Copyright

by

Nathan Charles Crook

2014

The Dissertation Committee for Nathan Charles Crook Certifies that this is the approved version of the following dissertation:

Novel Approaches for Metabolic Engineering of Yeast at Multiple Scales

Committee:

Hal Alper, Supervisor

Lydia Contreras

Andrew Ellington

George Georgiou

Jennifer Maynard

Novel Approaches for Metabolic Engineering of Yeast at Multiple Scales

by

Nathan Charles Crook, B.S.C.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2014

Dedication

To my family and friends

Acknowledgements

The success of the work presented here is due in large part to the support I received from several extraordinary individuals. Firstly, I would like to thank Hal for his unfailing enthusiasm about my work. He has supported my ideas (even when they are crazy), guided me through the process of getting them funded, and cheered me on when experiments failed. He has given thorough, helpful advice about my writing and presentations and has been incredibly valuable during my search for a postdoc. Importantly, he has put up with my personality quirks and was rather amused by his office getting filled with balloons and the sudden appearance of a throne made of pipette tip boxes. When I was deciding my PhD advisor, my friend gave me this advice: "Choose the type of advisor you want to be." I will have done very well if I end up like Hal.

I am deeply indebted to my colleages in the Alper Lab. Jie worked with me on the ICE project, and her work ethic, creativity, and attention to detail have been the direct cause of many of the breakthroughs during this work. She undertook the majority of the strain engineering, carried out the evolution experiments, and developed robust methods of isolating mutants from the evolved strains. She is a model researcher, and has taught me the majority of what I know about mentoring others. She became much more than a colleague to me, and has been my source of strength and joy when times got tough. Kate worked with me on the weak promoter engineering projects (she was engineering the NUP57 and TFC promoters) and on the nucleosome project (she did all the experiments, I did the computations), and her discussions with me about all manner of topics related to metabolic engineering have been productive and insightful. Kate is an outstanding

scientist and a true friend. Joe's optimism, dedication, and insight have buoyed the ICE project when things got tough. He improved the capabilities of ICE through reverse transcriptase engineering, gene overexpressions, and reduction of retroelement copy number. I am confident ICE will remain in good hands after I graduate.

No graduate student can complete the staggering amount of work required to complete a PhD on their own. I have had the privilege of mentoring some very talented undergraduates, who have not only been extremely helpful, but have also been a source of inspiration on the projects they've undertaken. Liz was my first undergrad, and was extremely helpful during the massive amount of cloning required for the MCS project despite my inexperience as a mentor. 4 years later, I still use a technique she developed for doing confirmation digestions using a minimum amount of reagents! Heming, Andrea, and Divya were very helpful and hardworking during the promoter engineering project and helped us more quickly understand what was going on with these weak promoter variants. Alex, as my last undergrad, enabled me to do many very large experiments during my last year to push my projects to completion. The IRES, RNA interference, and P2A projects have enormously benefited from his dedication and insight.

I have also been very fortunate to be a member of a community which was willing to provide me with techniques and facilities if I needed them for my projects. Johnny and Grant taught me how to do western blotting, Joe Taft gave me some very helpful advice with python, Dr. Marcotte graciously provided us with many of the knockouts needed for my work, and Dr. Iyer was a great help in figuring out what was going on in the initial stages of the MCS project.

My thesis committee has been an invaluable resource during my studies. I would like to especially thank Dr. Maynard for introducing me to the fascinatingly quirky 2A

site – the more I learn about it the more I want to engineer it! Dr. Contreras provided excellent advice on the construction of cDNA libraries for the RNAi project, and graciously opened her lab to me during the ICE and P2A projects. Drs. Georgiou and Ellington were not only extremely insightful during the conception of these projects, but they also gave me the impetus to complete the ICE project by betting me $100 each that it wouldn't work. It's probably the most hard-earned $200 I've ever made, and also the most rewarding.

Without the friends I made at UT, I probably would have gone crazy. Lynn, Erwan, Zach, and James have been my support when times were tough, and have made the good times amazing. Lynn and Erwan, our month in China was unforgettable. Let's do it again in France! Zach, at some point among the TNSRs, camping, bar golf, bingo, and riding our bikes on the upper deck of I-35 at 3am with a bunch of crazy hipsters, Austin became home. James, nothing takes care of frustrations in the lab like scoring a critical hit on a mindflayer in a party composed of a dinosaur with 10ft arms, a reluctant succubus, and an eco-terrorist druid.

Novel Approaches for Metabolic Engineering of Yeast at Multiple Scales

Nathan Charles Crook, PhD

The University of Texas at Austin, 2014

Supervisor: Hal Alper

Living systems contain enormous potential to solve many pressing engineering problems, including the production of usable energy, the synthesis and degradation of a variety of materials, and the treatment of disease. Metabolic engineering, as one approach to harness this potential, treats the behavior of a living system as the combined product of multiple interacting modules, each of which can be tuned to maximize performance. However, the scarcity of techniques for predictive or high-throughput engineering design of these modules, especially in eukaryotes, contributes to long strain development times and high research cost. In this work, we develop several new tools to expand our capabilities for predictive design and high-throughput engineering in yeast. At the transcriptional level, we develop a method which, for the first time, enables predictive strengthening endogenous yeast promoters and also the *de novo* design of strong synthetic promoters. At the translational level, we show that it is possible to exploit the context resulting from the arrangement of DNA parts in order to predictably increase or decrease gene expression. We also develop a powerful new approach for directed evolution of enzymes in yeast, termed *in vivo* continuous evolution, which enables the creation of library sizes orders of magnitude larger than can be obtained with

the current state of the art using significantly less labor. Finally, we harness the programmatic inhibitory potential of RNA interference to optimize and demonstrate a system for rapid strain engineering with minimal genomic editing. Taken together, this work provides new techniques which enable a significant reduction in the development time of new yeast strains and informs future development of new tools for metabolic engineering.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Living systems exhibit many useful phenotypes, including the synthesis of a wide variety of compounds (1), the extraction of energy from diverse substrates (2), and the degradation of toxins (3). In many cases, the capabilities and potential of living things to solve important problems have no equal in man-made, nonliving systems. Therefore, there has been significant interest in developing a greater understanding of the mechanisms by which organisms achieve these interesting behaviors as well as in designing modified organisms that exhibit improved functionality. As one way to meet this objective, metabolic engineering aims to develop novel phenotypes by applying engineering strategies and formalizations to living systems. These efforts have played a pivotal role in the development of strains which convert renewable substrates into high quantities of useful compounds (4-8) and which exhibit synthetic behaviors (9-12). Metabolic engineering treats organisms at multiple layers of complexity, each of which coordinately determine the behavior of the cell. These layers can be at the basic level of regulation of transcription or translation, but they can also represent more complicated systems, such as enzymatic activity or even networks of interacting proteins (**Figure 1-1**). A deep understanding and the ability to engineer of each of these layers is critical to the development of living systems which effectively solve humanity's problems. Here, current techniques for engineering cellular behavior at each of these levels will be briefly reviewed.

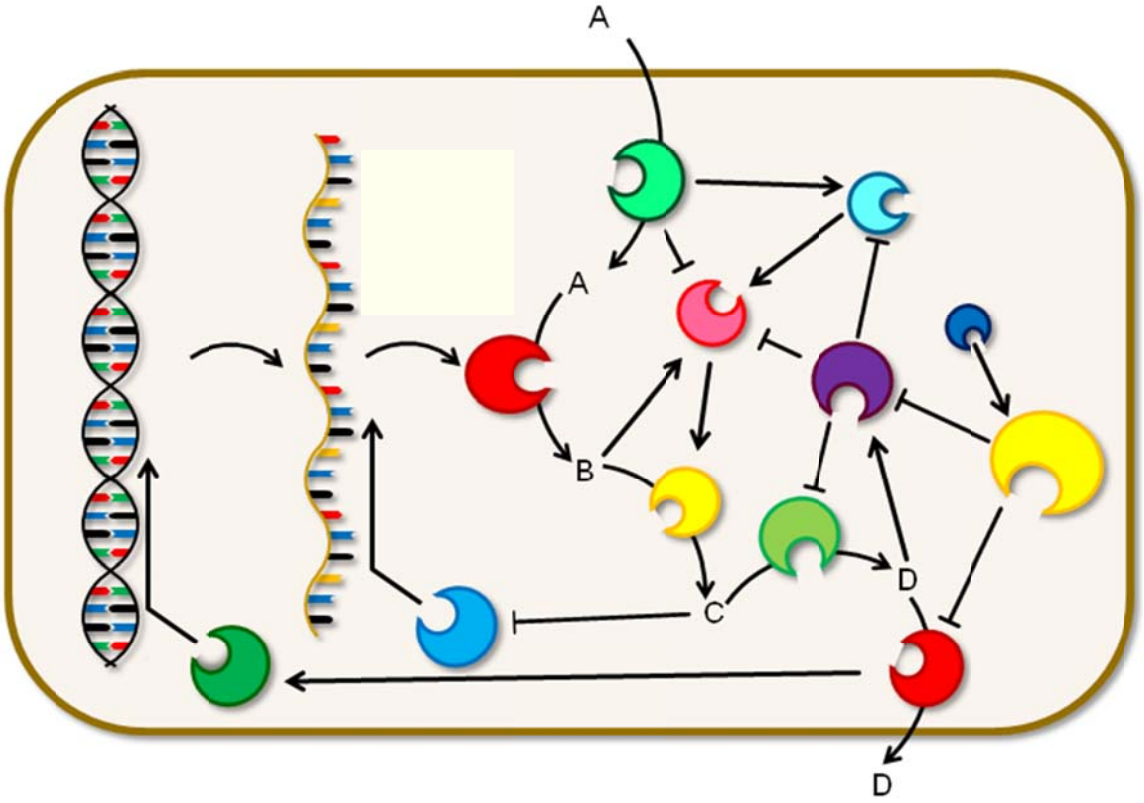


Figure 1-1: Multiscale metabolic engineering

The cell can be viewed as a hierarchical collection of interacting modules, each of which can be engineered to control cell phenotype. These modules can include transcription, translation, enzymatic activity, and strain background.

1.1 Engineering Transcriptional Machinery

As the first step towards the production of an enzymatic catalyst by a living system, the importance of understanding and controlling the transcription process cannot be overstated. Transcription rate and timing for each gene are controlled by promoters, which are DNA regions upstream of the gene's open reading frame. Therefore, well-characterized, robust transcriptional regulators/promoter elements with diverse behaviors are highly sought-after parts in the metabolic engineer's toolbox (13). At first approximation, transcriptional regulation is achieved through recruitment of transcription

factors and RNA polymerase. These transcription factors tend to dictate the amount of RNA produced as well as its timing in relation to environmental cues and cellular processes. Many transcription factors have known DNA binding sites, prompting the development of several databases linking sequence motifs to transcription factors in several organisms (14-18). These databases enable engineers to create novel promoters by combining transcription factor binding sites in a modular fashion. Indeed, Blazeck, et al. developed a series of synthetic hybrid promoters by successively adding upstream activating sequences to a core promoter region, and it was shown that the effect of subsequent additional upstream activating sequences is well-described by a cooperative hill function (19-21). Further, Amit, et al. developed a predictive model for the behavior of bacterial promoters through a statistical thermodynamics approach (22). In this study, every potential state of the promoter (transcription factors bound/unbound, enhancer proteins in proximity to the RNA polymerase, etc.) was assigned a weight corresponding to its free energy to compute a partition function and the Boltzmann distribution was used to compute the fraction of promoters in an active state. For this model, higher active fractions corresponded to more active promoters. In addition to work demonstrating the importance of the transcription factors themselves, previous studies have also demonstrated both the importance of chromatin structure in regulating the accessibility of transcription factor binding sites (23) as well as the capacity to alter transcription rates by modifying nucleosome binding sequences (24). These rational approaches, coupled with part-mining from natural systems (25) as well as diversification of native promoters through mutagenesis (26,27) have given researchers a large toolkit for developing new promoters, although well-characterized promoters of low strength have not yet been developed, and *de novo* design of promoters remains a challenge in eukaryotes. For model systems such as *E. coli,* formalizations have been developed which enable

researchers to measure and characterize promoters in a standardized fashion, leading to the development of online repositories of standard parts (28). Therefore, there are currently multiple avenues which a metabolic engineer can pursue in order to enable desired levels of transcript production, yet more work is needed to in order to develop large libraries of promoters which span the full range of gene expression in yeast.

1.2 Engineering Cellular Behavior at the Translational Level

In order to realize the promise of synthetic biology, researchers are constructing ever more complex genetic architectures. In these constructs, multiple DNA parts are assembled in order to achieve a desired outcome. As important as the gene products and their associated regulatory machinery are in these engineered systems, method by which these DNA parts are assembled into a functional construct is equally so. In particular, issues such as mRNA structure and construct homology can greatly impact the yield and functionality of synthetic constructs through effects at the translational level, thus prompting metabolic engineers to develop methods which mitigate or exploit these effects.

1.2.1 Regulating Translation through RNA Structure

RNAs often undergo intra- and intermolecular hybridization reactions to form structures which may inhibit or enhance gene expression through interference with the ribosome. In contrast to the complex DNA-protein and protein-protein interactions which determine promoter activity, these RNA structures are formed through nucleic acid hybridization, which has a mature modeling framework. Within this framework, it is necessary to enumerate nearly all possible secondary structures and their energies by making use of the well-known free energies of hybridization between nucleotides (29). This set of states may then be used to compute a partition function, enabling the

calculation of the most probable free energy structures and ensemble free energies. Several software packages enable prediction of DNA and RNA secondary structure as well as sequence design to achieve a user-defined secondary structure (30,31). However, these programs limit treatment of pseudoknotted structures to short nucleic acids due to the computational complexity of enumerating all structures for nucleotides above 100 bp in length. Nevertheless, these thermodynamic models can enable the design of nucleic acids with defined secondary structures, which may be used to occlude ribosome binding sites, impede ribosome scanning, or respond to the presence of small molecules in solution.

In addition, a myriad of RNA-based regulators have been developed (32). In particular, Isaacs, et al. (33) developed "riboregulators" which enable post-transcriptional tuning of gene expression through competitive binding to a ribosome binding site (RBS)-occluding stemloop. In this study, prediction tools such as mfold (34) enabled the design of riboregulators with stable secondary structures. By varying the concentration of a small trans-activating RNA, the expression of a gene encoded by the bound transcript could be modulated independently of promoter strength. Ribosome binding site occlusion by mRNA secondary structure represents a simple way to modulate gene expression. In fact, several groups have leveraged RNA structure prediction tools to develop programs which can design ribosome binding sites which are occluded by secondary structure to a prescribed extent, enabling a user-defined control of translation (35,36) in prokaryotes. However, gene expression modulation by mRNA secondary structure is not limited to this kingdom. As eukaryotic translation initiation is dependent upon ribosomal "scanning" along the 5' untranslated region (5'UTR) for the start codon, any secondary structure in this region will pose a barrier to gene expression. Thus, the same thermodynamic RNA structure prediction tools may be used in eukaryotes to design

cloning elements present in the 5'UTR to minimize this inhibitory secondary structure (37).

In addition to tuning levels of gene expression in a static sense, RNA structure may be used to regulate gene expression in response to the presence of compounds present inside the cell through RNA sensors called aptamers. These TNAs are capable of precisely binding a small molecule and subsequently undergoing a conformational change to elicit a corresponding change in the actuation domain of this RNA structure. These structures can regulate gene expression in response to a single input (38), or may perform Boolean logic operations on multiple inputs (39,40). Finally, thermodynamic models of RNA folding have been integrated with kinetic models of transcription and translation to develop prokaryotic promoters with defined dynamic behaviors (41). Although these kinetic models require extensive experimentation to fit unknown parameters, they enable design of promoters with time-dependent behaviors and are reminiscent of the models used to describe and predict the behavior of gene regulatory networks. Collectively, these strategies enable researchers to tune gene expression through modulation of translation rate.

1.2.2 Streamlining Eukaryotic Translation through the use of 2A peptides

2A peptides are 22 amino acid "self-cleaving" sequences discovered in picornaviruses which enable the production of physically separated protein products from genes which are encoded in the same open reading frame (42), thus enabling co-regulation of several gene products without the necessity for the addition of promoter and terminator sequences between each gene. The defining feature of a 2A site is a proline-glycine-proline sequence at the polypeptide cleavage site. This unique peptide sequence induces stalling of the translating ribosome during the synthesis of the glycyl-prolyl

peptide bond, presumably due to steric hindrance. This allows hydrolysis of the nascent polypeptide chain at the glycine-proline junction, releasing the first protein product and enabling the ribosome to continue translating the second (43). These sites have been used for engineering applications in mammalian cells and in plants (but interestingly, not in prokaryotes) (42), but reports of its use for biotechnological objectives in yeast are sparse (44).

1.2.3 Development of Polycistronic Expression Cassettes in Eukaryotes through Internal Ribosome Entry Sites

Internal ribosome entry sites (IRESs) are cis-encoded elements which direct ribosome binding to the interior of a polycistronic transcript. These elements can be derived from RNA viruses which infect mammalian hosts and enable translation of viral RNAs in a cap-independent fashion (45), or they can be derived from endogenous mRNAs which enable translation in a cap-independent manner, which is thought to be a mechanism to maintain protein production during stress (46). It is hypothesized that the ability to direct eukaryotic translation machinery to the interior of an mRNA is dependent on the unique secondary structure that an internal ribosome entry site adopts. In particular, it has been shown that viral IRESs adopt structures that resemble the canonical tRNA cloverleaf structure (as is the case with the encephalomyocarditis virus (47-49)) or even pseudoknotted structures that resemble the anticodon loop of a tRNA bound to its cognate mRNA substrate (as is the case for IRESs derived from the *dicistroviridae* family (50)). However, the sequence, structural, or mechanistic determinants of IRES function have yet to be definitively elucidated.

Nevertheless, the unique ability of an internal ribosome entry site to enable translation of a polycistronic mRNA in eukaryotes has spurred interest in its use as a tool for biotechnology. In mammalian cells, high-level production of a gene product is often

enabled by the use of a polycistronic expression cassette consisting of the gene of interest in tandem with a selection marker. In this scheme, culture in selective media permits only the growth of those cells which express the selection marker, and hence, the gene of interest. In addition, the ability of IRESs to initiate translation at an uncapped mRNA has enabled researchers to use dedicated bacterial RNA polymerases (such as T7) to transcribe a gene of interest in a non-prokaryotic host (51). Although bacterial RNA polymerases do not produce capped RNAs, the presence of an IRES enables translation to occur. Not only does this system enable high transcript production, but it is also orthogonally regulated, thus minimizing inhibition from the cell's native machinery. For metabolic engineering applications, the use of IRES elements to express a synthetic pathway would enable pathway components to be co-regulated in a facile manner, thus reducing waste associated with functionally redundant DNA as well as reducing the risk of homologous recombination-associated construct instability. Furthermore, the variable efficiency of IRES elements would enable tunable expression of each component of a synthetic pathway while each maintaining the same induction or repression responses. Therefore, there has been much interest in the development of an IRES for *S. cerevisiae*, as this organism is a platform for industrial biotechnology yet its potential to utilize known internal ribosome entry sites remains underdeveloped (52). Additionally, although several publications have reported the identification of internal ribosome entry sites which function (at least to a small extent) in yeast (53), these sites have not been utilized in a metabolic engineering context.

1.3 ENGINEERING PROTEINS

The properties of enzymatic machinery are of fundamental importance to the overall efficiency and economics of a bioprocess. Although maximum productivity can

be greatly enhanced through optimization of the quantity and timing of gene expression and the availability of substrate, yield of product per substrate consumed is a major factor driving cost, and increases to yield come through optimizing the properties of the biocatalyst. Catalytic rate, substrate promiscuity, product specificity, and cofactor requirements are of critical importance to the design of any catalyst, and enzymes are no exception. Hence, the development of tools for facile, rational enzyme engineering is an area of intense research (54-56). Several packages exist to assist in predicting the effects of small changes to an existing protein structure, the most famous of which is ROSETTA (57) which has found wide use suggesting potential routes for enzyme engineering (58-60). In addition, Fold-It is a user-friendly program for protein structure optimization, and also leverages humans' ability to recognize patterns by crowd-sourcing difficult folding problems (61). This approach has succeeded in determining increasing the Diels-Alderase activity of an enzyme (62). Despite the relative dearth of predictive models, several groups have reported outstanding successes in rationally re-engineering enzymes for novel functions and also developing enzymes *de novo*. Milestones include developing novel ligand-binding capabilities (63) and expanded activity on non-native substrates (64) through docking simulations, novel enzyme function through engineering a catalytic site to stabilize a highly divergent transition state complex (65,66), and completely novel enzymes by stabilizing metal ions on an artificial scaffold (67). It is hoped that increases in computational power will soon enable routine design of novel, specific, and active biocatalysts.

Although techniques for *ab initio* modeling (61,68), as well as semi-rational procedures such as site-directed mutagenesis (69-72) and domain shuffling (73-75) have seen outstanding success in the development of proteins with improved functionality, these techniques require detailed structural information. Additionally, accurate *de novo*

prediction of enzyme function requires exquisitely precise molecular modeling (and thus computational prowess outside the grasp of most synthetic biology labs), so enzymes are commonly engineered through "nonrational" procedures such as directed evolution (76). Classical directed evolution relying on error-prone PCR has seen remarkable successes in yielding improvements to enzymes (77), transcription factors (78), and regulatory elements (79). In addition, several techniques for *in vivo* library generation have been developed, most notably Phage-Assisted Continuous Evolution (PACE). PACE (80) exploits the high reproductive capacity of M13 phage to introduce mutations in a plasmid enabling pIII production in *E. coli*. This technique requires the use of a constant-flow bioreactor to apply selective pressure to the phage, and is generally limited to the evolution of biological parts which activate transcription. In addition, this technique is not applicable to the evolution of parts which specifically function in eukaryotic systems. Therefore, there is significant interest in developing a system for continuous evolution which avoids these shortcomings.

1.4 ENGINEERING METABOLIC NETWORKS

Metabolism represents the concerted effort of thousands of enzymes to synthesize thousands of molecules from substrates such as small, simple sugars and large, complex polypeptides. This highly complex, interconnected network is responsible for the diverse chemistries enabled by microbial catalysts. However, these networks contain thousands of enzymes whose behavior is often not fully characterized. Thus, synthetic biologists face significant challenges when predicting perturbations to microbial metabolism with the goal of increasing bioprocess productivity. As a result, several modeling frameworks have been developed to aid in the design of engineered metabolic networks, differing mainly in the level of detail they provide. Stoichiometric models tabulate all the

chemical reactions possible in metabolism, thermodynamic models contain information regarding the feasibility of each reaction, and kinetic models contain rate information for each enzyme. Models for a variety of organisms have been curated using portions of each of these frameworks, and several automated tools have been developed to aid in the construction and utilization of these models.

