histidine. Pictures of the plates were taken and colonies were counted using Benchling (https://benchling.com).

Multiplexed knockouts

1μg of the Cas9/sgRNA plasmid (~100ng/μL) and 5μg of linear repair DNA (~500ng/μL) were added to 300μL of transformation mix. For the no repair controls, 10μL of water was added in place of the DNA. After heat shock, cells were washed with 300μL of water, pelleted, and resuspended in 100μL of water and plated entirely onto synthetic media lacking uracil.

To screen for the knockout phenotype(s), 40 colonies were picked into 500μL of synthetic media lacking uracil in 96-deep-well blocks and grown at 30C in an ATR shaker, shaking at 750RPM. Saturated cultures were washed twice in 500μL of water, then diluted 1:100 into four different media, each lacking the appropriate amino acid (leucine, histidine, methionine (and cysteine), or tryptophan). These cultures were then incubated again at 30C at 750RPM, and we counted the number of clones that showed growth in the correct set of media.

Protospacer sequences for sgRNAs were designed using Benchling (see **Table 2-2** for a list). See **Figure 2-10** for details on the design of repair DNA.

2.6 References

1. Duportet, X. *et al.* A platform for rapid prototyping of synthetic gene networks in mammalian cells. *Nucleic Acids Res* **42,** 13440–13451 (2014).
2. Engler, C. *et al.* A golden gate modular cloning toolbox for plants. *ACS Synth. Biol.* **3,** 839–843 (2014).
3. Torella, J. P. *et al.* Rapid construction of insulated genetic circuits via synthetic sequence-guided isothermal assembly. *Nucleic Acids Res* **42,** 681–689 (2014).
4. Sun, Z. Z., Yeung, E., Hayes, C. A., Noireaux, V. & Murray, R. M. Linear DNA for rapid prototyping of synthetic biological circuits in an Escherichia coli based TX-TL cell-free system. *ACS Synth. Biol.* **3,** 387–397 (2014).
5. Smanski, M. J. *et al.* Functional optimization of gene clusters by combinatorial design and assembly. *Nat Biotechnol* **32,** 1241–1249 (2014).
6. Bonnet, J., Subsoontorn, P. & Endy, D. Rewritable digital data storage in live cells via engineered control of recombination directionality. *Proc Natl Acad Sci USA* **109,** 8884–8889 (2012).
7. Wen, M., Bond-Watts, B. B. & Chang, M. C. Y. Production of advanced biofuels in engineered E. coli. *Curr Opin Chem Biol* **17,** 472–479 (2013).
8. Tsai, C.-S., Kwak, S., Turner, T. L. & Jin, Y.-S. Yeast synthetic biology toolbox and applications for biofuel production. *FEMS Yeast Res* (2014). doi:10.1111/1567-1364.12206
9. Temme, K., Zhao, D. & Voigt, C. A. Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca. *Proc Natl Acad Sci USA* **109,** 7085–7090 (2012).
10. Martin, V. J. J., Pitera, D. J., Withers, S. T., Newman, J. D. & Keasling, J. D. Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat Biotechnol* **21,** 796–802 (2003).
11. Ro, D.-K. *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440,** 940–943 (2006).
12. Ajikumar, P. K. *et al.* Isoprenoid pathway optimization for Taxol precursor overproduction in Escherichia coli. *Science* **330,** 70–74 (2010).
13. Boeke, J. D., LaCroute, F. & Fink, G. R. A positive selection for mutants lacking orotidine-5'-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance. *Mol. Gen. Genet.* **197,** 345–346 (1984).
14. Tong, A. H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294,** 2364–2368 (2001).
15. Da Silva, N. A. & Srikrishnan, S. Introduction and expression of genes for metabolic engineering applications in Saccharomyces cerevisiae. *FEMS Yeast Res* **12,** 197–214 (2012).
16. Giaever, G. *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418,** 387–391 (2002).
17. Tong, A. H. Y. *et al.* Global mapping of the yeast genetic interaction network. *Science* **303,** 808–813 (2004).

18. Forster, J., Famili, I., Fu, P., Palsson, B. Ø. & Nielsen, J. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res* **13,** 244–253 (2003).
19. Cherry, J. M. *et al.* SGD: Saccharomyces Genome Database. *Nucleic Acids Res* **26,** 73–79 (1998).
20. Paddon, C. J. *et al.* High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature* **496,** 528–532 (2013).
21. Hong, K.-K. & Nielsen, J. Metabolic engineering of Saccharomyces cerevisiae: a key cell factory platform for future biorefineries. *Cell. Mol. Life Sci.* **69,** 2671–2690 (2012).
22. Buchholz, K. & Collins, J. The roots--a short history of industrial microbiology and biotechnology. *Appl Microbiol Biotechnol* **97,** 3747–3762 (2013).
23. Canton, B., Labno, A. & Endy, D. Refinement and standardization of synthetic biological parts and devices. *Nat Biotechnol* **26,** 787–793 (2008).
24. Arkin, A. P. & Fletcher, D. A. Fast, cheap and somewhat in control. *Genome Biol.* **7,** 114 (2006).
25. Sprinzak, D. & Elowitz, M. B. Reconstruction of genetic circuits. *Nature* **438,** 443–448 (2005).
26. Purnick, P. E. M. & Weiss, R. The second wave of synthetic biology: from modules to systems. *Nat Rev Mol Cell Biol* **10,** 410–422 (2009).
27. Shetty, R. P., Endy, D. & Knight, T. F. Engineering BioBrick vectors from BioBrick parts. *Journal of biological engineering* **2,** 5 (2008).
28. Casini, A. *et al.* One-pot DNA construction for synthetic biology: the Modular Overlap-Directed Assembly with Linkers (MODAL) strategy. *Nucleic Acids Res* **42,** e7–e7 (2014).
29. Litcofsky, K. D., Afeyan, R. B., Krom, R. J., Khalil, A. S. & Collins, J. J. Iterative plug-and-play methodology for constructing and modifying synthetic gene networks. *Nat Methods* **9,** 1077–1080 (2012).
30. Anderson, J. C. *et al.* BglBricks: A flexible standard for biological part assembly. *Journal of biological engineering* **4,** 1 (2010).
31. Sarrion-Perdigones, A. *et al.* GoldenBraid: an iterative cloning system for standardized assembly of reusable genetic modules. *PLoS ONE* **6,** e21622 (2011).
32. Weber, E., Engler, C., Gruetzner, R., Werner, S. & Marillonnet, S. A modular cloning system for standardized assembly of multigene constructs. *PLoS ONE* **6,** e16765 (2011).
33. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE* **3,** e3647 (2008).
34. Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS ONE* **4,** e5553 (2009).
35. Siddiqui, M. S., Choksi, A. & Smolke, C. D. A system for multilocus chromosomal integration and transformation-free selection marker rescue. *FEMS Yeast Res* **14,** 1171–1185 (2014).
36. Li, M. Z. & Elledge, S. J. Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nat Methods* **4,** 251–256 (2007).
37. Gibson, D. G. *et al.* Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science* **319,** 1215–1220 (2008).

38. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6,** 343–345 (2009).
39. de Kok, S. *et al.* Rapid and reliable DNA assembly via ligase cycling reaction. *ACS Synth. Biol.* **3,** 97–106 (2014).
40. Shao, Z., Zhao, H. & Zhao, H. DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res* **37,** e16–e16 (2009).
41. Curran, K. A. *et al.* Design of synthetic yeast promoters via tuning of nucleosome architecture. *Nat Commun* **5,** 4002 (2014).
42. Blazeck, J., Garg, R., Reed, B. & Alper, H. S. Controlling promoter strength and regulation in Saccharomyces cerevisiae using synthetic hybrid promoters. *Biotechnol Bioeng* **109,** 2884–2895 (2012).
43. Newman, J. R. S. *et al.* Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441,** 840–846 (2006).
44. Lee, M. E., Aswani, A., Han, A. S., Tomlin, C. J. & Dueber, J. E. Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay. *Nucleic Acids Res* **41,** 10668–10678 (2013).
45. Keren, L. *et al.* Promoters maintain their relative activity levels under different growth conditions. *Mol Syst Biol* **9,** 701–701 (2013).
46. Hawkins, K. M. & Smolke, C. D. The regulatory roles of the galactose permease and kinase in the induction response of the GAL network in Saccharomyces cerevisiae. *J Biol Chem* **281,** 13485–13492 (2006).
47. Hackett, E. A., Esch, R. K., Maleri, S. & Errede, B. A family of destabilized cyan fluorescent proteins as transcriptional reporters in S. cerevisiae. *Yeast* **23,** 333–349 (2006).
48. Jensen, N. B. *et al.* EasyClone: method for iterative chromosomal integration of multiple genes in Saccharomyces cerevisiae. *FEMS Yeast Res* **14,** 238–248 (2014).
49. Wingler, L. M. & Cornish, V. W. Reiterative Recombination for the in vivo assembly of libraries of multigene pathways. *Proc Natl Acad Sci USA* **108,** 15135–15140 (2011).
50. Dicarlo, J. E. *et al.* Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res* **41,** 4336–4343 (2013).
51. Ryan, O. W. *et al.* Selection of chromosomal DNA libraries using a multiplex CRISPR system. *Elife* **3,** (2014).
52. Bao, Z. *et al.* Homology-Integrated CRISPR-Cas (HI-CRISPR) System for One-Step Multigene Disruption in Saccharomyces cerevisiae. *ACS Synth. Biol.* (2014). doi:10.1021/sb500255k
53. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat Genet* **14,** 450–456 (1996).

Chapter 3. Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay†

3.1 Introduction

Metabolic engineering offers the promise of inexpensive and clean biosynthesis of both high value products such as pharmaceuticals[1,2], and commodity chemicals such as transportation fuel replacements[3,4]. As noted in a recent review of the field[5], standardized engineering frameworks will be key in enabling faster iteration of the "design-build-test" cycle, leading to more productive strains. Recent advances in DNA assembly[6-12] have dramatically improved our ability to efficiently build multi-gene pathway libraries where we can vary expression levels, enzyme homologs and mutants, and other attributes in a combinatorial fashion. Once assembled, the large size inherent to these combinatorial libraries demands high-throughput analysis to isolate a high-performance strain. However, the majority of target molecules cannot be measured in high-throughput, which places the natural inclination to approach optimization of multiple variables via library screening at odds with the strict requirement to minimize the number of measurements. Here we describe a strategy that overcomes this limitation by coupling regression modeling with multi-gene combinatorial libraries and show that sparse sampling of those libraries can be sufficient to optimize metabolic pathways.

To achieve efficient bioconversion, it is often crucial to balance the relative activity of each enzyme in a pathway to avoid detrimental effects from accumulated intermediate metabolites[13-15]. Additionally, it can be a burden on the cell to support a highly expressed foreign pathway[16,17], and, indeed, in some cases lowering expression of certain enzymes in a pathway has been shown to increase product titers[2,18], highlighting the importance of determining the right balance (**Figure 3-1A**).

Perhaps the most straightforward approach to balancing enzyme expression levels would be to begin at an arbitrary starting expression level and then iteratively adjust expression of each gene to identify its optimum. However, this approach is time-consuming, particularly as the number of genes to balance increases. A more elegant solution is to survey all possible expression levels combinatorially, which has the advantage of not only reducing the time cost, but also reveals the overall, multi-dimensional production landscape. To date, few production landscapes have been explored due to the prior difficulties both in building libraries as well as determining enzyme expression levels. Close inspection of two landscapes that have been explored, the isoprenoid pathway for taxadiene production in *Escherichia coli*[2] and xylose fermentation in *Saccharomyces cerevisiae*[19], show that iterative expression tuning could potentially fail to identify the true optimum depending on the order in which operons or enzymes were tuned. While combinatorial libraries enable researchers to avoid these traps, one major difficulty faced is the limited scale that can be practically surveyed. The library diversities in the aforementioned examples were

† Reproduced with permission from Lee, M. E., Aswani, A., Han, A. S., Tomlin, C. J. & Dueber, J. E. "Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay." *Nucleic Acids Res* 41, 10668–10678 (2013).

sixteen and eight combinations, respectively, allowing these libraries to be exhaustively sampled. Much larger libraries that include more expression levels, operons, or enzymes approach a limit where exhaustive sampling is not feasible.

A notable exception to this limit exists for pathways with a phenotype that can be assayed in high-throughput, such as growth rate or production of a colored molecule. Recently, another xylose-utilizing *S. cerevisiae* strain was isolated from a library of approximately one thousand combinations via selection on xylose as the sole carbon source[20]. In another study, expression levels of twenty-four genes involved in lycopene biosynthesis in *E. coli* were optimized using multiplex automated genome engineering (MAGE)[21]. These stunning examples of large-scale optimization demonstrate the power of combinatorial expression libraries; however, harnessing this enormous diversity required a high-throughput screen or selection to efficiently comb through the vast assortment of genetic variants. Unfortunately, the majority of small molecules of interest, including most biofuels and specialty chemicals, must be quantified using analytical methods such as HPLC, GC-MS, LC-MS, etc., which provide insufficient throughput to warrant the use of these emerging technologies for constructing massive libraries for combinatorial searches.

We propose that computational modeling can provide the necessary link between large searches and targets that are difficult to screen. If gene expression can be reliably controlled, the production landscape of a molecule can be discretized into a multi-dimensional grid of expression space, and, by sampling this space, we can fit a function that relates gene expression to product titer. To that end, we constructed and characterized a *S. cerevisiae* promoter library that exhibited robust control over gene expression. We

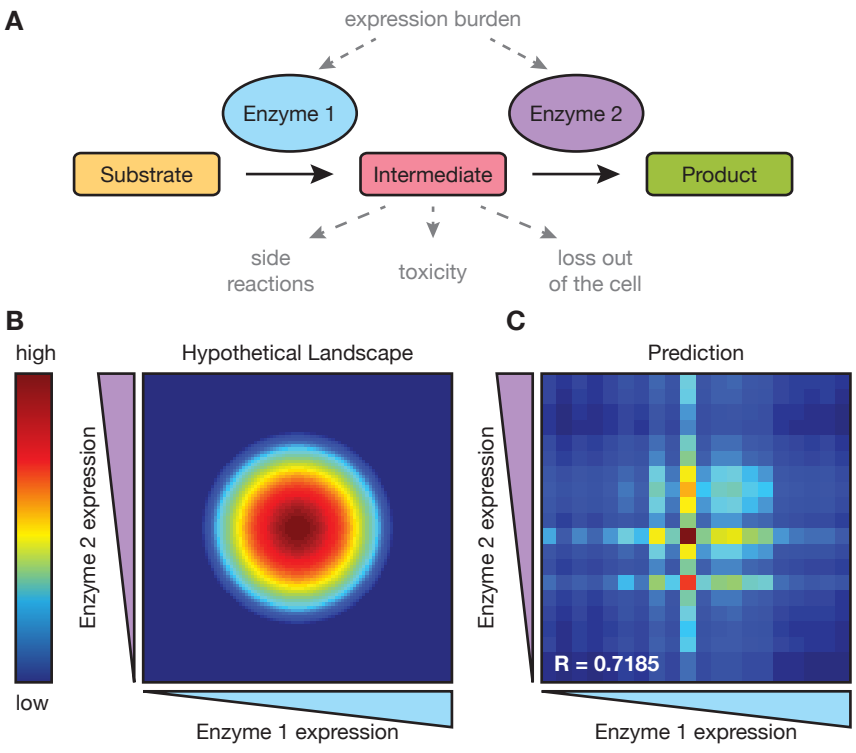


Figure 3-1. Metabolic enzyme expression balancing and modeling. (A) A hypothetical two-gene metabolic pathway. Overexpression of enzymes may cause a burden to the cell; the intermediate may be non-productive by reacting with alternative enzymes, causing cytotoxicity, or leaving the cell by active or passive transport. (B) The associated production landscape showing that for one or a combination of the above-stated scenarios, moderate gene expression is optimal. (C) An example of the landscape predicted by a regression model trained on 100 randomly sampled points from the true landscape. The correlation coefficient shown is between the predicted and true values in the 20-by-20 discretized space.

developed a standardized assembly strategy to build combinatorial libraries, as well as a rapid genotyping method to determine the promoter identity for each gene in a given library member. We then used linear regression to fit a model to the genotype and titer measurement data. As a challenging test of this relatively straightforward modeling approach, we examined the highly complex violacein biosynthetic pathway. This pathway exhibited several characteristics that commonly plague metabolic engineers: a branched pathway structure leading to off-target side reactions, both enzymatic and spontaneous; promiscuous enzymes that can recognize multiple intermediates as their substrate; and, being the first report to our knowledge to express the pathway in *S. cerevisiae*, uncertainty in enzyme activity in the heterologous host. Despite these traits, we successfully produced violacein in yeast, and we utilized a regression model to predict strains that selectively maximized production of any one of the four primary products in the pathway.

3.2 Results

3.2.1 Modeling a production landscape using linear regression

Modeling the intricate network of enzymes and metabolites of cell metabolism presents a daunting task. There are many parameters to be considered, such as enzyme kinetics and intracellular metabolite concentrations, but these data are often unavailable, especially for heterologously expressed genes. Additionally, gene clusters taken from exotic organisms may not be fully characterized, and even the order of the reactions and identity of the intermediates of the pathway could be unknown. Therefore, it can be advantageous to take a simpler modeling approach that is somewhat naïve to the complexities of biology.

We chose to use a linear regression model[22] trained on empirical data to relate expression level combinations to product titer. As an initial test, we generated a hypothetical production landscape of a two-gene pathway designed to mimic that of the taxadiene pathway described by Ajikumar *et al.*[2] where intermediate expression levels were optimal (**Figure 3-1B**). We discretized the continuous expression of each gene into twenty levels (*e.g.*, promoter strengths), and sampled one hundred random points from the resulting lattice. The model we trained is a categorical model, wherein the presence or absence of each promoter-gene combination is represented as an independent variable, rather than a relative expression level for each gene (see **Section 3.4** for more details). A predicted landscape representative of one hundred simulations is shown in **Figure 3-1C**. While the model is not perfect in accurately predicting every point, it is certainly able to determine that moderate expression of both enzymes is preferred over high or low expression.

The limitations of the model's predictive power are a consequence of the assumptions necessary to maintain simplicity. First, we assumed that each enzyme contributes to pathway flux independently. We know that this may be biologically inaccurate, due to potential interactions between enzymes or regulation that would result in co-dependence of two or more enzymes. However, including these non-linear interactions would greatly increase the amount of data necessary to train the model, a quantity we sought to limit. By extension, we assumed that production landscapes in general are relatively smooth

and contain a single peak. Naturally, these assumptions will prevent the model from successfully identifying the optimum for certain outstanding cases, but for most pathways, it should provide an excellent first-pass analysis of how the pathway responds to changing gene expression.

3.2.2 Constitutive promoters provide robust control over protein expression

In order to implement the modeling approach described above, we first needed control over protein expression, which we accomplished by varying promoters. We defined several criteria for designing a promoter library: a) a wide range of transcriptional strengths that are evenly distributed; b) minimal variation in strength with respect to different coding sequences; and c) orthogonal DNA sequences to minimize recombination and simplify genotyping. Thus, we avoided promoter mutagenesis, such as the common-

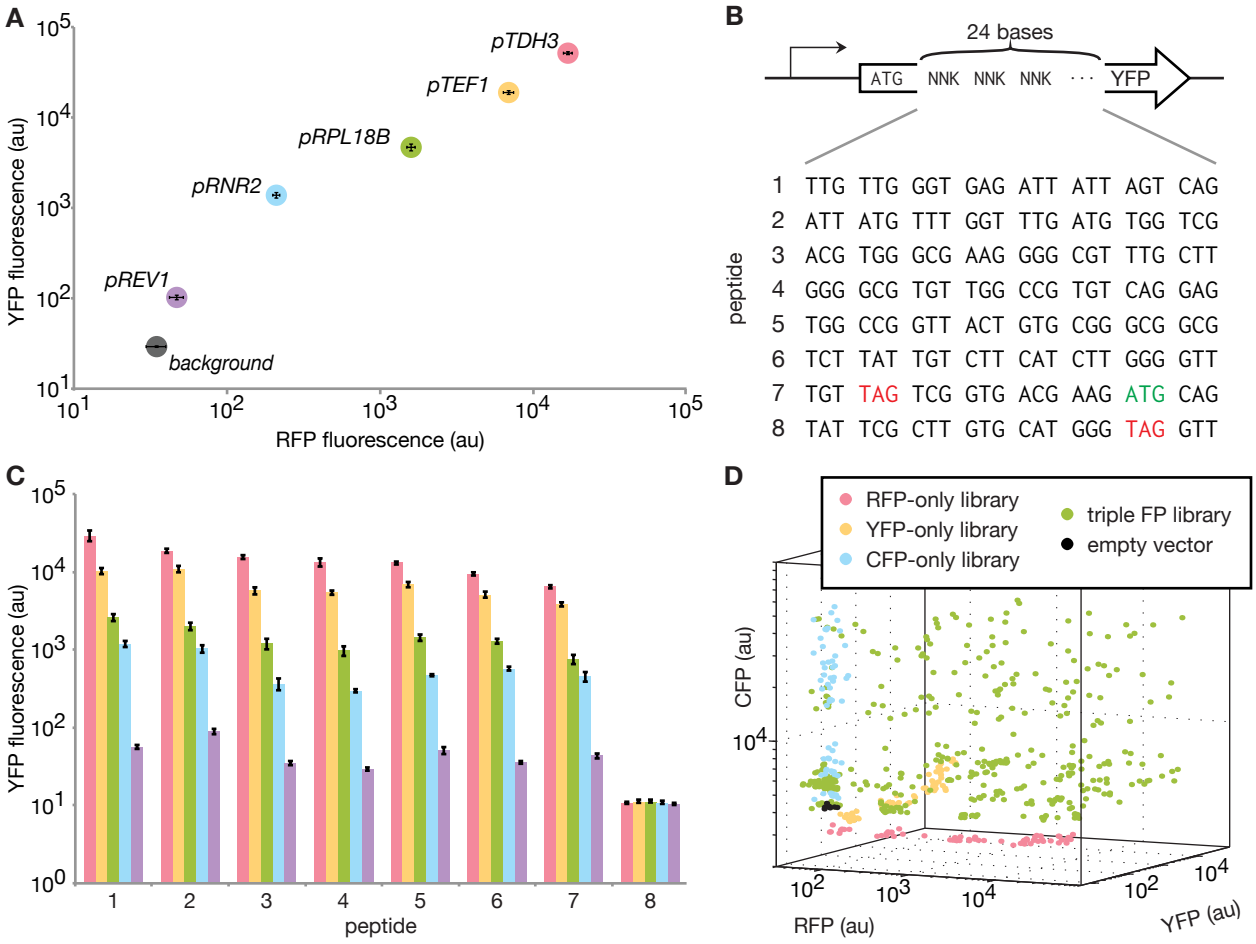


Figure 3-2. Characterization of yeast constitutive promoters. (A) Five promoter regions cloned from the yeast genome give consistent expression of two fluorescent reporters. (B) Twenty-four random nucleotides are fused to the 5' end of YFP. Notably, sequence 7 has an in-frame stop codon and a second start codon; sequence 8 has an in-frame stop codon. (C) Random nucleotide sequences do not appreciably alter YFP fluorescence (compare all bars of a single color). Additionally, the rank order of promoter strengths for a given coding sequence is maintained (compare all sets of bars for a given peptide). (D) Combinatorial assembly of promoter libraries: five promoters are combinatorially cloned in front of RFP, YFP, and CFP separately (red, yellow, and blue); RFP, YFP, and CFP libraries are combinatorially assembled (green); empty vector (black). Error bars in panels (A) and (C) indicate s.e.m. n=8.

ly-used *TEF1* library[23,24], because of the high degree of homology between those promoters and their relatively limited range of ten-fold expression. Instead, we collected a set of sequences taken from upstream of the translational start site of several yeast genes observed to have a broad range of expression levels[25], and we cloned 700bp as canonical "promoters" in front of three fluorescent reporters, mKate2 (RFP), Venus (YFP), and CFP, to test against our criteria.

We identified a set of five promoters—*pTDH3* (only 680bp), *pTEF1*, *pRPL18B*, *pRNR2*, and *pREV1*—that had all of our desired characteristics. The promoters spanned nearly three orders of magnitude in red and yellow fluorescence, with relatively even separation between members on a log-scale (**Figure 3-2A**). We were concerned that the strength of these promoters would be influenced by the downstream coding sequence, as is often observed in *E. coli* due to interactions with the ribosome binding site[26-28]. To address this, we cloned our library in front of a random sequence of twenty-four nucleotides fused to YFP, and saw that the relative rank order of promoters was remarkably well maintained (**Figure 3-2B and C**). Because we are only controlling transcription, we cannot ensure absolute protein levels, which may be influenced by other factors such as transcript and polypeptide length, folding, or translation rate; however, these effects are largely dependent on sequence, not concentration, and so high and low amounts of a given transcript should be affected equally, giving rise to the consistency of relative promoter strengths for a particular coding sequence.

In contrast to the simulated scenario, we decided to use these five promoters rather than twenty for practical reasons. First, while having more promoters would provide higher resolution of the landscape, it would also increase the total diversity of the library, thus

Table 3-1. Recombination rates of tandem expression plasmids.

A	RFP loss	YFP loss	intact
pTDH3	0	0	48
pTEF1	0	1	47
pRPL18B	0	0	48
pRNR2	1	0	47

B	RFP loss	YFP loss	intact
pTDH3	0	0	48
pTEF1	3	0	45
pRPL18B	0	0	48
pRNR2	0	0	48

C	RFP loss	YFP loss	CFP loss	intact
pTDH3	1	0	0	47
pTEF1	1	0	0	47
pRPL18B	0	0	0	48
pRNR2	0	1	0	47

Plasmids containing tandem expression cassettes of RFP and YFP (A), YFP and RFP (B), or RFP, YFP, and CFP (C) were transformed into yeast, and 48 colonies were picked for each construct. *pREV1* was omitted due to its low signal over background making it difficult to discern a loss of fluorescence.

requiring a higher sampling rate. Second, given the limits on the dynamic range that can be accessed by changing the promoter at single copy (approximately three orders of magnitude), having twenty promoters would mean that each promoter resulted in only 50% more protein than the next lowest promoter. It could prove difficult to deconvolve the contributions of these small differences in expression and noise in sample measurement. Rather, the larger, roughly 500% increments from a five-member library are more likely to provide meaningful data.