1.4.1 Models of Metabolism

Models of metabolism are essentially systematic enumerations of all metabolic reactions, detailing the ratios with which metabolites react to form products. At their heart is a large matrix which details the potential sources and sinks of each metabolite. Together with the flux through each metabolic pathway, this matrix is used to determine the change over time of the concentrations of each metabolite. In general, this framework results in a system of coupled differential equations for the concentrations of each metabolite. The distinguishing features of each modeling framework detailed below are the assumptions used to simplify and solve this often severely underdetermined system of equations.

1.4.1.1 Stoichiometric Models

Most studies employing stoichiometric models assume metabolism is at steady state, that is, the concentration of each metabolite does not change. Constraints on the permissible values of the flux through each reaction may also be specified. This results in a set of algebraic equations and inequalities which restricts the metabolic fluxes that may be observed. However, since the number of reactions in a model greatly exceeds the number of metabolites, these assumptions do not specify a unique flux profile. Thus, additional assumptions must be made to determine the state of the cell. Usually, these assumptions take the form of an objective function and optimization algorithms may be

used to solve for metabolic fluxes. Several of these objective functions have seen wide use, including the constraint that cell growth must be maximized (81) and the constraint that deviation of metabolic flux from some "base case" is minimized (82). The latter objective function is frequently used to analyze engineered strains by assuming their metabolic flux profile is similar to that of the parent strain (usually calculated by assuming maximization of growth). These objective functions are sufficient to define a unique metabolic state, including growth rate and flux through every pathway (83).

1.4.1.2 Thermodynamic Models

Models incorporating thermodynamics are similar to the stoichiometric models mentioned above, except that the free energy of each reaction is constrained to be negative (84,85). These free energies may be estimated for every reaction in the cell using the method of group contributions (86). This added constraint reduces the space of feasible fluxes and eliminates flux distributions which violate the laws of thermodynamics, such as internal flux cycles. The energy associated with moving solutes up a concentration gradient through a membrane is also considered in this analysis. Feasible flux distributions may be computed using linear optimization and the objective functions mentioned above. These modeling structures help further refine the solution space of this underdetermined system.

1.4.1.3 Kinetic Models

The most accurate description and prediction of cellular metabolism must include rate laws for each reaction in the organism. Inclusion of these rate laws transforms the algebraic system treated above into a system of coupled differential equations. It is also expected that these kinetic models will include the dynamic effects of regulatory proteins, which is not accounted for in either the stoichiometric or thermodynamic modeling

framework. Manual characterization of the kinetics associated with each metabolic component initially seems like a daunting task. However, several techniques currently exist to experimentally determine metabolic fluxes in the cell, enabling estimates of kinetic parameters to be developed. Metabolic flux analysis using radioactive tracer compounds is the most commonly used method to estimate intracellular fluxes and several algorithms to compute the resulting flux distribution have been developed (87-89). MASS (mass-action stoichiometric simulation) uses this metabolomics data to create estimates of rate constants (90), and similar techniques have been used to make estimates in yeast (91) and in mammalian cells (92). Estimates of these kinetic parameters will enable the classical technique of metabolic control analysis to be used on a genome scale (93,94) in addition to more detailed simulations of the temporal behavior of engineered cells (95).

1.4.2 Curated Models

The first models of metabolism were stoichiometric in nature and were created for common laboratory species such as *E. coli* (96) and *S. cerevisiae* (97) among other organisms (98,99) though literature search and manual annotation. Recently, metabolic models have been automatically generated and updated for many other organisms based on genome sequences and through analysis of reaction thermodynamics. Current models for *E. coli* take into account the thermodynamics of each reaction (100), and stoichiometric models of other model organisms such as *S. cerevisiae* (101) are nearing completion. For organisms which do not have a metabolic model, a collection of software packages have been developed to automate stoichiometric network reconstruction based on genomic data (102,103). In addition, these pathways can be refined by comparing *in silico* growth phenotypes to those determined experimentally

and adding or deleting the necessary reactions (104). These networks have, until recently, been curated in an *ad hoc* manner throughout the biological literature. As a consequence, the Systems Biology Markup Language (SBML) has been developed to describe metabolic networks in a standardized and machine-readable format (105).

1.4.3 Optimization of Metabolism

The utility of a metabolic model to a metabolic engineer lies in its capacity to enable the design of heterologous metabolism. Several algorithms and software packages have been developed which automate the process of strain design using the models mentioned above. Flux balance analysis enables the calculation of metabolic fluxes at equilibrium given either a stoichiometric or thermodynamic model. Thus, identifying promising engineering targets simply involves making perturbations to the host genome *in silico* and iteratively performing flux balance analysis until a specified production goal is achieved. Exhaustive search methods have been used to identify promising knockouts improving lycopene production in *E. coli* (106) and formic acid production in *S. cerevisiae* (107). A linear optimization approach to identifying promising knockouts (108) has been used to improve lactic acid (109) and 2,3-butaendiol titers (110). In addition, pathway databases such as MetaCyc (111,112), KEGG (113), BRENDA (114), and BiGG (115) have been used to suggest heterologous enzymes enabling improved 3-hydroxypropionate (116) and 1,4-butanediol production (117). Methods incorporating concepts from retrosynthesis have also been developed to assemble heterologous pathways (118-120).

Several user-friendly software packages exist to automate the process of searching for interesting metabolic perturbations. The well-known COBRA toolbox is an add-on to the MATLAB computing environment which performs flux balance analysis and

knockout identification (121,122). OptGene (123) utilizes a genetic algorithm to identify promising knockouts and has been used successfully to improve the production of vanillin (124) and cubebol (125). The Biomet toolbox automates heterologous pathway assembly by interfacing with databases of reactions (126) and CycSim (127) is a web-based tool that performs simple flux calculations in a user-friendly environment. Finally, suggested metabolic changes may be easily visualized on the metabolic network with tools such as iPATH2 (128) and GLAMM (129). Progress has moved rapidly in this area of metabolic modeling and has recently been demonstrated to be a potent approach for designing synthetic metabolic systems.

1.4.4 Engineering Metabolic Networks in a High-Throughput Manner

Applications in both synthetic biology and metabolic engineering often rely upon the ability to either partially or completely remove the activity of a gene product in order to provide living systems with the optimal catalytic repertoire to meet a certain goal. Thus, knockdown and knockout strategies which have been informed by the rational approaches mentioned above have been instrumental in rewiring microbial systems for the production of a wide variety of chemicals. However, it is often the case that the phenotype of interest is not directly linked to metabolism, and is rather a complex phenotype such as tolerance. These phenotypes are very difficult to model *in silico*, necessitating genome-wide screening *in vivo* (6,130-132). Among possible host organisms, yeasts have gained traction as a highly attractive system for the bioproduction of fuels and chemicals (133). However, current methods for elimination or reduction of endogenous gene activity in yeast remain laborious and necessitate highly sequential workflows in spite of the need to test (often many) different gene knockdown/out strategies. Strain choice can also significantly influence the yield and productivity of

industrial bioprocesses, but *a priori* strain selection is not always feasible (8). Therefore, although parallel processing of multiple yeast strains during the design-build-test cycle is highly desirable, the high cost associated with each genome modification limit the amount of possible parallelization. Furthermore, if a superior wild-type strain is identified late during process optimization, it may be extremely costly and difficult to transfer genomic modifications to the new strain using the same linearized workflow. Thus, there is a strong need for a synthetic methodology to quickly and cheaply introduce gene knockdown/outs into multiple strains of yeast in order to rapidly prototype within the design-build-test cycle.

Recent reports have demonstrated the use of small regulatory RNAs as a means for rapid, facile knockdowns in the bacterial system *Escherichia coli* (134). In higher eukaryotic systems, RNA interference (RNAi) is used to systematically target and reduce mRNA levels through the action of the RNA-induced silencing complex on double stranded RNA (135). Specifically, double stranded RNA (dsRNA) is cleaved by Dicer to form small guide RNAs, which are then used by Argonaute to recognize and degrade the corresponding mRNA in a programmatic manner (**Figure 1-2**). A major advantage of RNAi is that it is highly portable and only requires the requisite machinery (i.e. Argonaute, Dicer, and dsRNA) to be expressed– no genome engineering is explicitly required. Furthermore, the targeting dsRNA can be generated without prior knowledge of a host's genome using existing cDNA library techniques. RNAi thus enables rapid strain prototyping through synthetic import of this machinery into novel host strains. As a result, RNAi has been widely used for targeted loss-of-function studies and metabolic engineering in a wide variety of eukaryotic organisms (136-140). Despite its utility, a functional RNAi pathway is endogenously absent from common yeast hosts such as *S. cerevisiae*. Fortunately, the RNAi system can be introduced into *S. cerevisiae* through

the heterologous expression of Argonaute and Dicer from *S. castelli* (141). As a result, it is possible to use RNAi for the editing of metabolic networks in yeast.

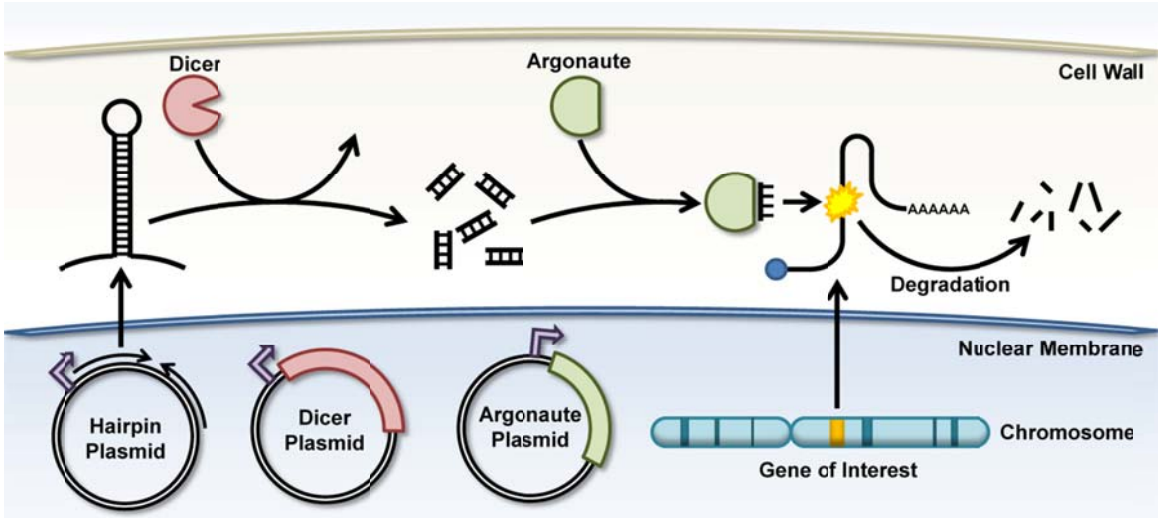


Figure 1-2: Schematic of RNA interference

Double-stranded RNA with homology to a target gene is degraded by Dicer. The resulting guide RNAs are then used by Argonaute to recognize and cleave the mRNA of the target gene.

1.5 CURRENT CHALLENGES IN METABOLIC ENGINEERING

Despite the substantial advances in our ability to engineer cells through the tools mentioned above, reliable, high-throughput methods are significantly lacking for platform eukaryotic hosts such as yeast, as evidenced by the high development times and costs associated with the construction of yeasts that produce useful compounds from renewable sources (4-8). In order to overcome this limitation, tools for strain engineering must be developed which enable 1) expedited learning and 2) quick implementation of engineering strategies informed by new knowledge. In this work, I will demonstrate the development of a myriad of new tools for engineering yeast at the level of transcription, translation, enzyme, and network level which collectively achieve these goals.

Chapter 2: Model-Based Design of Synthetic Yeast Promoters via Tuning of Nucleosome Architecture

2.1 INTRODUCTION

Synthetic biology design is ultimately constrained by our capacity to specify function of synthetic parts at the DNA sequence level. This capacity would redirect the field away from relying on a "parts-off-the-shelf" strategy and toward an approach marked by pure, synthetic design and customizable specification. Toward this end, great strides have been made to enable model-based design of cellular behavior (142) and to allow for rational design of small sequences (such as ribosome binding sites, transcription factors and enhancers) (13,19,20,22,33,35,143). Yet, pure *de novo* design of full promoters, one of the most fundamental components in synthetic circuits, remains difficult, especially in eukaryotic model organisms like yeast. Traditional approaches spanning the last decade of promoter engineering efforts (13) rely upon part-mining (144), mutagenesis strategies (26,27,145), and/or chimeric design (19,20) to identify promoter variants.

In contrast, here we present the first approach for DNA-level specification of promoter activity based on predicted nucleosome affinity. Based on previous studies demonstrating the importance of nucleosome occupancy on promoter activity (23,24), our overall hypothesis is that promoter activity can be predicted and controlled based on nucleosome architecture (**Figure 2-1**). To test this hypothesis, we made use of a previously-developed hidden Markov model to *de novo* predict nucleosome occupancy along an arbitrary DNA sequence (146). Our approach can enable both the redesign of

endogenous promoter sequences as well as the *de novo* design of synthetic promoters in a single design cycle.

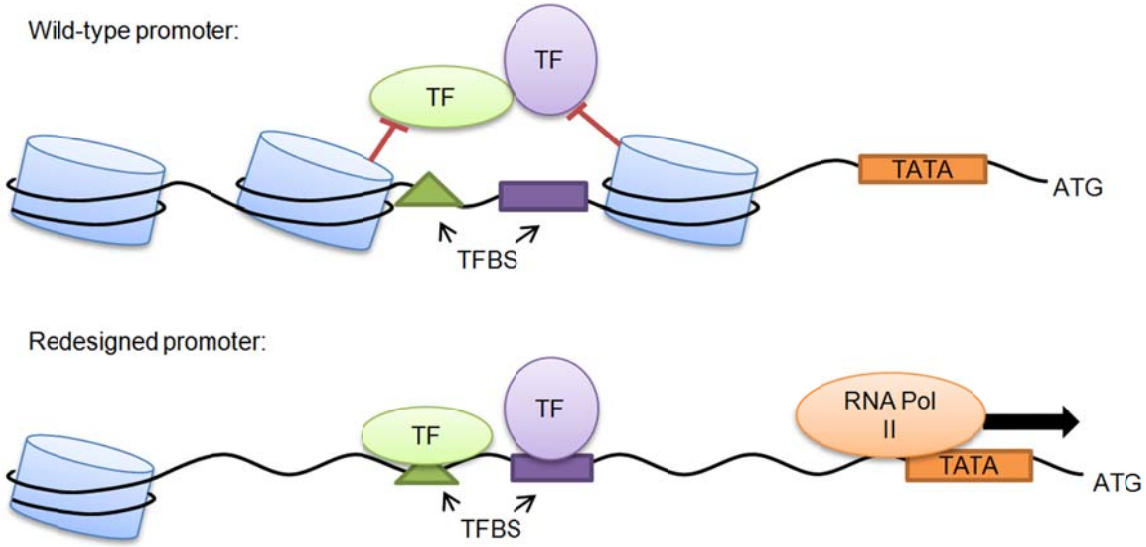


Figure 2-1: A model for promoter strength

Native promoters can be redesigned for increased strength by decreasing nucleosome affinity. Transcription factors are designated "TF" and binding sites are "TFBS."

2.2 RESULTS

2.2.1 Rational re-design of native yeast promoters

Our earliest efforts in yeast promoter engineering (26,27) relied upon large-scale mutagenesis and selection to generate a *TEF1* promoter library. This process clearly demonstrated that distributed point mutations in promoters can alter expression levels—although in most cases, lower expression than wild-type is obtained. Here, we sought to extract a design principle from this 15-member promoter library that collectively spans a 15-fold dynamic range in expression and encompasses between 5 and 71 mutations across 401 base-pairs. By evaluating predicted nucleosome affinity across the 15-member *TEF1* promoter library, we found that the cumulative sum of predicted

nucleosome affinity across the entire promoter (hereafter referred to as the "cumulative affinity score") is inversely proportional to promoter strength in a very robust, predictable manner, despite the great diversity of sequence and transcription factor binding site mutations **(Figure 2-2A,B)**. This strong correlation underpins the potential for nucleosome architecture to be used generically as a design principle for promoter engineering in yeast.

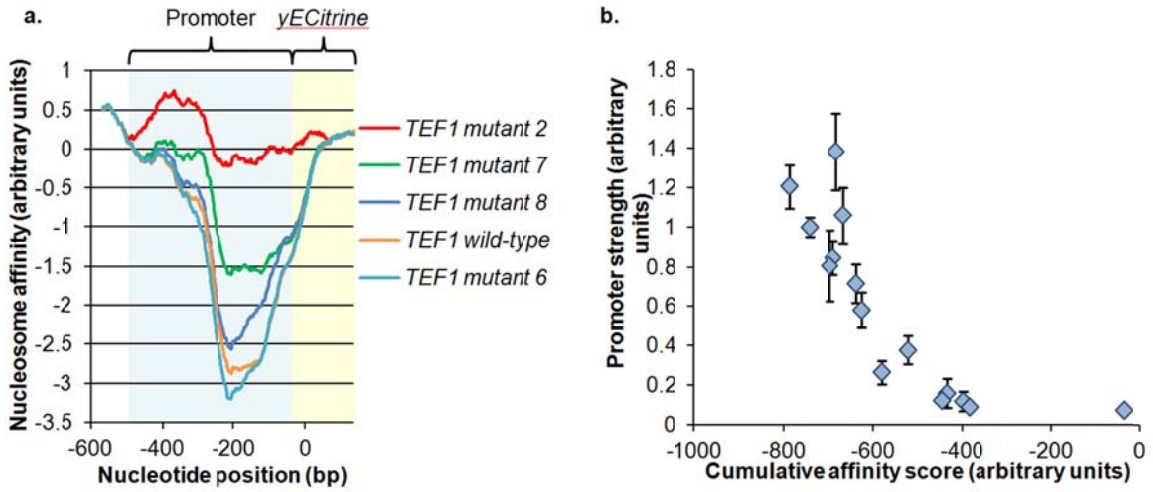


Figure 2-2: Nucleosome affinity correlates to mutant promoter strength

A) Computational nucleosome affinity profiles generated using a hidden Markov model (146) for several *TEF1* mutant promoters (26,27), with *TEF1 mutant 2* being the weakest and *TEF1 mutant 6* the strongest **B)** Experimental promoter strength as a function of cumulative affinity scores based on profiles in (**A**) for the *TEF1* mutant promoter library.

Using these results along with a computational exploration of sequence space, we established a framework to specify increased promoter strength at the DNA level by designing sequences with decreased predicted nucleosome affinity. Although this study focused on predictive increases in promoter activity, this approach may also be used to more generally decrease or otherwise tune promoter strength. Our nucleosome affinity minimization technique employed a greedy algorithm to minimize the cumulative affinity score over several rounds of optimization; in each round, all possible candidates differing

by a single base pair were computationally generated and the candidate with the smallest cumulative affinity score was used as an input for the next round. Importantly, this optimization was bounded by the sequence-based requirement to avoid the destruction or creation of well-known transcription factor binding sites (147,148) (see **Appendix B.1**). A greedy algorithm was chosen for computational convenience rather than for exhaustive nucleosome occupancy optimization. Moreover, we have validated this choice by finding that optimizing over all pairs of nucleotide substitutions in each round resulted in promoters with only slightly lower predicted nucleosome affinity although at a substantially increased computational cost (700 sec per mutation vs. 218,000 sec per pair of mutations in the case of *CYC1*) (**Figure 2-3**). Thus, the greedy algorithm is well-suited for the rapid identification of designer promoter sequences. Since each round of the greedy algorithm evaluated all candidates differing by single base pair changes (a space on the order of 10^3 for each promoter tested), and because our design cycle consisted of 50-100 rounds, this proof-of-concept demonstration corresponds to sequence space searches of upwards of 10^5 in a facile manner. The scope of this sequence space for the first round of the *CYC1* promoter optimization is depicted in **Figure 2-4**. This initial search illustrates hot-spots in sequence space that result in lower cumulative nucleosome affinity scores. For example, in **Figure 2-4A**, there are a series of variants clustered near the -100 base-pair position that show decreased cumulative nucleosome affinity scores when mutated to T, and higher scores when mutated to G or C. Furthermore, it should be noted there are examples where changing a particular nucleotide to an A or T does not result in the lowest predicted score for that position even though AT-rich regions are generally less likely to bind nucleosomes.

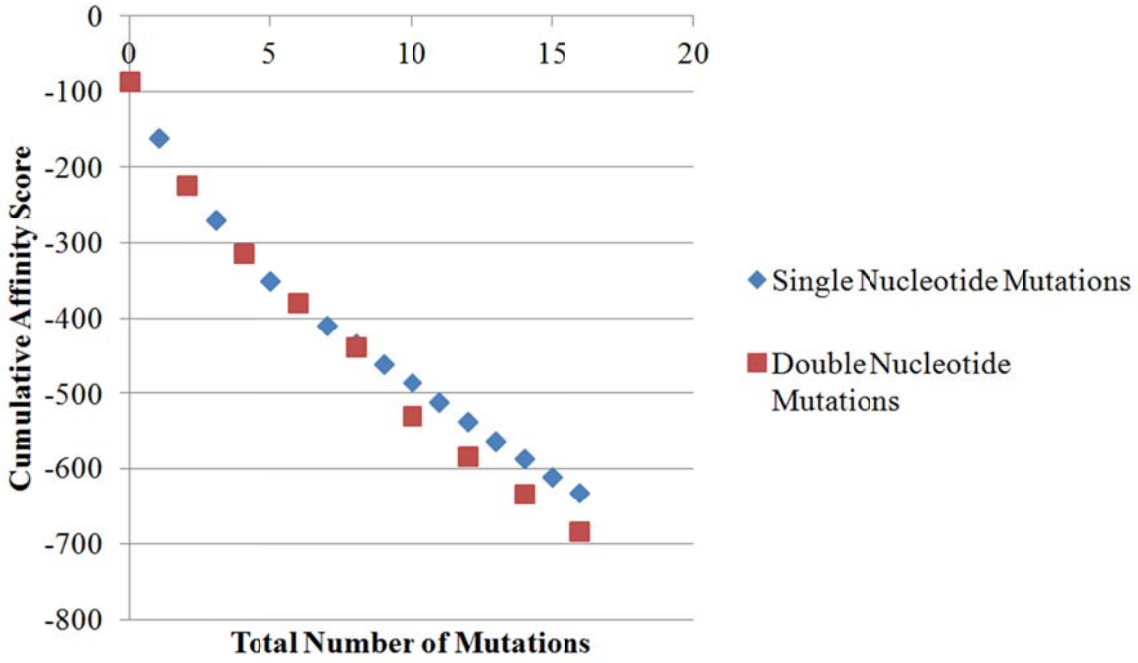


Figure 2-3: Comparison of greedy algorithm and algorithm considering double nucleotide substitutions.

Nucleosome affinity was minimized for the *CYC1* promoter using the simple greedy algorithm used in this study and a modified algorithm which takes into consideration double nucleotide substitutions. It can be seen that the algorithm considering double nucleotide substitutions performs slightly better than the simple greedy algorithm

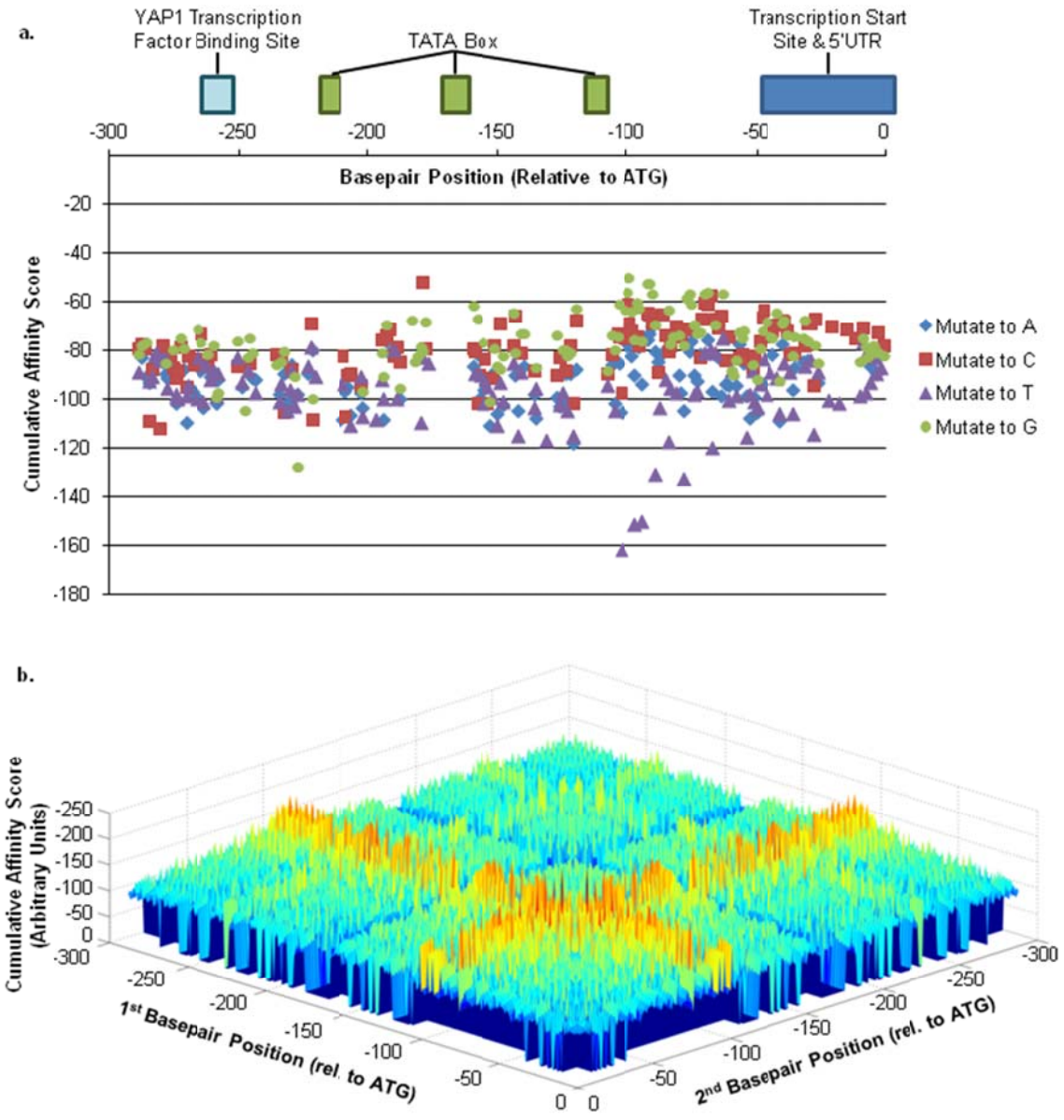


Figure 2-4: Computational candidates generated for one round of the CYC1 promoter redesign.

Each candidate queried for the *CYC1* promoter redesign was plotted for the first round of **A)** a greedy algorithm searching over all possible single base pair changes per round and **B)** a greedy algorithm searching over all possible double base pair changes per round. For the algorithm searching over all single base pair changes, known transcription factor binding sites, TATA boxes, and transcription start sites are annotated. For the algorithm searching over all pairs, each point on the surface represents the most favorable pair of mutations (out of 16 possibilities) for a particular pair of positions.

Using this approach, we successfully defined promoter sequences that experimentally increased the strength of four different native yeast promoters (*CYC1*, *HIS5*, *HXT7*, and *TEF1*) that natively span an order of magnitude in expression level (**Figure 2-5A,** see **Figure 2-6** for a comparison of wild-type promoter strengths and predicted nucleosome affinity profiles). In each of these cases, we used our approach to computationally redesign sequences for higher strength promoter variants by choosing the products of select rounds of optimization to synthesize, and then experimentally demonstrating improved transcriptional activity in a plasmid-based system. Furthermore, using the *CYC1* promoter as a test case, we showed that a variety of expression levels can be generated by synthesizing the products of varying rounds of optimization, with *CYC1v1* the product of an early round and *CYC1v3* the product of a late round (see **Appendix Table A1-1** for full promoter sequences). The greatest improvement in strength over wild-type for all of the redesigned promoters was 3.2-fold, exhibited by the *CYC1v3* promoter, which is the result of the 30th round of optimization. Subsequent measurement of transcript level using quantitative PCR confirmed that the redesigned promoters increased transcriptional expression over each corresponding wild-type promoter (**Figure 2-5B**).

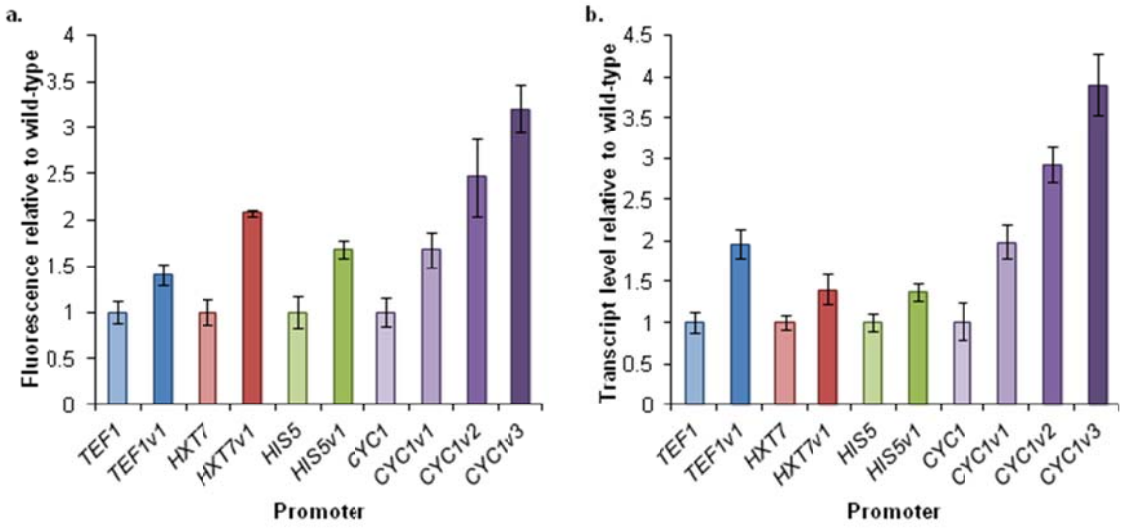


Figure 2-5: Redesign of native yeast promoters for increased expression by decreasing nucleosome affinity.

A) Computationally redesigned promoters exhibiting upwards of 3.2-fold increases in fluorescence over wild-type. Error bars represent standard deviation from three biological replicates. See **Figure 2-6** for a comparison of wild-type promoter strengths and predicted nucleosome affinity profiles. **B)** Relative transcript level as measured by quantitative PCR for the promoters shown in (**A**). Error bars represent standard deviation from three technical replicates.

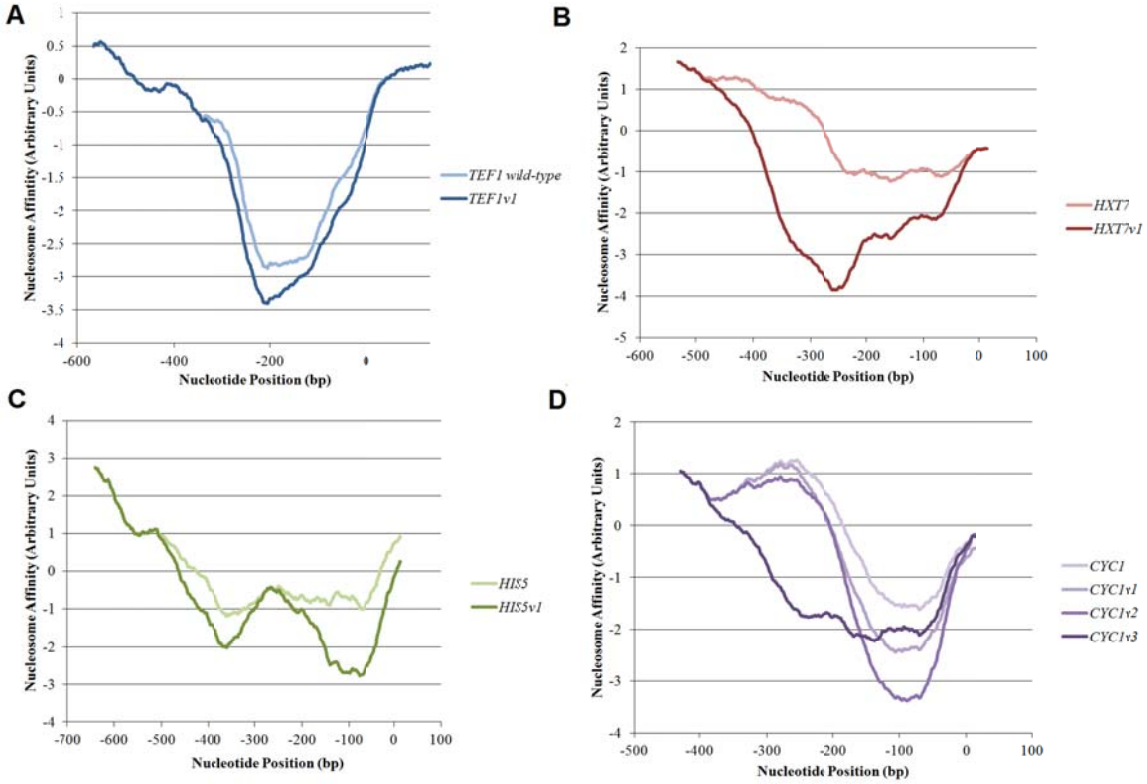


Figure 2-6: Computational nucleosome affinity profiles generated using a hidden Markov model (146)

A) *TEF1v1* and *TEF1* wild-type **B)** *HXT7v1* and *HXT7* wild-type **C)** *HIS5v1* and *HIS5* wild-type **D)** *CYC1v1-3* and *CYC1* wild-type.

It should be noted that nucleosome architecture did not appear to be as limiting among the absolute strongest native promoters in yeast (including *TDH3* and *GAL1*). While our previous work has demonstrated that these promoters have the capacity for increased expression through the use of chimeric hybrid promoters, no increase in expression was seen in this work (**Figure 2-7**), indicating that nucleosome architecture is likely evolutionarily optimized for these promoters. These two promoters represented the only cases in which false positives were identified by this algorithm. However, a nucleosome architecture approach could still likely be used to tune down the expression of these highest promoters.

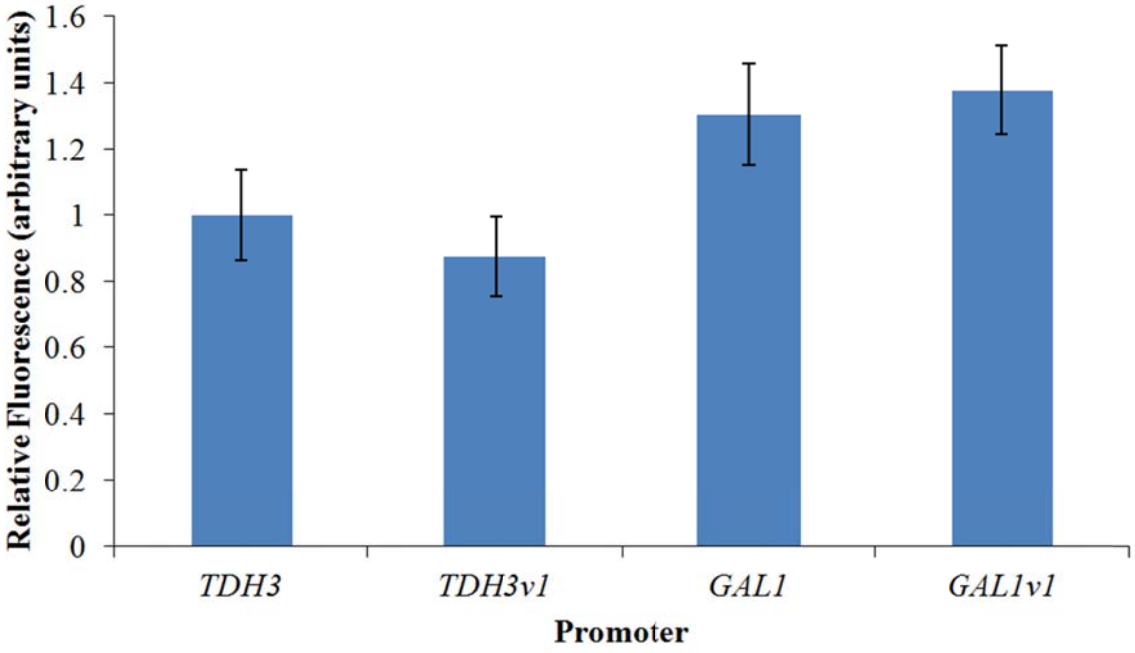


Figure 2-7: Relative fluorescence of TDH3, *GAL1*, and redesigned promoter constructs.

Both re-designed promoters had the same fluorescence relative to wild-type, demonstrating that nucleosome architecture is not limiting in these very strong promoters.

To confirm the biological underpinning of this design algorithm, nucleosome occupancy was measured via micrococcal nuclease digestion and quantitative PCR tiling array. This experiment demonstrated that nucleosome occupancy was reduced in *CYC1v3* relative to wild-type *CYC1*, as predicted by the model (**Figure 2-8**). These results clearly demonstrate that actual nucleosome occupancy was reduced in the redesigned promoter (**Figure 2-8A**). These results can be compared qualitatively to the computational predictions generated by the hidden Markov model (**Figures 2-8B,C**). Collectively, these results confirmed our hypothesis that promoter strength may be controlled by manipulating nucleosome occupancy and demonstrated that nucleosome architecture can be used to specify sequence-function relationships for yeast promoters.

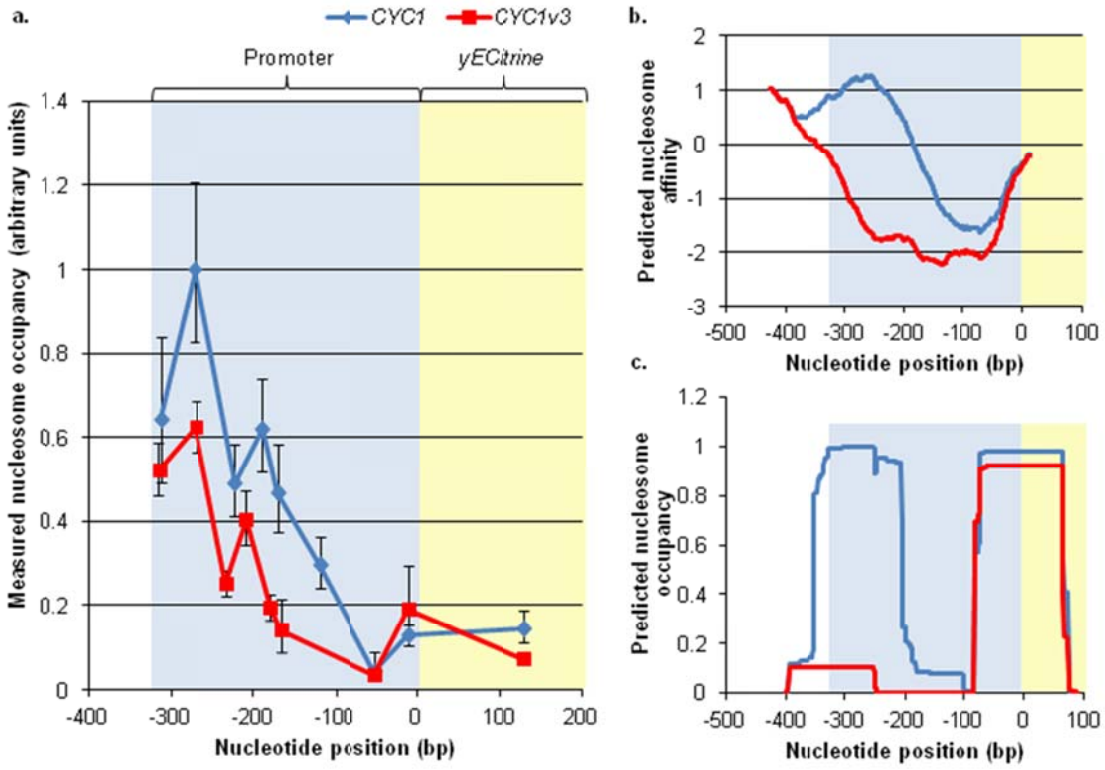


Figure 2-8: Nucleosome occupancy is decreased in the CYC1v3 promoter relative to the CYC1 promoter.

A) Relative abundance of nucleosomal DNA as measured by micrococcal nuclease assays in *CYC1* and *CYC1v3* promoters. After micrococal nuclease digestion, copy number was measured across the promoter using a quantitative PCR (qPCR) tiling array. Each point represents the relative copy number of the qPCR amplicon centered at that base-pair location. Relative copy number of each amplicon was calculated in comparison to a control amplicon in the ampicillin gene. Error bars represent standard deviation from three technical measurements of each amplicon and ampicillin gene. The redesigned *CYC1v3* promoter exhibits lower nucleosome occupancy in the promoter region than the wild-type version. **B)** Predicted nucleosome affinity profile for the *CYC1* and *CYC1v3* promoters using the hidden Markov model. **C)** Predicted nucleosome occupancy profiles for the *CYC1* and *CYC1v3* promoters using the hidden Markov model[15].

2.2.2 Re-designed promoters function in multiple genetic contexts

All of the above-described characterization was performed within a singular genetic context, namely a single plasmid design. Thus, we next sought to test the capacity for rationally designed promoters to function in alternative genetic contexts.

Specifically, alternative contexts can be used to test the importance of the predicted mutation to potentiate nucleosome architecture rearrangements independent of upstream and downstream DNA segments. Differences in the genetic contexts that surround the promoter, either due to the promoter's location in the genome or due to the particular gene being expressed, could result in changes to the local chromosomal architecture and could therefore influence the final expression level of the promoter. This phenomenon of genetic loci-dependent expression is well-documented for the yeast genome (149).

First, the *CYC1* series of re-designed promoters was evaluated with an alternative reporter gene. In this case, the *yECitrine* gene used in our previous experiments was replaced with a beta-galactosidase gene from *E. coli* (*LacZ*). Beta-galactosidase activity was detected and the relative increase in expression level using this reporter was similar to that from the *yECitrine* constructs (**Figure 2-9A**). In this case, the *CYC1v3* had a 3.8 fold higher relative expression compared to wild-type *CYC1*.

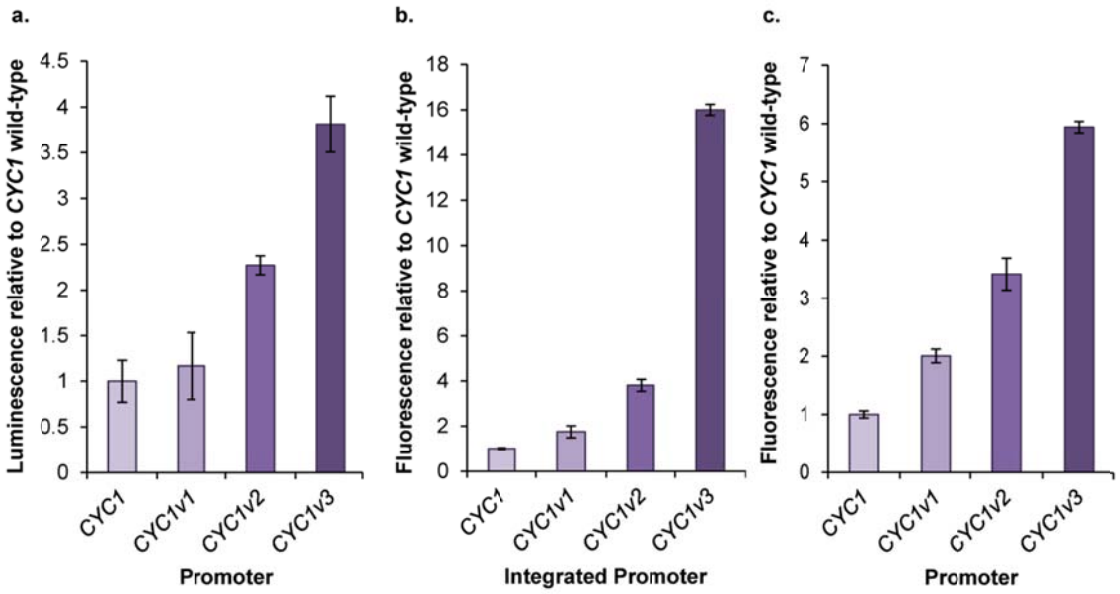


Figure 2-9: CYC1 promoter redesigns have consistently increased expression levels in different genetic contexts.