Although we avoided highly homologous sequences for each promoter, because we intended to use them in long pathways, we were still concerned about recombination since the same promoter could appear more than once in a single plasmid. Thus, we cloned RFP and YFP onto a single plasmid with both genes driven by the same promoter (*e.g.*, *pTDH3*-RFP-terminator-*pTDH3*-YFP-terminator), and used loss of fluorescence as an indicator of homologous recombination between repeated promoter or terminator sequences. We also reversed the order of the genes (YFP-RFP) and included CFP (RFP-YFP-CFP). Only approximately one percent of colonies lost a fluorescent reporter (**Table 3-1**), and in the absence of any selective pressure to recombine, transformants with fully intact plasmids remained stable after subculturing every twenty-four hours for five days, with zero clones out of forty-eight losing any of their reporters.

3.2.3 Construction of multi-gene libraries

Using this characterized set of promoter sequences, we sought to generate combinatorial libraries in which we simultaneously titrated the expression of all pathway genes. We designed standardized, modular cloning vectors for constructing multi-gene plasmids using Gibson assembly[6,29] (**Figure 3-3**), which allowed us to combine arbitrary combinations of genes and promoters easily. To test our cloning strategy, we took three separate fluorescent protein libraries (RFP, YFP, and CFP) and assembled them into a single plas-

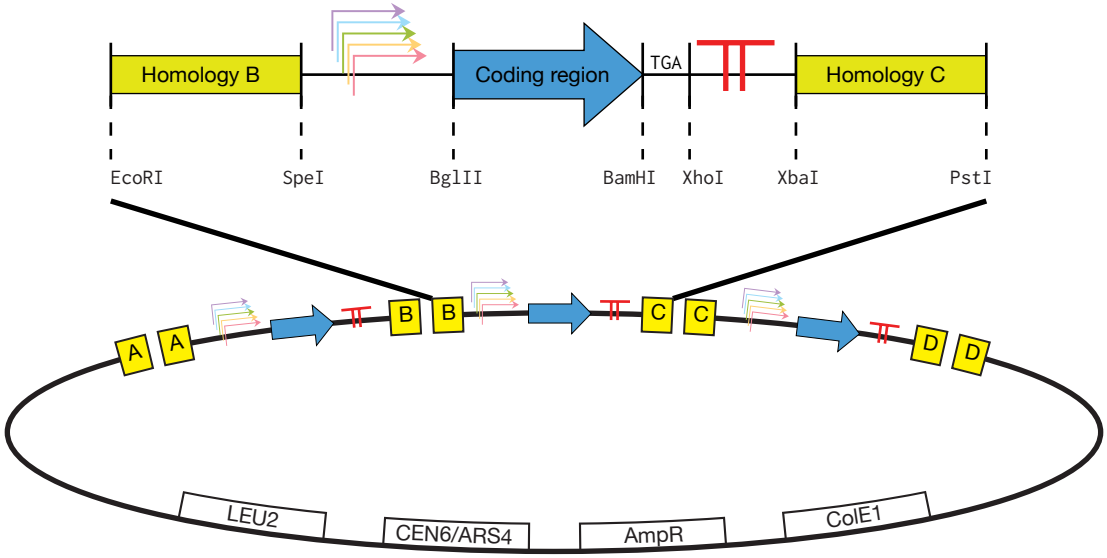


Figure 3-3. Gibson assembly of multi-gene constructs. Expression cassettes comprising of a promoter (library), gene, and transcriptional terminator are flanked by unique DNA homology sequences. Homology allows for specific assembly of multiple cassettes into a recipient vector backbone.

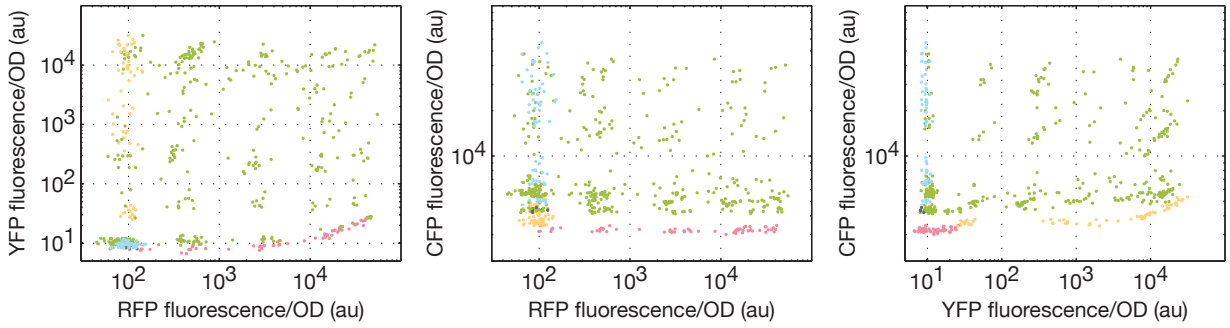


Figure 3-4. Combinatorial assembly of a fluorescent protein library. Two-dimensional projections of the data shown in **Figure 3-2D**. RFP-only library (red dots, *LEU2*), YFP-only library (yellow dots, *HIS3*), CFP-only library (blue dots, *URA3*), triple FP library (green dots, *MET15*), empty vector (black dots, *MET15*). *N.b.*, the individual libraries and triple library are expressed from plasmids carrying different auxotrophic markers, which may contribute to the lower baseline CFP fluorescence observed for the RFP and YFP libraries.

mid library (complexity of $5^3 = 125$ members). In comparing the fluorescence of colonies picked from the three-fluorescent protein library to that of colonies picked from each of the single fluorescent protein libraries, we saw that the triple-library roughly covered all of the three-dimensional "expression space" spanned by our promoters (**Figure 3-2D** and **Figure 3-4**). As can be seen in **Figure 3-4**, fluorescence of the triple-library clones clustered around the discrete intervals set by the promoters, occupying a lattice of points. We expect to see a similar pattern of coverage for the n-dimensional expression space of an n-gene system.

3.2.4 TRAC, a rapid genotyping assay

Although the goal of our modeling approach was to reduce the number of sample measurements, our cloning strategy was intended to be unrestrictive in the number of genes that could be expressed. Therefore, we anticipated a need for a rapid and inexpensive method for identifying the unknown promoters driving each gene for a given clone isolated from the library. The turnaround time compared to purification and sequencing of plasmids or polymerase chain reaction (PCR) products could be considerably reduced with an assay that directly determined promoter identity.

Since our promoter sequences were highly orthogonal, we were able to easily adapt the TaqMan method used in quantitative real-time PCR and allelic discrimination. For our assay, ***T***aqMan ***R***apid ***A***nalysis of ***C***ombinatorial assemblies (TRAC), we designed five

Table 3-2. Sequences of TRAC duplex probes.

Target	Dye strand (5'-3')	Quencher strand (5'-3')
pTDH3	[6-FAM]-ACACAAGGCAATTGACCCACG-(P)	TGGGTCAATTGCCTTGTGT-[IABkFQ]
pTEF1	[Cy3]-ACAACAGAAAGCGACCACCCAAC-(P)	GGTGGTCGCTTTCTGTTGT-[IABkFQ]
pRPL18B	[Cy5.5]-TCACGCCCAAGAAATCAGGC-(P)	CTGATTTCTTGGGCGTGA-[IAbRQSp]
pRNR2	[6-ROXN]-AAGCACGGGCAGATAGCACC-(P)	GCTATCTGCCCGTGCTT-[IAbRQSp]
pREV1	[Cy5]-ATGCCGCGTTCACAGATTCC-(P)	CTGTGAACGCGGCAT-[IAbRQSp]

Dye strands are labeled on their 5' end with a fluorescent dye, indicated in brackets, and on their 3' end with a phosphate (P). Quencher strands are labeled on their 3' end with Iowa Black® FQ or RQ quenchers, indicated in brackets.

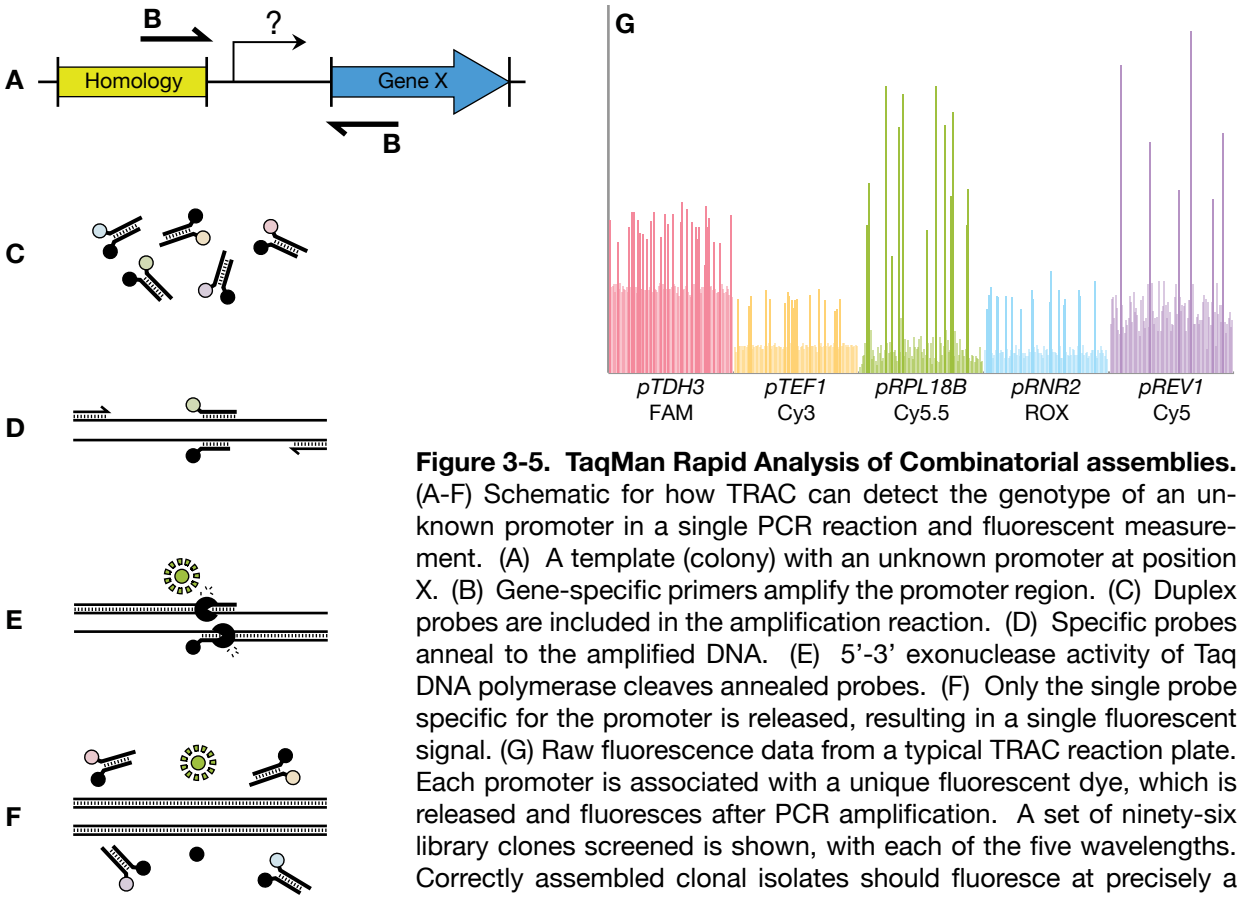


Figure 3-5. TaqMan Rapid Analysis of Combinatorial assemblies. (A-F) Schematic for how TRAC can detect the genotype of an unknown promoter in a single PCR reaction and fluorescent measurement. (A) A template (colony) with an unknown promoter at position X. (B) Gene-specific primers amplify the promoter region. (C) Duplex probes are included in the amplification reaction. (D) Specific probes anneal to the amplified DNA. (E) 5'-3' exonuclease activity of Taq DNA polymerase cleaves annealed probes. (F) Only the single probe specific for the promoter is released, resulting in a single fluorescent signal. (G) Raw fluorescence data from a typical TRAC reaction plate. Each promoter is associated with a unique fluorescent dye, which is released and fluoresces after PCR amplification. A set of ninety-six library clones screened is shown, with each of the five wavelengths. Correctly assembled clonal isolates should fluoresce at precisely a single wavelength.

orthogonal DNA oligonucleotide duplex probes[30], specific for each of the five promoter sequences and labeled with spectrally distinct, fluorescent dyes and Förster resonance energy transfer (FRET) quenchers (**Table 3-2**). When these probes were included in a PCR reaction with gene-specific primers amplifying an unknown promoter, only one fluorescent dye was released, which corresponded to the promoter present at that locus (**Figure 3-5**). This fluorescent signal could be read on a standard plate reader, which simplified the genotyping process by eliminating the need for a downstream gel, purification, or sequencing reaction. Not only was the time required for genotype identification low, but also the additional cost of oligonucleotide probes added only cents per reaction.

Because the specificity of the gene is determined by the PCR primers and not the fluorescent probes, this genotyping method is scalable to any number of genes. However, we were curious whether we could expand the number of unique sequences that could be identified by TRAC, in case a larger set of promoters were needed for future applications. There is a limit to the number of probes that can be used simultaneously due to overlapping excitation and emission spectra of the dyes. However, by designing sequences that contained either complementary or non-complementary sequences for all five probes in a row, we were able to detect thirty-two (2^5) unique "TRAC barcodes" (**Figure 3-6**). A sixth fluorescent dye, Alexa Fluor® 750, available from Integrated DNA Technologies, has excitation and emission spectra that do not overlap with our current set of five, although we have not tested it. If it proved to be compatible, it would enable detection of up to

six unique sequences by standard TRAC, or up to sixty-four (2^6) unique TRAC barcodes.

3.2.5 Violacein biosynthesis as a model pathway

With the tools in hand to construct and analyze metabolic pathways, we demonstrated our approach using the five-gene violacein biosynthetic pathway (*vioABEDC*) from *Chromobacterium violaceum*[31] (**Figure 3-7**). The primary reason we chose this pathway as a model system was not for its final product, but rather for the interesting characteristics of the pathway itself. First, the pathway is highly branched, leading to several potential products. This would allow us to raise the question of whether regression modeling can be used to predict strains that preferentially direct flux down a particular branch. Second, the enzyme encoded by *vioC* is known to act on two pathway intermediates (protoviolaceinic acid and protodeoxyviolaceinic acid) as substrates. Finally, not only had this pathway not been previously expressed in yeast, but also, much of the pathway was only very recently characterized[31,32] and some of the side pathway reactions have yet to be determined. Together, we felt these traits made violacein a challenging pathway for our strategy and one that was representative of many metabolic engineering efforts.

We transformed yeast with the assembled combinatorial pathway library ($5^5 = 3{,}125$ combinations), and the resulting colonies had a wide range of colors and intensities (**Figure 3-8**). Although the pathway's products exhibit a color phenotype, we recognized that the majority of chemical compounds are not colored. Therefore, we decided to forego a colorimetric screen in favor of HPLC, a low-throughput analytical method that would be more representative of other pathways of interest (**Figure 3-9**). HPLC analysis revealed that when the pathway was expressed in yeast, four major compounds—violacein, deoxyviolacein, proviolacein, and prodeoxyviolacein—were produced in significant quantities, while only trace amounts of deoxychromoviridans, chromoviridans, and oxychromoviridans were detected in some samples. The reaction mechanism for the formation of the chromoviridans compounds has not previously been determined, nor is it clear why that reaction would be inefficient in yeast.

3.2.6 Model predictions of the violacein pathway

We sampled ninety-one random transformants from our expression library, identified their promoter genotypes using TRAC, and measured their production titers for each

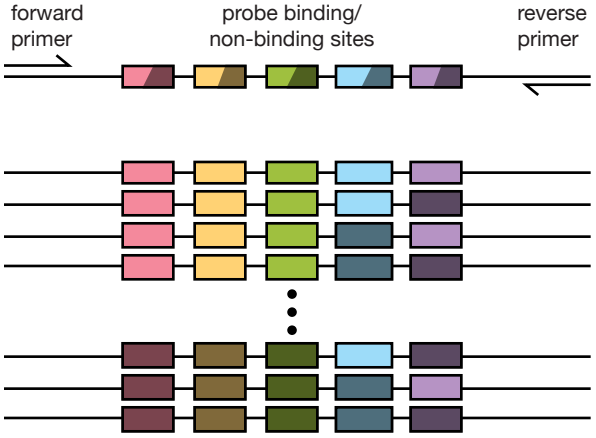


Figure 3-6. "TRAC barcode" design. Barcodes were cloned to include either a complementary or non-complementary sequence for all five TRAC probes (2^5 = 32 possible sequences) and flanking PCR primer binding sites. When a TRAC reaction was performed, combinations of zero to five fluorescent dyes were cleaved depending on whether the complementary sequence for a particular probe was present in the template. Fluorescence was measured on a plate reader as per a typical TRAC reaction, and all thirty-two unique barcodes were successfully identified.

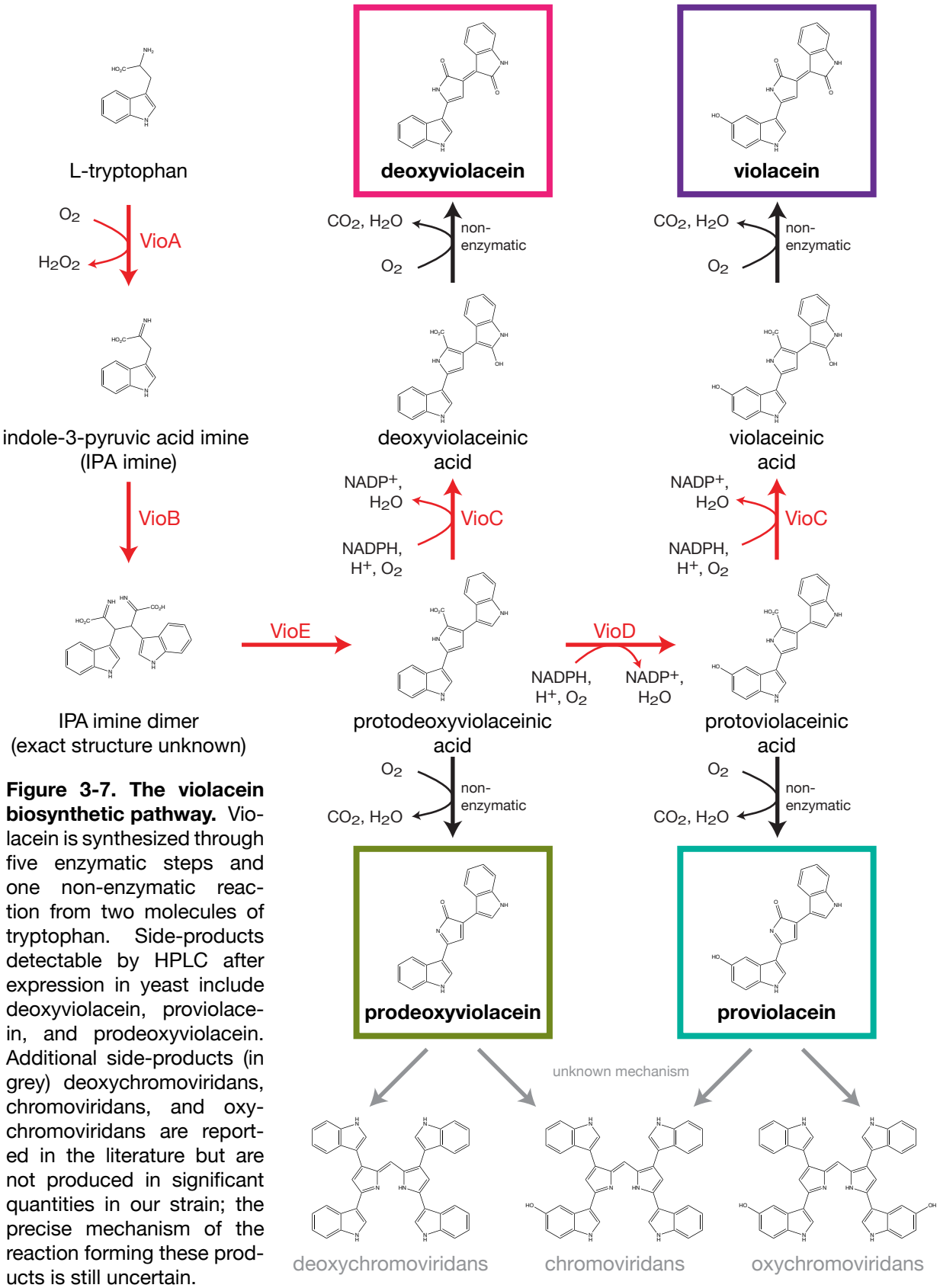


Figure 3-7. The violacein biosynthetic pathway. Violacein is synthesized through five enzymatic steps and one non-enzymatic reaction from two molecules of tryptophan. Side-products detectable by HPLC after expression in yeast include deoxyviolacein, proviolacein, and prodeoxyviolacein. Additional side-products (in grey) deoxychromoviridans, chromoviridans, and oxychromoviridans are reported in the literature but are not produced in significant quantities in our strain; the precise mechanism of the reaction forming these products is still uncertain.

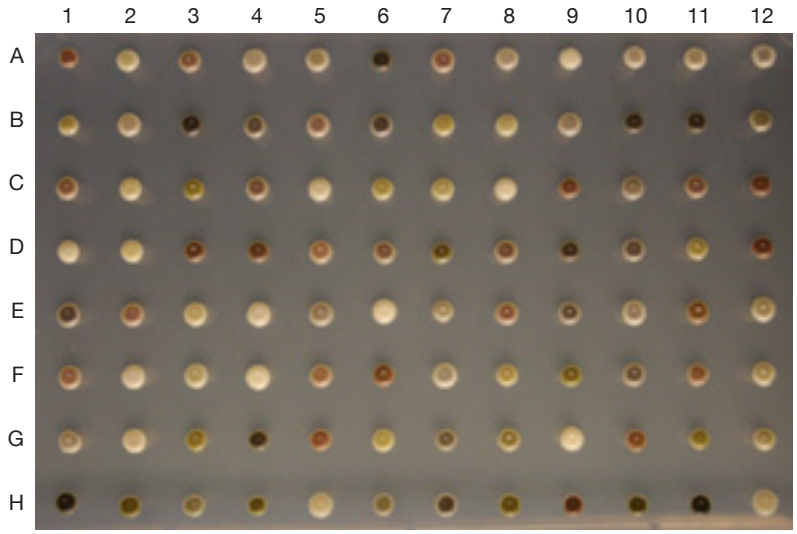


Figure 3-8. Violacein biosynthesis expression library. Ninety-six unique clones from a combinatorial expression library of the violacein biosynthetic pathway. The first ninety-one (A01-H07) were used as training data for the regression model. The last five are controls containing *pTDH3* driving: *vioABE* (H08), *vioABEC* (H09), *vioABED* (H10), *vioABEDC* (H11), empty vectors (H12).

of the four primary products. Using these data, we trained four models—one for each target—and then tested them against a test set of ninety-six additional, unique, random clones. Despite the complexities of the pathway, we found the correlation between the models' predictions and our empirical measurements were high (Pearson correlation coefficients were 0.80 for violacein, 0.77 for deoxyviolacein, 0.83 for proviolacein, and 0.92 for prodeoxyviolacein) (**Figure 3-10**). To test the effect of training set size on predictive power, we took random subsets of the original training set and measured correlation between the resulting models' predictions and the full, ninety-six-member test set data (**Figure 3-11**). We repeated this experiment one hundred times for subsets of size: five, ten, twenty, and fifty (and ninety-one). Interestingly, beyond the initial dramatic increase in correlation coefficient, only modest improvements were seen when increasing the training set to fifty or ninety-one samples. This suggests that a relatively low sampling rate (in this case, between one and two percent) may be sufficient for generating a predictive model.

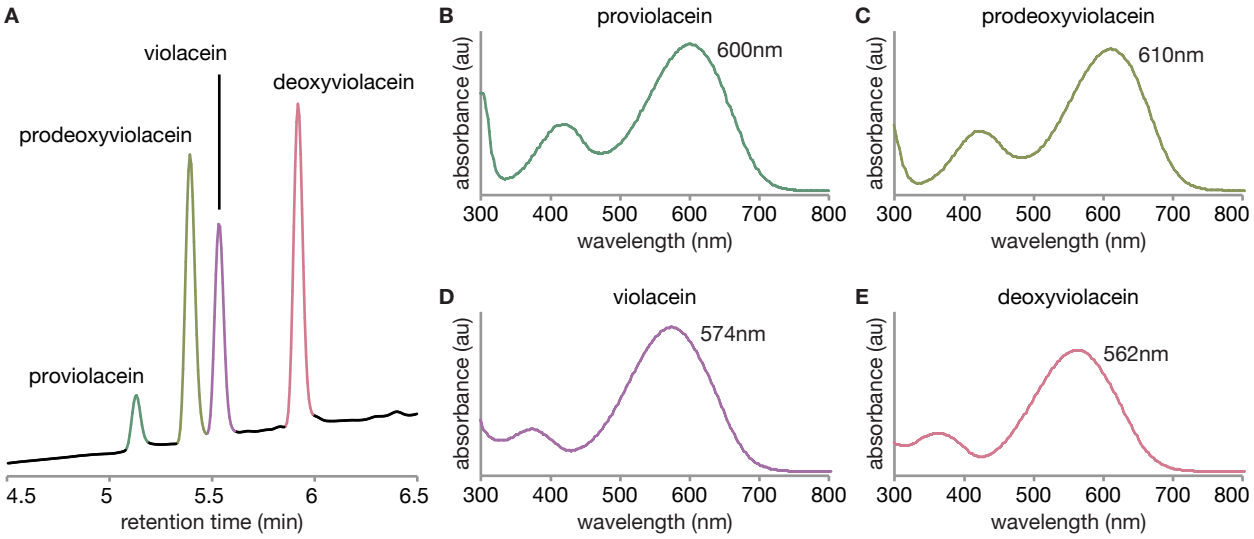


Figure 3-9. Chromatogram and absorbance spectra of violacein extractions. (A) Chromatogram for absorbance at 565nm. (B-E) Absorbance spectra for the four main products. Maximum absorbance wavelength is indicated.