A) Relative expression level from the *CYC1* promoter variants expressing the beta-galactosidase gene *LacZ* as measured by a chemiluminescent assay. Background luminescence from a strain not expressing *LacZ* was negligible. **B)** Relative expression level from the *CYC1* promoter variants expressing *yECitrine* and integrated into the *TRP1* locus of the BY4741 genome. **C)** Relative expression level from the *CYC1* promoter variants expressing *yECitrine* with the *K. lactis URA3* gene integrated upstream of the promoter. These plasmid constructs were the basis for the integration cassette used to create the strains measured in (**B**). Error bars represent standard deviation from biological triplicate.

Second, the *CYC1* series of re-designed promoters was evaluated using genome integration rather than from a plasmid. In this case, the *K. lactis URA3* gene was cloned upstream of each *CYC1* promoter variant as a marker gene, and the entire cassette was integrated into the *TRP1* locus in the genome of *S. cerevisiae* BY4741. Expression of *yECitrine* was measured using flow cytometry (**Figure 2-9B**). The trend and rank order of increased expression level along this series was the same as for the plasmids (both for *yECitrine* and *LacZ*). However, the relative fold-change in expression level was significantly higher for the integrated constructs than for the plasmids, with the highest increase from wild-type being 16-fold for *CYC1v3*. To determine whether this difference

was due to the move from the plasmid to the genome or due to the *URA3* marker gene integrated upstream of the promoter, a set of plasmids containing the *URA3* marker gene were also assayed for *yECitrine* expression (**Figure 2-9C**). Interestingly, the fold-change in expression level for these constructs was intermediate between the original plasmid constructs and the integrated constructs, with the highest increase being 5.9-fold for *CYC1v3*. It is therefore likely that both the addition of the marker gene and the integration of the cassette resulted in local repositioning of nucelosomes that changed the final ultimate nucleosome architecture of the expression cassette. Regardless, the re-designed promoters consistently increased expression level, indicating that these rational mutations are able to potentiate a decrease in the nucleosome occupancy of yeast promoters in a variety of genetic contexts, thereby increasing expression level in a general manner.

2.2.3 Design and creation of synthetic yeast promoters

As a second proof-of-concept, we sought to demonstrate that a model-guided approach can be used to create *de novo* promoters for synthetic biology without requiring the use of a native promoter as a scaffold. Previous attempts to create synthetic *S. cerevisiae* promoters usually relied upon hybrids of multiple promoter parts (20), synthetic zinc finger transcription factor binding sites inserted into a scaffold of a native promoter (143,150), the use of synthetic TALE transcription factors (151), or random libraries and screening (152). A purely synthetic, *de novo* designed promoter created merely upon the arrangement of desired transcription factors has not been previously demonstrated. Specifically, our goal in this proof-of-concept was to demonstrate that, even without information related to promoter architecture rules, it is possible to computationally specify active promoter sequences. To use our design and search

strategy to create such a synthetic promoter, we specified two arrangements as initial lead scaffolds for the promoter design. To do so, we utilized common glycolytic transcription factor binding sites embedded in random spacer sequences as the lead designs for our algorithm (**Figure 2-10A**, see **Table 2-1** for comparison to native promoters). This approach resulted in two synthetic base scaffolds: *Psynth1*, and a shorter version *Psynth2*, which were both used as inputs to our nucleosome affinity minimization techniqueThree synthetic promoters were designed for *Psynth1* and *Psynth2*: one version from the sixth round of optimization, a second version from the 50^{th} or 30^{th} round, and a third version from the 98^{th} or 59^{th} round, respectively. As a result, a total search space of 10^5 was evaluated over the entire design cycle for each base scaffold. The result was six DNA-specified promoters that were subsequently characterized. All six designs were found to be active promoters *in vivo* (**Figure 2-10B**) that span nearly a 20-fold dynamic range with most of them being similar or higher in strength to the *CYC1* promoter—a promoter representative of the mean expression level of native yeast promoters (153). The power of our affinity minimization technique to increase promoter activity is especially evident in the case of *Psynth1*. *Psynth1v1* is only marginally higher in expression than the negative control, whereas *Psynth1v2* is 3.5-fold higher and approaches the strength of *CYC1*. *Psynth1v3* is nearly 20-fold higher than *Psynth1v1* and is on par with the strength of a commonly used promoter, the *HXT7* promoter. Moreover, the substantial transcriptional capacity of this purely synthetic promoter places it in the 6^{th} percentile of expression when compared to endogenous yeast promoters (153). Furthermore, it should be noted that each of these synthetic promoters is quite distinct on a sequence level from native *S. cerevisiae* promoters. In fact, the most significant homology consisted of a 39 base-pair sequence surrounding the TATA box of *Psynth1* (E-value =0.48). Thus, our *Psynth* promoters are not enriched with native sequences and are therefore pure, *de novo*

synthetic designs. Moreover, these *de novo* designed promoters did not require native spacing between transcription factors nor did they require the need to exactly mimic any given native promoter sequence as a scaffold.

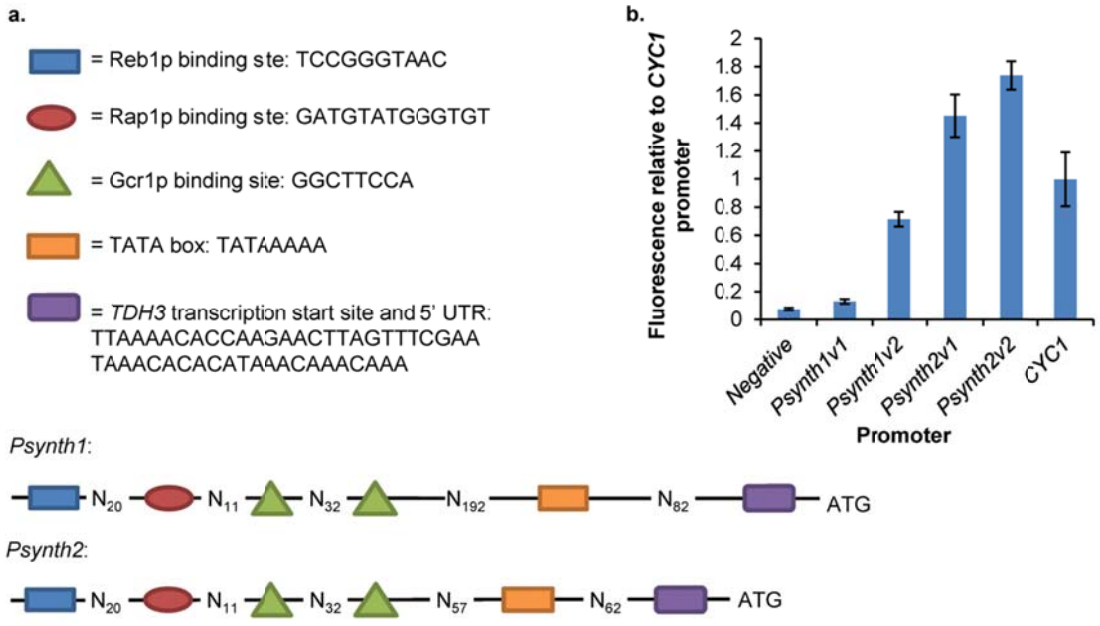


Figure 2-10: Model-guided creation of *de novo* synthetic promoters.

A) Two synthetic lead sequences, each containing prescribed transcription factor binding sites and randomized linker sequences, were used for de novo promoter design. B) Two computationally derived versions of each synthetic promoter were tested, one with only a few mutations relative to the starting random sequence, and one with many. Psynth1v1 and Psynth1v2 both have six mutations relative to their starting random sequence while Psynth1v2 has 50 and Psynth2v2 has 30. Expression levels of the re-designed synthetic promoters spanned a nearly 14-fold range and all were functional. Error bars represent standard deviation of three biological replicates.

Promoter	1 Gcr1	2 Rap1	3 Reb1	4 5' UTR	5 TATA -> TSS	6 Gcr1 -> TATA	7 Gcr1 -> Gcr1	8 Rap1 -> Gcr1	9 Rap1 -> Reb1
TDH3	2	1	1	41	93	309	38	27	31
FBA1	3	1	0	11	105	255	42	N/A	N/A
TPI1	2	0	1	31	140	167	40	N/A	N/A
ADH1	2	1	0	39	82	207	N/A	N/A	N/A
PGK1	2	1	1	42	104	279	17	24	88
CDC19	5	1	1	29	163	64	47	12	170
TDH2	4	0	1	33	93	170	45	N/A	N/A
GPM1	2	1	0	11	128	196	16	37	N/A
Average	**2.75**	**0.75**	**0.625**	**29.6**	**113.5**	**205.9**	**35**	**25**	**96.3**
Minimum	**2**	**0**	**0**	**11**	**82**	**64**	**16**	**12**	**31**
Psynth1	2	1	1	41	100	200	40	24	30
Psynth2	2	1	1	41	80	65	40	24	30

Table 2-1: Glycolytic promoter architecture and design of Psynth1 and Psynth2.

The positions and lengths between various transcription factors in a collection of yeast glycolytic promoters were catalogued in order to design Psynth1 and Psynth2. All lengths refer to the distance in basepairs between the start of each binding site. Column Descriptions: 1. Number of Gcr1p binding sites. 2. Number of Rap1p binding sites. 3. Number of Reb1p binding sites. 4. 5' UTR length. 5. Length between TATA box and transcription start site. 6. Length between Gcr1p binding site and TATA box. 7. Length between two Gcr1p binding sites when they occur in close proximity to a Rap1p or Reb1p binding site. Values of N/A mean that the sites did not occur in a pair. A value of 40 bp was chosen for the synthetic promoters because PGK1 and GPM1 were identified as outliers in this category. 8. Length between Rap1p binding site and Gcr1p binding site when they occur in close proximity to each other. Values of N/A mean that the sites did not occur close together or there was no Rap1p binding site. 9. Length between Rap1p and Reb1p binding sites. Values of N/A mean that the promoter did not have both a Rap1p site and a Reb1p site. The minimum distance was chosen in this category because the three values had a large distribution and some of the promoters that lack a Reb1p binding site have an Abf1 binding site in a similar position.

2.3 DISCUSSION

Taken together, these results present the first DNA-level specification of promoter strength for yeast promoters based on a nucleosome architecture model. We have demonstrated the potential of this approach for (1) the re-design of endogenous promoter scaffolds and (2) the design of *de novo* synthetic promoters.

Specifically, native yeast promoters were redesigned into highly homologous sequences with promoter strengths up to 16-fold higher than their wild-type sequences. For each of the four promoter case studies, we improved activity by first interrogating ~10^5 promoter variants *in silico* (10^3 candidates were queried per round when searching over all possible single base pair changes, and 10^6 could be queried per round when searching over doubles, see **Figure 2-3**) then characterizing the products of selected rounds of the greedy algorithm *in vivo*. For the case of the *CYC1* promoter, we chose the products of three different rounds of optimization to synthesize. This approach stands in stark contrast to the generation of large mutagenic libraries followed by screening. The extent of expression level increase did not always correlate with the absolute number of base pairs changed, as increases obtained in *TEF1v1* required only five rounds of optimization (see **Appendix Table A1-1** for full sequences). However, the utility of the greedy algorithm to sequentially identify increasingly optimal sequences was upheld for each case tested. Regardless, each of the redesigned promoters required multiple rounds (i.e. basepair changes) to significantly increase expression, underscoring that these specific high-strength-potentiating combinations would be undetectable in random mutant libraries. Additionally, we confirmed that these improvements were indeed due to decreased nucleosome occupancy in the case of the *CYC1v3* promoter. Finally, we showed that these rationally designed promoters consistently display increased expression in a variety of genetic contexts, demonstrating that these directed changes are able to potentiate a decrease in nucleosome occupancy despite variation in the surrounding chromosomal architecture.

Further, we created several fully synthetic yeast promoters which attain a variety of strengths and have minimal homology to any native sequence. The base promoter scaffolds for these synthetic promoters were only very loosely based on the native

glycolytic promoters in yeast, demonstrating that close homology to native promoters may not be necessary for synthetic eukaryotic promoters. Given this surprising result, the range of synthetic promoter design possibilities is unbounded by traditional promoter architecture design rules inferred from native promoter structures. Furthermore, one of our synthetic promoters, *Psynth1v3*, is on par with a commonly used promoter for metabolic engineering purposes, the *HXT7* promoter, and resides among the top six percent of native yeast promoters in regards to strength (153).

This work confirms that nucleosome occupancy is an important, causative factor limiting the strength of native yeast promoters and is likely an evolutionary mechanism for controlling transcriptional strength (154). This method significantly advances the state-of-the-art in a field currently entrained in mutation and chimeric library construction by enabling the predictable specification of synthetic parts in single design-build-test cycles rather than by the generation of large libraries. Thus, this method opens the door to the rational design and creation of synthetic eukaryotic promoters as well as expands our capacity for pure synthetic biology design.

Chapter 3: Fine-Tuning Transcriptional Control through Weak Promoters

3.1 INTRODUCTION

In order to efficiently direct metabolism towards the production of a desired product, synthetic and endogenous cellular machinery must be expressed at precise levels to ensure maximal productivity while minimizing metabolic waste. Although well-characterized libraries of strong promoters have been curated and developed (25-27), there remains a need for promoters which exhibit weak levels of expression (i.e. of a strength lower than the promoter driving *CYC1*, which is of average strength in the yeast transcriptome (153)). For example, it is often the case that endogenous biosynthetic machinery competes with the pathway of interest for metabolite flux. In these cases, it is often desirable to knock out the competing pathway. However, if the competing pathway is essential to cell growth, its expression must be optimally downregulated so as to balance the needs of cell growth with productivity. As a second example, graded expression of a dominant mutant has been shown to yield more detailed information regarding gene function than knockout alone (155). In this application, weak promoters must be used to span the full range of dominant mutant expression. A common strategy to develop weakened promoters is through random mutagenesis of a promoter template followed by screening to identify altered variants. This strategy has been effective at creating an attenuated library of the strong *TEF1* promoter in yeast (26,27). Here we use random mutagenesis followed by screening in order to develop well-characterized variants of the *CYC1* promoter (an average-strength promoter in yeast (153)) which exhibit very low levels of expression.

3.2 RESULTS

3.2.1 Screening Methodology

In order to identify pCYC1 variants of weak activity, pCYC1 mutant libraries were used to drive the expression of *URA3p*. Mutants of low but nonzero activity could then be isolated through exposure to media containing sublethal amounts of 5-fluoro-orotic acid (5-FOA) and low concentrations of uracil. In this scheme, mutants which are too strong will result in cell lethality due to the presence of 5-FOA, while mutants which are not strong enough will also prevent cell growth by failing to compensate for the low levels of uracil. To identify optimal 5-FOA/uracil ratios for maximal enrichment of promoters weaker than pCYC1, wild-type pCYC1, pNUP57 (40% activity relative to pCYC1), and pTFC1 (30% activity relative to pCYC1) promoters were separately used to drive *URA3* expression and growth rates were measured at 30 different selective conditions (0-0.4 g/L 5-FOA and 0-5 mg/L uracil). Those conditions which yielded the largest growth advantage of pNUP57 and pTFC1 were found to be 0.75 mg/L uracil and 0.2 g/L 5-FOA (**Table 3-1**). Therefore, pCYC1 mutant libraries were plated on media containing 0.3, 0.4, and 0.5 g/L 5-FOA and 0.75 mg/L uracil, and large colonies were picked to identify weaker variants of pCYC1. In total, 30 colonies were picked and sequenced. To reduce the possibility of false positives, those promoters which had mutated to form a start codon in the 5'UTR were discarded. The remaining promoters are listed in **Appendix Table A2-5**. These promoters were then characterized as described below.

TFC1 growth advantage (h^{-1})		5-FOA (g/L)				
		0	0.5	0.1	0.2	0.4
Uracil (mg/L)	0	-0.02726	0.013311	0.032763	0.037059	0
	0.25	-0.01602	0.020889	0.025389	0.027616	0
	0.5	-0.01546	0.01205	0.035186	0.03247	0
	0.75	-0.01254	0.014216	0.024039	0.039503	0
	1	-0.0045	0.025246	0.035632	0.028183	0.007384
	5	0.008656	0.012585	0.031115	0.036332	0
	50					

NUP57 growth advantage (h^{-1})		5-FOA (g/L)				
		0	0.05	0.1	0.2	0.4
Uracil (mg/L)	0	-0.0102	0.028138	0.03775	0.032935	0
	0.25	0.014061	0.027226	0.030414	0.017106	0
	0.5	-0.00746	0.024193	0.037623	0.002495	0
	0.75	0.003338	0.018786	0.030508	0.045039	0.00695
	1	-0.00358	0.034552	0.033212	0.028546	0
	5	0.008614	0.01098	0.02649	0.032768	0
	50					

Table 3-1: Growth advantage of weak promoters in 5-FOA/uracil screen

pCYC1, pTFC1, and pNUP57 were used to drive the expression of *URA3p* in BY4741. Yeast cells expressing each cassette were grown in varying concentrations of 5-FOA and uracil in order to determine the best conditions to use to select pCYC1 mutants of reduced strength.

3.2.2 Characterization of Isolated Mutants

Promoter variants identified above were then used to drive the expression of yECitrine (a yeast codon-optimized yellow fluorescent protein) in a high copy vector. As a control, a vector was also constructed which did not contain a promoter. These constructs were then analyzed via flow cytometry (**Figure 3-1**). Excitingly, the pCYC1 variants spanned a wide range of gene expression, indicating that the *URA3*/5-FOA screen is well-suited for the identification of weak-expression constructs. Interestingly, however, it was also observed that many of the constructs enabled a lower expression of YFP than a construct lacking a promoter. This implied that in this context, these promoters acted more like terminators and enabled cell survival during screening by reducing the ability of upstream transcriptional noise to activate expression.

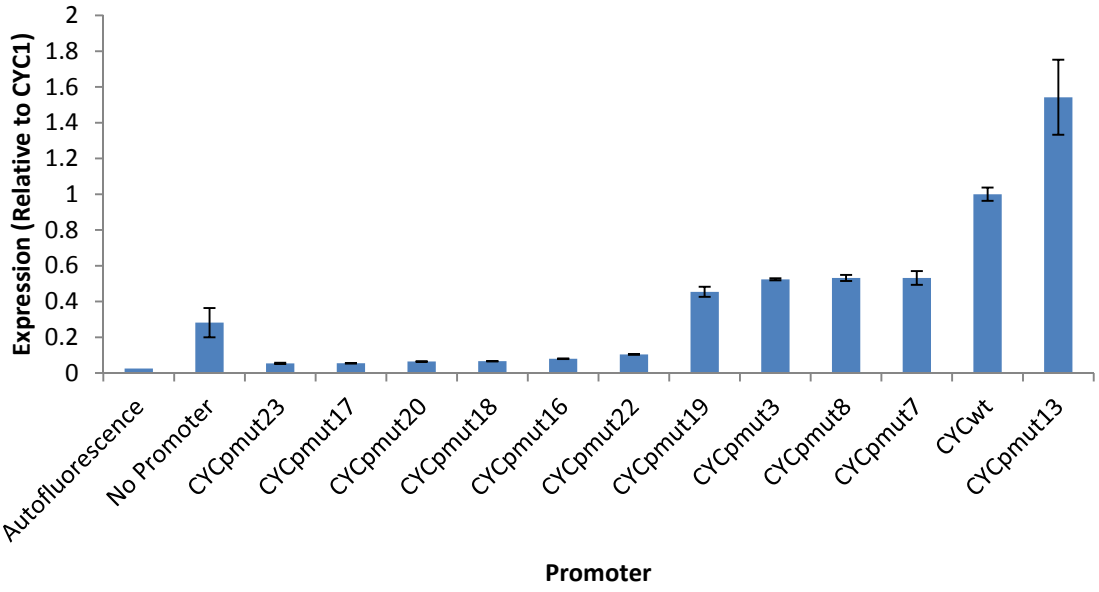


Figure 3-1: Expression level attained by mutant promoters.

3.3 DISCUSSION

During this work, we were largely successful at developing promoter variants which enabled a wide range of expression in our screening construct. This demonstrated that random mutagenesis followed by screening using 5-FOA and uracil is an efficient method to develop attenuated promoter variants. However, it was observed that several promoters identified by this work would be better characterized as terminators, as illustrated by their ability to reduce the baseline level of gene expression enabled by the expression vector we were using. Surprisingly, this baseline level of gene expression was rather high: 30% the strength of pCYC1. This baseline expression was also significantly higher than yeast's autofluorescence, implying that this limitation was not related to instrument sensitivity, but rather represented an innate level of noise and context dependence in biological systems. One of several conclusions may be drawn from this result. Firstly, if the level of background we observed with our expression vector is characteristic of that found in the genome, then many promoters for lowly-expressed genes may contain sequences which function as insulators in order to enable more precise levels of gene expression. Alternatively, if the expression vector we used enabled an unusually high level of background expression, then it may not be suitable for the assembly and implementation of phenotypes dependent on very low levels of gene expression. As a final possibility, if level of transcriptional noise inherent to the yeast genome varies by genomic location, then the regulatory machinery controlling lowly-expressed genes may function in a context-dependent manner and may perform differently when implemented in a synthetic construct. Taken together, these results emphasize the importance of context during the design of low-expression synthetic constructs and demonstrate the need for the development of synthetic insulators to ensure

that engineered elements perform reliably in multiple background transcriptional contexts.

Chapter 4: Tuning Translational Efficiency in the Context of Multicloning Sites

4.1 INTRODUCTION

Expression vectors with pre-defined multiple cloning sites (MCSs) are among the most common tools employed in molecular biology and genetics. These vectors have enabled the facile expression and cloning of recombinant genes and have recently ushered in the era of synthetic biology (156). The flexibility of restriction enzyme sites in MCSs facilitate easy cloning of genes of interest for diverse applications from genetic analysis to creation of biofuels-producing strains. Common improvements to vectors containing MCSs are focused at controlling transcript levels (via promoter replacement/engineering (79), transcription machinery engineering (78), or copy number manipulations (157)) or translation rate (e.g. by improving codon bias (158) or by reducing expression noise (159)). In all these applications, multiple cloning sites are thought to be benign, non-interacting elements that exist for mere convenience. However, a promoter element is usually placed upstream of the MCS. As a result, several base pairs (or even multiple restriction sites) will appear in the 5' untranslated region (5'UTR) of the mRNA of the cloned gene depending on the restriction site chosen. Thus, it is conceivable that the composition of these sites can significantly influence translation efficiencies of the downstream gene. Here, we demonstrate the first performance-based assessment of multiple cloning sites and develop a novel theoretical framework enabling the prediction of a MCS's effect on translation. Furthermore, we apply this understanding to rationally redesign these sites for improved function and reduced variability associated with restriction enzyme choice. We posit that this phenomenon of 5'UTR structure inhibition is most pronounced when using shorter, codon-optimized genes.