We observed that a large number of samples in both the training and test sets had production levels below the limit of detection of our extraction and measurement protocols. Because of this, the models were trained on inherently flawed data on the low-production end, and therefore could not be expected to be as successful in predicting that range. However, it is encouraging that the models show much better correlation for highly productive strains (Pearson correlation coefficients were 0.84 for violacein, 0.90 for deoxyviolacein, 0.88 for proviolacein, and 0.92 for prodeoxyviolacein) (**Figure 3-10**), suggesting that the models' predictive power could be easily improved with higher quality data.

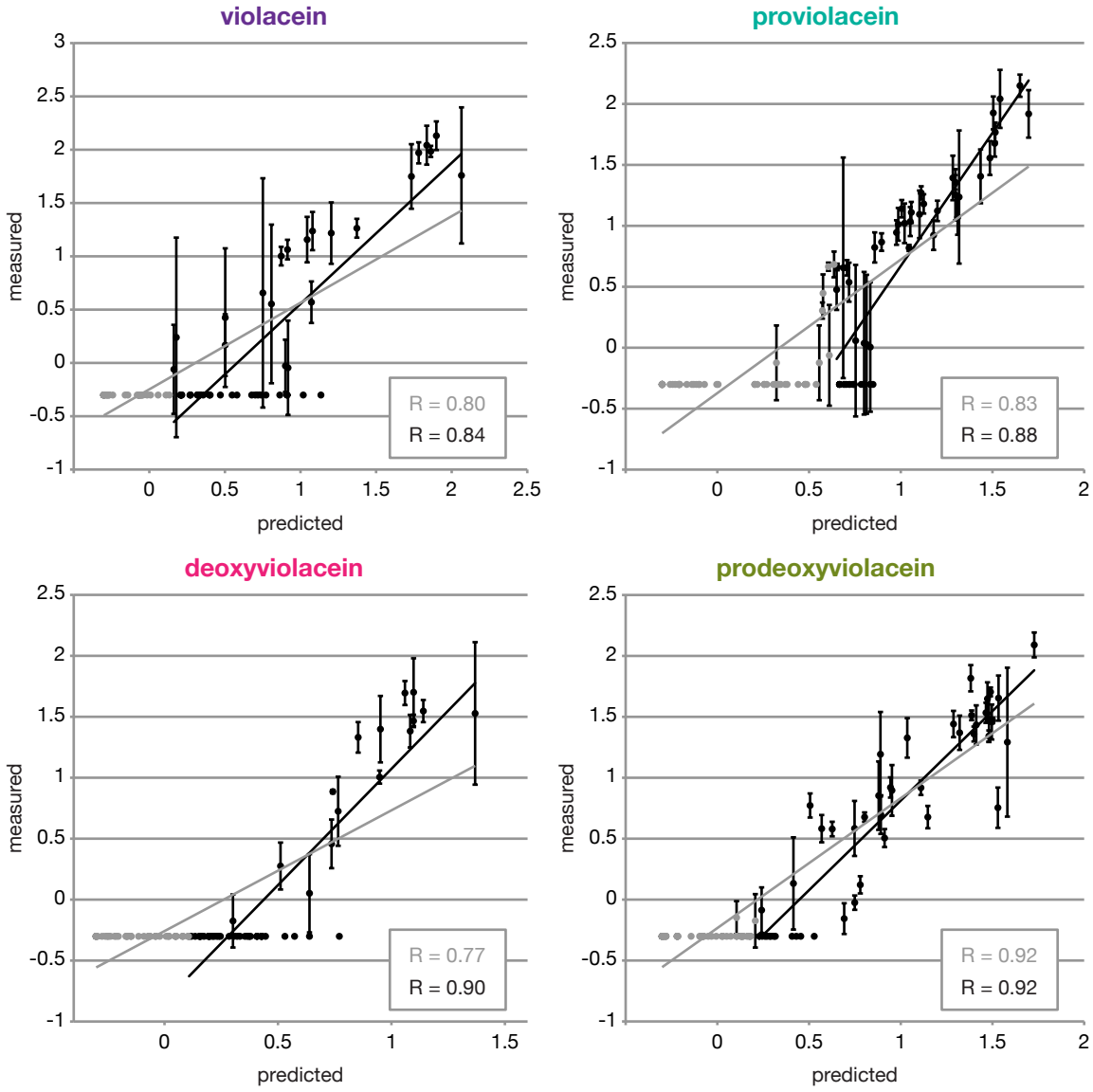


Figure 3-10. Model predictions. Comparison of model predictions with empirical measurements for a test set of ninety-six unique combinations. Black circles indicate the upper forty-eight combinations sorted by predicted titer for each respective product; grey circles indicate the lower forty-eight. The grey lines and correlation constants were calculated using all ninety-six data points; the black lines and correlation constants were calculated using only the upper forty-eight data points (*i.e.*, to roughly omit data that could be below the limit of detection). Axes are the logarithm of the titer, where titer is measured by the HPLC peak area in arbitrary units, *n.b.*, negative values indicate a titer less than 1au, not a net negative production; error bars indicate s.d. *n*=3.

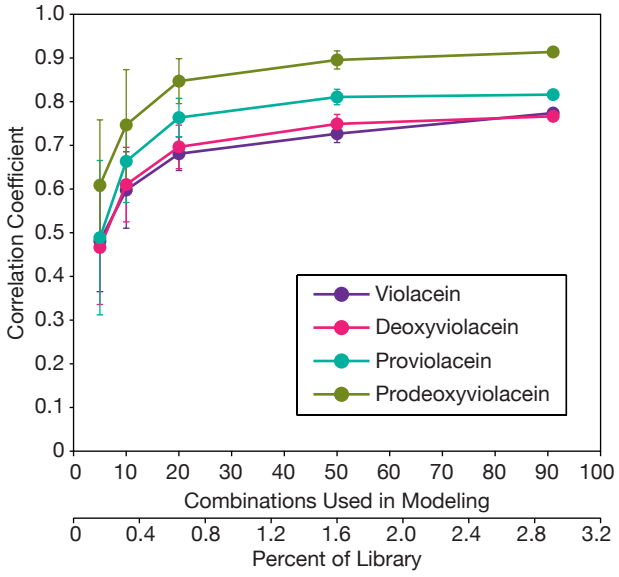


Figure 3-11. Effect of training set size on model accuracy. The original training set included ninety-one combinations; random subsets (of size 5, 10, 20, 50, and the full 91) were used to retrain the model, and the mean correlation coefficient of one hundred trials is shown. Larger training sets improve correlation, although the marginal benefit decreases as the number of samples increases. Error bars indicate s.d., n=100.

A far more interesting test of a model is whether it can forward-predict strains that would result in a desired phenotype. In this case, we took advantage of the branched nature of the violacein pathway and considered whether the models could predict genotypes that direct flux down any particular branch. We cloned the top five predicted combinations for each of the four products, and measured product titers from the resulting strains. We found the models were able to accurately capture the behavior of the pathway and provided predictions that preferentially produced one out of the four possible products (**Figure 3-12** and **Table 3-3**). For a given product queried, the predicted strains produced greater amounts of the desired target while minimizing the amount of off-target production as compared to strains predicted for any of the other three compounds.

3.3 Discussion

Synthetic biology strives to engineer biological systems to meet desired specifications using rigorously tested parts and models to achieve predictable behaviors. Given our incomplete understanding of the cell and its metabolism, but bolstered by our knowledge that metabolic flux is highly impacted by enzyme concentrations, systematically varying expression provides a promising approach for increasing production titers. The utility of these expression libraries can be augmented by using robust, well-characterized promoters that enable researchers to infer expression phenotype from genotype, and consequently gain insight into the design principles of a particular pathway. The promoters we constructed reliably span a wide range of expression strengths while maintaining their relative rank ordering irrespective of the coding sequence of the expressed gene. The steps of library construction and analysis are designed to be both generalizable to other pathways and scalable to increased numbers of enzymes to enable researchers to adopt the approach with relative ease. Additionally, this strategy need not be limited to this set of promoters or even to *S. cerevisiae*, as the only data required are production output and a measure for gene expression. The ability to link expression levels directly to the DNA sequence by using reliable, well-characterized control elements—whether they be transcriptional, translational, or post-translational—is essential for rapidly gathering data from many clones on several genes. For example, a newly developed expression architecture termed "bicistronic design" could provide robust control in *E. coli*, where it had previously been lacking[28]. Certainly, manipulating expression alone cannot be expected

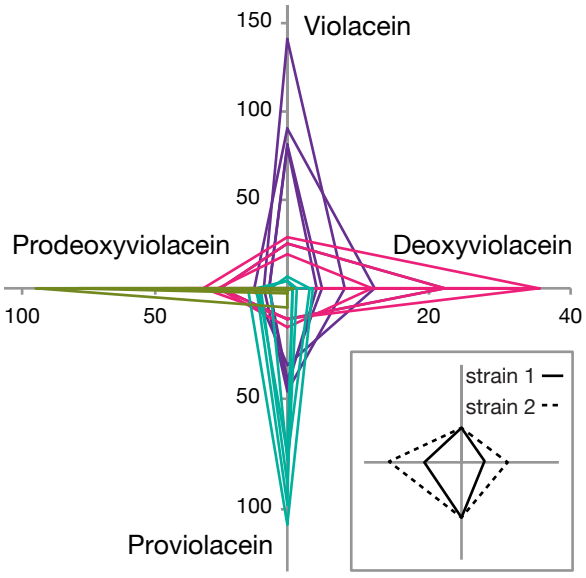


Figure 3-12. Strains with directed flux. The top five predicted combinations for each product and their associated relative titers. The five predictions for a given product are grouped by color: purple for violacein, pink for deoxyviolacein, teal for proviolacein, green for prodeoxyviolacein. Each predicted group shows preferential production of one product over the other three. Axes are relative product titer (HPLC peak area) in arbitrary units; each point is an average of four biological replicates (error bars not shown for clarity, but values are provided in **Table 3-3**). Each closed loop represents a single strain and the vertices indicate the titers of the four products. For example, strain 1 (solid line) produces equal amounts of violacein and proviolacein as strain 2 (dotted line), but half as much deoxyviolacein and prodeoxyviolacein (inset).

to yield a perfect strain, but these combinatorial libraries are compatible with classical methods such as protein engineering and directed evolution.

A recent review of metabolic engineering proposed that the goal of new engineering frameworks is to gain as much information as possible from a small number of experiments in order to allow researchers to hone in on the relevant areas and directions to ex-

Table 3-3. Strains with directed flux raw data.

Strain	Violacein	Deoxyviolacein	Proviolacein	Prodeoxyviolacein
V1	141.0 ± 14.5	8.1 ± 1.1	46.5 ± 8.5	8.7 ± 0.4
V2	81.5 ± 12.1	4.9 ± 0.7	41.3 ± 3.0	7.0 ± 0.5
V3	80.1 ± 8.1	4.1 ± 0.2	45.0 ± 6.3	6.9 ± 0.4
V4	25.4 ± 11.4	22.2 ± 10.8	14.1 ± 5.1	25.4 ± 8.6
V5	90.5 ± 18.5	12.3 ± 1.9	34.8 ± 3.3	12.5 ± 1.3
DV1	25.4 ± 11.4	22.2 ± 10.8	14.1 ± 5.1	25.4 ± 8.6
DV2	0 ± 0	32.5 ± 23.2	0 ± 0	38.7 ± 12.8
DV3	28.6 ± 4.4	35.5 ± 4.4	13.7 ± 3.3	31.7 ± 5.3
DV4	0 ± 0	35.6 ± 23.9	0 ± 0	47.5 ± 14.3
DV5	19.4 ± 1.6	11.9 ± 1.6	17.6 ± 2.3	26.2 ± 2.5
PV1	4.4 ± 1.7	0.8 ± 1.5	97.8 ± 5.0	11.8 ± 1.2
PV2	4.1 ± 2.4	1.3 ± 1.5	88.0 ± 10.5	10.2 ± 0.6
PV3	0 ± 0	0.7 ± 1.4	74.6 ± 6.5	11.1 ± 1.5
PV4	0 ± 0	3.6 ± 0.7	106.7 ± 5.6	15.0 ± 2.4
PV5	6.7 ± 1.4	3.1 ± 2.7	77.8 ± 52.2	6.7 ± 7.7
PDV1	0 ± 0	0 ± 0	0.9 ± 1.1	78.4 ± 34.6
PDV2	0 ± 0	0 ± 0	0 ± 0	79.3 ± 11.4
PDV3	0 ± 0	0 ± 0	2.3 ± 0.2	82.7 ± 2.6
PDV4	0 ± 0	0.8 ± 1.7	1.0 ± 1.2	73.2 ± 13.0
PDV5	0 ± 0	0 ± 0	8.7 ± 1.0	94.6 ± 15.9

Raw data represented in **Figure 3-12**. Strains designated V#, DV#, PV#, and PDV# are strains predicted by the model to produce high amounts of violacein, deoxyviolacein, proviolacein, and prodeoxyviolacein, respectively. Values shown are the average titer (as measured by HPLC peak area) of four biological replicates and the standard deviation.

plore[5]. Our modeling strategy is very much aligned with this idea, as it only requires only a modest number of measurements, unlike traditional combinatorial library approaches, which necessitate a high-throughput screen or selection. While linear regression may appear to be an overly simplified representation of a metabolic pathway, this is not the first time that linear regression has been used to describe a highly complex biological phenomenon. Linear regression was used in protein engineering to great effect in order to improve activity of a halohydrin dehalogenase[33] and predict thermostability in engineered cytochrome P450s[34].

Although protein-folding energy landscapes are commonly thought to be highly irregular due to the numerous semi-stable conformations that a protein may access, we believe that metabolic production landscapes are generally smoother. While it has been previously shown that moderate gene expression can sometimes be optimal[2,20], it isn't clear whether the inverse is possible—a multi-peaked landscape where moderate expression is detrimental and *both* high and low expression are beneficial. The model would likely be incapable of accurately describing a landscape containing multiple peaks, depending on the relative size and sharpness of the peaks and the sampling bias in the training set. However, we would expect this type of scenario to be rare. A more likely occurrence is a pathway that produces a toxic intermediate, where the relationship between enzymes (*e.g.*, the ratio or the sum of activities) must be maintained, which we suspect would result in a ridge-like topology. These ridges would still present a challenge to the model since they are incongruent with our assumption of enzyme independence, and so depending on the particular shape of the ridge (*e.g.*, a shallow slope along the top of the ridge), the model may not succeed in identifying the true optimum. The objective for this modeling strategy is to provide an estimation of the production landscape for newly engineered pathways, and, as such, failure to accurately describe and predict expression-level dependent performance, while problematic, actually highlights the possible presence of interesting biology to investigate in more detail.

In conclusion, we have developed a novel approach for optimizing enzyme expression for an engineered metabolic pathway that integrates combinatorial libraries with regression modeling to guide the researcher with a map of the production landscape. A major advantage of this strategy is that it requires no knowledge of absolute protein or metabolite levels, enzyme kinetics or thermodynamics, or even the order of the reactions. As such, the method is particularly useful when engineering new pathways that are not fully characterized, for example, gene clusters mined from metagenomic studies, or pathways with enzymes that have not been or cannot be easily purified and biochemically characterized. The results from an initial modeling attempt could be used as a starting point to investigate other avenues of optimization, be they as simple as further expression optimization or as involved as mutagenesis and directed evolution. In concert with these and other established metabolic engineering techniques, our strategy should dramatically accelerate the development of highly optimized strains as a sustainable replacement for chemical production.

3.4 Materials and Methods

Strains and growth media

The base *S. cerevisiae* strain for all experiments in this paper was BY4741 (*MAT*a *his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*). Wild-type yeast cultures were grown in YPD (10g/L Bacto Yeast Extract; 20g/L Bacto Peptone; 20g/L Dextrose). Yeast transformed with plasmids containing the *MET15*, *HIS3*, *LEU2*, or *URA3* auxotrophic markers were selected and grown on synthetic complete media (6.7g/L Difco Yeast Nitrogen Base w/o Amino Acids; 2g/L Drop-out Mix Synthetic Minus appropriate amino acids, w/o Yeast Nitrogen Base (US Biological); 20g/L Dextrose).

Yeast expressing the violacein pathway were grown on selective media for 48 hours at 30ºC. Cells grown on solid media containing 2% agar often took an additional 24-48 hours (at 4ºC) for color to develop fully.

Restriction cloning reactions were transformed in TG1 and DH10B chemically competent *E. coli*. Gibson assembly reactions were transformed in TransforMax EPI300 (Epicentre) electrocompetent *E. coli*. Transformed cells were selected on LB containing antibiotics ampicillin or kanamycin.

Standard yeast cloning vectors

Yeast cloning vectors derived from pRS316 were constructed to include unique restriction sites that flank each modular region of an expression cassette as well as allow for BglBrick-style cloning of protein fusions (using BglII, BamHI, and XhoI)[35] and BioBrick-style idempotent cloning of entire cassettes (using EcoRI, SpeI, XbaI, and PstI)[36]. Cloning vectors are listed in **Table 3-4**.

Yeast fluorescent protein measurement

Yeast transformed with plasmids expressing one or more fluorescent proteins were grown to saturation shaking in 96-deep-well blocks at 30ºC. Cell density (OD_{600}) and fluorescence were measured using a TECAN Safire2.

Violacein biosynthetic pathway

Genes for the violacein biosynthetic pathway were amplified from plasmid BBa_K274002 obtained from the Registry of Standard Biological Parts (partsregistry.org). A list of primers used for cloning the violacein genes are listed in **Table 3-5**, and a list of plasmids expressing those genes are listed in **Table 3-4**.

One-step isothermal assembly

Standard vectors were constructed, flanked by pairs of homology sequences derived from yeast barcodes[37] at the ends of each expression cassette. We reasoned that since these barcode sequences were designed to be orthogonal, they could serve a dual purpose of reducing the probability of mis-annealing and dictating the assembly order of multiple

Table 3-4. List of plasmids used in this study.

Plasmid	SynBERC Registry ID	Description	Yeast auxotrophic marker
pJED101	SBa_000896	Yeast cloning vector	*MET15*
pJED102	SBa_000897	Yeast cloning vector	*HIS3*
pJED103	SBa_000898	Yeast cloning vector	*LEU2*
pJED104	SBa_000899	Yeast cloning vector	*URA3*
pAH056	SBa_000900	*pTDH3*-RFP-*tADH1*	*LEU2*
pAH002	SBa_000901	*pTEF1*-RFP-*tADH1*	*LEU2*
pAH007	SBa_000902	*pRPL18B*-RFP-*tADH1*	*LEU2*
pSL030	SBa_000903	*pRNR2*-RFP-*tADH1*	*LEU2*
pAH005	SBa_000904	*pREV1*-RFP-*tADH1*	*LEU2*
pML167	SBa_000913	GibA-*pTDH3*-RFP-*tADH1*-GibB	*LEU2*
pML168	SBa_000914	GibB-*pTDH3*-YFP-*tADH1*-GibC	*HIS3*
pML159	SBa_000915	GibC-*pTDH3*-CFP-*tADH1*-GibD	*URA3*
pML203	SBa_000916	GibA-GibD vector	*MET15*
pML223	SBa_000917	GibA-GibD vector (KanR)	*URA3*
pML242	SBa_000891	GibA-*pTDH3-vioA-tADH1*-GibC	*LEU2*
pML243	SBa_000892	GibC-*pTDH3-vioC-tADH1*-GibD	*URA3*
pML244	SBa_000893	GibA-*pTDH3-vioB-tADH1*-GibB	*LEU2*
pML245	SBa_000894	GibB-*pTDH3-vioD-tADH1*-GibC	*HIS3*
pML246	SBa_000895	GibC-*pTDH3-vioE-tADH1*-GibD	*URA3*
pML256	SBa_000918	*vioAC* overexpression plasmid	*MET15*
pML258	SBa_000919	*vioBDE* overexpression plasmid (KanR)	*URA3*

All plasmids contain a ColE1 *E. coli* replication origin, carry an ampicillin resistance gene (unless otherwise indicated), and contain a CEN6/ARS4 *S. cerevisiae* replication origin. Annotated plasmid sequences can be found at the SynBERC Registry (registry.synberc.org). Sequences of plasmids not listed in this table (e.g., the series of YFP plasmids) can be determined simply by replacing the appropriate genes or promoters.

cassettes. *vioA* was flanked by an "A" and "C" homology sequence; *vioB* by "A" and "B"; *vioC* by "C" and "D"; *vioD* by "B" and "C"; *vioE* by "C" and "D"; backbone vectors contained "A" and "D" receiving sequences. Entire 5'homology-promoter-gene-terminator-3'homology cassettes were amplified by PCR; backbone vectors were also amplified by PCR or double-digested using SpeI/XbaI (*n.b.*, Taq DNA Ligase in the Gibson enzyme mix does not ligate compatible 4bp overhangs). Thus, *vioAC* and *vioBDE* plasmids were assembled using the compatible homology regions as the overlapping sequences for one-step isothermal assembly, which were performed as described in Gibson, et al[6]. See **Tables 3-4** and **3-5** for a list of plasmids and amplification primers.

There were some instances of mis-assembly where one or more cassettes may not be incorporated; however, this represented a relatively low percentage in three-gene assemblies (~25-33%), and even lower for two-genes (~8%). Additionally, in many of these cases of mis-assembly, homology of the inserts with the middle of the vector backbone resulted in the loss of the yeast replication origin and/or selection marker, such that upon transformation into yeast, the fraction of correctly assembled constructs that propagated in yeast was considerably higher.

Table 3-5. List of primers used in this study.

Primer	Sequence
vioA cloning forward	gcatAGATCTatgaaacattcttccgatat
vioA cloning reverse	atgcCTCGAGttaGGATCCcgcggcgatacgctgcaaca
vioB cloning forward	gcatAGATCTatgagcattctggatttccc
vioB cloning reverse	atgcCTCGAGtcaGGATCCggcctcgcggctcagtttgc
vioC cloning forward	gcatAGATCTatgaaacgtgcgattatcgt
vioC cloning reverse	atgcCTCGAGtcaGGATCCattcacgcgaccaatcttgt
vioD cloning forward	gcatAGATCTatgaagattctggtcattgg
vioD cloning reverse	atgcCTCGAGtcaGGATCCgcgctgcaaagcataacgca
vioE cloning forward	gcatAGATCTatggagaaccgtgagccacc
vioE cloning reverse	atgcCTCGAGtcaGGATCCgcgcttggccgcgaaaaccg
Gibson A amplification forward	ggtacagacactgcgacaac
Gibson A amplification reverse	gtattgcgacgaattgccacgttgtcg
Gibson B amplification forward	gggtcatcacggctcatc
Gibson B amplification reverse	agctgtgttgacatctggc
Gibson C amplification forward	ggtgatccgctgactcct
Gibson C amplification reverse	ggctcacgtcttatttgggc
Gibson D amplification forward	cacaaggtcagggcactcatgcgac
Gibson D amplification reverse	tgcatcgagttgattgtcgc
Gibson A TRAC forward	gccgataattgcagacg
Gibson B TRAC forward	ccagatgtcaacacagctac
Gibson C TRAC forward	acacactggcttaaggagac
vioA TRAC reverse	caatgcagatatcggaagaatg
vioB TRAC reverse	aagtggatacgcgggaaatc
vioC TRAC reverse	gacgtgcacttcgtagcc
vioD TRAC reverse	gtcattcttctccacgatgtca
vioE TRAC reverse	tcgggctccaataagagacata

Library plasmid purification

Libraries constructed by restriction or one-step isothermal assembly were transformed and plated on LB-agar plates containing antibiotic. After colonies appeared, plates were scraped, and the pooled collection of colonies was used for plasmid purification.

Extraction of pathway products

Yeast clones were grown in 1mL of synthetic media split into two wells in a 96-deep-well block in an ATR shaker at 30ºC for 48 hours. Cultures were recombined and pelleted in a microcentrifuge for three minutes at 14,000rpm. The pellets were resuspended in 500μL of methanol and boiled at 95ºC for 15 minutes, vortexing halfway through. Resuspensions were pelleted and the supernatant (extract) was transferred to new microcentrifuge tubes and pelleted to remove remaining cell debris. Final extracts were transferred to glass vials for analysis on HPLC.

HPLC analysis of pathway products

Ten microliters of extract were run on an Agilent Rapid Resolution SB-C18 column (30x2.1mm, 3.5μm particle size) on an Agilent 1200 Series LC system with the following method (Solvent A is 0.1% formic acid in water; Solvent B is 0.1% formic acid in acetoni-

trile): start at 5% B; hold at 5% B for 1.5min; 16.9%/min to 98% B; hold at 98% B for 2min; 3.1%/sec to 5% B; hold at 5% B for 2.5min. The column temperature was 30ºC and absorbance was measured with a UV/VIS detector. All measurements presented here reflect the peak area at a specified elution time and wavelength (5.5min/565nm for violacein; 5.9min/565nm for deoxyviolacein; 5.1min/600nm for proviolacein; 5.4min/610nm for prodeoxyviolacein) (see **Figure 3-9** for sample chromatogram and absorbance spectra). Pure standards for our target compounds were commercially unavailable, and therefore absolute mass measurements were not possible; a mixed extract of violacein/deoxyviolacein could be purchased (Sigma-Aldrich), and we estimate that a peak area of 150au corresponds to approximately 10mg/L violacein.

TRAC

A slightly modified version of the TaqMan protocol described in Kong, *et al*[30] was used to identify each unique promoter. A list of probes and their sequences (labeled oligonucleotides provided by Integrated DNA Technologies) are available in **Table 3-2**, and a list of amplification primers are listed in **Table 3-5**. A universal probe mix (2μM each dye-strand, 2.4μM each quencher-strand) was prepared in water. Template for PCR was prepared by resuspending a 1mm-diameter yeast colony in 25μL of 20mM NaOH or by pelleting and resuspending a saturated yeast culture in 2.5 volumes of 20mM NaOH, then boiling for ten minutes, pelleting and recovering the supernatant. A 25μL TRAC reaction included: 2.5μL of 10x PCR buffer (100mM Tris-HCl, 500mM KCl, 15mM $MgCl_2$, pH 8.3 @ 25ºC), 0.5μL of 10mM dNTP mix, 1μL of each 10μM PCR primer, 0.75μL of probe mix, 2.5μL of template, 0.5μL of Taq DNA polymerase, and 16.25μL of water. PCRs were run as follows: initial denaturing at 94ºC for 5 min, 50 amplification cycles (94ºC for 10 sec, 50ºC for 30 sec, 68ºC for 1 min), and a final elongation at 68ºC for 10 min. 20μL of the reaction were diluted with 80μL of water and loaded onto a Costar 96-well flat bottom polystyrene assay plate and measured for fluorescence using a TECAN Safire2.

For a sufficiently large number of randomly sampled colonies, fluorescence measurements for each channel segregated into two distinct clusters corresponding to background (quenched) and positive hits (released) (**Figure 3-5**).

Regression model implementation details

Although we assume independence of each enzyme's contribution, we also posit that the relationship between enzyme expression and product titer is not necessarily monotonic. Therefore, one natural framework for building the model is through the use of categorical variables that represent the presence or absence of a particular promoter in front of each gene. Thus, using a log-linear model (the training data were skewed towards zero, and we found a log transform of the data improved performance), the product titer t as a function of the promoter-gene combinations is modeled as

$$t = \exp\left(\beta_{00} + \sum_{i\in\{1,\ldots,\#E\}} \sum_{j\in\{1,\ldots,\#P_i\}} \beta_{ij} x_{ij}\right)$$

where β_{ij} are the unknown coefficients of the model, and $x_{ij} = 1$ if the j-th promoter is driving the i-th gene and 0 otherwise. Because only one promoter can be in front of each

gene, the independent variables x_{ij} are constrained such that

$$\sum_{j \in \{1,\ldots,\#P_i\}} x_{ij} = 1$$

for the i-th gene. In the case of five genes and five promoters for each gene, $\#E = 5$ and $\#P_1 = \#P_2 = \#P_3 = \#P_4 = \#P_5 = 5$.

For N experimental measurements, we define the vector of response variable (titer) measurements as

$$T = \begin{bmatrix} t^1 \\ \vdots \\ t^N \end{bmatrix}$$

where the superscript notation t^k denotes the measurement from the k-th experiment. Similarly, we define the matrix of promoter combinations as

$$X = \begin{bmatrix} x_{11}{}^1 & \cdots & x_{\#E\#P_{\#E}}{}^1 \\ \vdots & \ddots & \vdots \\ x_{11}{}^N & \cdots & x_{\#E\#P_{\#E}}{}^N \end{bmatrix}$$

where each row represents the genotype of the k-th sample. The vector of unknown coefficients is

$$B = \begin{bmatrix} \beta_{11} \\ \vdots \\ \beta_{\#E\#P_{\#E}} \end{bmatrix}$$

Thus, the model can be succinctly represented as $\log(T) = \beta_{00} + XB$. Because the logarithm of zero is negative infinity, we set entries of T that are zero to 0.5, because this is the smallest amount that we can experimentally measure. To train this model, we obtained $N = 182$ measurements (*i.e.*, ninety-one clones in duplicate).

Identification of the unknown β_{ij} coefficients in the model is challenging because of the high-dimensional nature of the problem. We used the previously described Exterior Derivative Estimator (EDE) method[22] to identify the coefficients of the model because it can better protect against overfitting than traditional methods (for the violacein pathway, using ordinary least squares regression resulted in a model with almost no correlation in the test set: Pearson R-values of -0.01, 0.06, -0.02, and 0.01 for violacein, deoxyviolacein, proviolacein, and prodeoxyviolacein, respectively). EDE protects against overfitting by learning constraints that the data obeys, and then it uses these constraints to reduce the degrees of freedom in the regression. More specifically, the coefficients estimated by EDE are given by

$$\hat{B} = \arg\min_B \| \log(T) - XB - \beta_{00} \|^2 + \lambda \| PB \|^2$$

where $P = UU^T$, and U is a matrix whose columns are the m smallest principal components of the covariance matrix X^TX. λ and m are tuning parameters that are chosen in a data-driven manner using cross-validation.

In general, the rows of the matrix X form a manifold, and the projection matrix P enforces

that the regression coefficients lie close to the manifold formed by X. This methodology is motivated by differential geometry, which says that the exterior derivative of a function on an embedded submanifold lies in the cotangent space[38].

3.5 References

1. Paddon, C. J. *et al.* High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature* **496,** 528–532 (2013).
2. Ajikumar, P. K. *et al.* Isoprenoid pathway optimization for Taxol precursor overproduction in Escherichia coli. *Science* **330,** 70–74 (2010).
3. Steen, E. J. *et al.* Metabolic engineering of Saccharomyces cerevisiae for the production of n-butanol. *Microb Cell Fact* **7,** 36 (2008).
4. Steen, E. J. *et al.* Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* **463,** 559–562 (2010).
5. Yadav, V. G., De Mey, M., Lim, C. G., Ajikumar, P. K. & Stephanopoulos, G. The future of metabolic engineering and synthetic biology Towards a systematic practice. *Metab Eng* **14,** 233–241 (2012).
6. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6,** 343–345 (2009).
7. Gibson, D. G. *et al.* One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic Mycoplasma genitalium genome. *Proc Natl Acad Sci USA* **105,** 20404–20409 (2008).
8. Shao, Z., Zhao, H. & Zhao, H. DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res* **37,** e16–e16 (2009).
9. Li, M. Z. & Elledge, S. J. Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nat Methods* **4,** 251–256 (2007).
10. Quan, J. & Tian, J. Circular polymerase extension cloning of complex gene libraries and pathways. *PLoS ONE* **4,** e6441 (2009).
11. Zhang, Y., Werling, U. & Edelmann, W. SLiCE: a novel bacterial cell extract-based DNA cloning method. *Nucleic Acids Res* – (2012). doi:10.1093/nar/gkr1288
12. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE* **3,** e3647 (2008).
13. Pitera, D. J., Paddon, C. J., Newman, J. D. & Keasling, J. D. Balancing a heterologous mevalonate pathway for improved isoprenoid production in Escherichia coli. *Metab Eng* **9,** 193–207 (2007).
14. Zhu, M. M., Skraly, F. A. & Cameron, D. C. Accumulation of Methylglyoxal in Anaerobically Grown Escherichia coli and Its Detoxification by Expression of the Pseudomonas putida Glyoxalase I Gene. *Metab Eng* **3,** 218–225 (2001).
15. Kristensen, C. *et al.* Metabolic engineering of dhurrin in transgenic Arabidopsis plants with marginal inadvertent effects on the metabolome and transcriptome. *Proc Natl Acad Sci USA* **102,** 1779 (2005).
16. Glick, B. R. Metabolic load and heterologous gene expression. *Biotechnol. Adv.* **13,** 247–261 (1995).

17. Neubauer, P., Lin, H. Y. & Mathiszik, B. Metabolic load of recombinant protein production: inhibition of cellular capacities for glucose uptake and respiration after induction of a heterologous gene in Escherichia coli. *Biotechnol Bioeng* **83,** 53–64 (2003).
18. Pfleger, B. F., Pitera, D. J., Smolke, C. D. & Keasling, J. D. Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat Biotechnol* **24,** 1027–1032 (2006).
19. Lu, C. & Jeffries, T. W. Shuffling of promoters for multiple genes to optimize xylose fermentation in an engineered Saccharomyces cerevisiae strain. *Applied and Environmental Microbiology* **73,** 6072–6077 (2007).
20. Du, J., Yuan, Y., Si, T., Lian, J. & Zhao, H. Customized optimization of metabolic pathways by combinatorial transcriptional engineering. *Nucleic Acids Res* (2012). doi:10.1093/nar/gks549
21. Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460,** 894–898 (2009).
22. Aswani, A., Bickel, P. & Tomlin, C. Regression on manifolds: Estimation of the exterior derivative. *Annals of Statistics* **39,** 48–81 (2011).
23. Alper, H. S., Fischer, C., Nevoigt, E. & Stephanopoulos, G. Tuning genetic control through promoter engineering. *Proc Natl Acad Sci USA* **102,** 12678–12683 (2005).
24. Nevoigt, E. E. *et al.* Engineering of promoter replacement cassettes for fine-tuning of gene expression in Saccharomyces cerevisiae. *Applied and Environmental Microbiology* **72,** 5266–5273 (2006).
25. Newman, J. R. S. *et al.* Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441,** 840–846 (2006).
26. Qi, L. L., Haurwitz, R. E. R., Shao, W. W., Doudna, J. A. J. & Arkin, A. P. A. RNA processing enables predictable programming of gene expression. *Nat Biotechnol* **30,** 1002–1006 (2012).
27. Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* **27,** 946–950 (2009).
28. Mutalik, V. K. *et al.* Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat Methods* **10,** 354–360 (2013).
29. Ramon, A. & Smith, H. O. Single-step linker-based combinatorial assembly of promoter and gene cassettes for pathway engineering. *Biotechnol Lett* **33,** 549–555 (2011).
30. Kong, D. *et al.* Duplex probes: a new approach for the detection of specific nucleic acids in homogenous assays. *Analytica Chimica Acta* **491,** 135–143 (2003).
31. Hoshino, T. Violacein and related tryptophan metabolites produced by Chromobacterium violaceum: biosynthetic mechanism and pathway for construction of violacein core. *Appl Microbiol Biotechnol* (2011). doi:10.1007/s00253-011-3468-z
32. Balibar, C. J. & Walsh, C. T. In vitro biosynthesis of violacein from L-tryptophan by the enzymes VioA-E from Chromobacterium violaceum. *Biochemistry* **45,** 15444–15457 (2006).
33. Fox, R. J. *et al.* Improving catalytic function by ProSAR-driven enzyme evolution. *Nat Biotechnol* **25,** 338–344 (2007).

34. Li, Y. *et al.* A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat Biotechnol* **25,** 1051–1056 (2007).
35. Anderson, J. C. *et al.* BglBricks: A flexible standard for biological part assembly. *Journal of biological engineering* **4,** 1 (2010).
36. Shetty, R. P., Endy, D. & Knight, T. F. Engineering BioBrick vectors from BioBrick parts. *Journal of biological engineering* **2,** 5 (2008).
37. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat Genet* **14,** 450–456 (1996).
38. Lee, J. *Introduction to Smooth Manifolds*. (Springer, 2003).

Chapter 4. Employing a combinatorial expression approach to characterize xylose utilization in *Saccharomyces cerevisiae*†

4.1 Introduction

Biological synthesis of liquid transportation fuels provides an attractive route for sustainably meeting the growing demands of an increasingly expanding global economy[1]. Bulk production of commodity biofuels requires engineered microbes to efficiently convert inexpensive biomass-derived substrates. The first generation of biofuel production has relied on fermentation of the glucose, often sourced from sugar cane and corn; however, as the demand for biofuels grows, it will become increasingly critical to utilize lignocellulosic feedstocks[2,3]. Lignocellulose is primarily comprised of lignin, a complex aromatic polymer, and two sugar biopolymers: cellulose, a polymer of glucose molecules, and hemicellulose, a heterogeneous polymer representing approximately a third of biomass by dry weight (33% for corn stover, 27% for *Miscanthus*, and 32% for hardwoods) of which the pentose xylose is the major constituent[4]. Compared to cellulose, hemicellulose is more readily hydrolyzed to its component monosaccharides[2], but its use is limited by the metabolism of most microbes, which have not evolved to rapidly utilize xylose (or at all, in some cases). Thus, rapid xylose utilization is an important engineering target for efficient and commercially viable microbial conversion of diverse feedstocks into various compounds, such as sustainable biofuels.

S. cerevisiae has historically been the consensus choice as a production host for biofuels for a number of reasons: it has a high tolerance to toxic intermediates produced during most lignocellulose pretreatments[5]; it can naturally ferment an isomer of xylose, xylulose[6]; it exhibits high tolerance to low pH and the fermentation product ethanol[7]; and it has well-developed large-scale fermentation protocols. While there are many yeast species that can natively utilize xylose[8], *S. cerevisiae* lacks this capability. To confer the ability to convert xylose into fermentable xylulose in *S. cerevisiae*, two enzymes, xylose reductase (XR) and xylitol dehydrogenase (XDH) must be heterologously expressed (**Figure 4-1A**). These are often derived from the natural xylose fermenting yeast, *Scheffersomyces stipitis*[9]. Although redox balanced, the cofactor usage of these two enzymes is asymmetric, with XR preferring NADPH and XDH exclusively using NAD^+. Mutant XR and XDH enzymes with altered cofactor preference have been developed in an attempt to resolve this asymmetry, but the effects of these mutations have been confounded by the simultaneous alteration of cofactor usage and enzyme kinetics[10-15]. A third enzyme, xylulokinase (XK), is usually overexpressed to convert xylulose into the pentose phosphate pathway (PPP) intermediate xylulose-5-phosphate[16-21], which is further metabolized by native PPP enzymes into substrates for glycolysis.

Despite vigorous engineering efforts over the past few decades, xylose catabolism in *S. cerevisiae* has not reached industrially viable efficiencies, in part because the ideal combi-

†Reproduced with permission from Latimer, L.N., Lee, M.E., Medina-Cleghorn, D., Kohnz, R.A., Nomura, D.K., & Dueber, J.E. "Employing a combinatorial expression approach to characterize xylose utilization in *Saccharomyces cerevisiae*." *Metab Eng* 25, 20–29 (2014).

nation of enzyme expressions in any particular strain is unclear. Most previous studies have used arbitrary overexpression of a subset of enzymes chosen from the heterologous xylose utilization enzymes (XR, XDH, and XK) and pentose phosphate enzymes ribulose-5-phosphate epimerase (RPE), ribulose-5-phosphate isomerase (RKI), transketolase (TKL), and transaldolase (TAL)[22]. Improvements were often achieved with overexpression of many of these enzymes individually as well as in combination, suggesting that not only are high amounts of the heterologous xylose utilization enzymes required, but also increased PPP enzyme activity[23-28]. This might be expected since the PPP, especially in *S. cerevisiae*, has not evolved to handle the elevated flux required when xylose is used as the sole carbon source[29,30]. In addition, conflicting findings for the optimal expression of some of these enzymes, such as xylulokinase, has resulted in confusion in understanding pathway design principles. In some studies, the highest overexpression of XK produced the fastest growing strains[31]; however, in other studies, intermediate expression levels were determined to be ideal, suggesting toxicity, perhaps due to ATP depletion[32,33]. It is quite possible that intermediate expression levels of other pathway enzymes would be optimal as well. To this end, combinatorial expression engineering has been applied in two studies to various enzymes in the xylose utilization pathway. Lu and Jeffries combinatorially sampled two expression levels of three *S. cerevisiae* genes—the PPP genes *TKL1* and *TAL1* and the glycolytic enzyme *PYK1*—and observed the best ethanol titers when *TKL1* and *PYK1* were expressed with the stronger promoter and *TAL1* with the weaker promoter[28]. Du and colleagues placed the upstream genes—XR, XDH, and XK—under control of three promoter mutant libraries to identify fast-growing genotypes in both a laboratory and industrial strain. Interestingly, in this study, the optimal ratio of enzyme activities changed in the two strain backgrounds[34]. This finding is highly representative of an issue that this field faces: different strain backgrounds, growth media, and conditions used all impact the metabolic context of this pathway, which complicates the integration of knowledge from various studies for strain engineering.

Here we present a study where we simultaneously titrated expression of eight genes involved in xylose utilization. In accordance with much of the previous work on this pathway[20,21,35-37], we chose to include the three heterologous enzymes XR, XDH, and XK from *S. stipitis* in our library. Because the PPP does not typically need to support high flux in *S. cerevisiae*[30], we included additional copies of these enzymes (RPE, RKL, TKL, TAL). Like the xylose catabolic enzymes, we elected to express *S. stipitis* homologs of these genes under the assumption that they have evolved to support high flux in the natively xylose-consuming yeast. Finally, although we did not expect glycolytic enzymes to be metabolically limiting, we also included *S. stipitis* pyruvate kinase (PYK) based on previous characterization of the *S. cerevisiae* homolog, *PYK1*, which was shown to determine the glycolytic rate when driven by a weak promoter[38] and improved xylose fermentations when expressed highly during combinatorial expression experiments of *TKL1*, *TAL1* and *PYK1*[28].

Using growth on xylose as a selection, we probed the role of both intrinsic variables – changes made directly to the starting strain such as number and variants of pathway genes expressed – and extrinsic variables – changes to the external selection pressures

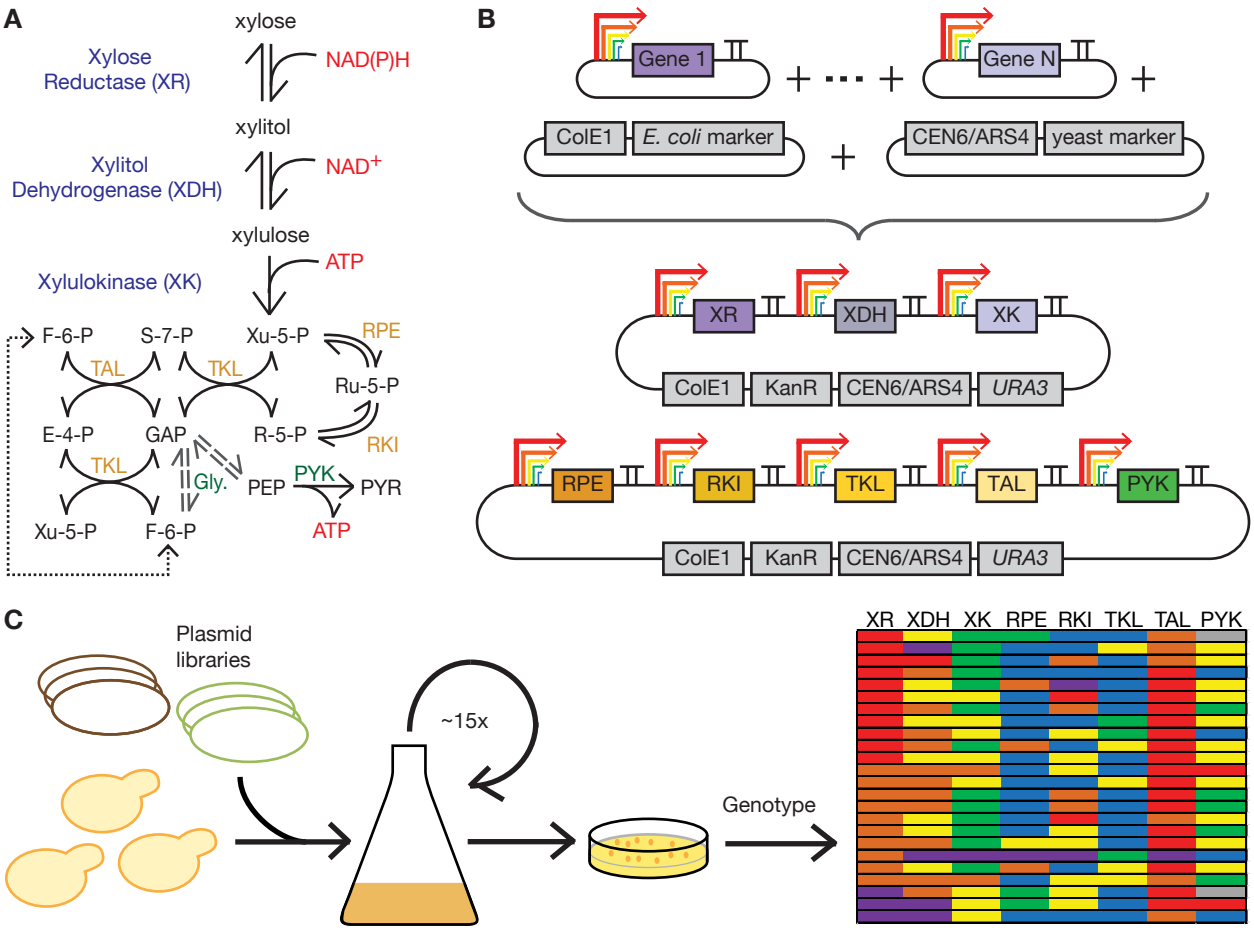


Figure 4-1. Experimental design for optimization of xylose metabolism using a combinatorial promoter library. (A) Xylose catabolism to pyruvate. The fungal catabolic pathway (blue) feeds into glycolysis (green) via the non-oxidative pentose phosphate pathway (gold). Nomenclature is in accordance with the Saccharomyces Genome Database (http://www.yeastgenome.org). (B) Plasmid-based promoter library assembly scheme. A mixture of backbone and cassette plasmids with various promoters and *tADH1* are assembled in a one-pot golden gate reaction. (C) Cartoon depicting the library enrichment protocol. Yeast cells are transformed with the promoter library plasmids and grown on xylose as the sole carbon source in the desired conditions. After iterative dilutions, individual colonies are isolated and genotyped. Genotypes for colonies after 29 days of aerobic growth are shown (sorted by XR promoter), where red represents *pTDH3*, orange represents *pTEF1*, yellow represents *pRPL18B*, green represents *pRNR2*, blue represents *pREV1*, grey indicates no detected promoter and purple indicates a mixed signal in the TRAC sequencing reaction.

applied to the strain such as oxygenation – on optimal gene expression. We included such a large number of genes to investigate the possibility of local optima when only a fraction of the pathway enzymes are optimized. In this way, we were able to identify important factors in xylose utilization.

4.2 Results

4.2.1 Construction and enrichment of combinatorial pathway expression libraries

To determine the optimal expression profile of enzymes involved in xylose utilization in the BY4741 laboratory *S. cerevisiae* strain, we chose to employ a combinatorial approach

to simultaneously modulate expression of each pathway enzyme using a set of five previously characterized constitutive promoters[39]. These promoters sample evenly-spaced transcriptional strengths over approximately three orders of magnitude of expression space and include *pTDH3*, the strongest constitutive promoter in *S. cerevisiae*[40]. In contrast to a promoter mutagenesis-based approach, our library is not limited by the availability of multiple high-strength promoters, which allows each gene to sample the highest possible expression. Additionally, these promoters enable rapid genotyping of library members using the previously described TRAC method[39].

For the eight *S. stipitis* genes, libraries of all possible promoter-gene combinations were constructed using Golden Gate assembly[41] (**Figure 4-1B**). To allow for facile inclusion or exclusion of the PPP in our enrichments, the pathway libraries were constructed on two CEN6/ARS4 plasmids: 1) the xylose utilization genes XR, XDH, and XK and 2) the PPP genes and PYK. Expression of each gene was regulated by any of the five promoters to yield a $5^8 = 390{,}625$ member library. The vector backbones used separated the yeast marker and origin from the bacterial marker and origin to prevent bleed-through of unassem-

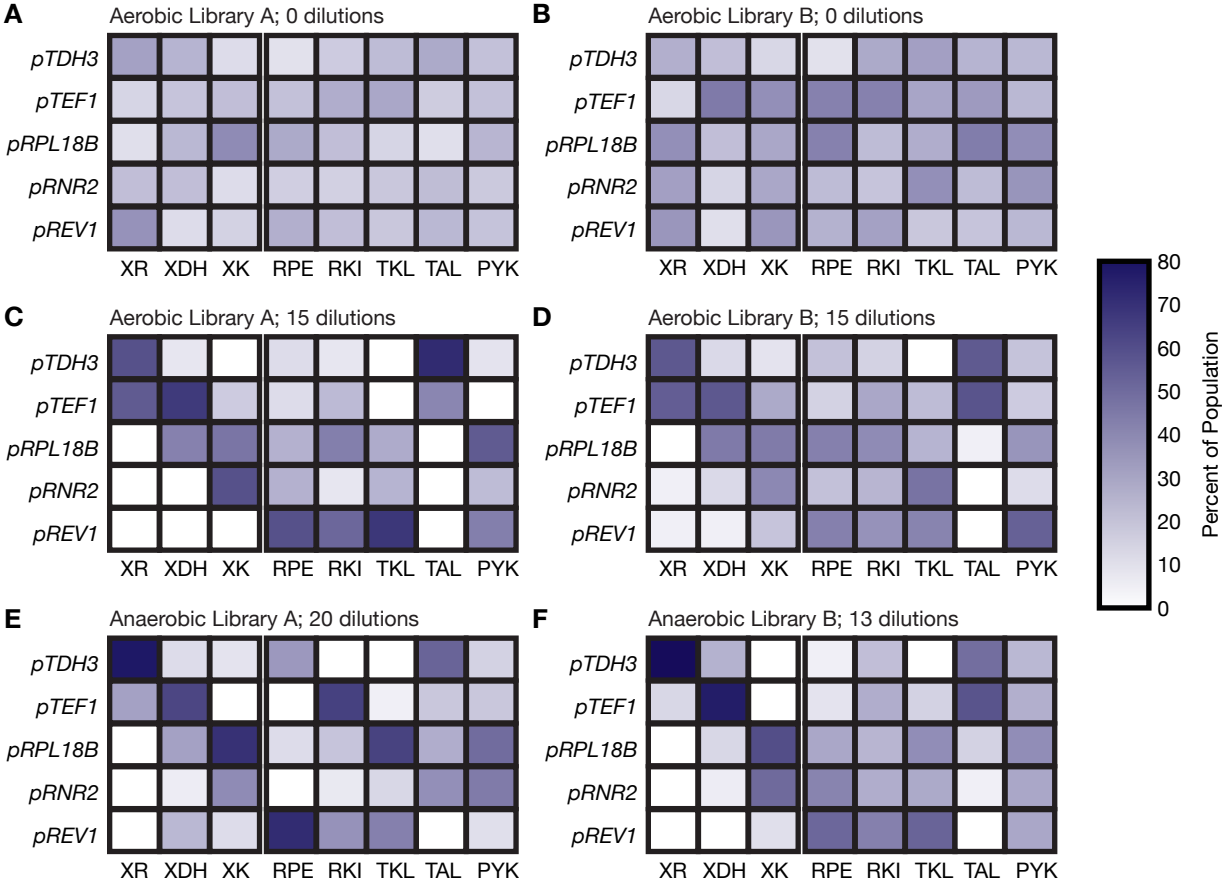


Figure 4-2. Enrichment profiles. Duplicate experiments show consistent enrichment profiles. (A,B) Initial distribution of promoters before enrichment. (C,D) Distribution after 15 rounds of enrichment under aerobic conditions. (E,F) Distribution after 20 or 13 rounds of enrichment under anaerobic conditions. Heatmaps were generated from genotyping 48 colonies from biological duplicate library enrichments on 2% synthetic xylose dropout media under either aerobic (baffled flask, 200 rpm) or anaerobic conditions (serum vial, 100 rpm). The heatmap colors correspond to the percentage of colonies that had a given promoter in front of the corresponding gene.

bled bacterial vector into the subsequent yeast transformation. *E. coli* transformations of the library assembly reactions, which were pooled and purified to isolate plasmid libraries, yielded approximately 40,000 colonies, providing over 10 fold coverage of the $5^3 = 125$ and $5^5 = 3{,}125$ member pathway libraries.