Secondary structure in the 5'UTR of messenger RNA has been found to affect expression in both prokaryotes (160,161) and eukaryotes (162-167) at the translational level. In prokaryotes, translation is initiated by the assembly of the 70S initiation complex on the ribosome binding site (RBS), normally within a few base pairs of the start codon, and it is thought that RNA secondary structure can inhibit translation by occluding the RBS (160,168). In fact, predictive models of RBS performance explicitly treat the inhibitory effect of 5'UTR secondary structure (160). Due to the differences in translation initiation in prokaryotes, the design criteria of prior methods would be of little use in highly relevant eukaryotic systems such as S. cerevisiae. Hence, a novel modeling approach resulting from a distinct theoretical framework is needed to address the issue of 5'UTR secondary structure for yeast systems. In eukaryotes, the 43S initiation complex must scan along the 5'UTR before commencing translation at the start codon, often 50 bp or more from the 5' cap structure (168). It has been hypothesized that the presence of secondary structure in these organisms decreases the rate of translation initiation by impeding ribosome scanning (165). Multicloning sites impose distance (and therefore a high likelihood of structure) between a promoter and the gene of interest in a restriction site-dependent manner, leading to the hypothesis that cloning location affects protein expression, especially in eukaryotes. In several cases, irreproducible or conflicting results have been explained by differences in restriction site usage (169,170). However, most attempts at mitigating translation-inhibiting secondary structure in eukaryotes result in "quick fixes" such as point mutations which are only applicable for the precise gene construct under consideration (171-175). Moreover, no prior work has successfully minimized secondary structure to optimize a genetic component of such widespread importance as the multicloning site or to develop a system which achieves nearly context-independent levels of protein expression, both of which are of critical significance to

obtaining high titers of heterologous proteins in eukaryotes and to enabling precise control of genetic circuits. Therefore, due to their enormous utility and widespread use for heterologous gene expression, the characterization and optimization of MCSs to minimize the effects of mRNA structure in a more general context represents a promising and novel avenue toward improving protein titers and controlling protein production.

A variety of algorithms exist for the prediction of RNA secondary structure (34,176,177). A common approach is to compute the free energy of the strand of interest through a partition function, using empirically-determined base-stacking energies to weight each possible conformation (178,179). One limitation of this approach is that enumeration of all possible conformations becomes impractical for large strands, so certain classes of folds (e.g. pseudoknots) are commonly ignored, though are possibly significant. It is important to note that a strand's free energy of folding computed in this manner is not a simple function of its composition. Since MCSs must additionally contain certain sequence motifs, any attempt to rationally design MCSs based on minimized free energy is prohibitively difficult, necessitating the use of a metaheuristic such as a genetic or hill-climbing algorithm. This difficulty is exacerbated by the requirement that designed MCSs refrain from folding regardless of where the gene of interest is inserted, highlighting the potential rarity of desirable MCSs.

In this study, we establish the variations in downstream protein translation imparted by multicloning sites and isolate the effect of secondary structure-based inhibition especially in cases of short, codon-optimized genes. This effect is demonstrated using the MCS of a common yeast vector system (180,181). Due to the unacceptably large variance found along the cloning site, a predictive model was developed to redesign multiple cloning sites with minimized secondary structure and thus

improved mRNA translation. These models led to promoter-specific, re-designed multiple cloning sites that outperform standard constructs.

4.2 RESULTS

4.2.1 Performance-based Assessment of the pBLUESCRIPT SK multiple cloning site in yeast

To gain a quantitative performance assessment of a commonly used multiple-cloning site in yeast, we inserted an optimized YFP fluorescent protein, yECitrine (182), after each restriction site in the p416 vector (181). This base vector is derived from the commonly used pRS yeast shuttle vector (180) and contains the popular pBLUESCRIPT SK MCS. Three common, distinct yeast promoters were chosen to drive expression of these cassettes. Protein output (as measured by fluorescence of YFP) changes significantly and exhibits drastic decreases as a function of position along the MCS (**Figure 4-1**). These results demonstrate that the choice of restriction site is not benign and can significantly influence performance. Moreover, this phenomenon is not strictly controlled by spacing/length as the relative fluorescence at each site depends strongly on the promoter being used to drive transcription. Additionally, it is clear that there exist promoter-specific effects beyond what would be expected from strength differences. Indeed, if the fluorescence trend was simply scaled by promoter strength, the graphs shown in **Figure 4-1** would be identical. It is also worthy of note that the fluorescence trends are not monotonically decreasing, implying that any predictor function of MCS performance must not vary monotonically with the length of mRNA between the end of the promoter and the start codon. To determine whether decreased efficiency across the multicloning site was due to translation or transcription limitations, yECitrine transcript levels resulting from the series developed above were measured relative to Alg9, a known

reference gene in yeast (183). The transcript levels of yECitrine were not found to significantly vary and thus do not correlate with measured fluorescence (**Figure 4-2**). As a result, it was concluded that the observed restriction site-dependent performance was predominantly a translation-level effect. Therefore, transcriptional effects such as the presence/absence of transcription factor binding sites do not lead to the phenomenon seen in **Figure 4-1**. Based on this characterization, it is therefore imperative that any studies relying on the precise quantity of protein (e.g. promoter strength assays or comparative enzyme assays) consider and report the intervening nucleotides between the promoter and the gene of interest, as they can confound measurements of gene expression or activity.

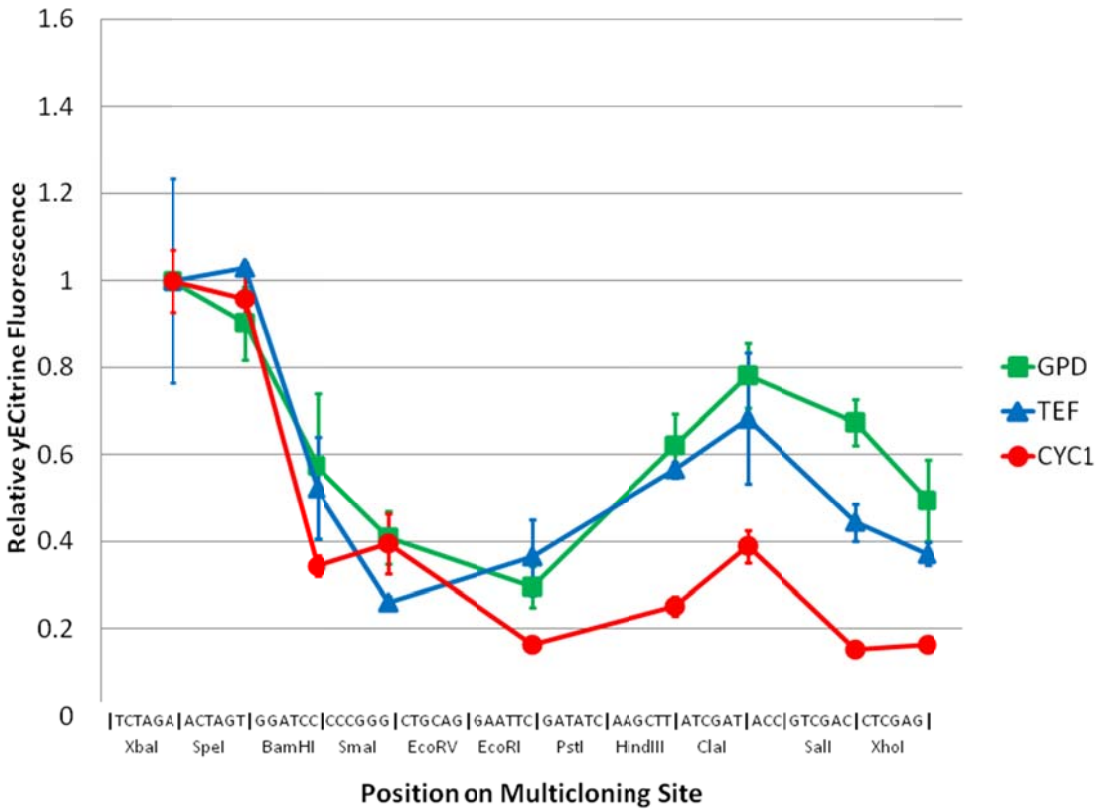


Figure 4-1: Performance Assessment of the pBLUESCRIPT SK Multicloning Site

Three promoters (TEF, CYC1, GPD) were used to drive yECitrine inserted at each available restriction site of the pBLUESCRIPT SK MCS in the p416 vector. Each series has been scaled to unity at the first restriction site. Unscaled fluorescence values for $pGPD_01YFP$, $pTEF_01YFP$, and $pCYC1_01YFP$ are 1050, 611, and 30.2, respectively (**Appendix Table A3-4**). Error bars represent the standard deviation in fluorescence observed across biological triplicates. Fluorescence is seen to vary in a promoter specific manner across each of the sites in the MCS.

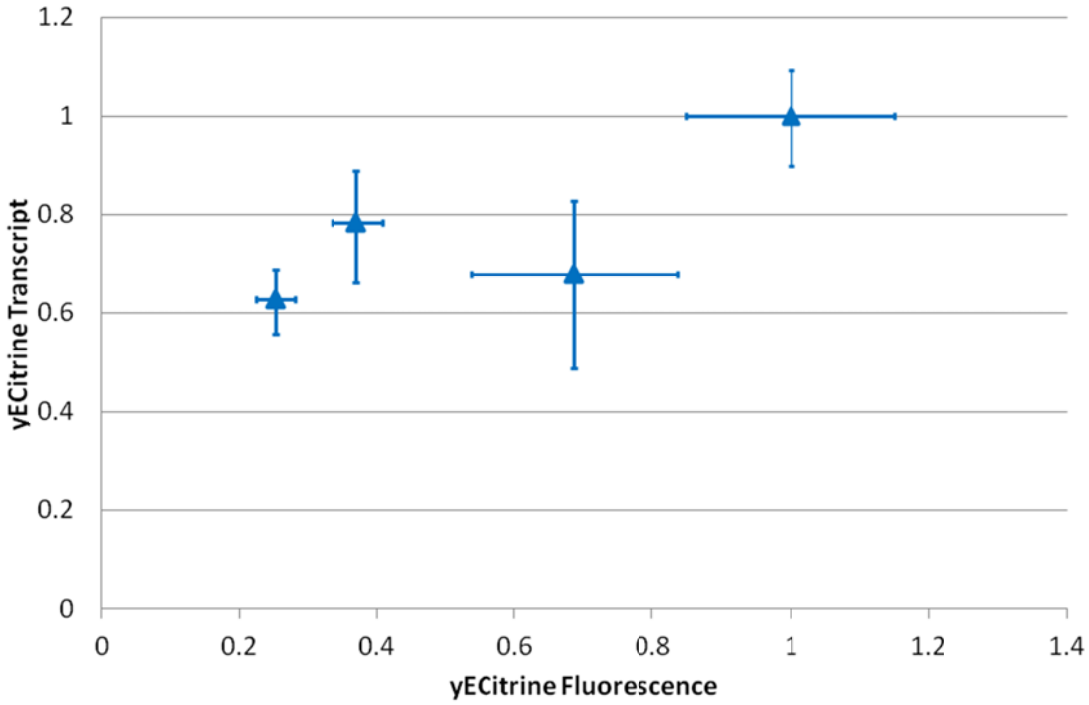


Figure 4-2: yECitrine Transcript Levels vs yECitrine Fluorescence in the p416-TEF multicloning site series

yECitrine transcript levels were quantified in $pTEF_01YFP$, $pTEF_04YFP$, $pTEF_07YFP$, and $pTEF_09YFP$ (**Appendix Table A3-4**) and compared with fluorescence values obtained for that construct. Error bars in transcript level correspond to the standard deviation resulting from three measurements of the same RNA extract, and error bars in fluorescence level correspond to those shown in **Figure 1**. There is a non-correlation between transcript level and fluorescence level, thus suggesting translational inhibition.

4.2.2 Determination of possible correlates of 5'UTR-dependent translational inhibition

Given evidence that the restriction site-dependent inhibition is a translation-level effect, several physical characteristics of mRNA were considered as possible correlates of yECitrine fluorescence. Initially, both 5'UTR GC content and length were evaluated

using an expanded data set consisting of the TEFpmut5 promoter (79,184) and various intervening sequences (**Appendix Table A3-1**). This dataset represented the first instance in which we observed this translational inhibition, inspiring a more complete characterization of this effect in the wild-type, canonical *TEF1*, *GPD1*, and *CYC1* promoters in subsequent experiments. TEFpmut5 is almost identical to pTEF1, containing 8 point mutations and retaining 95% of TEF's promoter activity, indicating that the two promoters are comparable. Relative fluorescence was plotted against length and GC content for these TEFpmut5 constructs (**Figure 4-3A,B**), and no clear relationship was observed in either variable. However, upon plotting the computed thermodynamic folding energy of the 5'UTR (a more direct predictor of secondary structure) against yECitrine expression (**Figure 4-3C**), a clear monotonic downward trend was observed, consistent with earlier reports that significant 5'UTR secondary structure can inhibit gene expression (164-166). Since RNA transcription begins in the 3' end of the promoter, different promoters will yield different base pair compositions (and hence differing secondary structure) in the 5'UTR. This result partially explains the promoter-specific impact of MCS found in **Figure 4-1**. Therefore, it was hypothesized that restriction site-dependent inhibition in the pBLUESCRIPT SK multicloning site was best explained by the thermodynamic free energy of folding of the 5'UTR.

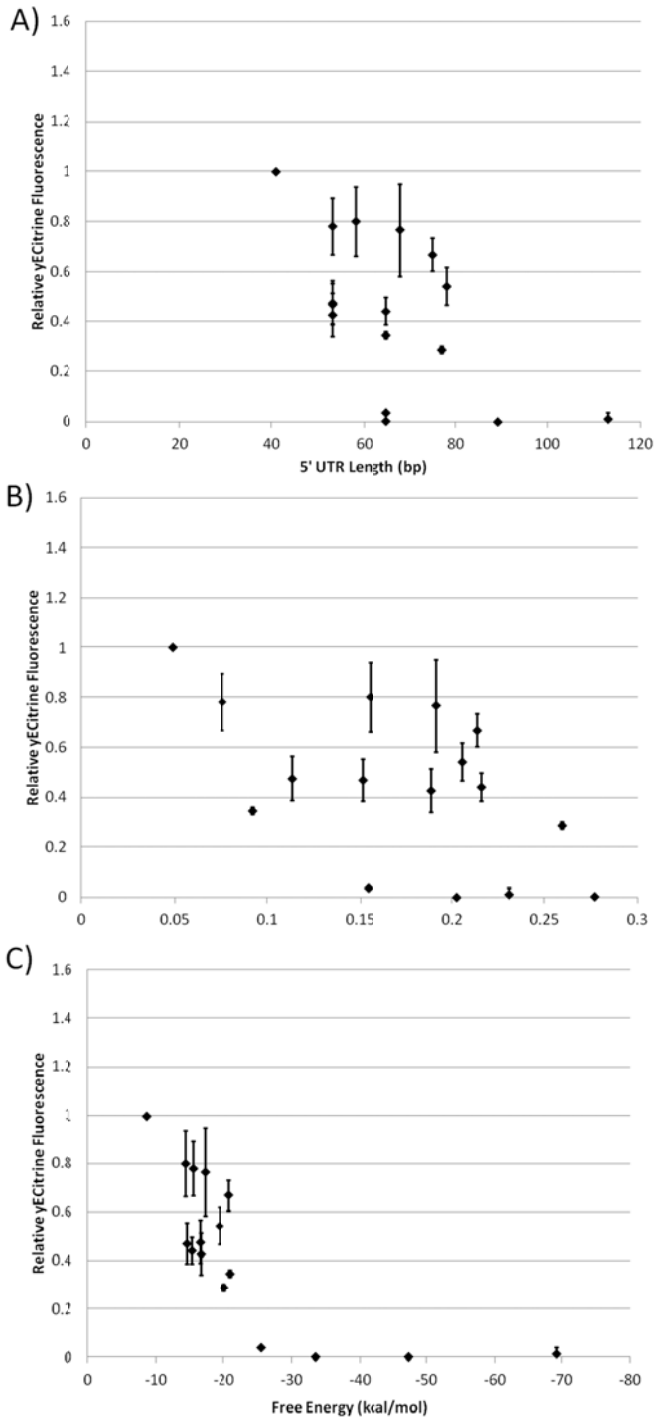


Figure 4-3: Prospective Correlates of Expression in the TEFpmut5 Insert Series

yECitrine expression levels were measured in each of the TEFpmut5 constructs listed in **Appendix Table A3-1** and plotted against **(A)** 5'UTR length, **(B)** GC content, and **(C)** folding energy. Each plot has been scaled relative to the fluorescence of pT5Y. Error bars represent the standard deviation in fluorescence observed across biological triplicates. Fluorescence is seen to monotonically vary with free energy level, thus suggesting 5'UTR secondary structure as the leading cause of this phenomenon.

4.2.3 Comparing the impact of 5'UTR structure to codon usage and gene length

Beyond 5' UTR structure, gene-specific traits such as length and codon usage can impact translation. To this end, genes for β-galactosidase and an *E. coli* optimized green fluorescent protein (GFP) (185) were inserted into the MCS of p416-TEF and performance was compared with yECitrine. The codon adaptation index (CAI) (186), a common measure of codon optimality, for both β-galactosidase and GFP in yeast are quite low. In addition, β-galactosidase is relatively long (>3 kbp), whereas the lengths of GFP and yECitrine are almost identical (~700 bp) (**Table 4-1**). In the case of yECitrine, a short, codon-optimal gene, 5'UTR structure dominated as reporter output varied greatly as a function of cloning position (**Figure 4-4**). In contrast, as the gene of interest becomes longer and uses progressively rarer codons (as with β-galactosidase and GFP), the effects of gene length or codon biases become the rate-limiting steps in translation, dwarfing the effects of secondary structure. As a result, the restriction site-dependent effects of mRNA secondary structure are substantially muted by poor codon usage and/or large size (**Figure 4-4**). Therefore, the effect documented here of 5'UTR structure inhibition in MCSs is extremely relevant to synthetic biology in which codon-optimized genes are routinely being synthesized and used.

Gene	Length (bp)	Codon Adaptation Index
yECitrine	717	0.519
eGFP	756	0.0888
LacZ	3075	0.0570

Table 4-1: Genetic parameters for yECitrine, eGFP, and LacZ

Codon Adaptation Indices were computed with JCat (187) in *S. cerevisiae* and gene lengths are reported.

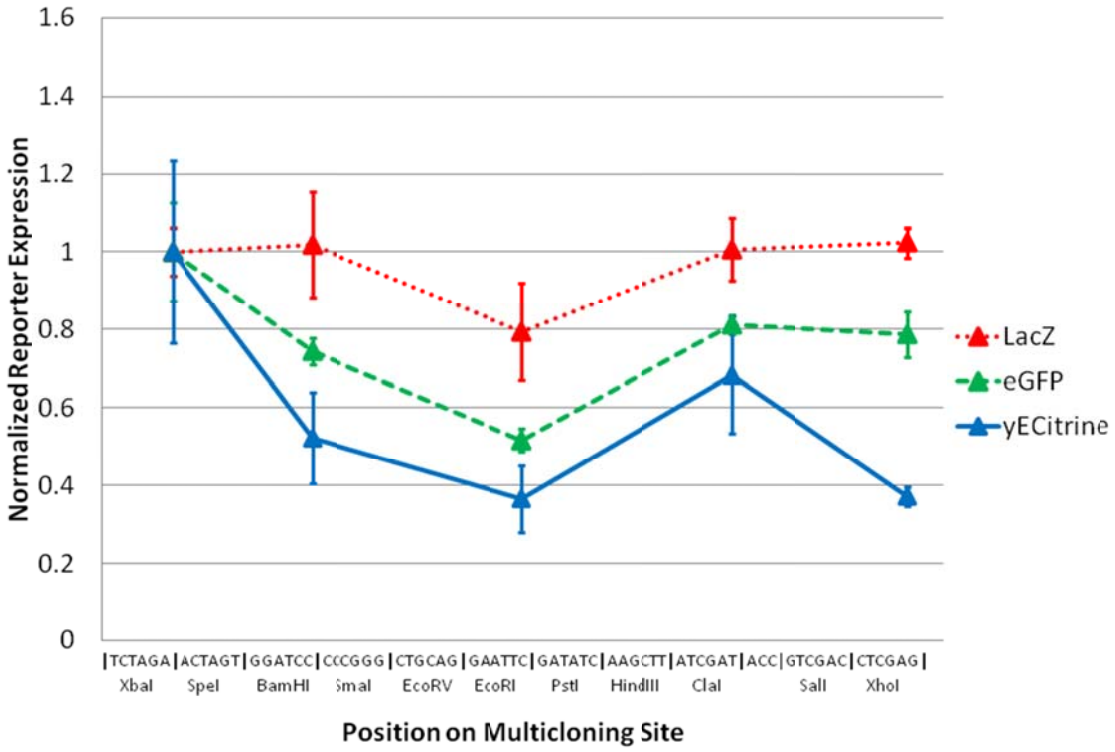


Figure 4-4: Effect of gene length and codon usage on translational inhibition. LacZ and eGFP

LacZ and *eGFP* expression levels were measured in each of the constructs listed in **Appendix Tables A3-5** and **A3-6**, respectively, and compared with data for yECitrine. Each series has been normalized to unity at the first restriction site. Position on the MCS has been measured according to the unique restriction sites in the p416 vector. Error bars represent the standard deviation in reporter output observed across three biological replicates. The impact of 5'UTR inhibition is most pronounced in short, codon optimized genes.

4.2.4 Initial Multicloning Site Design

Our initial hypothesis was that expression inhibition was due to the folding energy and structure of the entire 5'UTR. Therefore, promoter-specific MCSs were redesigned with the aim of increasing the ensemble folding energy of the 5'UTR (See Materials and Methods). Both reordering of enzyme sites and insertion of base-pairs to remove secondary structure were allowed. The resulting MCSs, dubbed TEF_1 and $CYC1_1$ for use after the *TEF1* and *CYC1* promoters, respectively, are shown in **Appendix Table A3-2**.

yECitrine was inserted at each restriction site as for the pBLUESCRIPT SK MCS, and the results are shown in **Figure 4-5**. Despite the crudeness of this original model for MCS performance, both TEF_1 and $CYC1_1$ showed improved desirable performance.

The redesigned MCS TEF_1, using the *TEF1* promoter, is remarkable in that it maintains a narrow range of reporter expression between the 2nd and 9th restriction sites. In this region, the expression from $pTEF_1$xYFP ranges between 0.69 and 0.42, whereas the expression from pBLUESCRIPT SK ranges between 1.03 and 0.26 in the same region. This property makes TEF_1 more appropriate for applications in which consistency in expression across varying sites within the MCS is desired.