Assembled libraries were transformed into yeast, recovered for one hour in YPD, and then transferred to selective dropout media with xylose as the sole carbon source. Dilution platings of the transformations routinely showed at least 10^6 transformants, corresponding to multiple-fold library coverage, such that nearly every promoter-gene combination should be present at the start of enrichment. We verified the diversity and assembly of the library by genotyping 48 colonies from each library transformation using a Taqman-based method previously developed in our lab termed TRAC (Taqman Rapid Analysis of Combinatorial assemblies)[39]. We observed little promoter bias and 80% of colonies had a clear, single promoter driving a given gene.

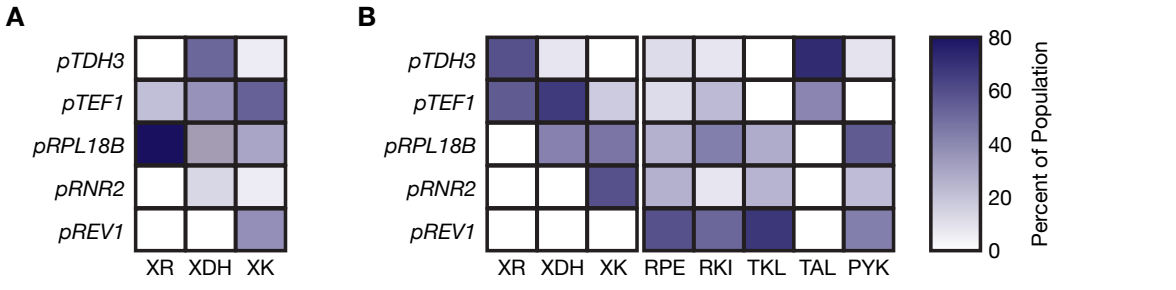
Following transformation into yeast, we employed a selection for growth on xylose to enrich for strains with superior xylose utilization, as shown in **Figure 4-1C**. Cultures were enriched by repeated growth to late log phase (at least 10^9 cells), but not saturation, with subsequent dilution into fresh media with 10^7 cells. Periodically, culture aliquots were plated to isolate individual strains, and the promoter driving expression of each heterologous gene was identified by TRAC. Serial enrichments were performed until the population converged to a consensus combination of promoters, which we term "enrichment profiles." These enrichment profiles allowed us to track global expression trends within the library culture. Holistic understanding of promoter enrichment for each gene is useful because, based on the promoter driving a given protein, we can predict relative expression of that protein between strains where different promoters are present. Although absolute abundances will differ from protein to protein, the relative rank-ordering of expression from this characterized promoter set remains consistent irrespective of the coding sequence[39]. To verify the reproducibility of our enrichments, libraries containing all eight enzymes were enriched in duplicate from independent transformations, and the enrichment profiles were similar between replicates (**Figure 4-2**).

4.2.2 Aerobic enrichments and strain characterizations

Combinatorial expression optimization of pathway enzymes has been shown to be dependent on the strain background[34]. A corollary hypothesis is that optimal enzyme expression is dependent on which genes are expressed and purposefully varied within a strain. Thus, we sought to investigate the dependence of enrichment profiles on which heterologous enzymes are included during the optimization. Our initial experiments were performed under aerobic conditions to better compare findings to previous combinatorial expression engineering efforts for xylose utilization[28,34]. Similar to these previous experiments, we tested a library of only the three xylose utilization enzymes (i.e., XR, XDH, and XK), which we refer to as the "partial pathway," as well as a second library, expanded to include additional *S. stipitis* copies of the pentose pathway enzymes (i.e., RKI, RPE, TKL, and TAL) and the glycolytic enzyme PYK, which we term the "full pathway."

Expression optimization of the partial pathway results in a local optima

We hypothesized that including or omitting the downstream pathway enzymes in our library would alter the optimal expression levels of the upstream xylose utilization genes. Thus, we enriched both the partial and full versions of the pathway under aerobic conditions and compared the resulting enrichment profiles. Indeed, we found the enrichment profile of the promoters for XR, XDH and XK to be considerably different between the full and partial libraries (**Figure 4-3A and B**). Strains expressing the partial pathway strong-



C

Strain	Strain description	XR	XDH	XK	RPE	RKI	TKL	TAL	PYK
pTDH3 full	full pathway all *pTDH3* driven	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*
LL110A	full pathway aerobically enriched	*pTDH3*	*pRPL18B*	*pRPL18B*	*pREV1*	*pRPL18B*	*pREV1*	*pTDH3*	*pRPL18B*
pTDH3 ØPPP	partial pathway all *pTDH3* driven	*pTDH3*	*pTDH3*	*pTDH3*					
LL100A	partial pathway aerobically enriched	*pRPL18B*	*pTDH3*	*pTEF1*					

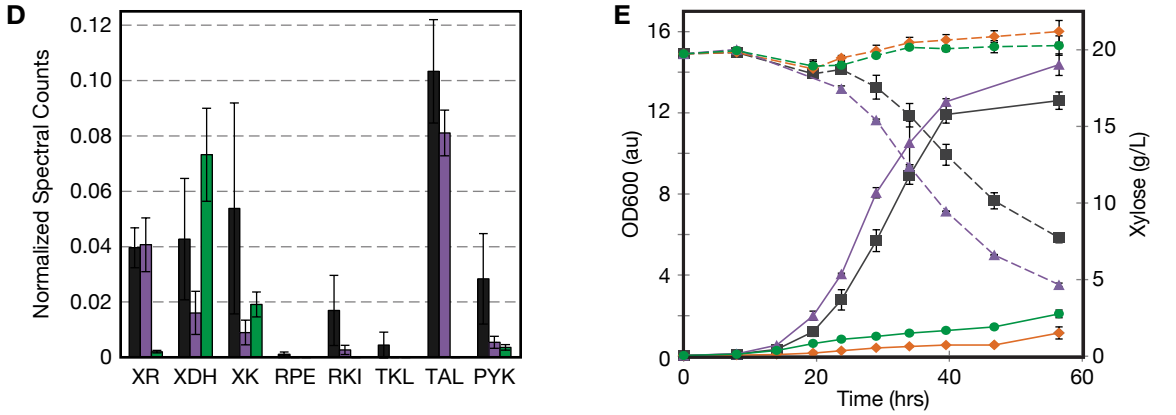


Figure 4-3. Inclusion of the PPP enzymes produces superior growth and xylose consumption rates with optimal expression being not all genes driven by *pTDH3*. (A) Heatmap representation of the enrichment profile generated from screening 32 colonies isolated from an aerobically enriched promoter library regulating the first three enzymes in the xylose metabolic pathway. Heatmap colors represent the percentage of colonies with a given promoter regulating the corresponding gene. (B) Enrichment profile generated from screening 48 colonies isolated from an aerobically enriched promoter library regulating all eight enzymes. (C) Genotype and description of reference strains and enriched genotypes identified from the library enrichments. (D) Shotgun proteomic data indicating relative protein abundance in strains described in (C) when grown aerobically to mid-log phase on xylose. Protein spectral counts were normalized to total endogenous spectral counts. There is no statistical significance between total endogenous protein expression between samples. (E) Aerobic growth curve for yeast strains indicated in (C): OD600 (——); Xylose (– – –). Error bars represent SD of biological triplicates.

ly enrich for the medium-strength *pRPL18B* promoter driving expression of XR, higher (*pTDH3* or *pTEF1*) expression of XDH, and variable XK expression (**Figure 4-3A**). Such expression enrichment is consistent with previous optimizations in microaerobic conditions[23,27,34]. However, when the full pathway is expressed, XR is enriched exclusively for the stronger *pTEF1* or *pTDH3* promoters, while XDH and XK are driven by somewhat lower strength promoters (moderate to high). These results for the full pathway are more consistent with what would be expected from the *in vitro* catalytic efficiencies of the pathway, where XR is predicted to be the limiting enzyme (**Table 4-1**). These drastic differences in expression profiles between partial and full pathways are intriguing as they suggest that the optimal expression profile is highly dependent on the other enzymes that are also varied.

To experimentally validate the predicted lower XR expression in the partial pathway compared to the full pathway, we chose to regenerate strains with two highly enriched genotypes, LL100A and LL110A, by specifically assembling those plasmids and transforming them into the original strain background (**Figure 4-3C**). These two strains were grown on xylose and analyzed by shotgun proteomics to determine relative protein abundances (**Figure 4-3D**). This technique has the advantage of directly measuring relative protein abundance rather than mRNA levels, an attractive feature since measuring mRNA amounts would miss any potential translational effects on final protein concentration. This is particularly useful as one might suspect that as the number of overexpressed enzymes increases, there is a possibility of saturating shared cellular machinery. Comparison of XR and XDH expression between LL110A and LL100A shows the expected change in XR and XDH peptide abundance, indicating a major reduction in XR expression and increase in XDH expression in LL100A (**Figure 4-3D**). From these findings, we conclude that optimization of the xylose utilization enzymes alone results in a local utilization optimum of low XR expression, whereas under conditions where portions of the PPP are also overexpressed, far more XR is required to reach a higher, more global optimum.

Table 4-1. Reported enzyme kinetics for purified xylose catabolism and PPP enzymes.

Enzyme	Organism	First Substrate	Second Substrate	K_m (mM)	kcat (s^{-1})	kcat/K_m ($M^{-1}s^{-1}$)	Reference
XR	*S. stipitis*	Xylose	NADPH	32.4	10	3.40E2	Chen et al., 2012
wtXDH	*S. stipitis*	Xylitol	NAD^+	21.7	1050	4.84E4	Watanabe et al., 2005
mutXDH	*S. stipitis*	Xylitol	$NADP^+$	72.6	3840	5.29E4	Watanabe et al., 2005
XK	*S. stipitis*	Xylulose	ATP	0.27	24.8*	4.77E4*	Chen et al., 2012
RPE	*S. cerevisiae*	Ru-5-P	-	1.5	3340*	2.22E6*	Bär et al., 1996
RKI	*S. cerevisiae*	R-5-P	-	1.6	1140*	7.15E5*	Reuter et al., 1998
TKL	*S. stipitis*	Xu-5-P	R-5-P	0.72	85	1.18E5	Chen et al., 2012
TAL	*S. stipitis*	F-6-P	E-4-P	0.25	7.1	2.84E4	Chen et al., 2012
PYK	*S. cerevisiae*	PEP	ADP	2.76	188	6.81E4	Collins et al., 1995

K_m values correspond to affinity for the first substrate.
*These values are estimated based on reported enzyme activities and molecular masses.

Transaldolase activity is limiting for aerobic xylose consumption

By simultaneously titrating multiple genes' expression, we hoped to identify those genes whose activities were limiting growth on xylose. When we enriched the full pathway library, we observed weaker promoters driving most of the PPP genes, predominantly the lowest strength *pREV1*, suggesting that either these enzymatic activities are not limiting growth on xylose or their overexpression is detrimental to the cell. Only TAL enriched for the strongest promoter, *pTDH3*, suggesting that its activity is limiting. Other rational and inverse metabolic engineering efforts in various yeast backgrounds along with *in vitro* enzyme kinetics have also identified transaldolase activity to be significantly limiting for aerobic and microaerobic xylose utilization (**Table 4-1**)[24,42,43]. Considering these studies, we were able to corroborate TAL's role as a limiting component for xylose utilization in the presence of oxygen.

Aerobic expression optimization yields small improvements in xylose utilization

Since most studies conventionally use high-strength promoters in a mostly arbitrary manner, we were curious how the enriched strains compared to a naïve overexpression strain in their ability to utilize xylose. Again, we used strains LL110A and LL100A, as well as two strains that simply use our strongest promoter to drive all genes ("*pTDH3* full" and "*pTDH3* ØPPP") (**Figure 4-3C**). These latter two strains mimic the arbitrarily high expression of each gene that would be naively done as opposed to balanced expression[16,19,44,45]. As expected, the strains expressing only XR, XDH, and XK grow poorly on xylose compared to the strains expressing extra copies of the PPP[19,24,27,43]. Strain LL110A grows at a maximum growth rate of 0.25 ± 0.01 hr^{-1}, making it the fastest aerobically growing strain in 2% xylose reported in literature to date (**Figure 4-3D**)[46,47]. Despite this, both LL110A and LL100A show only modest improvements in growth and xylose consumption compared to their respective all *pTDH3*-driven reference strain (**Figure 4-3E**). This is surprising as we expected a flux imbalance or protein burden in the *pTDH3* full pathway strain to more drastically reduce growth and consumption of xylose. By comparing proteomic data between LL110A and *pTDH3* full pathway, we observe no change in expression of genes regulated by *pTDH3* in both strains (e.g. XR and TAL), but for all other genes expressed with lower strength promoters we observe the expected decreases in expression in the enriched strain (**Figure 4-3D**). Having verified reductions in protein expression, we theorize that aerobic xylose consumption is primarily determined by XR and TAL expression.

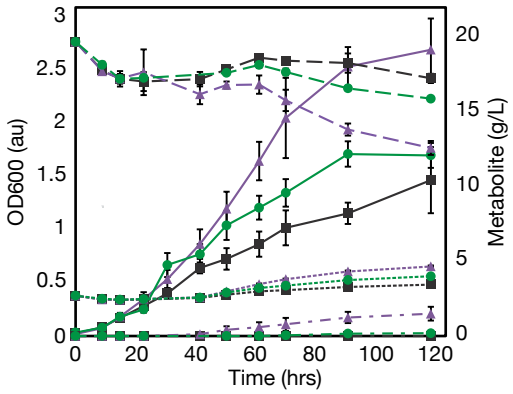
4.2.3 Anaerobic enrichments and strain characterizations

A major advantage of the described combinatorial expression strategy is that we can readily select under different conditions to characterize how enrichment profiles change depending on the growth environment. Accordingly, we grew our libraries anaerobically because, given the enormous fermentation volumes that are required for large-scale biofuel production, the process would need to be fully anaerobic.

A

Strain	*pTDH3* full	LL110B	LL111A
Description	all *pTDH3* driven	aerobic enriched	anaerobic enriched
XR	*pTDH3*	*pTEF1*	*pTDH3*
XDH	*pTDH3*	*pTEF1*	*pRPL18B*
XK	*pTDH3*	*pTEF1*	*pRPL18B*
RPE	*pTDH3*	*pTDH3*	*pREV1*
RKI	*pTDH3*	*pTEF1*	*pREV1*
TKL	*pTDH3*	*pRNR2*	*pRPL18B*
TAL	*pTDH3*	*pTDH3*	*pTDH3*
PYK	*pTDH3*	*pRPL18B*	*pRNR2*

B



C

Strain	*pTDH3* ØPPP	LL100A	LL101A
Description	all *pTDH3* driven	aerobic enriched	anaerobic enriched
XR	*pTDH3*	*pRPL18B*	*pTDH3*
XDH	*pTDH3*	*pTDH3*	*pTEF1*
XK	*pTDH3*	*pTEF1*	*pRPL18B*
RPE			
RKI			
TKL			
TAL			
PYK			

D

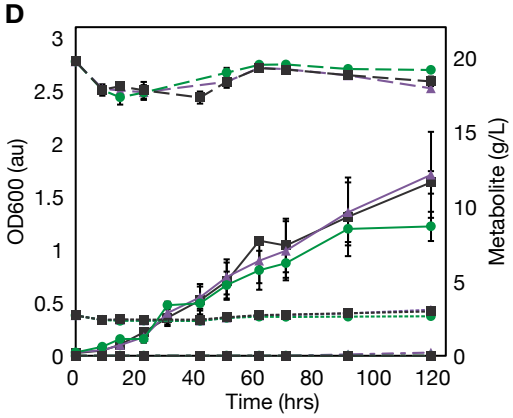


Figure 4-4. Selection conditions are important for strain optimization. (A) Genotype and description of partial pathway reference strains and enriched strains identified from aerobic and anaerobic library enrichments. (B) Anaerobic fermentations in SX media supplemented with 0.01g/L ergosterol, 0.43g/L Tween 80 and 2.8 g/L ethanol for yeast strains indicated in (A): OD600 (——); Xylose (– – –); Ethanol (- - -); Xylitol (– - –).(C) Genotype and description of full pathway reference strains and enriched strains identified from aerobic and anaerobic library enrichments. (D) Anaerobic fermentations as described in (B) for yeast strains indicated in (C). Error bars indicate the SD of biological triplicates.

Inclusion of the PPP only improves anaerobic xylose utilization upon expression optimization

Unlike the aerobic libraries, the enrichment profiles for XR, XDH and XK for the anaerobically enriched partial (XR, XDH, and XK only) and full (XR, XDH, XK, RPE, RKI, TKL, TAL, and PYK) pathways were notably similar with XR enriching for the strongest promoter and the promoter rank ordering following XR ≥ XDH ≥ XK (**Figure 4-2**). Anaerobically enriched strains were regenerated and characterized in comparison to the *pTDH3* reference strains to determine the effects of expression optimization (**Figure 4-4**). Unlike in aerobic conditions, where only small differences were observed between enriched and reference strains, the full pathway LL111A significantly outperforms all other strains in biomass accumulation, xylose consumption, and ethanol production (**Figure 4-4B**). Both of the *pTDH3* reference strains and the enriched partial strain LL101A all perform comparably, with the *pTDH3* full pathway consuming slightly more xylose. Interestingly, inclusion of the downstream enzymes with naïve *pTDH3* expression results in minimal to no improvement in growth and xylose consumption compared to overexpressing only the partial pathway.

Expression optimization is sensitive to oxygenation conditions

Comparing the aerobic and anaerobic full pathways, the enrichment profiles were similar between conditions with many of the strains having comparable genotypes (**Figure 4-2 and 4-5A and B**). However, some of the individual strain genotypes varied between the aerobic and anaerobic libraries, such as strains LL111A with LL110B (**Figure 4-4A**). We examined the anaerobic performance of strains enriched both aerobically and anaerobically to ascertain the extent to which selection conditions result in strain specialization (**Figure 4-4**). Despite the similarity of the consensus enrichment profiles between the full aerobic

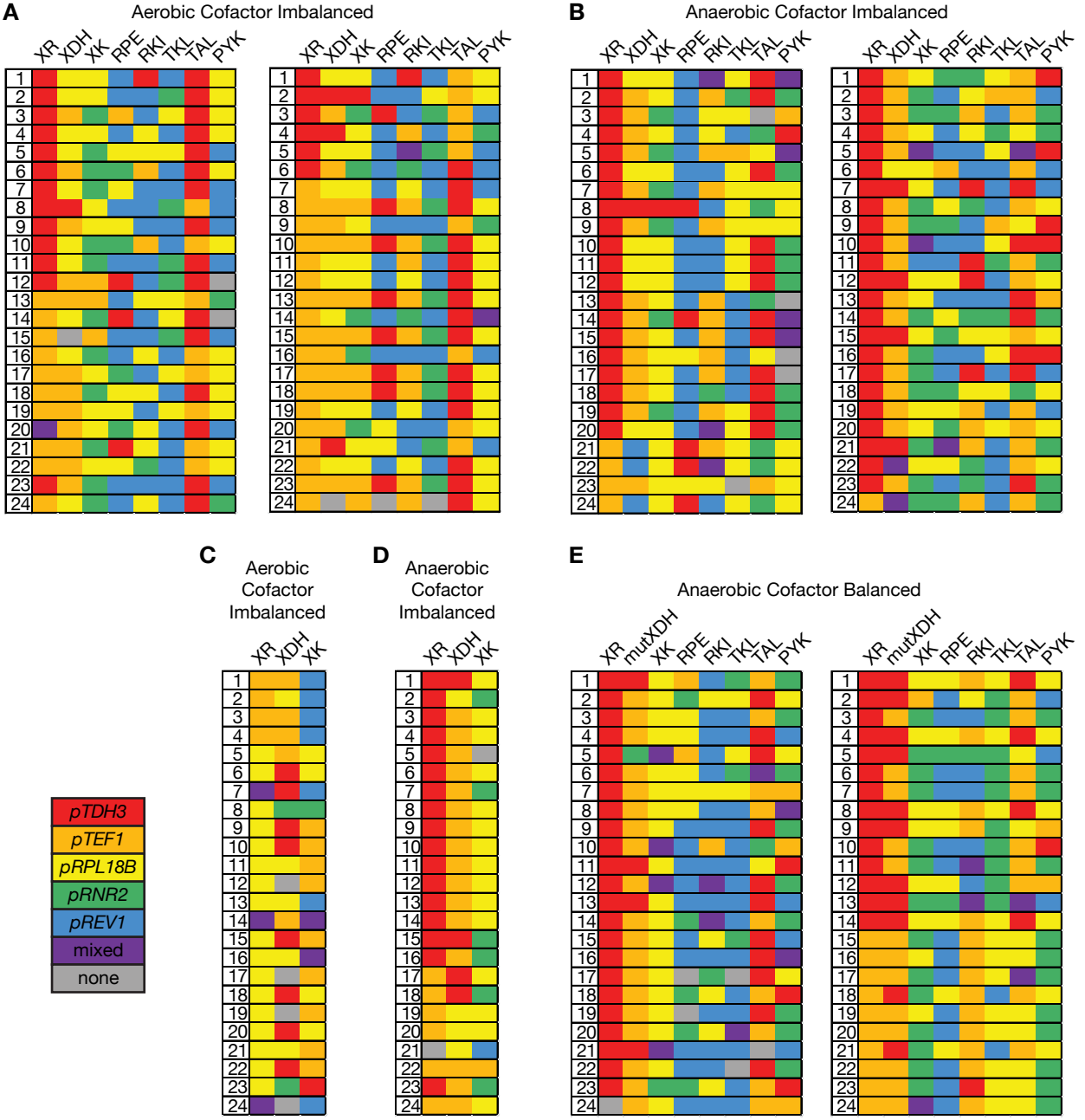


Figure 4-5. Library enrichment genotypes. Genotype sequencing results from 24 individual strains isolated from indicated enriched promoter libraries, where colored squares indicate the detected promoter(s) for each gene.

and anaerobic libraries, the two aerobically enriched strains LL110B and LL100A, underperform compared to their anaerobically enriched counterparts, LL111A and LL101A, under anaerobic growth conditions (**Figure 4-4B and D**). Therefore, it is important during strain optimization to not only use the correct genetic context (i.e., full vs. partial pathway), but also to enrich under the target conditions because high-performance strains may not translate well from one condition to another.

4.2.4 Expression optimization with a mutant, cofactor-balanced xylitol dehydrogenase

The net cofactor imbalance in the fungal xylose pathway caused by the NADPH preference of XR and the strict NAD^+ requirement for XDH has long been implicated as a limitation in achieving high ethanol yields during xylose fermentation using these enzymes[48,49]. Numerous attempts to address this problem have included engineering these enzymes to have switched cofactor usage[10-13]. Despite their promise, reports on the effectiveness of these mutant enzymes in improving xylose consumption have been mixed[14,25,50]. One limitation when assessing mutated pathways has been the inescapable complication of altered enzyme activity, which changes simultaneously with the cofactor affinities upon mutation of XR or XDH[15]. Consequently, fermentation differences observed between cofactor balanced and imbalanced pathways could be attributed to either switched cofactor preference or altered balance of enzymatic activities. With our ability to optimize expression of the entire xylose utilization pathway, we can better separate the role of cofactor balancing from altering catalytic activity levels to address this long-standing question.

Expression optimized anaerobic cofactor balancing improves growth, but not fermentation yields

We assembled alternative XR-XDH-XK promoter library plasmids where xylitol dehydrogenase was substituted with the ARSdR mutant XDH developed by Watanabe *et al*, which almost exclusively uses $NADP^+$ instead of NAD^+[13]. Mutant libraries were anaerobically enriched and genotyped as described previously. The expression profiles for mutant libraries show a somewhat higher mutant XDH expression levels than the native

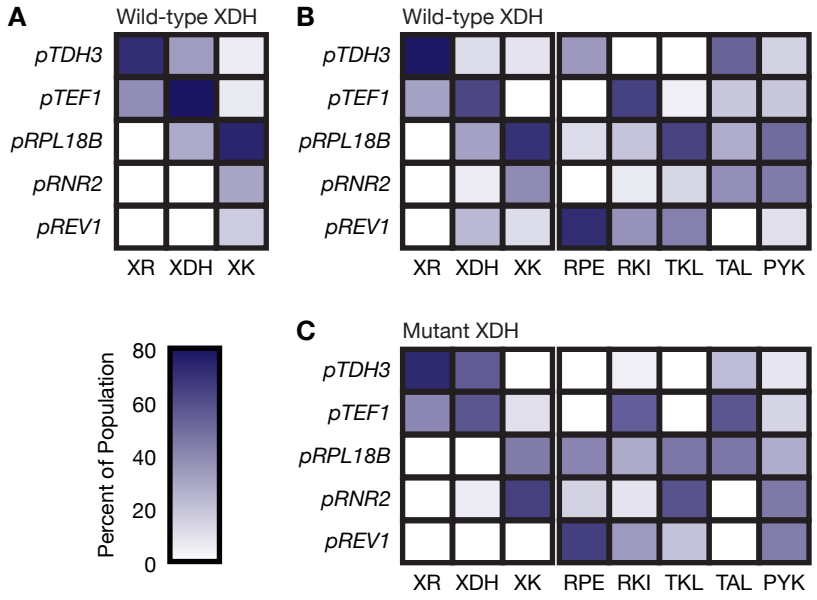


Figure 4-6. Anaerobic library enrichment profiles for wild-type and mutant XDH. Heatmaps were generated from genotyping colonies isolated from enriched libraries grown on 2% synthetic xylose dropout media (serum vial, 100 rpm). The heatmap colors correspond to the percentage of colonies having a given promoter in front of the corresponding gene.

XDH and slightly lower TAL and PYK expression levels (**Figure 4-6**). Highly-enriched genotypes from both mutant and wild-type XDH libraries along with *pTDH3* reference strains were chosen for comparison to assess the differences resulting from mutating XDH (**Figure 4-7A**).

Comparing the *pTDH3* mutXDH and *pTDH3* wtXDH strains verified previous work with the mutant XDH: the mutant pathway shows slightly improved growth (**Figure 4-7B**) and a 47% increase in xylose utilization and ethanol titer compared to the wild-type pathway (**Figure 4-7C**)[14,51]. These fermentation differences could be the result of changes in both catalytic activity and cofactor preference. Surprisingly, this increase in utilization upon mutating XDH was reduced between the expression optimized strains. Strains LL111A and LL211A consume nearly identical amounts of xylose, and concurrently produce nearly the same amounts of xylitol and ethanol (**Figure 4-7C**).

4.3 Discussion

Lignocellulosic fermentation promises a sustainable route for converting enormous plant biomass resources into biofuels. Xylose utilization is a critical step towards that goal, but it is difficult to understand how a multiple heterologous enzyme, high-flux pathway can be integrated into the metabolism of the host cell. Although there have been numerous

A

Strain	Strain description	XR	XDH	XK	RPE	RKI	TKL	TAL	PYK
pTDH3 wtXDH	full pathway all *pTDH3* driven	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*
pTDH3 mutXDH	full pathway all *pTDH3* driven	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*	*pTDH3*
L111A wtXDH	full pathway anaerobically enriched	*pTDH3*	*pRPL18B*	*pRPL18B*	*pREV1*	*pREV1*	*pRPL18B*	*pTDH3*	*pRNR2*
LL211A mutXDH	full pathway anaerobically enriched	*pTEF1*	*pTEF1*	*pRNR2*	*pREV1*	*pTEF1*	*pRPL18B*	*pRPL18B*	*pRNR2*

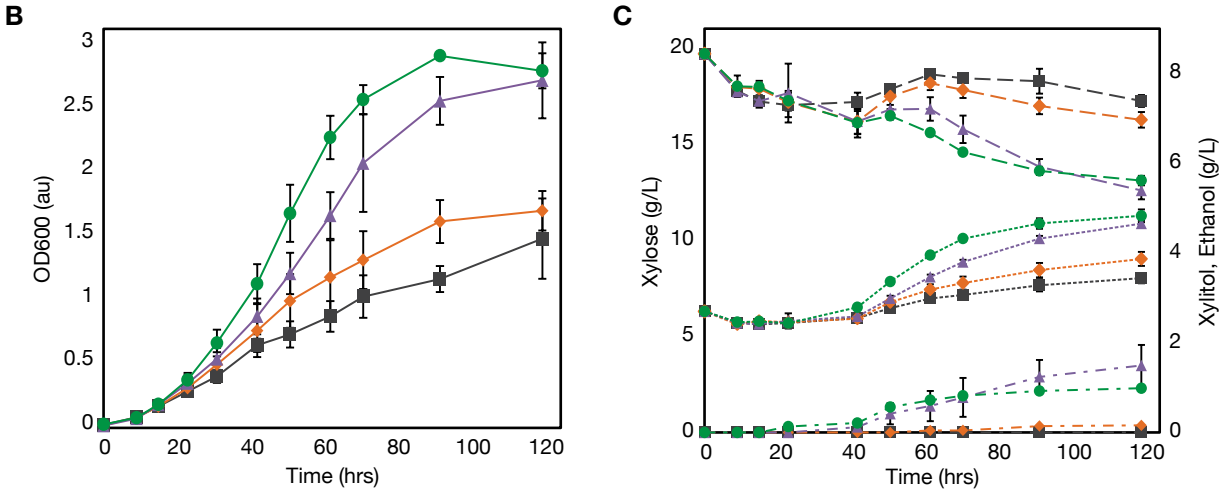


Figure 4-7. Differences in xylose fermentation due to mutating XDH for the full pathway are reduced with expression optimization. (A) Genotype and description of reference strains and enriched strains identified from anaerobic library. (B) Anaerobic fermentations in SX media supplemented with 0.01g/L ergosterol, 0.43g/L Tween 80 and 2.8 g/L ethanol for yeast strains indicated in (A): OD600 (——) (C) Extracellular metabolites for the fermentation shown in (B). Xylose (– – –); Ethanol (- - -); Xylitol (– - –). Error bars indicate the SD of biological triplicates.

studies inspecting various aspects of the xylose utilization pathway in *S. cerevisiae*, comparing the findings from these reports[22] is complicated by the different strain backgrounds and growth conditions used. Therefore, it is imperative that the pathway is examined as a whole in order to find the optimal balance of catalytic activities for a given host and growth condition. Accordingly, we employed a recently described combinatorial expression strategy[39] to simultaneously vary the expression levels of the full *S. stipitis* xylose utilization and pentose phosphate pathways, which allowed us to better understand the importance of both intrinsic (genotypic) and extrinsic (growth conditions) variables for xylose utilization in our engineered yeast. One major advantage of this particular approach is that every gene can access the maximal range of transcriptional strength irrespective of the pathway size. Other strategies that rely on promoter mutagenesis are limited by a lack of unique, high-strength promoters that can be used for each gene. While repeated use of promoter sequences in a single construct may result in gene loss through homologous recombination, we have not observed frequent instances of recombination in direct tests[39] or difficulties in either cloning or repeated cycles of growth of our all-*pTDH3* strains (up to eight genes driven by the same promoter sequence). With our experimental design, we easily changed whether the partial or full pathway was expressed and which version of XDH was included. We also tested both aerobic and anaerobic growth conditions. This unbiased, systematic survey allowed us to glean some insight into the aggregate of prior research, which at times is unclear.

Genotyping many colonies from a library allowed us to determine convergent enzyme expression patterns, for example, bottleneck enzymes were expected to enrich for the highest-strength promoters. By comparing the enrichment profiles in different genetic contexts, we revealed interplay between the upstream xylose utilization pathway and the downstream pentose phosphate pathway. Depending on the number of enzymes included in the library, we observed markedly different enrichment profiles and dramatically different growth capabilities, showing that omission of genes in a pathway library can select for local optima that fail to perform as well as the optima achieved with co-expression and optimization of more enzymes. For example, under aerobic conditions, our full pathway libraries enriched for strong promoters driving XR, but the partial pathway libraries almost exclusively enriched for the medium-strength *pRPL18B* promoter (**Figure 4-3A and B**); the full pathway strains also grow substantially faster (**Figure 4-3E**). Our hypothesis for the cause of these enrichment differences is that for our strain background, enzymatic activity downstream of the first three enzymes is limiting, likely in the form of transaldolase activity based on the strong enrichment for *pTDH3* expression of TAL in the full pathway, along with previous metabolic engineering efforts by others[24,52]. Under this hypothesis, excess expression of upstream pathway enzymes beyond endogenous TAL activity may cause depletion of NADPH and accumulation of NADH[53], and depletion of ATP[33]. Two possible solutions, both of which were enriched in our library selections, are either to reduce upstream expression or increase downstream capacity.

During their expression optimization of the first three enzymes, Zhao and coworkers observed that lower XR expression relative to XDH improved their laboratory strain's microaerobic fermentation while a high XR:XDH ratio was best for their industrial strain[34].

Our aerobic enrichment expression profiles of XR, XDH, and XK between the partial and full pathways (**Figure 4-3A and B**) mirror their observed trends between the lab and industrial strains, respectively, suggesting that their laboratory strain may be limited in downstream expression and that a contributing factor to the better performance of their industrial strains could be increased PPP activity. This hypothesis could also explain the observation by Sedlak and coworkers that overexpression of the PPP resulted in almost no improvement in their industrial strain, as these activities may not be limiting in this strain background[25].

The discrepancies in the literature regarding the possible benefit from overexpression of the PPP for xylose fermentation could also be caused by variability in strain construction[21,25,26,52]. As an example, the combinatorial optimization of *S. cerevisiae* TKL1, TAL1, and PYK1 by Lu and Jeffries found optimal xylose consumption when endogenous TAL1 was recombinantly expressed with a weaker promoter rather than a stronger one, while we observed higher TAL expression during our enrichments, particularly under aerobic conditions[28]. This discrepancy could be the result of expressing different versions of TAL – *S. stipitis* in our study and *S. cerevisiae* in the Lu study. Indeed, toxicity associated with *S. cerevisiae* TAL1 overexpression has been reported[24]. As another example within our study, benefits of heterologous expression of downstream genes under anaerobic conditions are only evident upon expression optimization. In fact, strains overexpressing the PPP at arbitrary high levels show no improved performance over the partial pathway (**Figure 4-4**). Similar effects may be present in a subset of these other studies. Particularly under anaerobic conditions, there is likely a trade-off between high expression of requisite enzymes for xylose utilization and an expression burden to the cell, which would be more apparent upon further stresses from PPP overexpression.

Repeated use of a promoter may decrease transcription from this promoter due to saturation of cellular expression machinery. Of particular concern would be saturation of the strongest promoter, *pTDH3*, which also regulates the expression of the glycolytic enzyme, *GAPDH*. Indeed, shotgun proteomics revealed a 1.4-fold decrease ($p < 0.05$) in GAPDH protein between the *pTDH3* reference strain with eight copies of *pTDH3* and optimized strains containing only a couple copies. This GAPDH reduction appears to have minimal impact aerobically as the *pTDH3* full pathway grows similarly to LL110A (**Figure 4-3E**). Thus, reduced GAPDH has either a minor contribution to differences in the *pTDH3* reference strains and optimized strains or its effects are condition dependent. Whatever the cause, arbitrarily high expression of a large number of pathway enzymes can produce a multitude of potential problems that can be solved, or at least largely alleviated, by combinatorial expression engineering, where complications such as these will be selected against by having non-limiting enzymes expressed with a weaker promoter.

Another interesting aspect of xylose utilization is the asymmetric cofactor usage, which has led to the hypothesis that resolving this imbalance would improve fermentation performance[15,48,54]. However, in mutating XDH to switch its cofactor preference[13], its activity is unavoidably also altered, thereby convoluting the contributions from cofactor balancing and potential activity balancing with other pathway enzymes. By incorporating this mutant into our library, we were able to empirically optimize the balance of enzymatic

activities to investigate the importance of cofactor balance. Surprisingly, the mutant XDH only slightly improved fermentative yields in expression-optimized strains (**Figure 4-7B**), certainly not as dramatic as would be expected if cofactor imbalances were a major flux impediment in our engineered system. The mutant XDH strains do show a modest improvement in growth, which implies cofactor balancing of this pathway aids in biomass accumulation, perhaps due to the role of NADPH in anabolic processes[30].

Finally, as expected, changing external variables led to differing expression profiles. Here we optimized under both aerobic and anaerobic conditions. Some of the aerobically enriched strains, such as LL110B, had notably inferior anaerobic fermentation performances (**Figure 4-4**). The overall use of stronger promoters in LL110B may be affecting this strain's performance in an oxygen-dependent manner due to a flux imbalance or metabolic stress, such as XK toxicity[33]. These findings, together with the observation that expression optimization under aerobic conditions had a fairly small impact on strain growth over arbitrarily high expression of all pathway genes, suggest that expression balancing is more critical under the more stringent anaerobic conditions.

In conclusion, we applied combinatorial expression engineering for the optimization of the full *S. stipitis* xylose utilization pathway in the favored production host, *S. cerevisiae*. We observed, in both aerobic and anaerobic conditions, dramatically improved performance of strains with the full pathway optimized over complementing *S. cerevisiae* with only the minimal, requisite activities. Expression optimization was also used to separate the effects of activity balancing from cofactor balancing and effects from altered oxygenation. Based on our findings, it would be prudent to include more metabolic genes – perhaps from glycolysis, gluconeogenesis, and/or the oxidative PPP – in further studies of this pathway. Additionally, this described strategy can be employed toward other pathways with outputs amenable to high-throughput screening or selection.

4.4 Materials and Methods

Strains and media

Single gene (cassette) plasmids were transformed in chemically competent TG1 cells grown in LB containing spectinomycin (50mg/L). All multi-gene plasmid assemblies were transformed into TransforMax EPI300 (Epicentre) electrocompetent *E. coli*. Transformed cells were selected on LB plates containing antibiotics chloramphenicol (34mg/L) or kanamycin (25mg/L). The *S. cerevisiae* strain background for all experiments in this paper was BY4741 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0). Dry cell weight conversion for this strain was determined to be 0.19 +/- 0.01g cell/L at OD_{600} = 1 as described previously[55]. In all instances, *S. cerevisiae* strains were grown at 30C. Wild-type yeast cultures were grown in YPD (10g/L Bacto Yeast Extract; 20g/L Bacto Peptone; 20g/L Dextrose). Excluding library transformations, yeast transformed with plasmids containing the *LEU2* and *URA3* auxotrophic markers were selected and grown on synthetic drop-out media (6.7g/L Difco Yeast Nitrogen Base w/o Amino Acids; 2g/L Drop-out Mix Synthetic minus Leucine and Uracil, w/o Yeast Nitrogen Base (US Biological); 20g/L Dextrose or 20g/L Xylose).

Table 4-2. Plasmids used in this study.

Plasmid	SynBERC Registry ID	Description
pML530	SBa_001115	Golden Gate Cassette Vector for SsXK and SsRPE
pML531	SBa_001116	Golden Gate Cassette Vector for SsXR and SsRKI
pML532	SBa_001117	Golden Gate Cassette Vector for SsTKL
pML533	SBa_001118	Golden Gate Cassette Vector for SsTAL
pML534	SBa_001119	Golden Gate Cassette Vector for SsPYK
pML557	SBa_001120	Golden Gate Cassette Vector for SsXDH and mutSsXDH
pML635	SBa_001121	CamR Golden Gate Backbone
pML636	SBa_001122	Yeast Origin, *LEU2* Golden Gate Backbone
pML637	SBa_001123	KanR Golden Gate Backbone
pML638	SBa_001124	Yeast Origin, *URA3* Golden Gate Backbone
pLNL52L	SBa_001125	Assembled SsXR-SsXDH-SsXK library plasmid
pLNL53L	SBa_001126	Assembled SsXR-SsXDH(ARSdR)-SsXK library plasmid
pLNL54L	SBa_001127	Assembled SsRPE-SsRKI-SsTKL-SsTAL-SsPYK library plasmid
pML660	SBa_001149	All promoters *pTDH3*, SsRPE-SsRKI-SsTKL-SsTAL-SsPYK
pML661	SBa_001150	All promoters *pTDH3*, SsXR-SsXDH-SsXK
pML662	SBa_001151	All promoters *pTDH3*, SsXR-SsXDH(ARSdR)-SsXK
pLNL64	SBa_001152	*pTDH3*-SsXR-*pRPL18B*-SsXDH-*pRPL18B*-SsXK
pLNL66	SBa_001153	*pTEF1*-SsXR-*pTEF1*-SsXDH-*pTEF1*-SsXK
pLNL67	SBa_001154	*pREV1*-SsRPE-*pRPL18B*-SsRKI-*pREV1*-SsTKL-*pTDH3*-SsTAL-*pRPL18B*-SsPYK
pLNL69	SBa_001155	*pTDH3*-SsRPE-*pTEF1*-SsRKI-*pRNR2*-SsTKL-*pTDH3*-SsTAL-*pRPL18B*-SsPYK
pLNL74	SBa_001156	*pREV1*-SsRPE-*pREV1*-SsRKI-*pRPL18B*-SsTKL-*pTDH3*-SsTAL-*pRNR2*-SsPYK
pLNL78	SBa_001157	*pTEF1*-SsXR-*pTEF1*-SsXDH(ARSdR)-*pRNR2*-SsXK
pLNL101	SBa_001158	*pREV1*-SsRPE-*pTEF1*-SsRKI-*pRPL18B*-SsTKL-*pRPL18B*-SsTAL-*pRNR2*-SsPYK
pLNL105	SBa_001159	*pRPL18B*-SsXR-*pTDH3*-SsXDH-*pTEF1*-SsXK
pLNL107	SBa_001160	*pTDH3*-SsXR-*pTEF1*-SsXDH-*pRPL18B*-SsXK

All plasmids contain a ColE1 *E. coli* replication origin. Annotated plasmid sequences can be found at the SynBERC Registry (registry.synberc.org). Sequences of plasmids not listed in this table (e.g., the series of cassette plasmids) can be determined simply by replacing the appropriate genes or promoters.

Pathway construction and combinatorial library assembly

All pathway enzymes, except the mutant *SsXDH* (D207A/I208R/F209S/N211R), were cloned by PCR from the *S. stipitis* genome. Four of the genes: *SsXK, SsRPE, SsRKI,* and *SsTKL* were cloned by SOEing PCR[56] to introduce silent mutations to mutate internal BglII or BsmBI sites. The mutant XDH was synthesized by GenScript and included flanking BglII and XhoI sites. Cassette plasmids were assembled through standard restriction-ligation subcloning using restriction enzymes BglII and either SpeI or XhoI into backbones (pML530-4 and pML557). Plasmid libraries or plasmids with specific promoter genotypes were assembled in a BsmBI golden gate reaction using 20 fmol of each cassette plasmid

template and two backbone plasmids (pML634-638)[41]. Plasmid information is summarized in **Table 4-2** and backbone vector construction details are available upon request.

For library plasmids, golden gate reactions were electroporated, cells were recovered in LB for one hour, plated on 241mm x 241mm plates and grown overnight. The resulting colonies (over 35,000) were scraped into 15 mL of ddH_2O and treated as a liquid culture for plasmid purification with a HiSpeed Plasmid Maxi Kit (Qiagen) following manufacturer's instructions. Plasmid libraries were test digested to confirm correct assembly. Individual plasmids were also isolated from single colonies and confirmed by test digest.

Eight gene library transformations into *S. cerevisiae* followed standard Lithium acetate transformation protocol[57] scaled to a 500 mL culture and transformed using 100 μg of each library plasmid. Following heatshock, cells were recovered in 50 mL of YPD (250 mL baffled flask, 200 rpm) for one hour before being pelleted, washed in SX-LU and then resuspended in 500 mL SX-LU under either aerobic (2L baffeled flask) or anaerobic (media supplemented with 0.01g/L ergosterol, 0.43g/L Tween 80 and 2.8 g/L ethanol; 1L Erlenmeyer screw cap flask flushed with $N_{2(g)}$) conditions with an aliquot plated on SD-LU for sampling of initial library coverage and diversity. Three gene library transformations followed the above procedure using an 'empty' plasmid containing markers and origins only, instead of the five-gene PPP plasmid; 1/100 the listed masses and volumes were used in 50 mL culture tubes or serum vials.

Library enrichments

Transformed libraries were grown 4-7 days on SX-LU until the OD_{600} of the library increased twofold over the initial OD_{600}. Library cultures were subsequently diluted into 50 mL of fresh SX-LU carrying over at least $1x10^7$ cells into either a 250 mL baffled Erlenmeyer flask (aerobic, 200 rpm) or 250 mL serum vial (anaerobic, 100 rpm). After this, cells were similarly diluted in late log-phase (OD_{600} 5-10, aerobic; 1.5-2.5, anaerobic) into fresh media every 1-4 days, depending on growth rate. During select dilutions, aliquots of cells were also plated on SD-LU, grown for 2-3 days and genotyped according to the TaqMan-based TRAC protocol previously developed in our lab[39]. Primer sequences used for TRAC genotyping reactions are listed in **Table 4-3**.

Table 4-3. Primers used in this study.

Primer	Sequence
GibB-for (XR, RKI)	CCAGATGTCAACACAGCTAC
XR-rev	GTTCCACAACTTGGAGGTAA
GibC-for (XDH, TKL)	ACACACTGGCTTAAGGAGAC
XDH-rev	GTAGAAGTGGATGTCGGAAC
GibA-for (XK, RPE)	GCCGATAATTGCAGACG
XK-rev	AGCTTATCTGGAGCATCAAA
RPE-rev	AGATGGACGGAGAGATGATA
RKI-rev	TTCAGCTACGTAAACGACAG
TKL-rev	CCTTAGGGTTGAATCTCATCT
GibD-for (TAL)	AATAAAGCTCCACACAGTCG
TAL-rev	CGTATTCAGGCTTCTTAGCA
GibE-for (PYK)	TATGGGCACAGACAACCTA
PYK-rev	GACCAAGACTTCGACAGAGT

Growth curves and fermentations

Following plasmid transformation into yeast, cells were grown on SD-LU agar plates for 2-3 days. Colonies were picked into SD-LU (3 mL), grown for 24 hours and then diluted

into SX-LU (3 mL) by transferring an aliquot of grown cells into fresh SX-LU. After 48 hours of aerobic growth in 24-well blocks, OD_{600} was measured and cells were diluted into a larger volume of SX-LU and shaken at either 100 or 200 rpm depending on the growth conditions. At designated time points, aliquots were taken from the cultures to measure OD_{600} and media was stored at -20C for later analysis by HPLC.

Metabolite quantification

Media aliquots were pelleted, and supernatant was transferred to GC/MS vials for sampling. From each sample, 10 μL was analyzed by refractive index on a Shimadzu LC20AD HPLC equipped with a Rezex RFP-fast acid H^+ column (100x7.8mm, 55C) run with 1 mL/min 0.01N H_2SO_4 mobile phase. Metabolite concentrations were determined by comparing to a standard curve.

Shotgun proteomics

Cells were grown following the same procedure used during a growth curve. After 24 (aerobic) or 48 (anaerobic) hours of growth in the final, larger SX-LU culture, OD_{600} was measured for each culture and 10 OD units (where 1 OD unit = the cells in 1 mL of culture at 1 OD_{600}) of cells were pelleted, washed in 1 mL PBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 2 mM KH_2PO_4), and immediately frozen at -80C until later use.

Cell lysates (100 μg total protein) were precipitated in 20% TCA at -80C overnight, pelleted at 4C, washed three times with ice cold 0.01 N HCl/90% acetone, dried at 25C and then resuspended in 8 M urea. ProteaseMax (0.1%, Promega) was added, vortexed and the reaction was diluted to 100 μL with NH_4HCO_3 (70 mM). Protein was reduced by incubated with TCEP (10 mM) for 30 min at 55C and then alkylated with Iodoacetamide (12.5 mM) for 30 min in the dark with shaking. The reaction was brought up to a final volume of 234 μL with PBS and ProteaseMax (0.03%) and treated with trypsin (0.5μg/μL, Promega) overnight at 37C. The digested peptides were acidified with formic acid (5%), spun at max speed for 30 min and the supernatant was stored at -80C until further use. The resulting tryptic peptides were loaded onto inline filters and peptides were chromatographically separated a C18 nanospray column. Peptides were analyzed by an LTQ-XL. The resulting data were analyzed against the *S. stiptis* and *S. cerevisiae* tryptic proteomes using Integrated Proteomics Pipeline. ms2 spectra data were searched using the SEQUEST algorithm (Version 3.0)[58]. SEQUEST searches allowed for oxidation of methionine residues (16 Da), static modification of cysteine residues (57 Da-due to alkylation), no enzyme specificity and a mass tolerance set to ± 1.5 Da for precursor mass and ± 0.5 Da for product ion masses. The resulting ms2 spectra matches were assembled and filtered using DTASelect (version 2.0.27)[59]. A quadratic discriminant analysis was used to achieve a maximum peptide false positive rate of 1% as previously described[60,61].

4.5 References

1. Balat, M. & Balat, H. Recent trends in global production and utilization of bio-ethanol fuel. *Applied Energy* **86,** 2273–2282 (2009).

2. Hayes, D. J. An examination of biorefining processes, catalysts and challenges. *Catalysis Today* **145,** 138–151 (2009).
3. Somerville, C. Biofuels. *Curr Biol* **17,** R115–9 (2007).
4. Pauly, M. & Keegstra, K. Cell-wall carbohydrates and their modification as a resource for biofuels. *Plant J.* **54,** 559–568 (2008).
5. Olsson, L. & Hahn-Hägerdal, B. Fermentative performance of bacteria and yeasts in lignocellulose hydrolysates. *Process Biochemistry* **28,** 249–257 (1993).
6. Deng, X. X. & Ho, N. W. Xylulokinase activity in various yeasts including Saccharomyces cerevisiae containing the cloned xylulokinase gene. Scientific note. *Appl Biochem Biotechnol* **24-25,** 193–199 (1990).
7. Yomano, L. P., York, S. W. & Ingram, L. O. Isolation and characterization of ethanol-tolerant mutants of Escherichia coli KO11 for fuel ethanol production. *J Ind Microbiol Biotechnol* **20,** 132–138 (1998).
8. Skoog, K. & Hahn-Hägerdal, B. Xylose fermentation. *Enzyme and Microbial Technology* **10,** 66–80 (1988).
9. Kötter, P., Amore, R., Hollenberg, C. P. & Ciriacy, M. Isolation and characterization of the Pichia stipitis xylitol dehydrogenase gene, XYL2, and construction of a xylose-utilizing Saccharomyces cerevisiae transformant. *Curr Genet* **18,** 493–500 (1990).
10. Jeppsson, M. *et al.* The expression of a Pichia stipitis xylose reductase mutant with higher K(M) for NADPH increases ethanol production from xylose in recombinant Saccharomyces cerevisiae. *Biotechnol Bioeng* **93,** 665–673 (2006).
11. Bengtsson, O., Hahn-Hägerdal, B. & Gorwa-Grauslund, M. F. Xylose reductase from Pichia stipitis with altered coenzyme preference improves ethanolic xylose fermentation by recombinant Saccharomyces cerevisiae. *Biotechnol Biofuels* **2,** 9 (2009).
12. Khoury, G. A. *et al.* Computational design of Candida boidinii xylose reductase for altered cofactor specificity. *Protein Sci.* **18,** 2125–2138 (2009).
13. Watanabe, S., Kodaki, T. & Makino, K. Complete reversal of coenzyme specificity of xylitol dehydrogenase and increase of thermostability by the introduction of structural zinc. *J Biol Chem* **280,** 10340–10349 (2005).
14. Watanabe, S. *et al.* Ethanol production from xylose by recombinant Saccharomyces cerevisiae expressing protein engineered NADP+-dependent xylitol dehydrogenase. *J Biotechnol* **130,** 316–319 (2007).
15. Krahulec, S. *et al.* Fermentation of mixed glucose-xylose substrates by engineered strains of Saccharomyces cerevisiae: role of the coenzyme specificity of xylose reductase, and effect of glucose on xylose utilization. *Microb Cell Fact* **9,** 16 (2010).
16. Johansson, B., Christensson, C., Hobley, T. & Hahn-Hagerdal, B. Xylulokinase overexpression in two strains of Saccharomyces cerevisiae also expressing xylose reductase and xylitol dehydrogenase and its effect on fermentation of xylose and lignocellulosic hydrolysate. *Applied and Environmental Microbiology* **67,** 4249–4255 (2001).
17. Kim, S. R. *et al.* Rational and evolutionary engineering approaches uncover a small set of genetic changes efficient for rapid xylose fermentation in Saccharomyces cerevisiae. *PLoS ONE* **8,** e57048 (2013).