The redesigned MCS $CYC1_1$, using the *CYC1* promoter, yields yECitrine expression equal to or greater than the pBLUESCRIPT SK for all but one of the available restriction sites, making this multicloning site desirable. Furthermore, the 2nd, 5th, and 6th sites attain the same level of expression as 1st, allowing more cloning possibilities without decreasing effective promoter strength. It is interesting to note that increases in expression can be attained by adding nucleotides to the 5'UTR (exemplified by $pCYC1_1$5YFP and $pCYC1_1$6YFP), illustrating that MCS inhibition is not simply due to length. This observation also required a different model for inhibition, as free energy of folding of the 5'UTR always decreases as more base pairs are added to the MCS. The assumption that the entire 5'UTR produces translation-inhibiting secondary structure was therefore incorrect, and so the model was reevaluated to create better multiple cloning sites for each of the three promoters.

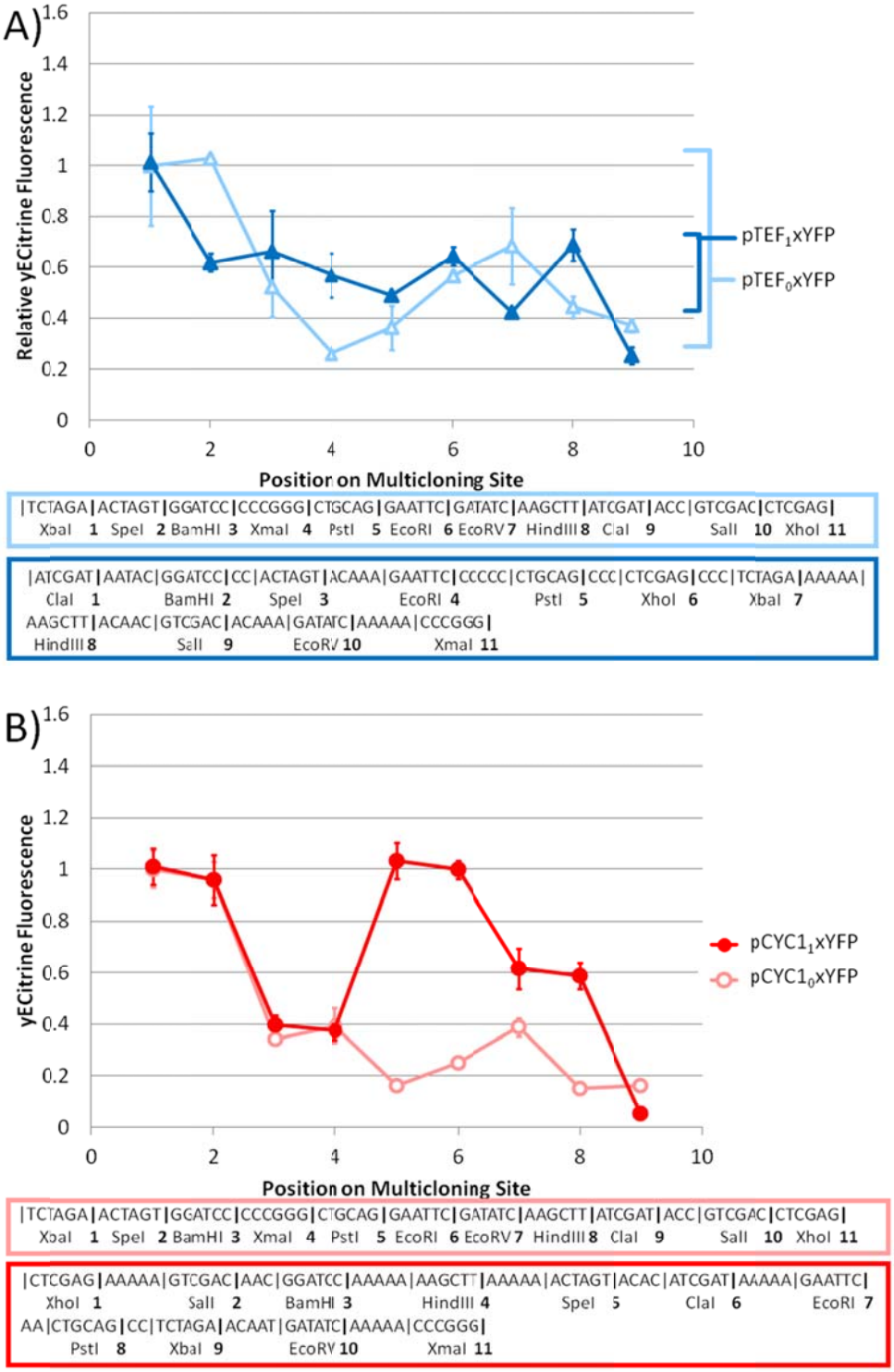


Figure 4-5: Performance of designed multicloning sites (A) TEF_1 and (B) $CYC1_1$

Two MCSs were designed to minimize the ensemble free energy of the 5'UTR when placed after TEF or CYC1, respectively. Data in (A) has been scaled to the fluorescence of $pTEF_0$1YFP and in (B) to $pCYC1_1$1YFP. Position on the MCS has been measured according to the unique restriction sites in the p416 vector. Error bars represent the standard deviation in fluorescence observed across biological triplicates.

4.2.5 Re-engineering Multicloning Sites for Function and Convenience

Given the substantial effect MCSs can have on protein production, we sought to redesign these elements by mitigating secondary structure inhibition. An initial, crude model based on complete minimization of secondary structure across the entire 5'UTR enabled the design of improved MCSs: TEF_1 and $CYC1_1$. However, this model is fundamentally limited as it suggested that protein output always decreased as a function of length across the 5'UTR. Counterexamples to this feature were found in our dataset. Due to this shortcoming, GPD_1 was not constructed and a more accurate model framework was developed to redesign multicloning sites for all promoters.

To address the observation that adding specific sets of nucleotides between the promoter and the start codon can yield increases in translational efficiency, a new model framework was developed incorporating two (or more) regions whose free energy of folding correlates with protein production (**Figure 4-6**). Such a model is grounded in the fundamental biology of the process. Successful initiation requires the presence of eIF4a, an ATP-dependent helicase which unwinds mRNA in preparation for ribosome loading. In addition, scanning through a structured 5'UTR requires ATP, though the enzyme responsible is unknown (168). Thus, the initiation complex can be modeled as a particle passing through several states (**Figure 4-6**), each separated by a free energy of folding, before reaching the start codon (See Materials and Methods). The models which best explained the available data (CYCModel1, TEFModel1, and GPDModel1) are shown in **Table 4-2**. It is important to note that in no model was the presence of mRNA structure beneficial for reporter expression. To validate these models, a second set of promoter-

specific MCSs were generated: TEF_2, $CYC1_2$, and GPD_2, detailed in **Appendix Table A3-2**. It is important to note that this design process was nontrivial due to the large number of sequence constraints which must be satisfied, in addition to the requirement that the designed MCSs refrain from folding in a variety of genetic contexts, in contrast to attempts at structure minimization in other systems for which the number of sequence constraints is relatively low and applicability is restricted to a specific gene construct (160). Furthermore, the promoters for which these MCSs are designed differ in transcriptional output by up to two orders of magnitude from one another, providing an excellent test of our framework's applicability in multiple transcriptional contexts. yECitrine was cloned at each restriction site for the three MCSs, and the fluorescence measurements are shown in **Figure 4-7**.

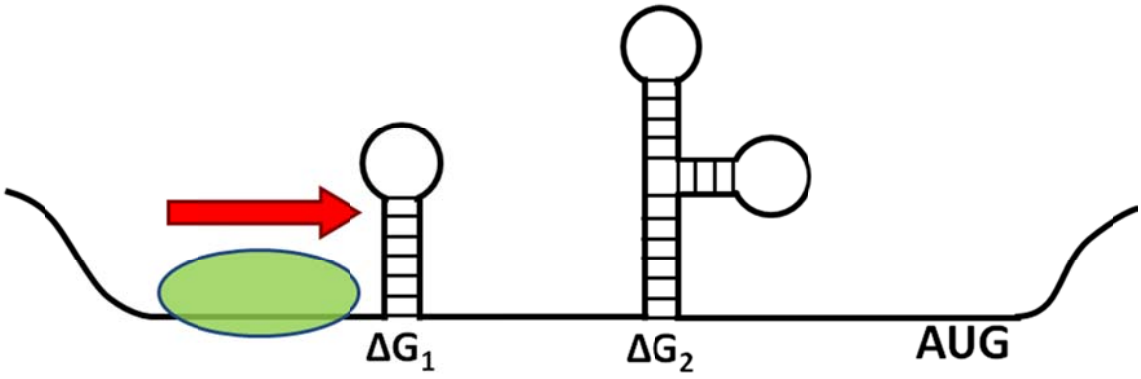


Figure 4-6: Model of translation inhibition by secondary structure in the 5' untranslated region.

The pre-initiation complex (green) scans in the 3' direction and is impeded by one or more regions of mRNA structure, decreasing the rate of translation initiation. To capture this effect, a model was created that allowed for two or more regions of secondary structure that can influence translational efficiency.

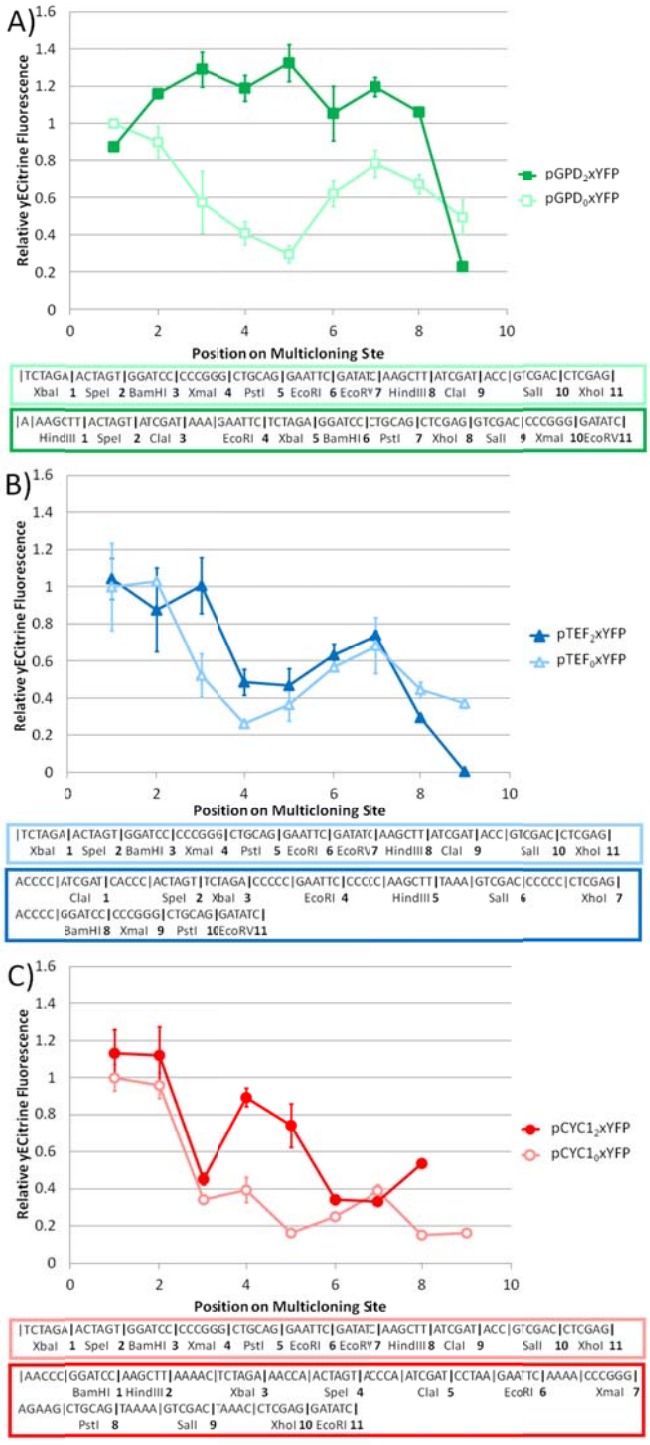


Figure 4-7: Performance of designed multicloning sites.

Performance of **(A)** pGPD$_2$xYFP, **(B)** pTEF$_2$xYFP, and **(C)** pCYC1$_2$xYFP are depicted. Three MCSs were designed with the aid of the models listed in **Table 3-2** and inserted after GPD, TEF, or CYC1, respectively. Data in (A) has been scaled to the fluorescence of pGPD$_0$1YFP, in (B) to pTEF$_0$1YFP, and in (C) to pCYC1$_0$1YFP. The scaling for each series within each graph are identical. Position on the MCS has been measured according to the unique restriction sites in the p416 vector. Error bars represent the standard deviation in fluorescence observed across three biological replicates. These MCSs had improved performance compared with pBLUESCRIPT SK.

Name	Region 1	Region 2	Model	Correlation Coefficient	Predicted Residual Sum of Squares	Residual Sum of Squares
CYC1Model1	[-166,-45]	[-37,6]	$\ln(f) = 0.0986 * \Delta G_1 + 0.1253 * \Delta G_2 + 0.5004$	0.7809	0.3016	0.01701
TEFModel1	[-137,-7]	[-6,-1]	$\ln(f) = 0.1042 * \Delta G_1 + 41.5185 * \Delta G_2 - 0.6856$	0.5922	1.5128	1.5226
GPDModel1	[-115,-98]	[-53,19]	$\ln(f) = 2.3378 * \Delta G_1 + 0.1227 * \Delta G_2 - 1.4524$	0.8340	1.2294	0.02174
CYC1Model2	[-105,-95]	[-53,-5]	$\ln(f) = 1.1331 * \Delta G_1 + 0.0936 * \Delta G_2 - 0.1545$	0.8600	0.1904	
TEFModel2	[-93,-87]	[-32,-8]	$\ln(f) = 106.9974 * \Delta G_1 + 0.3197 * \Delta G_2 + 0.4363$	0.9100	0.2278	
GPDModel2	[-126,-99]	[-76,-4]	$\ln(f) = 0.6411 * \Delta G_1 + 0.1221 * \Delta G_2 + 1.2860$	0.9536	0.2264	

Table 4-2: Computational Models of yECitrine Fluorescence based on 5'UTR structure

Indicated regions are measured relative to the first nucleotide of the start codon. The correlation coefficient was computed for all data available at the time of model training. The Predicted Residual Sum of Squares was computed with the hat matrix after regression. The Residual Sum of Squares was computed for CYCModel1, TEFModel1, and GPDModel1 with the natural log of the data from pCYC1$_2$xYFP, pTEF$_2$xYFP, and pGPD$_2$xYFP, respectively.

The redesigned MCS for the *GPD1* promoter exhibited superior performance over the original, unoptimized MCS (**Figure 4-7A**). This new MCS, GPD$_2$, shows negligible multicloning site inhibition for the first eight restriction sites which, coupled with high levels of yECitrine expression, makes this the ideal MCS for this strong promoter (**Figure 4-7A**). Furthermore, this trend was predicted by GPDModel1, lending support to the hypothesis that protein expression is influenced by secondary structure in a few key regions of the 5'UTR (**Figure 4-8A**). The excellent agreement between model and observation suggests that secondary structure may be the only significant translational rate-limiting step in protein expression for this extraordinarily strong promoter with a short, codon-optimized protein.

In further extension of this approach, the TEF-promoter-specific MCS TEF$_2$ shows improved performance over pBLUESCRIPT SK or TEF$_1$, exhibiting similar or increased expression levels across the sites in the MCS (**Figure 4-7B**). Furthermore, the observed expression trend was predicted remarkably well by TEFModel1, showing that

mRNA structure is also major limiting factor in this promoter (**Figure 4-8B**), albeit not as limiting as in the GPD promoter case.

Applying this approach for a yet weaker promoter (pCYC1), a new MCS, $CYC1_2$, was designed that provides better, more consistent performance across the first four restriction sites than $CYC1_1$ or pBLUESCRIPT SK (**Figure 4-7C**). However, $CYC1_1$ (**Figure 4-5**) provides better performance than $CYC1_2$ or pBLUESCRIPT SK when cloning after the fourth restriction site. The measured performance of $CYC1_2$ was well predicted by CYC1Model1, validating its predictive ability (**Figure 4-8C**).

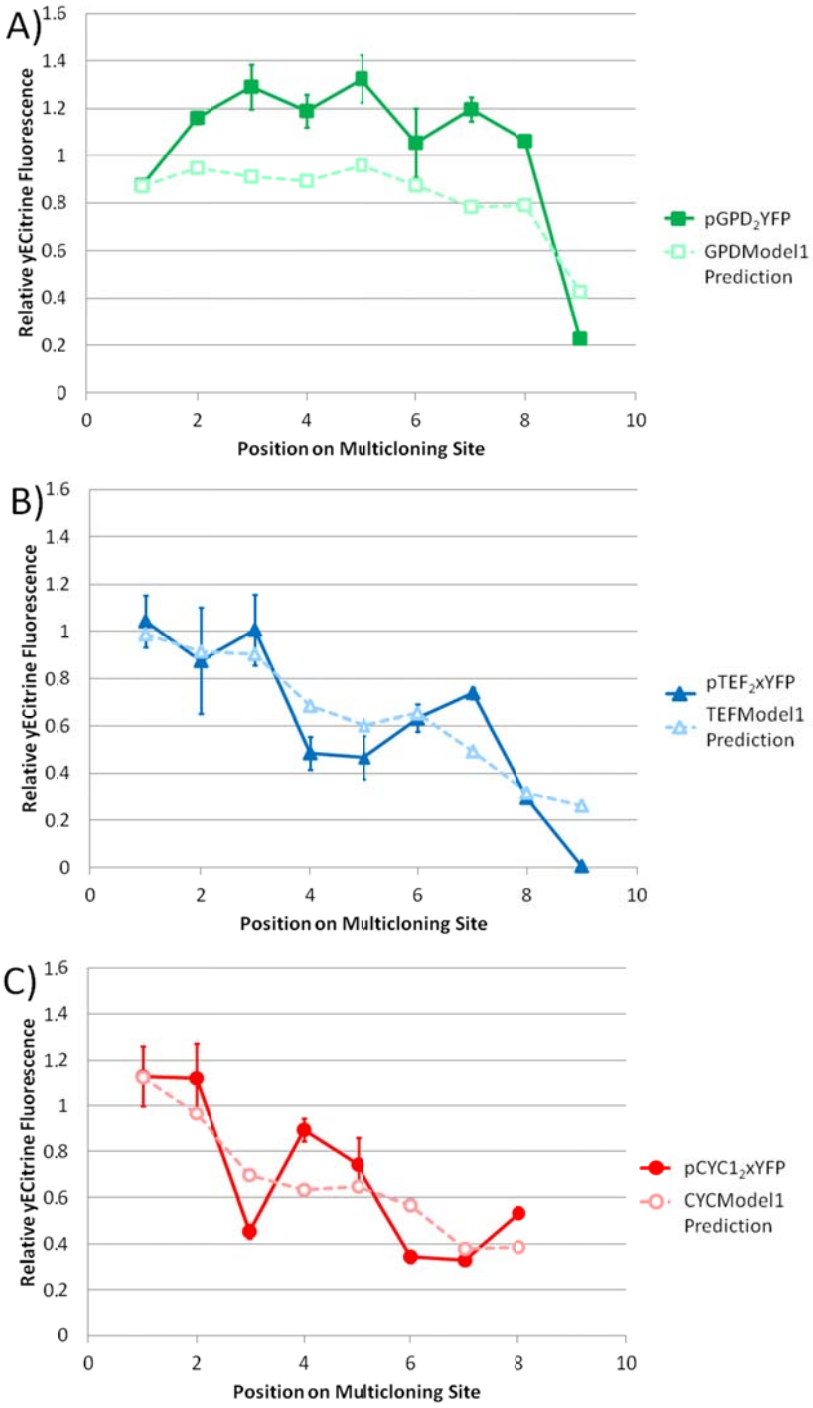


Figure 4-8: Predicted Performance of Designed Multicloning Sites (A) GPD_2, (B) TEF_2, and (C) $CYC1_2$

Prospective MCSs were designed according to the procedures described in Materials and Methods. Observed values for the expression levels allowed by each designed multicloning site are plotted as in **Figure 4-7**. Position on the MCS has been measured according to the unique restriction sites in the p416 vector. Error bars represent the standard deviation in fluorescence observed across three biological replicates. Designed multicloning sites show good agreement with model predictions.

Taken together, these results indicate that the expression-inhibiting effects of multicloning sites can be substantially mitigated in a variety of transcriptional contexts through minimization of 5'UTR secondary structure. In addition, no designed MCS elicited a significant change in gene expression noise, indicating that these constructs are ideal for development of precisely controlled gene networks **(Figure 4-9)**. However, it should be noted that neither TEF_2 nor $CYC1_2$ matched the outstanding performance of GPD_2, either due to random errors in the modeling process or due to the manifestation of other rate-limiting steps in expression not accounted for in our simplistic structure-based model of expression. As pTEF and pCYC1 are both substantially weaker promoters than pGPD, the presence of additional rate-limiting factors (possibly stemming at the transcriptional level) is not surprising. Finally, all data collected above was used to upgrade the weighting factors and relevant 5'UTR regions in our models (**Table 4-2**). These upgraded models are expected to give researchers more accurate predictions of 5'UTR structure-based inhibition of protein expression in yeast.

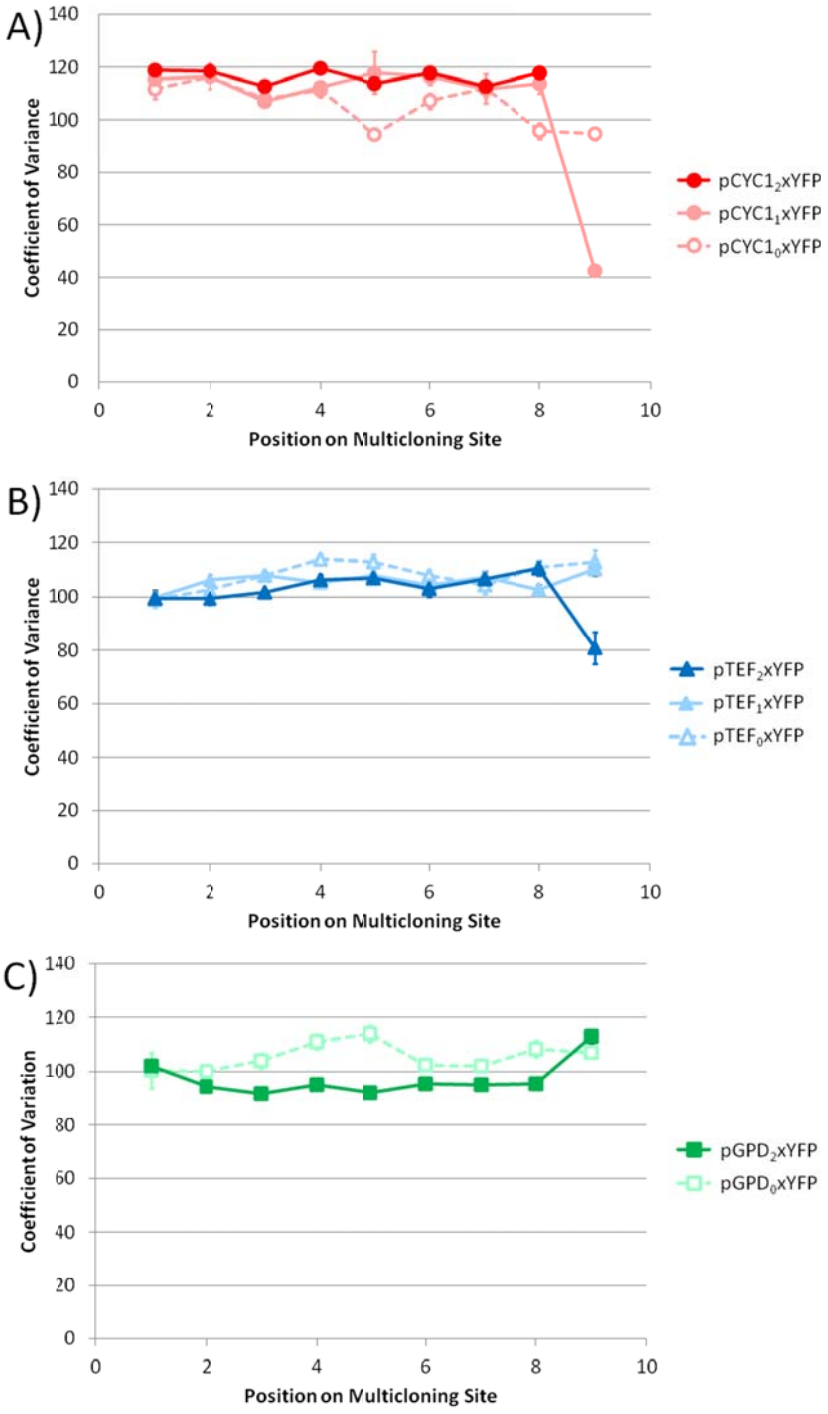


Figure 4-9: Effects of Designed MCSs on Expression Noise in (A) CYC1 MCSs, (B) TEF MCSs, and (C) GPD MCSs.

Expression noise is seen to be largely invariant with respect to restriction site and MCS. pCYC$1_1$9YFP and pTEF$_2$9YFP had fluorescence near the detection limits of our flow cytometer, leading to the decreased coefficients of variants seen in these constructs.

4.3 DISCUSSION

We have demonstrated that simplistic models of 5'UTR RNA secondary structure can be used to predict and rationally design multicloning site performance. The approach defined here is novel and significant for several reasons: (1) most modeling and prediction efforts in this area have examined prokaryotic systems (esp. for ribosome binding sites) whereas this work utilizes yeast, a eukaryotic system. The mechanics of eukaryotic translation are sufficiently different and require a novel mechanistic approach. (2) Most prior studies evaluate the impact of 5' hairpin loops and their inhibitory effect on translation, especially when sequestering the start AUG. In contrast, our work demonstrates that the observed translation inhibition by structure was highly dependent on the position of the secondary structure, and not always a set distance from the transcription initiation site. (3) Most prior studies evaluate the impact of specific point mutations that can change secondary structure. No prior work has successfully predicted and achieved a global redesign of a genetic circuit of such widespread importance as a multicloning site.

In contrast to prior studies, this method of prediction and optimization of 5'UTR structure is valid in a general context, enabling significant increases in expression despite the implementation of a diverse set of promoters and restriction sites. This aspect of translation-level control seems to be most strongly pronounced when expressing short, codon-optimized gene products. Moreover, this effect exhibits a promoter-specific nature implying that individual components of gene expression cassettes cannot be designed in isolation. It is also important to note that this phenomenon is not a generic effect of 5'UTR length, as indicated by (1) the significant increases in expression observed upon

adding length to the 5'UTR, and (2) the inability of one-part folding models to predict the behavior of TEF_1 and $CYC1_1$. Although this effect was first experimentally characterized here for pBLUESCRIPT SK, it is expected that other MCSs will behave similarly in yeast and perhaps other eukaryotes. In particular, 5'UTR based folding models predict that significant secondary structure issues can arise in other common MCSs such as the one present in pUC. As a result, it is important to understand and appreciate this impact especially when attempting to compare experiments or genes cloned into distinct sites.

Optimization of 5'UTR secondary structure therefore represents a facile and cost-effective way to increase protein expression and product titers in eukaryotic bioprocesses, especially when it is undesirable to change promoters. Designed MCSs were found to be superior to the multicloning site found in the commonly used pBLUESCRIPT SK plasmids, and in the case of GPD_2 showed negligible activity reduction along the MCS. This experiment shows not only that MCSs have a significant effect on translation, but also that MCSs can be rationally engineered to mitigate this effect. Such a model-based optimization approach is unprecedented for this ubiquitous genetic component and highlights the importance of rational design in synthetic biology. It is expected that a similar approach can be undertaken for other eukaryotic expression vectors. Control of 5'UTR secondary structure also represents an alternative to promoter engineering, allowing protein expression to be controllably weakened by up to an order of magnitude without altering the dynamics of its regulation.

As we have demonstrated, optimization of 5'UTR secondary structure is context-specific, making the performance of each multicloning site highly dependent on the upstream promoter. It is not unreasonable to expect that the nucleotides of the open reading frame could also participate in translation-inhibiting secondary structure.

Therefore, in cases where inhibition due to secondary structure is significant (i.e. in highly codon optimized genes), the assumption of interchangeability of promoter, MCS, and gene becomes highly questionable. These results go against several of the tenets of synthetic biology, especially with respect to the assumption of completely interchangeable, non-interacting parts, and are part of a growing body of work indicating the non-modularity of genetic components (188). Yet, as the cost of gene synthesis decreases, these results demonstrate that it is more desirable to create entire self-sustained transcriptional/translational units—from promoter to terminator. This paradigm is in contrast to the widespread assumption that two arbitrary sequences, when attached, will not generate translation-inhibiting secondary structure.

These results have significant implications beyond redesign of gene expression cassettes. Expression vectors with multiple cloning sites have seen widespread use across the field of functional genetics and basic cloning. Given the strong difference in performance across sites in the MCS, experiments and conclusions will be highly dependent on these sites. Therefore, conclusions about gene impact, function, or activity as well as promoter strength analysis will depend highly on the cloning sites used. As a result, many conflicting results and conclusions may be attributed to this phenomenon. More broadly, this research shows that the secondary structure inherent to the 5'UTR has significant impacts upon the efficiency of translation initiation. Any mRNA, whether it has been derived from a natural system or designed synthetically, will contain a 5'UTR with this regulatory potential. Therefore, it is imperative for metabolic engineers to design synthetic constructs with this initiation efficiency in mind, especially because this will be the rate-limiting step for the production of codon-optimized genes which are otherwise translationally optimal.

In conclusion, we have demonstrated the first performance-based analysis of multiple cloning sites in yeast systems. Following this, we have shown that a simplistic model of 5'UTR secondary structure with two regions can predict this phenomenon when it is the most dominant determinant of protein translation. Under these conditions, we have for the first time successfully redesigned multiple cloning sites for function rather than simple convenience. It is anticipated that this work can be extended to other vectors and potentially to other organisms, both eukaryotic and prokaryotic alike. The capacity to design MCSs with consistent performance across multiple cloning sites will greatly impact the ease and utility of recombinant cloning and genetic analysis.

Chapter 5: Development of Operons in Yeast through 2A Peptides

5.1 INTRODUCTION

For metabolic engineering or synthetic biology applications, it is often desired to co-regulate the expression of several gene products such that the pathway or circuit of interest functions in a coordinated and efficient manner. Although the number of well-characterized promoters is growing rapidly for many organisms, there remains a lack of non-homologous promoters which enable co-regulation (i.e. perform the same regulatory function). Therefore, co-regulation necessarily introduces instability to yeast vectors, as regions of high sequence homology are a prime target for recombination-induced excision. However, even if a collection of non-homologous co-regulatory DNA parts existed, the structure of yeast expression cassettes poses a challenge. In particular, the extra DNA space needed to encode separate promoters and terminators for each gene poses a significant synthesis cost (sometimes as much as 50% of the total cost of the construct) to metabolic engineers. Although translational fusions can overcome these issues for applications in which co-localization of pathway enzymes is appropriate (189), there remains a need for a general strategy for co-regulation of gene products in yeast which is not susceptible to homologous recombination-induced construct instability.

In prokaryotes, co-regulation can be achieved in a facile manner through the use of operons, in which a single promoter controls the transcription of multiple genes. Translation rates of each gene are controlled by the strength of the ribosome binding site which precedes each open reading frame. Although the rules regarding ribosome binding and translation are distinct in eukaryotic systems, this organizational paradigm provides an attractive alternative to yeast expression cassettes in their current implementation.

In order to enable to facile co-regulation of multiple genes in yeast, we endeavored to characterize the activity of known 2A sites in yeast and implement them

for the facile expression of multigene pathways. This capability would enable significant cost savings in terms of DNA synthesis and assembly (2A sites are only ~60bp whereas promoters and terminators in yeast often measure 400bp and 200bp, respectively) as well as reduce the likelihood of homologous-recombination associated instability. Finally, the implementation of 2A sites in yeast would enable facile co-regulation of multiple genes. Although a publication demonstrating the utility of 2A sites as engineering tools was released as this research was being undertaken (190), we endeavored to complete this research in order to lay the foundation for future work developing a more highly optimal 2A site system.

5.2 RESULTS

5.2.1 Characterization of a Panel of 2A Sites

Although many viruses have been shown to contain 2A sites, sites from three viruses are most commonly used in studies characterizing function and demonstrating efficacy: equine rhinitis A virus, porcine teschovirus-1, and *Thoseaasigna* virus (42). Codon-optimized versions (191) of these 2A sites (E2A, P2A, and T2A) were cloned into a bicistronic reporter construct to test 2A site function, in which mStrawberry comprised the first cistron and YFP comprised the second. In this method, a functional 2A site would enable a high level of mStrawberry and YFP production in the presence of galactose, but not in the presence of glucose. We also included a his-tag to the c-terminus of YFP to facilitate characterization by western blotting. To identify putatively functional 2A sites, cells expressing bicistronic cassettes containing either 2A site were analyzed with flow cytometry (**Figure 5-1**). It can be seen that, unlike the other sites, T2A is nonfunctional in this context, exhibiting background levels of expression of the second cistron upon galactose induction, and reduced expression of the first cistron.

Protein extracts from cells expressing the remaining 2A sites (E2A and P2A) were then analyzed through western blotting (**Figure 5-2**). It can be seen that while the cleaved product is visible in the case of P2A, no cleavage can be seen in the case of E2A. This indicates that while E2A enables translation of the entire bicistronic construct, it does not enable separation of the two gene products. However, cleavage enabled by P2A appears to be quite efficient, with minimal production of the uncleaved product. Interestingly, these results stand in direct contradiction to work indicating that T2A was functional and that P2A was nonfunctional in a related strain of yeast (190). Nevertheless, this data confirms that the P2A site is functional in BY4741, thus opening the door to the use of operon architecture for the construction of pathways in this system. Furthermore, it is exciting to note that through the use of a 2A site, inducible co-expression of two genes was enabled in a facile manner, which until this point required the expression of riboswitches (38).

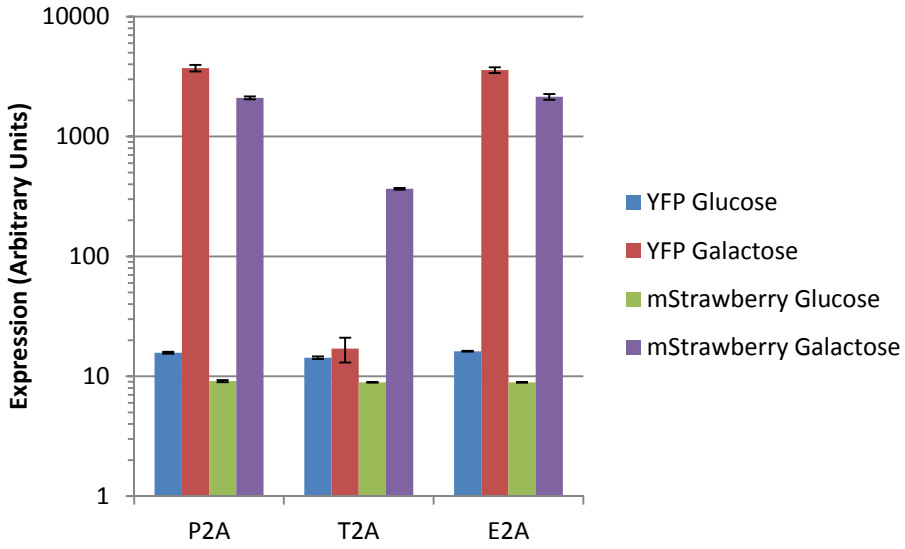


Figure 5-1: Characterization of a panel of 2A sites using flow cytometry

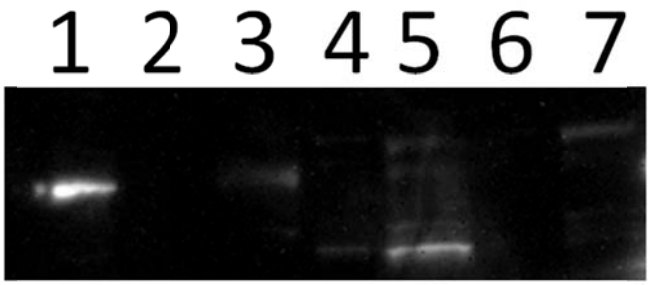


Figure 5-2: Characterization of E2A and P2A activity with western blotting.

Total protein from cells expressing the bicistronic reporter cassette with either P2A or E2A were analyzed with western blotting. Lane 1: 47 kDa standard, Lane 2: 10x diluted extract of P2A in glucose, Lane 3: extract of P2A in glucose, Lane 4: 10x diluted P2A in galactose, Lane 5: P2A in galactose, Lane 6: 10x diluted E2A in galactose, Lane 7: E2A in galactose. Fusions of YFP and mStrawberry are 56 kDa in size, whereas YFP alone is 23 kDa.

5.2.2 Generation of P2A Variants

In order to enable the construction of operons with greater than two open reading frames with minimum propensity for homologous recombination-associated instability, it is necessary to have several functional variants of P2A with low homology to each other for placement at the junction between each gene. Because P2A functionality is thought to be the result of peptide sequence, generation of alternative P2A sites can be readily accomplished by altering codon usage. In this way, P2Av2 was constructed, which differs from P2A by 22 of its 66 nucleotides. We also wished to construct a P2A variant which would serve as a negative control in which no cleavage was enabled. The use of this nonfunctional P2A variant would serve to illustrate the performance of the pathway of interest when its constituent genes were translationally fused. As fusion proteins have, in some cases, shown markedly improved performance relative to when the proteins are expressed separately (189), it is important to show that any improvements enabled through the use of a 2A site could not be obtained through construction of a simple fusion protiein. This nonfunctional P2A variant, P2Ad, was identical to P2A with the substitution of a phenylalanine for proline at the terminal position. This substitution has

been previously shown to result in the formation of a defective 2A site (192). To confirm the activity of these variant 2A sites, extracts from cells expressing the appropriate bicistronic reporter construct were analyzed through western blotting (**Figure 5-3**). Although P2Ad functioned as expected, interestingly P2Av2 showed very low cleavage ability despite having an identical amino acid sequence to P2A. This implied that codon usage may have an effect on the ability of 2A sites to enable ribosome stalling.

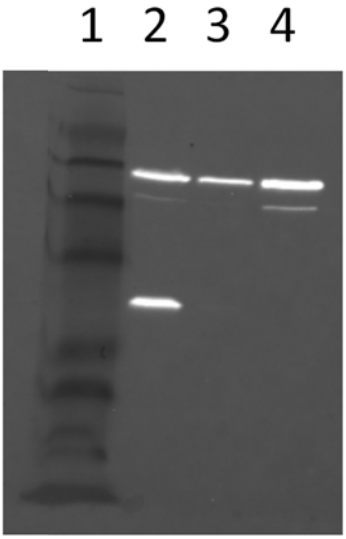


Figure 5-3: Characterization of P2A, P2Av2, and P2Ad with western blotting.

Total protein from cells expressing the bicistronic reporter cassette with P2A, P2Av2, or P2Ad were analyzed with western blotting. Lane 1: ColorPlus™ Prestained Protein Marker, Broad Range (NEB), Lane 2: P2A, Lane 3: P2Av2, Lane 4: P2Ad. Fusions of YFP and mStrawberry are 56 kDa in size, whereas YFP alone is 23 kDa.

5.3 DISCUSSION

In this work, 2A sites from several viruses were tested in yeast and it was shown that of these, only the site from porcine teschovirus-1 was functional. It is reasonable that 2A sites would demonstrate some host specificity, as polypeptide cleavage is enabled by ribosome stalling, and both ribosome and tRNA structures are known to vary among organisms. It is interesting to note that other studies have found P2A to be nonfunctional

in a related strain of *S. cerevisiae* (CEN.PK2) and T2A, rather, to be functional. The ability to use 2A sites in yeast is very exciting, not only enabling significant cost savings in terms of DNA synthesis, but also reducing the likelihood of homologous recombination-associated construct instability. Furthermore, this site also enables facile co-regulation of gene products, which has not heretofore been easily achieved for yeast. It is also interesting to note that generation of alternative 2A sites through re-coding was not completely successful – use of a P2A with alternative codons completely abolished polypeptide cleavage. This indicates that peptide sequence, although important, may not be the only factor influencing 2A site functionality. In addition, translation rate or tRNA structure may also have an effect. This finding may shed light on our observation that P2A was functional in yeast while T2A was not, contradicting the results of an earlier study. In fact, the 2A sites in other studies used an alternative coding for P2A and T2A. This apparent contradiction emphasizes the importance of elucidating the precise requirements for 2A site functionality and efficiency. Once these rules have been identified, it will be feasible to use 2A sites as a generic tool for the facile construction and regulation of multi-gene pathways in yeast.

Chapter 6: Tuning Translation through Internal Ribosome Entry in Yeast

6.1 INTRODUCTION

As an alternate method to enable polycistronic gene expression in eukaryotes, viruses have developed unique RNA structures which enable internal ribosome entry. In this work, we have attempted to use a combination of random and site-directed mutagenic approaches to develop efficient IRESs in yeast. In the process, we have characterized most IRESs previously reported to function in this organism as well as attempted to identify translational machinery which may be inhibitory towards the functionality of IRESs in this organism. Despite our best efforts, this approach was unsuccessful at identifying a definitively improved IRES, as detailed below.

6.2 RESULTS

6.2.1 Initial IRES Library

Our first attempt at developing an IRES in yeast focused on random mutagenesis of the EMCV IRES (193). In addition, randomized 50-nucleotide DNA segments were screened for IRES activity, following an earlier report claiming that such templates contained functional yeast IRESs (194). In order to detect IRES activity, we used a bicistronic reporter cassette consisting of the GPD promoter driving the expression of HIS3 in the first cistron and YFP in the second cistron. These two open reading frames were separated by a library of IRES candidates and screened for high YFP expression using flow cytometry. Four libraries were constructed with varying levels of mutagenesis: High, Medium and Low, and Very Low. The library sizes obtained for this experiment were as follows: High: 124k, Medium: 124k, Low: 160k, Very Low: 166k, 50N: 11k. After screening through fluorescence activated cell sorting, several variants

were isolated which enabled higher YFP expression than EMCV (**Figure 6-1**). These variants were sequenced and the unique variants were re-transformed into yeast and characterized (**Figure 6-2**). Unfortunately, no re-characterized variants displayed increased activity compared with EMCV. It was hypothesized that the inability to isolate functional variants was due to insufficient library size, so yeast homologous recombination was used to expedite the library construction process in the next library.

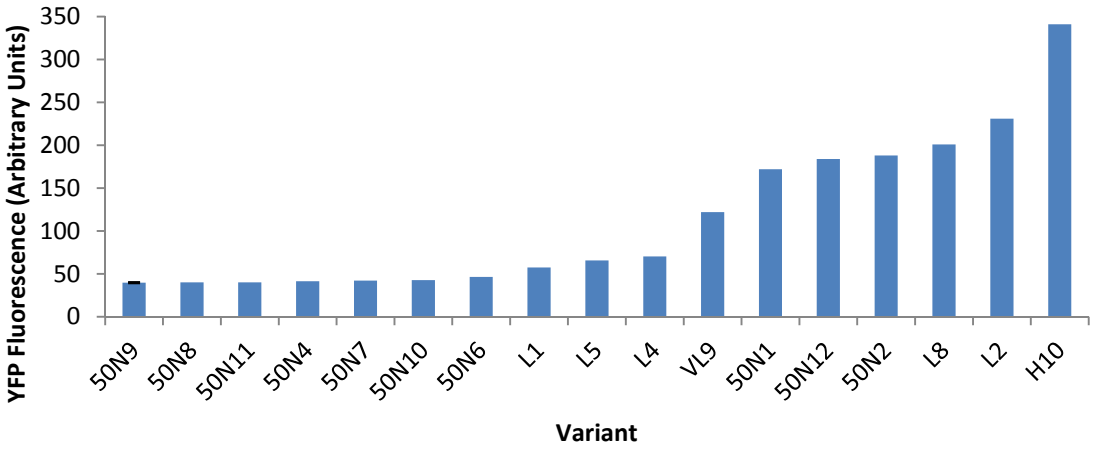


Figure 6-1: EMCV and 50N Isolates obtained from IRES Library 3.

Units of YFP fluorescence for this and all following figures have been arbitrarily defined by the flow cytometer upon which the measurements were taken. In most cases, yeast autofluorescence has a value between 20 and 30 arbitrary units across experiments.