18. Van Vleet, J. H., Jeffries, T. W. & Olsson, L. Deleting the para-nitrophenyl phosphatase (pNPPase), PHO13, in recombinant Saccharomyces cerevisiae improves growth and ethanol production on D-xylose. *Metab Eng* **10,** 360–369 (2008).
19. Scalcinati, G. *et al.* Evolutionary engineering of Saccharomyces cerevisiae for efficient aerobic xylose consumption. *FEMS Yeast Res* **12,** 582–597 (2012).
20. Ho, N. W., Chen, Z. & Brainard, A. P. Genetically engineered Saccharomyces yeast capable of effective cofermentation of glucose and xylose. *Applied and Environmental Microbiology* **64,** 1852–1859 (1998).
21. Karhumaa, K., Hahn-Hägerdal, B. & Gorwa-Grauslund, M.-F. Investigation of limiting metabolic steps in the utilization of xylose by recombinant Saccharomyces cerevisiae using metabolic engineering. *Yeast* **22,** 359–368 (2005).
22. Matsushika, A., Inoue, H., Kodaki, T. & Sawayama, S. Ethanol production from xylose in engineered Saccharomyces cerevisiae strains: current state and perspectives. *Appl Microbiol Biotechnol* **84,** 37–53 (2009).
23. Kim, S. R., Ha, S.-J., Kong, I. I. & Jin, Y.-S. High expression of XYL2 coding for xylitol dehydrogenase is necessary for efficient xylose fermentation by engineered Saccharomyces cerevisiae. *Metab Eng* **14,** 336–343 (2012).
24. Jin, Y.-S., Alper, H. S., Yang, Y.-T. & Stephanopoulos, G. Improvement of xylose uptake and ethanol production in recombinant Saccharomyces cerevisiae through an inverse metabolic engineering approach. *Applied and Environmental Microbiology* **71,** 8249–8256 (2005).
25. Bera, A. K., Ho, N. W. Y., Khan, A. & Sedlak, M. A genetic overhaul of Saccharomyces cerevisiae 424A(LNH-ST) to improve xylose fermentation. *J Ind Microbiol Biotechnol* **38,** 617–626 (2010).
26. Kuyper, M. *et al.* Evolutionary engineering of mixed-sugar utilization by a xylose-fermenting Saccharomyces cerevisiae strain. *FEMS Yeast Res* **5,** 925–934 (2005).
27. Karhumaa, K., Fromanger, R., Hahn-Hägerdal, B. & Gorwa-Grauslund, M. F. High activity of xylose reductase and xylitol dehydrogenase improves xylose fermentation by recombinant Saccharomyces cerevisiae. *Appl Microbiol Biotechnol* **73,** 1039–1046 (2007).
28. Lu, C. & Jeffries, T. W. Shuffling of promoters for multiple genes to optimize xylose fermentation in an engineered Saccharomyces cerevisiae strain. *Applied and Environmental Microbiology* **73,** 6072–6077 (2007).
29. Fiaux, J. *et al.* Metabolic-flux profiling of the yeasts Saccharomyces cerevisiae and Pichia stipitis. *Eukaryotic Cell* **2,** 170–180 (2003).
30. Blank, L. M., Lehmbeck, F. & Sauer, U. Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts. *FEMS Yeast Res* **5,** 545–558 (2005).
31. Parachin, N. S., Bergdahl, B., van Niel, E. W. J. & Gorwa-Grauslund, M. F. Kinetic modelling reveals current limitations in the production of ethanol from xylose by recombinant Saccharomyces cerevisiae. *Metab Eng* **13,** 508–517 (2011).
32. Rodriguez-Peña, J. M., Cid, V. J., Arroyo, J. & Nombela, C. The YGR194c (XKS1) gene encodes the xylulokinase from the budding yeast Saccharomyces cerevisiae. *FEMS microbiology letters* **162,** 155–160 (1998).

33. Jin, Y.-S., Ni, H., Laplaza, J. M. & Jeffries, T. W. Optimal growth and ethanol production from xylose by recombinant Saccharomyces cerevisiae require moderate D-xylulokinase activity. *Applied and Environmental Microbiology* **69,** 495–503 (2003).
34. Du, J., Yuan, Y., Si, T., Lian, J. & Zhao, H. Customized optimization of metabolic pathways by combinatorial transcriptional engineering. *Nucleic Acids Res* (2012). doi:10.1093/nar/gks549
35. Eliasson, A., Hahn-Hägerdal, B., Christensson, C. & Wahlbom, C. F. C. Anaerobic xylose fermentation by recombinant Saccharomyces cerevisiae carrying XYL1, XYL2, and XKS1 in mineral medium chemostat cultures. *Applied and Environmental Microbiology* **66,** 3381–3386 (2000).
36. Wahlbom, C. F., van Zyl, W. H., Jönsson, L. J., Hahn-Hägerdal, B. & Otero, R. R. C. Generation of the improved recombinant xylose-utilizing Saccharomyces cerevisiae TMB 3400 by random mutagenesis and physiological comparison with Pichia stipitis CBS 6054. *FEMS Yeast Res* **3,** 319–326 (2003).
37. Matsushika, A. & Sawayama, S. Efficient bioethanol production from xylose by recombinant saccharomyces cerevisiae requires high activity of xylose reductase and moderate xylulokinase activity. *Journal of Bioscience and Bioengineering* **106,** 306–309 (2008).
38. Pearce, A. K. *et al.* Pyruvate kinase (Pyk1) levels influence both the rate and direction of carbon flux in yeast under fermentative conditions. *Microbiology (Reading, Engl.)* **147,** 391–401 (2001).
39. Lee, M. E., Aswani, A., Han, A. S., Tomlin, C. J. & Dueber, J. E. Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay. *Nucleic Acids Res* **41,** 10668–10678 (2013).
40. Kuroda, S., Otaka, S. & Fujisawa, Y. Fermentable and nonfermentable carbon sources sustain constitutive levels of expression of yeast triosephosphate dehydrogenase 3 gene from distinct promoter elements. *J Biol Chem* **269,** 6153–6162 (1994).
41. Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS ONE* **4,** e5553 (2009).
42. Jeppsson, M., Johansson, B., Hahn-Hägerdal, B. & Gorwa-Grauslund, M. F. Reduced oxidative pentose phosphate pathway flux in recombinant xylose-utilizing Saccharomyces cerevisiae strains improves the ethanol yield from xylose. *Applied and Environmental Microbiology* **68,** 1604–1609 (2002).
43. Walfridsson, M., Hallborn, J., Penttilä, M., Keränen, S. & Hahn-Hagerdal, B. Xylose-metabolizing Saccharomyces cerevisiae strains overexpressing the TKL1 and TAL1 genes encoding the pentose phosphate pathway enzymes transketolase and transaldolase. *Applied and Environmental Microbiology* **61,** 4184–4190 (1995).
44. Zhou, H., Cheng, J.-S., Wang, B., Fink, G. R. & Stephanopoulos, G. Xylose isomerase overexpression along with engineering of the pentose phosphate pathway and evolutionary engineering enable rapid xylose utilization and ethanol production by Saccharomyces cerevisiae. *Metab Eng* (2012). doi:10.1016/j.ymben.2012.07.011

45. Matsushika, A., Inoue, H., Murakami, K., Takimura, O. & Sawayama, S. Bioethanol production performance of five recombinant strains of laboratory and industrial xylose-fermenting Saccharomyces cerevisiae. *Bioresour Technol* **100,** 2392–2398 (2009).
46. Runquist, D., Hahn-Hägerdal, B. & Bettiga, M. Increased ethanol productivity in xylose-utilizing Saccharomyces cerevisiae via a randomly mutagenized xylose reductase. *Applied and Environmental Microbiology* **76,** 7796–7802 (2010).
47. Bengtsson, O. *et al.* Identification of common traits in improved xylose-growing Saccharomyces cerevisiae for inverse metabolic engineering. *Yeast* **25,** 835–847 (2008).
48. Petschacher, B. & Nidetzky, B. Altering the coenzyme preference of xylose reductase to favor utilization of NADH enhances ethanol yield from xylose in a metabolically engineered strain of Saccharomyces cerevisiae. *Microb Cell Fact* **7,** 9 (2008).
49. Hahn-Hägerdal, B., Karhumaa, K., Jeppsson, M. & Gorwa-Grauslund, M. F. Metabolic engineering for pentose utilization in Saccharomyces cerevisiae. *Adv Biochem Eng Biotechnol* **108,** 147–177 (2007).
50. Cai, Z., Zhang, B. & Li, Y. Engineering Saccharomyces cerevisiae for efficient anaerobic xylose fermentation: reflections and perspectives. *Biotechnol J* **7,** 34–46 (2012).
51. Matsushika, A. *et al.* Efficient bioethanol production by a recombinant flocculent Saccharomyces cerevisiae strain with a genome-integrated NADP+-dependent xylitol dehydrogenase gene. *Applied and Environmental Microbiology* **75,** 3818–3822 (2009).
52. Johansson, B. & Hahn-Hägerdal, B. The non-oxidative pentose phosphate pathway controls the fermentation rate of xylulose but not of xylose in Saccharomyces cerevisiae TMB3001. *FEMS Yeast Res* **2,** 277–282 (2002).
53. Hector, R. E. *et al.* Saccharomyces cerevisiae engineered for xylose metabolism requires gluconeogenesis and the oxidative branch of the pentose phosphate pathway for aerobic xylose assimilation. *Yeast* **28,** 645–660 (2011).
54. Kötter, P. & Ciriacy, M. Xylose fermentation by Saccharomyces cerevisiae. *Appl Microbiol Biotechnol* **38,** 776–783 (1993).
55. Sonderegger, M. & Sauer, U. Evolutionary engineering of Saccharomyces cerevisiae for anaerobic growth on xylose. *Applied and Environmental Microbiology* **69,** 1990–1998 (2003).
56. Horton, R. M. PCR-mediated recombination and mutagenesis. SOEing together tailor-made genes. *Mol. Biotechnol.* **3,** 93–99 (1995).
57. Gietz, R. D. & Woods, R. A. Yeast transformation by the LiAc/SS Carrier DNA/PEG method. *Methods Mol Biol* **313,** 107–120 (2006).
58. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989 (1994).
59. Cociorva, D. & Yates, J. R., III. *DTASelect 2.0: improving the confidence of peptide and protein identifications*. (54th ASMS Annual Meeting Proceedings, 2006).
60. Cociorva, D., L Tabb, D. & Yates, J. R. Validation of tandem mass spectrometry database search results using DTASelect. *Curr Protoc Bioinformatics* **Chapter 13,** Unit 13.4–13.4.14 (2007).

61. Tabb, D. L., McDonald, W. H. & Yates, J. R. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1,** 21–26 (2002).
62. Chen, S.-H. et al. Engineering transaldolase in Pichia stipitis to improve bioethanol production. ACS Chem. Biol. 7, 481–486 (2012).
63. Bär, J., Naumann, M., Reuter, R. & Kopperschläger, G. Improved purification of ribulose 5-phosphate 3-epimerase from Saccharomyces cerevisiae and characterization of the enzyme. Bioseparation 6, 233–241 (1996).
64. Reuter, R., Naumann, M., Bär, J., Miosga, T. & Kopperschläger, G. Ribose-5-phosphate isomerase from Saccharomyces cerevisiae: purification and molecular analysis of the enzyme. Bioseparation 7, 107–115 (1998).
65. Collins, R. A., McNally, T., Fothergill-Gilmore, L. A. & Muirhead, H. A subunit interface mutant of yeast pyruvate kinase requires the allosteric activator fructose 1,6-bisphosphate for activity. Biochem J 310 (Pt 1), 117–123 (1995).

Chapter 5. Engineering a xylose-specific transporter for fermentation of lignocellulosic biomass.

5.1 Introduction

As discussed in Chapter 4, engineering xylose utilization in *Saccharomyces cerevisiae* is an important step towards enabling sustainable biofuel production from lignocellulosic (LC) biomass feedstocks. However, an additional challenge remains even once the utilization pathway has been optimized: carbon catabolite repression. When a mixture of sugars—such as glucose and xylose, the primary sugars found in LC feedstocks—is present in the media, yeast will preferentially consume glucose first, before consuming the xylose. This is problematic for both batch and continuous fermentations. In a batch fermentation, while separately consuming the two sugars is not itself an issue, the diauxic shift that occurs when the yeast's metabolism switches between the two sugars creates a period of inactivity during which the reactor is idle, reducing the overall productivity. Second, a strain that consumes the sugars at different times actually precludes the possibility of a continuous fermentation. Therefore, irrespective of the preferred reactor conditions, engineering simultaneous consumption of both (and all other) sugars is critical for efficient LC fermentation.

There are a number of possible mechanisms by which the cell could preferentially consume glucose over xylose. First, intracellular glucose could bind transcription factors or allosterically inhibit xylose catabolic enzymes. This was shown not to be the case in two previous studies where yeast were grown on a mixture of xylose and either maltose[1] or cellobiose[2] (the latter required expression of a heterologous cellodextrin transporter). In these studies, the rate of xylose consumption was not inhibited by the presence of intracellular glucose and was even enhanced in the cellobiose case. The second possible mechanism is that extracellular glucose could regulate transcription via glucose sensors on the membrane. Unfortunately, it is difficult to determine whether this is the case without a transporter that can specifically import xylose into the cell. This leads to the third possibility, which is that the two sugars compete for transport. This last mechanism is the most likely, given that 1) it is known that xylose is imported via the native hexose transporters[3,4]; and 2) the only enzymes exclusive to xylose metabolism are typically heterologously expressed and would not be expected to be under native transcriptional regulation. Thus, a xylose-specific transporter would either enable simultaneous uptake and consumption of both sugars, or would reveal that transport alone is not the inhibitory mechanism.

Prior to this study, a bioprospecting approach identified two xylose-specific transporters found naturally in *Scheffersomyces stipitis* and *Neurospora crassa*, but those transporters did not have high enough activity to exhibit a measurable effect when expressed in wild-type yeast or support growth when expressed in a transporter knockout strain[5]. Over the course of this work, three studies were published with the goal of engineering xylose specificity. The first identified a conserved, G-G/F-XXX-G, motif in several transporters across many species that appeared to dictate sugar specificity[6]. By mutating this motif

in *GXS1* from *Candida intermedia*, *RGT2* from *S. stipitis*, and *HXT7* from *S. cerevisiae*, the authors were able to abolish glucose transport activity while maintaining xylose uptake ability. Unfortunately, these mutants were inhibited by glucose and unable to transport xylose when the two sugars were both present in the media. The second study used a strain that was both incapable of transporting sugars (all native hexose transporters were knocked out) and unable to metabolize glucose (hexose kinases were knocked out)[7]. A native transporter was reintroduced, and the strain was selected on a mixture of the two sugars, reasoning that because the strain could only metabolize xylose, it would evolve specificity of its uptake. While the authors claimed to have identified mutants of *GAL2* and *HXT7* that appeared to be uninhibited by glucose, a key experiment directly demonstrating specific uptake was not shown. Finally, a third group performed an evolution experiment on a strain with its hexose kinases knocked out, but its transporters still present, hoping that having the full array of native transporters would give more potential options for evolving xylose-specificity[8]. The result of this work was a chimeric *HXT3*/*HXT6* with an additional single residue mutation (in fact, the same residue identified in *GAL2* and *HXT7* in the second study). When this mutant was expressed in a strain lacking its native transporters, the strain was able to co-ferment glucose and xylose. In this case, rather than identifying a purely xylose-specific transporter, a dual-functioning transporter was chosen. Unfortunately, it is as yet unclear what the effects of this transporter would be in a strain that still has its native transporters; that is, whether xylose uptake rates would be comparable to glucose, although it is a promising lead towards co-consumption.

Ultimately, despite these recent advances, the question of whether a strain of yeast can be engineered to consume glucose and xylose simultaneously and efficiently enough to support biofuel production remains unanswered. In this work, we attempt to engineer a native yeast transporter, *HXT5*, to specifically transport xylose. *HXT5* was chosen for its relatively high rate of xylose uptake compared to other members of the Hxt family[3]. Although *HXT7* exhibited a higher uptake rate, the gene could not be stably cloned due to significant toxicity in *Escherichia coli*, and with the goal of constructing high-diversity libraries, was discarded in favor of *HXT5*. We use a number of rational and random mutagenesis strategies, some of which draw from findings in the aforementioned studies. In order to provide a clean background with no basal transport of either sugar, we use an assay strain that has had eight of its transporters knocked out, which in this strain is sufficient to abolish uptake activity. We select against glucose import using the non-metabolizable analog, 2-deoxy-D-glucose (2DG). Additionally, we perform whole-genome sequencing on a number of evolved strains developed during this study, which may provide additional insight into sugar metabolism.

5.2 Results

5.2.1 Engineering an *HXT5*-dependent xylose-consuming strain

In order to directly assay *HXT5* activity *in vivo*, we required a strain that had no basal transport activity so that any observed uptake would be the result of *HXT5* alone. Two such strains exist in the literature: EBY.VW4000[9] and KY73[10]. EBY.VW4000 is a CEN.PK2-1C background with all seventeen *HXT* genes, *GAL2*, and four additional transporter genes

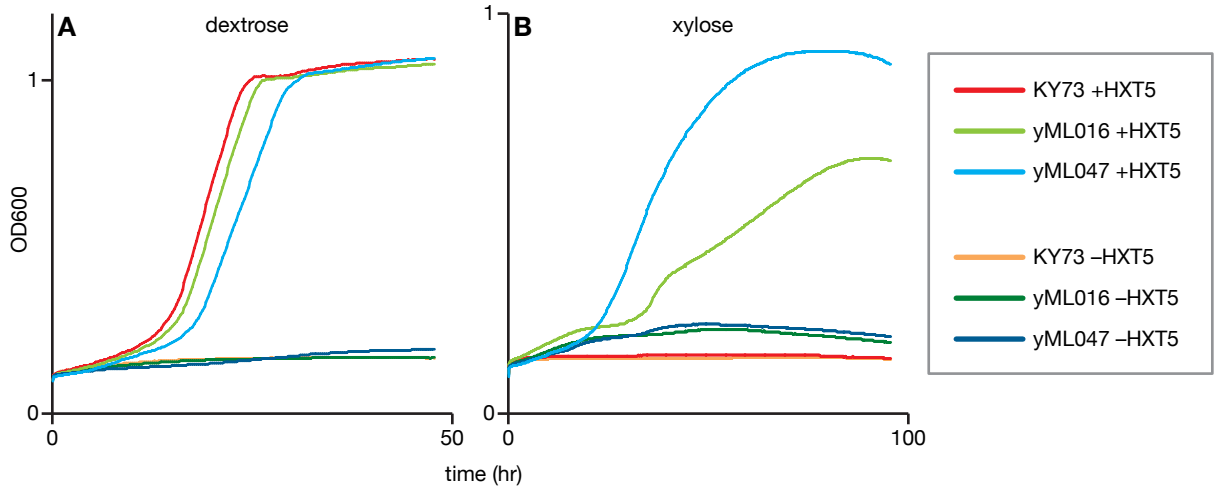


Figure 5-1. *HXT5*-dependent growth on dextrose and xylose. (A) Expression of *HXT5* is required for growth on dextrose of knockout strain KY73 and its derivatives, yML016 and yML047. (B) KY73 cannot grow on xylose without the xylose utilization genes. yML016 has the xylose utilization genes expressed, but grows weakly in xylose when *HXT5* is expressed. yML047 has robust growth in xylose when *HXT5* is expressed.

knocked out, and this strain is unable to grow on glucose as the sole carbon source. Unfortunately, this strain could not be acquired for use in this study. The second strain, KY73, is a MC996A background with only *HXT1-7* and *GAL2* knocked out. However, even with only these eight genes knocked out, the strain is unable to grow on glucose (**Figure 5-1A**), likely due to differences in the background compared to CEN.PK2-1C. We obtained this strain and used it as the basis for all subsequent experiments.

We knocked into KY73, the xylose utilization genes from *S. stipitis*—XR, XDH, and XK—as well as an additional copy of *TAL* (also from *S. stipitis*), to generate strain yML016. Due to the lack of any transporter expression, the preferred carbon source for this strain is maltose. Upon transforming a plasmid expressing *HXT5*, we did not immediately observe robust growth when cultures were transferred from maltose to xylose media. However, after an extended incubation at 30C with agitation for one week, some, but not all, cultures grew dense. These cultures were diluted into fresh maltose media, grown to saturation, and diluted back into fresh xylose media. These cultures grew immediately in xylose, without the long lag that was previously observed, suggesting an adaptation that conferred the ability to grow on xylose more rapidly. To confirm this, we cured one of these cultures of the *HXT5* plasmid (to generate strain yML047), and upon retransformation of a fresh *HXT5* plasmid, the strain was again immediately able to grow on xylose. yML047 alone remains incapable of growing on xylose, indicating that whatever background mutations arose were not enabling transport of xylose, and that growth on xylose was still dependent on expression of *HXT5* (**Figure 5-1B**).

5.2.2 Inhibition of transport using 2-deoxy-D-glucose

In order to select mutants for xylose transport that are uninhibited by glucose, we devised a strategy that utilizes the non-metabolizable glucose analog, 2-deoxy-D-glucose (2DG). 2DG causes cytotoxicity, and therefore, we can observe that 2DG uptake is mediated by

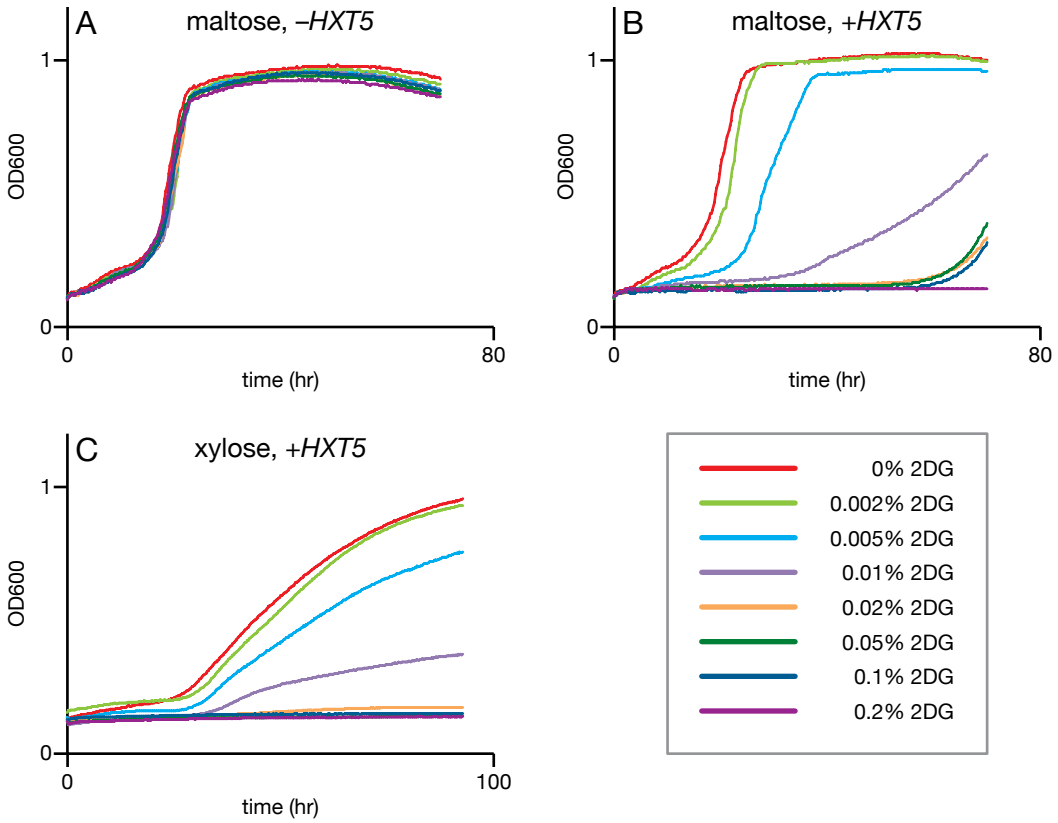


Figure 5-2. 2-deoxy-D-glucose growth inhibition. (A) Without *HXT5* expressed, 2DG does not cause toxicity to cells grown in maltose. (B) When *HXT5* is expressed, 2DG causes toxicity at concentrations as low as 0.005% and completely inhibits growth at 0.02%. Cells can adapt to toxicity over time, typically by mutating *HXT5* to block 2DG uptake. (C) 2DG inhibits growth on xylose, likely by competing for transport.

HXT5 (**Figure 5-2A and B**). Escape mutants can sometimes arise when grown on maltose by inactivating *HXT5* or overexpressing an endogenous phosphatase, *DOG1* or *DOG2*[11]. However, if xylose is provided as the sole carbon source rather than maltose, *HXT5* must remain active for cells to grow, and the mode of 2DG inhibition of growth becomes direct competition for transport (**Figure 5-2C**). Thus, our selection strategy is to enrich libraries of *HXT5* mutants on xylose as the sole carbon source while increasing the concentration of 2DG.

5.2.3 Site-directed mutagenesis of *HXT5*

We took two approaches to generating sequence diversity of *HXT5*, the first of which was rational, site-directed mutagenesis. While there is no crystal structure available for *HXT5*, the structure of a homolog, *E. coli xylE*, has been solved, bound to both glucose and xylose as substrates[12]. Based on the *xylE* structure, we identified four residues that coordinated glucose but not xylose binding (Q168, I171, F383, and G388). We therefore targeted the homologous residues in *HXT5* for site-saturation mutagenesis (Q230, I233, F461, and A466). Additionally, based on the G-G/F-XXX-G motif identified by Young et al, we targeted residues F98, V99, and F100 for mutagenesis. We generated libraries targeting all or a subset of these seven residues using degenerate NNK primers.

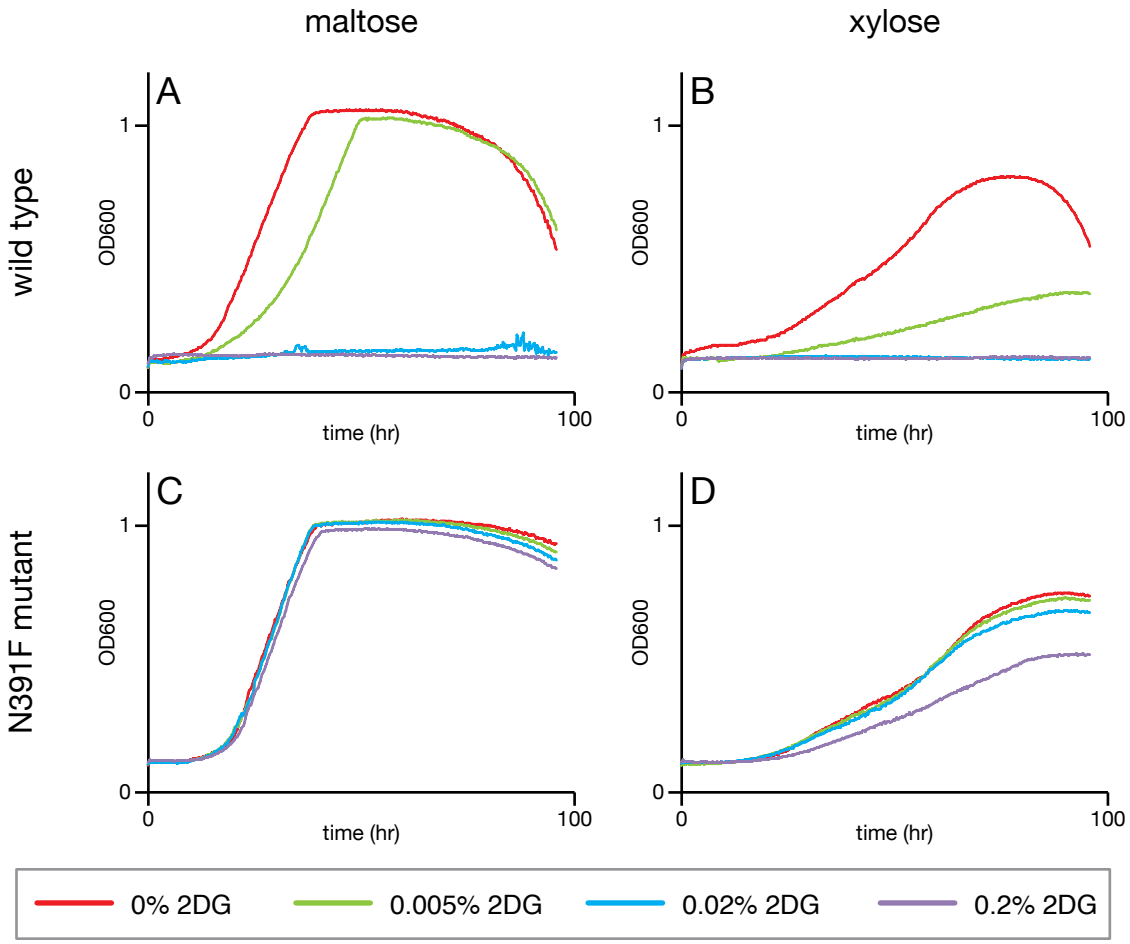


Figure 5-3. A single residue mutation in *HXT5* partially relieves 2DG growth inhibition. When wild type *HXT5* is expressed, toxicity can be observed at low concentrations of 2DG and growth is completely inhibited at 0.02% in both maltose (A) and xylose (B). A N391F mutant of HXT5 can tolerate >0.2% 2DG in maltose (C), and is only partially inhibited at 0.2% 2DG in xylose (D).

Unfortunately, technical difficulties of working with strain yML047 led to inconclusive results from these libraries. First, the strain (and its parent KY73) has very low transformation efficiency, resulting in low coverage of library diversity. Second, growth of the transformants in xylose was very slow, and cultures were often contaminated by bacteria and other fungi. Therefore, very few mutants from these libraries were screened, and none exhibited any xylose-specific transport activity.

We then looked at the two residues identified by Farwick et al, T234 and N391 in *HXT5*, and performed site-saturation mutagenesis on those two residues separately. In contrast to the original authors' results, we found that no mutants at the T234 position, including the T234S mutant isolated in the study, exhibited any specificity and were able to grow in both xylose media and glucose media. At the N391 position, although the reported N391T mutation also did not display sugar specificity, N391F and N391P mutants were unable to grow on glucose and only grew on xylose. The former substitution is consistent with the homologous N376F mutant found in *GAL2*. We tested these two mutants for their ability to transport 2DG and found that the N391P mutant did still allow for 2DG uptake and caused toxicity in maltose media. The N391F mutant, however, was resistant to 2DG, con-

firming the specificity of the transporter. Unfortunately, when a strain expressing the mutant was grown in xylose and 2DG, growth was still inhibited, suggesting that although the transporter was incapable of importing 2DG, it was still inhibited by it (**Figure 5-3**). A similar result was observed when testing the N376F mutant of *GAL2*. While this mutant of *HXT5* did not completely eliminate 2DG inhibition, it was a promising improvement over the wild type, and was the basis of further evolution.

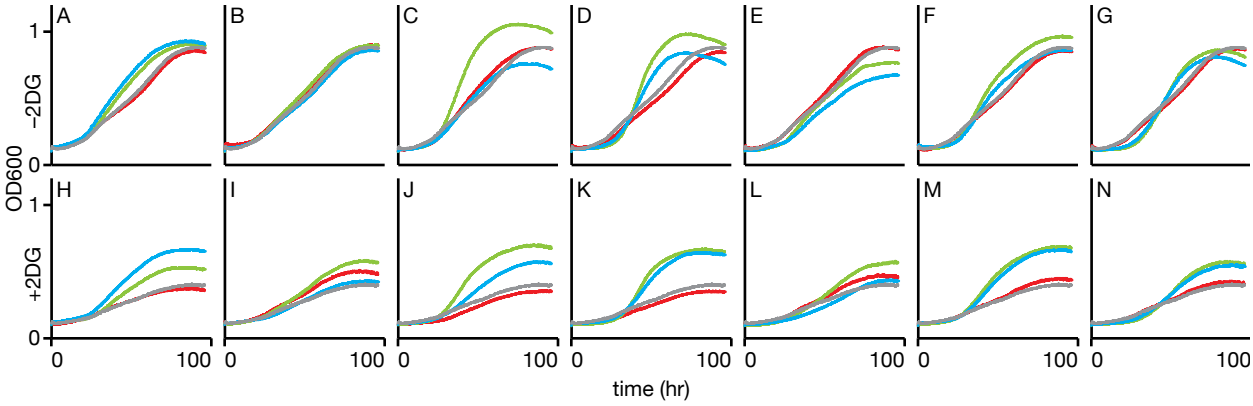
5.2.4 Random mutagenesis of the N391F *HXT5* mutant

We wanted to further improve the N391F mutant of *HXT5* to be truly uninhibited by 2DG. Given that the N391 residue itself was identified by evolution and random rather than rational mutagenesis, we elected to continue this approach to identify other important residues. We performed error-prone PCR on the mutant, targeting 1-3 mutations per clone, and transformed the resulting library into yML047. We grew the library first in maltose media, and then transferred the culture to xylose media without 2DG. We performed serial enrichments in increasing concentrations of 2DG (0.2%, 0.5%, 1%, 2%), and we isolated clones from the final enrichment at 2% 2DG for individual analysis.

We grew 186 clones in a TECAN Sunrise plate reader to compare growth in xylose media with and without 2DG. Several of these clones exhibited less inhibition by 2DG compared to the original N391F mutant (data not shown). We extracted the *HXT5* plasmid from 53 clones and sequenced the gene to identify the mutations.

We observed many of the same mutations repeatedly, and chose seven unique mutants to study further. When we retransformed these mutants back into yML047, we did not observe an appreciable difference in the growth phenotype compared to N391F (**Figure 5-4**, compare red and grey lines). We suspected that the strains themselves had mutated during the enrichments in 2DG. We took the strains from which we extracted the mutant plasmids, cured them of the plasmids, and verified that they were unable to grow on glucose or xylose without a heterologously expressed transporter (strains yML128-134). We then retransformed the strains with their respective mutant transporter plasmids or the N391F plasmid.

In some cases, the evolved strain grew noticeably faster than yML047, for example, yML128 grows much faster in 2DG, and in fact, the mutant transporter is slower than the N391F transporter in this strain (**Figure 5-4A and H**). In other cases, such as yML129, the best combination was to use the mutant transporter and evolved strain. Another unique phenotype we observed was that in yML130, the mutant transporter grew markedly faster in media without 2DG, but was comparable in growth rate to the N391F transporter when 2DG was added. It is unknown how, if at all, the mutations in the transporters contribute to these various phenotypes. One surprising mutation was in the aforementioned case of yML129; the G593R substitution occurs in the glycine-serine linker at the very C-terminal end of the protein that is the result of a cloning artifact and not part of the original protein sequence. This mutation is also present in two other mutants, and it seems to have an effect in one (yML132) but not in the other (yML133), so its role is unclear.



strain	yML128	yML129	yML130	yML131	yML132	yML133	yML134
plasmid	pML1463	pML1466	pML1470	pML1481	pML1502	pML1506	pML1507
nucleotide change	T1769A, A2016G	G2661A	G2010A	G1016A, A1054T, T2318G	C1921T, G2661A	G971A, G2661A	A1906G
amino acid change	S296T, E378G	G593R	G376S	A45T, E57D, L479V	T346I, G593R	G30R, G593R	Q341R

Figure 5-4. Mutated strains and transporters after evolution in 2DG. Seven strains and mutant transporters were isolated after multiple rounds of enrichment in xylose media supplemented with 2DG. Growth in xylose media with 0% (A-G) or 2% (H-N) 2DG is shown. Both the mutant and original transporter were transformed into both the evolved and original strain. Differences in growth rate can be attributed to either the strain background or the transporter depending on the variant. The specific nucleotide and amino acid changes for each variant are also shown.

5.2.5 Whole genome sequencing of evolved strains

Because it became apparent that the mutations in *HXT5* alone were not the only cause of phenotypic changes, we wanted to identify the background mutations that were occurring in the strains. Additionally, we were hoping to identify the genetic cause of the change between strains yML016 and yML047 that enabled growth on xylose. Therefore, we sequenced the genomes of KY73, yML016, yML047, and yML128-133.

Preliminary identification of single nucleotide polymorphisms (SNPs) revealed only a few unique mutations in the sequenced strains. There were three potentially interesting mutations found in strain yML047 and not in yML016; and there were two additional unique SNPs found in yML131 and yML133.

In yML047, one SNP was found between *SSA1*, a chaperone in the HSP70 family and *EFB1*, a translation elongation factor (the SNP was located 655bp upstream of *SSA1* and 78bp upstream of *EFB1*). A second SNP introduced an early stop codon at residue Y229 in *RPP0*, an essential gene encoding the P0 ribosomal stalk protein. A third SNP introduced a frame shift mutation at residue S192 in *ULP1*, an essential protease involved in sumoylation. Whether or not any of these three SNPs is responsible for enabling growth on xylose may be the subject of further study.

In the six adapted strains, yML128-133, only two unique SNPs were identified. In strain yML131, a frame shift mutation was found in *STO1*, a nuclear cap-binding complex protein involved in mRNA degradation. In strain yML133, a different frame shift mutation was found in *SGF73*, a subunit of the deubiquitinating module of the SAGA and SLIK complexes. Again, the potential role of these mutations in the changes to growth on xylose and xylose with 2DG is unknown.

5.3 Discussion and Future Directions

Engineering sugar transport specificity is a difficult problem, evidenced both by the challenges faced during this project and the numerous efforts made by other groups. Although some progress has been made, we are still far from realizing a highly efficient and robust co-fermentation of glucose and xylose. Although we and other groups have designed elegant selection strategies, it is naïve to expect any selection to be perfect and that cells will not find their way around the pressures we apply. Instead, we must extract what knowledge we can from these evolutions to inform the next iteration of study, even if it deviates from the original course. Thus, in addition to continuing to evolve the transporter itself, it may be prudent to look elsewhere in the genome for mutations that could contribute to improved transport and/or metabolism.

Based on previous work by others, and our own confirming experiments, it is clear that residue N391 is critical for dictating sugar specificity of *HXT5* (and the homologous residues in other transporters). Given the enormity of the effect that this single residue change has on specificity, there are two potential—though not mutually exclusive—routes to take moving forward. First, apply the same mutation to a suite of transporters, not just from *S. cerevisiae* but also from other organisms. How does this mutation affect transporters with different basal specificities and activities? Second, using the mutant as the new baseline, perform site-saturation mutagenesis at each individual residue to try and identify other hotspots in the gene. It is possible that synergistic mutations may not have been evident without the initial N391F mutation, and it is unlikely that a double mutant would arise in a short evolutionary timescale.

It is understandably tempting to isolate transport as the sole bottleneck in co-fermentation of glucose and xylose, and it was in fact the premise of our original study. However, mutations are constantly occurring throughout the genome, and exploring these SNPs further may prove to be more interesting scientifically than evolving the transporter itself. Repeating the adaptation and sequencing the resulting strains may validate some SNPs if the same ones arise independently. Additionally, rationally examining the SNPs we already identified by reintroducing them into a fresh strain would demonstrate their effects, if any. The three SNPs found in yML047 are of particular interest.

The first SNP in the non-coding region between *SSA1* and *EFB1* may alter the expression of either or both of those genes. *SSA1* is a member of the HSP70 family of chaperones, and a change to its expression may be a general response to the stress experienced by the cell during the adaptation. *EFB1* is a translation elongation factor, and changes to its expression could impact global protein expression.

The P0 ribosomal stalk protein encoded by *RPP0* is the core of a pentameric complex with P1α/P2β and P1β/P2α heterodimers. The truncation introduced by the SNP eliminates the binding site for one of these heterodimers[13], but it was previously shown that cells with this truncation are still viable, although they have a considerable growth defect[14]. It is possible that this SNP was actually deleterious and only carried through by chance; in which case, reverting back to wild type may increase the growth rate of the strain considerably. However, strains lacking the P1/P2 proteins and only containing the (full-length) P0 protein have been shown to have differentially expressed genes[15], suggesting this complex has a role in translational regulation. Thus, it may instead be that this SNP acts as a global regulator of gene expression that improves xylose metabolism, albeit with a general reduction in growth rate.

The most interesting mutation is the frameshift in *ULP1*. *ULP1* is an essential gene, required for progression through the cell cycle[16], so it is surprising that the strain is viable with this disruption. However, a previous study disrupted part of the N-terminal nuclear pore complex (NPC) targeting domain and found that cells were still viable[17]. This same study was investigating the effects of Ulp1p localization on regulation of *GAL1* transcription. The authors found that by delocalizing Ulp1p away from the NPC (either by removing the targeting domain or deleting the binding partners, *MLP1* and *MLP2*), the derepression kinetics of *GAL1* when switching from glucose to galactose was increased by 7-fold (although steady-state levels of *GAL1* expression were similar irrespective of Ulp1p localization). Interestingly, while the frameshift introduced by our SNP would truncate the protein due to an early stop codon, it also enables a previously out-of-frame ATG to now be used as an alternative start codon. If this were to occur, 50 nonsense residues would be translated at the N-terminus, followed by the normal residues from 192-621, the result of which would be a disrupted NPC targeting domain, but intact nuclear export sequence and catalytic domain. Therefore, it is possible that the SNP is causing delocalization of Ulp1p, enabling faster responses in transcription when changing carbon sources. This effect should be relatively straightforward to test as Texari et al found that the *ULP1* targeting mutant had a gain-of-function dominant effect, and so simply transforming an additional, truncated copy of *ULP1* into yML016 may reproduce the phenotype. Furthermore, this mutant could be tested in an optimized BY4741-derived xylose utilizing strain to see whether the length of the diauxic shift from glucose to xylose can be shortened.

Finally, it is unclear how the mutations in *STO1* and *SGF73* can so dramatically improve the growth on xylose with 2DG of strains yML131 and yML133, respectively. Both genes are involved in regulating gene expression and could therefore have global effects, which would make it difficult to identify the causative gene(s) whose change in expression enhances growth. Additionally, while there are clear differences in growth rate for the remaining adapted strains, no other unique SNPs were identified. A second pass at SNP-calling with different parameters may be warranted to find the missing mutations.

In summary, we have developed a strain of yeast that can grow on xylose as a sole carbon source in a transporter-dependent fashion. We used this strain to try to evolve a native hexose transporter to alter its specificity towards xylose, but there remains work to be done in that effort. Finally, we sequenced the genomes of a number of these evolved

strains and identified SNPs that arose during the evolutions. It is unknown whether and how these mutations contribute to improved consumption and/or transport of xylose. This preliminary work should inform further investigation into strategies for engineering a yeast strain that is capable of efficiently co-consuming glucose and xylose as a step towards lignocellulosic fermentation for the production of biofuels.

5.4 Materials and Methods

Strains and growth media

The base *S. cerevisiae* strain used in this work is KY73 (*MAT*a *hxt1*Δ*::HIS3::*Δ*hxt4 hxt5::LEU2 hxt2*Δ*::HIS3 hxt3*Δ*::LEU2::hxt6 hxt7::HIS3 gal2*Δ*::DR* ura3-52 his3-11,15 leu2-3,112 MAL2 SUC2 GAL MEL*). Rich media was used for preparing cells for transformation: 1% (w/v) Bacto Yeast Extract (Fisher Scientific), 2% (w/v) Bacto Peptone (Fisher Scientific), 2% (w/v) Maltose (Amresco). Rich media was supplemented with 200mg/L G418 to select for the xylose utilization pathway in constructing strain yML016. Transformants carrying transporter plasmids were selected on 2% agar plates using synthetic media containing 2% (w/v) Maltose, 0.67% (w/v) Yeast Nitrogen Base without amino acids (VWR International), 0.2% (w/v) Drop-out Mix Synthetic Minus Uracil w/o Yeast Nitrogen Base (US Biological).

Growth experiments were conducted in synthetic media containing 2% (w/v) sugar (one of Maltose, Dextrose (Fisher Scientific), or Xylose (Sigma)), 0.67% (w/v) Yeast Nitrogen Base without amino acids, 0.2% (w/v) Drop-out Mix Synthetic Minus Uracil w/o Yeast Nitrogen Base, 1X Penicillin/Streptomycin (Amresco).

Growth experiments and library enrichments using 2-deoxy-D-glucose were conducted in synthetic media as described above, supplemented with 0.002%-2% (w/v) 2-deoxy-D-glucose (Sigma).

Counterselection against transporter plasmids for curing strains was performed by growing in rich maltose media, then plating onto 2% agar plates containing 2% (w/v) Maltose, 0.67% (w/v) Yeast Nitrogen Base without amino acids, 0.2% (w/v) Drop-out Mix Complete w/o Yeast Nitrogen Base (US Biological), 1g/L 5-fluoroorotic acid (Zymo Research Company).

All standard cloning was performed using chemically competent TG1 *E. coli*. Library cloning was performed using TransforMax EPI300 electrocompetent *E. coli* (Epicentre). Transformed cells were selected on Lysogeny Broth (LB) with antibiotics (ampicillin, chloramphenicol, or kanamycin).

Yeast transformations

Individual yeast colonies were grown to saturation for 36-48hrs in YPM, then diluted 1:100 in 50mL of fresh media and grown to OD600~0.8 (between 8-12hrs). Cells were pelleted and washed once with water and twice with 100mM Lithium Acetate (Sigma). Cells were then mixed by vortexing with 2.4mL of 50% PEG-3350 (Fisher Scientific), 360μL 1M

Lithium Acetate, 250μL salmon sperm DNA (Sigma), and 500μL water. DNA was added to 350μL of the transformation mixture and incubated at 42C for 25min. When selecting for uracil prototrophy, the transformation mixture was pelleted, resuspended in water, and plated directly onto solid agar plates. When selecting for G418 resistance, the transformation mixture was pelleted, resuspended in YPM, incubated at 30C for 3-4hrs with shaking, pelleted and washed with water, then plated onto solid agar plates.

Library transformations were performed by scaling up the amount of DNA and transformation mixture used by ten-fold. Only a fraction was plated onto solid media, and the bulk of the transformation was transferred into 100mL synthetic maltose media lacking uracil and grown until saturated. The library culture was then diluted into xylose media.

Higher transformation efficiencies could be attained by isntead using a Frozen-EZ Yeast Transformation II Kit (Zymo Research Company).

Growth curves

Individual yeast colonies were picked and grown in 400μL of synthetic maltose media in 96-deep-well blocks at 30C in an ATR shaker, shaking at 750RPM until saturated. Cultures were diluted 1:100 in 150μL of fresh media (maltose, glucose, or xylose) in 96-well CELLSTAR Tissue Culture Plates (Greiner Bio-One) and sealed with Breathe-Easy film (USA Scientific). Plates were incubated at 30C with shaking in a TECAN Sunrise plate reader for 24-96hrs and absorbance measurements at 600nm were taken every 15 minutes.

Error-prone PCR

Error-prone PCR of *HXT5* was performed using a GeneMorph II Random Mutagenesis Kit (Agilent Technologies). To increase template concentrations (in order to reduce the error rate to ~1-3 per clone), we first performed standard, high-fidelity PCR using Q5 DNA polymerase (NEB). The error-prone PCR product was cloned into a vector, transformed into EPI300 cells, and plated on large agar plates (Nunc Low Profile BioAssay Dish 241mm, Fisher Scientific). Colonies were scraped, pooled, and maxi prepped (Qiagen). The mutation rate was estimated by retransforming the library into fresh *E. coli*, miniprepping individual clones, and sequencing the gene.

Library enrichments

Yeast transformed with *HXT5* libraries were grown in synthetic xylose media lacking uracil until saturated. Once saturated, cultures were diluted 1:100 into fresh media containing 0.2% 2DG. Enrichment cycles were repeated, increasing the 2DG concentration to 0.5%, 1%, and finally 2%.

Genome sequencing library preparation

Genomic preps of yeast strains were performed using a YeaStar Genomic DNA Kit (Zymo Research Company).

Table 5-1. List of barcodes used for genome sequencing.

Genome sample	Barcode Sequence
KY73	CGATGT
yML016	TTAGGC
yML047	TGACCA
yML128	ACAGTG
yML129	GCCAAT
yML130	ACTTGA
yML131	TAGCTT
yML132	CTTGTA
yML133	AGTCAA

Genomic preps were prepared for next generation sequencing using a NEBNext Ultra DNA Library Prep Kit for Illumina (NEB) with an average fragment size of 650bp. The adapter sequence used was AGATCGGAAGAGCACACGTC, and the barcode sequences for each genome sample can be found in **Table 5-1**. Sequencing was performed on a HiSeq 2500 (Illumina) in Rapid Run Mode for 100bp paired-end reads. Identification of SNPs was performed as described previously[18].

5.5 References

1. Subtil, T. & Boles, E. Competition between pentoses and glucose during uptake and catabolism in recombinant Saccharomyces cerevisiae. *Biotechnol Biofuels* **5,** 14 (2012).
2. Ha, S.-J. *et al.* Engineered Saccharomyces cerevisiae capable of simultaneous cellobiose and xylose fermentation. *Proc Natl Acad Sci USA* **108,** 504–509 (2011).
3. Sedlak, M. & Ho, N. W. Y. Characterization of the effectiveness of hexose transporters for transporting xylose during glucose and xylose co-fermentation by a recombinantSaccharomyces yeast. *Yeast* **21,** 671–684 (2004).
4. Saloheimo, A. *et al.* Xylose transport studies with xylose-utilizing Saccharomyces cerevisiae strains expressing heterologous and homologous permeases. *Appl Microbiol Biotechnol* **74,** 1041–1052 (2006).
5. Du, J., Li, S. & Zhao, H. Discovery and characterization of novel d-xylose-specific transporters from Neurospora crassa and Pichia stipitis. *Mol. BioSyst.* (2010). doi:10.1039/c0mb00007h
6. Young, E., Alper, H. S., Tong, A., Bui, H. & Spofford, C. Rewiring yeast sugar transporter preference through modifying a conserved protein motif. *Proc Natl Acad Sci USA* 201311970 (2013). doi:10.1073/pnas.1311970111
7. Farwick, A., Bruder, S., Schadeweg, V., Oreb, M. & Boles, E. Engineering of yeast hexose transporters to transport D-xylose without inhibition by D-glucose. *Proc Natl Acad Sci USA* **111,** 5159–5164 (2014).
8. Nijland, J. G. *et al.* Engineering of an endogenous hexose transporter into a specific D-xylose transporter facilitates glucose-xylose co-consumption in Saccharomyces cerevisiae. *Biotechnol Biofuels* **7,** 168 (2014).
9. Wieczorke, R. *et al.* Concurrent knock-out of at least 20 transporter genes is required to block uptake of hexoses in< i> Saccharomyces cerevisiae</i>. *FEBS Lett* (1999). doi:10.1016/S0014-5793(99)01698-1
10. Kruckeberg, A. L., YE, L., BERDEN, J. A. & van DAM, K. Functional expression, quantification and cellular localization of the Hxt2 hexose transporter of Saccharomyces cerevisiae tagged with the green fluorescent protein. *Biochem J* **339,** 299–299 (1999).
11. Randez Gil, F., Blasco, A., Prieto, J. A. & Sanz, P. DOGR1 and DOGR2: Two genes from Saccharomyces cerevisiae that confer 2-deoxyglucose resistance when overexpressed. *Yeast* **11,** 1233–1240 (1995).
12. Sun, L. *et al.* Crystal structure of a bacterial homologue of glucose transporters GLUT1-4. *Nature* **490,** 361–366 (2012).

13. Krokowski, D. *et al.* Yeast ribosomal P0 protein has two separate binding sites for P1/P2 proteins. *Mol Microbiol* **60,** 386–400 (2006).
14. Santos, C. & Ballesta, J. P. The highly conserved protein P0 carboxyl end is essential for ribosome activity only in the absence of proteins P1 and P2. *J Biol Chem* **270,** 20608–20614 (1995).
15. Remacha, M. *et al.* Ribosomal acidic phosphoproteins P1 and P2 are not required for cell viability but regulate the pattern of protein expression in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **15,** 4754–4762 (1995).
16. Li, S. J. & Hochstrasser, M. A new protease required for cell-cycle progression in yeast. *Nature* **398,** 246–251 (1999).
17. Texari, L. *et al.* The nuclear pore regulates GAL1 gene transcription by controlling the localization of the SUMO protease Ulp1. *Mol Cell* **51,** 807–818 (2013).
18. Ryan, O. W. *et al.* Selection of chromosomal DNA libraries using a multiplex CRISPR system. *Elife* **3,** (2014).