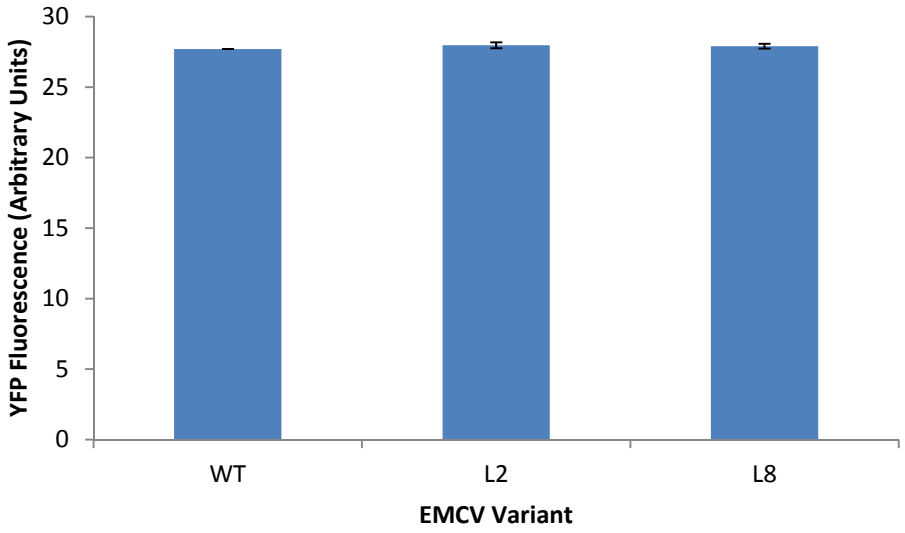


Figure 6-2: Re-characterization of EMCV isolates obtained from IRES Library 3.

6.2.2 IRES Screening on a High Copy Vector

EMCV was subjected to random mutagenesis and screening as before, with the exception that variant libraries were generated in a high copy number vector. Four libraries were constructed with varying levels of mutagenesis: High, Medium and Low, and Very Low. In addition, a library containing a randomized 50bp sequence was constructed in the same vector. The library sizes obtained for this experiment were as follows: High: 406k, Medium: 389k, Low: 372k, Very Low: 321k, 50N: 53k. After screening, several variants were isolated which enabled higher YFP expression than EMCV (**Figure 6-3**). These variants were then sequenced and the unique variants were re-transformed into yeast and characterized (**Figure 6-4**). We were pleased to observe that many IRESs maintained significantly higher YFP expression than wild-type. Therefore, we replaced the GPD promoter in these constructs with a terminator in order to test promoter activity of these putative IRESs. Of these, four (50NB4, 50NB8, 50ND3, and 50ND7, which were each derived from the 50N library) showed similar

promoter activity to wild-type EMCV (**Figure 6-5**). This was very exciting, as these variants were the best-performing candidates identified to date.

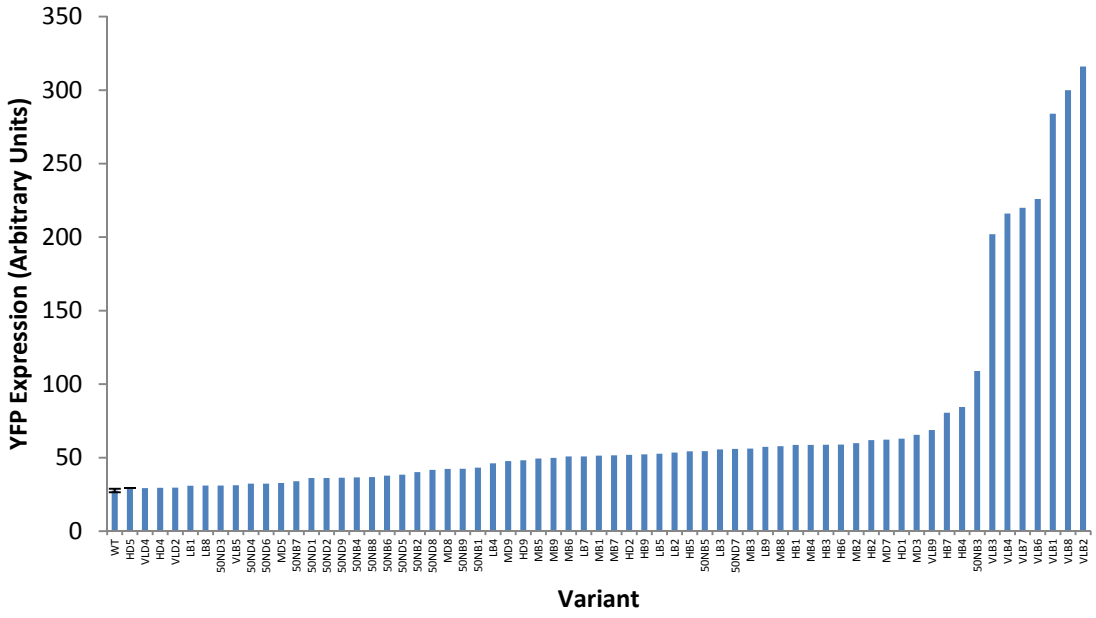


Figure 6-3: EMCV and 50N Isolates obtained from IRES Library 5

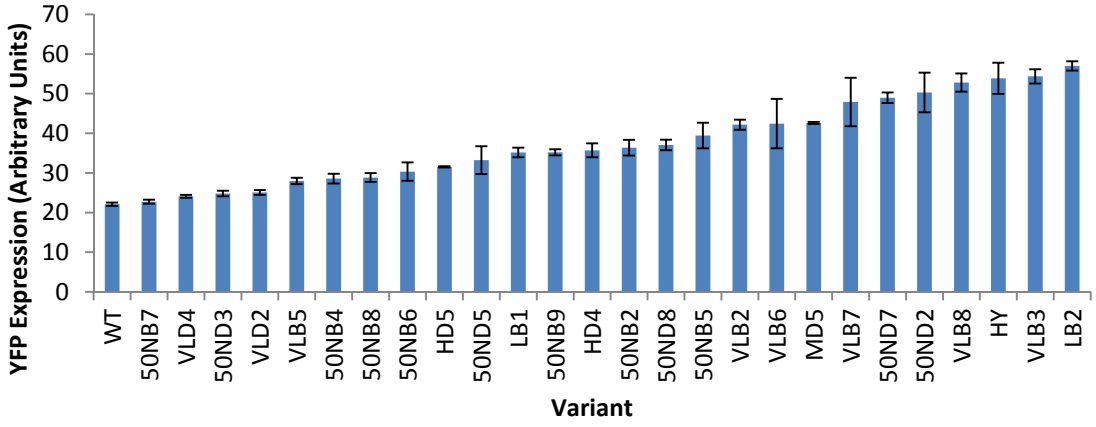


Figure 6-4: Re-characterization of isolates obtained from IRES Library 5

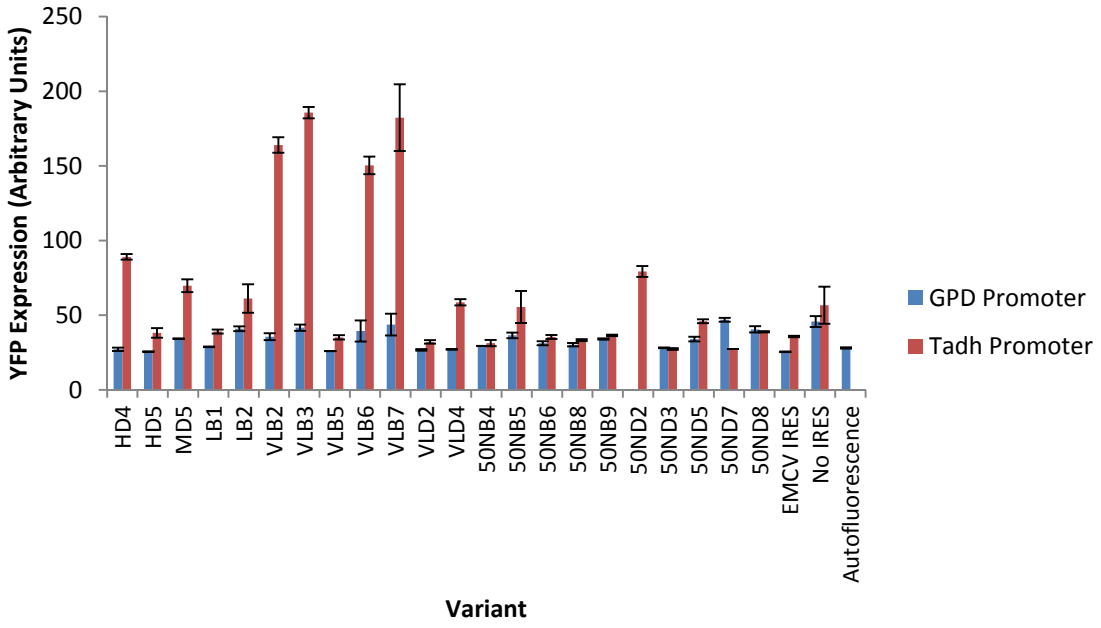


Figure 6-5: Characterization of promoter activity enabled by IRES Library 5 isolates

It was at this time that we noticed a shortcoming with our screening vector. We observed that the HIS3 gene which was placed upstream of YFP had substantial promoter activity. This phenomenon could be due to the fact that HIS3 is located only 300bp upstream of DED1 in the yeast genome and thus may contain sequences necessary for the transcriptional regulation of that gene. Nevertheless, we assayed the promoter activity of a panel of fluorescent proteins in order to identify a substitute (**Figure 6-6**). mStrawberry was chosen to replace HIS3 in our screening cassette, as it exhibits minimal promoter activity and negligible homology to YFP. We then constructed two variants of this vector: one with GPD replaced by a terminator and one with a stemloop placed after GPD (**Figure 6-7**). These three screening vectors would enable us to test each candidate's promoter activity as well as their propensity to form translational fusions in a consistent manner.

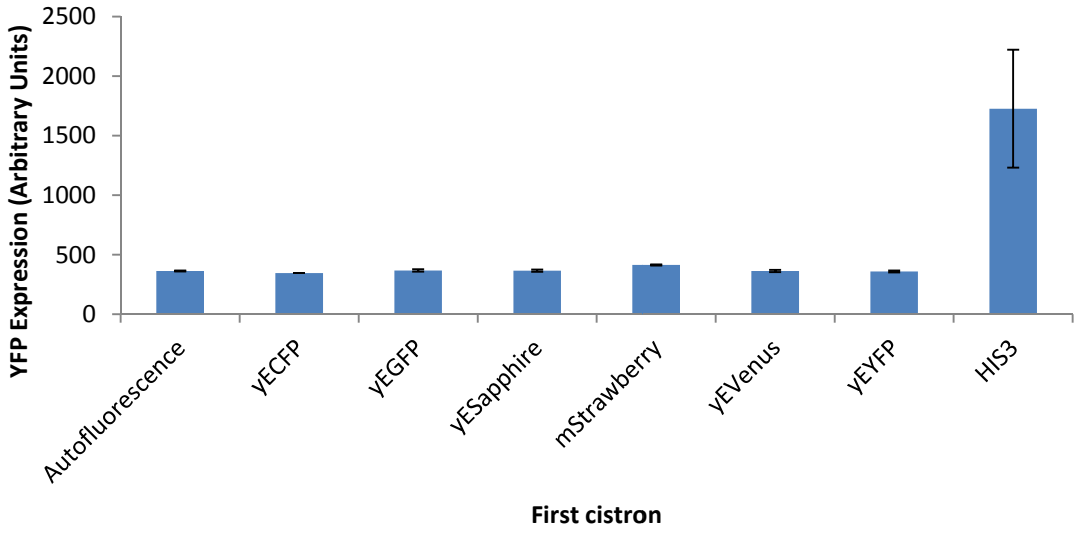


Figure 6-6: Measurement of promoter activity conferred by several reporter genes

The indicated reporter gene was cloned upstream of YFP in the IRES screening vector and fluorescence was measured with flow cytometry.

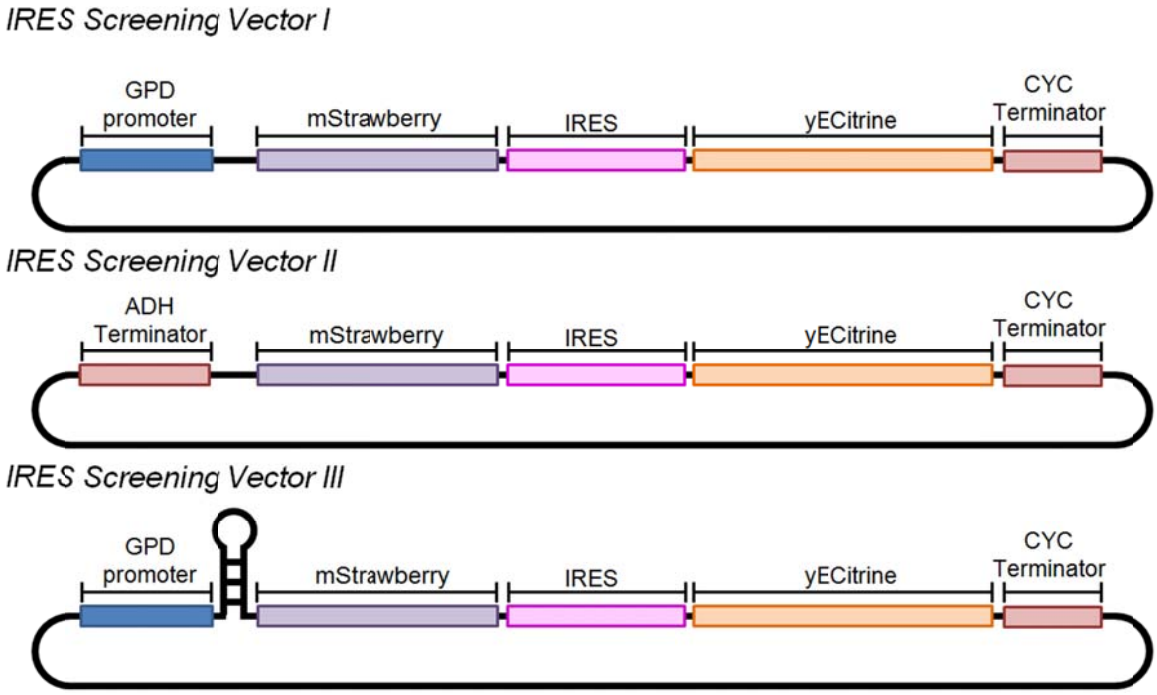


Figure 6-7: Updated screening vectors for characterization of IRES activity

With this updated set screening vector, we wished to confirm the IRES functionality of the candidates mentioned above. In addition, we assayed a panel of IRESs from the *dicistroviridae* family (PSIV, HIPV, and CrPV) recently shown to have some activity in yeast (53) (**Figure 6-8**). Unfortunately, it was observed that of the four IRES candidates isolated above, the activity shown by 50NB4, 50NB8, 50ND3 and 50ND7 was due to substantial promoter activity. The promoter activity of these contstricts was not identified using the previous screening construct because of interference from the high promoter activity of HIS3. This data also showed that EMCV contained substantial promoter activity, whereas the *dicistroviridae* IRESs contained almost undetectable amounts of promoter activity, with slight levels of IRES functionality. It is important to note that becase the *dicistroviridae* IRESs enable translation in an AUG-independent manner, the start codon of YFP was removed for these constructs in order to reduce promoter-derived YFP expression (50). We also undertook a characterization of many other reported IRESs using this screening system. It has been reported that the 5' untranslated regions of YAP1 and p150 contain IRESs which function in yeast (195). Therefore, we tested these putative cellular IRESs along with several viral IRESs (whitespot syndrome baculovirus IRES (SWSS) (196), turnip vein clearing virus IRES (crTMV) (197), and the IRES from the gypsy transposon (gypsy) (198)) (**Figure 6-9**). Through this analysis, it was concluded that the YAP1 IRES functioned as a strong promoter, the p150 IRES contained negligible IRES activity, SWSS enabled a small amount of IRES activity, and both crTMV and gypsy functioned as strong promoters as well. Taken together, these results indicated that the more promising starting points for future development of IRES functionality may be the three *dicistroviridae* IRESs characterized here as well as SWSS.

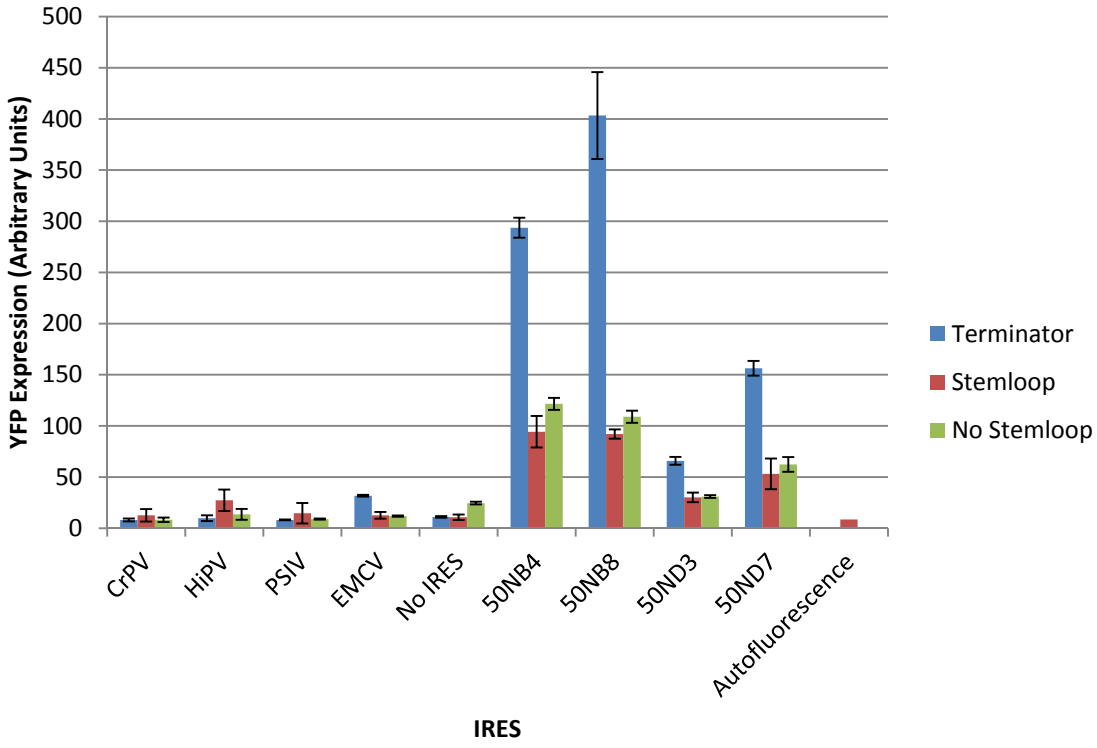


Figure 6-8: Characterization of *Dicistroviridae*, EMCV, and isolated IRESs using updated screening vectors

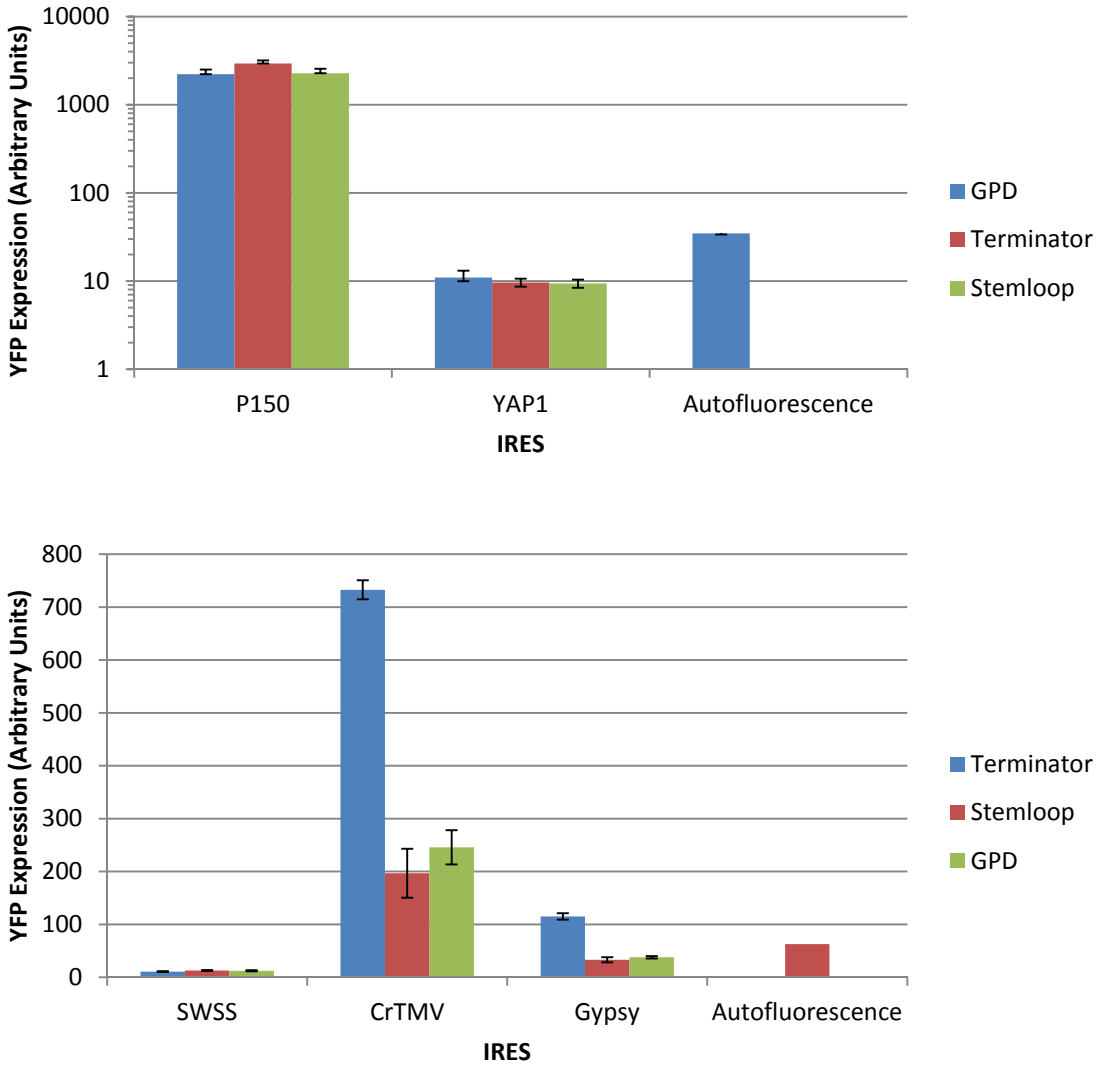


Figure 6-9: Characterization of alternative IRESs using updated screening system

6.2.3 Engineering *Dicistroviridae* IRESs

HIPV, PSIV, CrPV, and SWSS were subjected to random mutagenesis and screening in the new screening vectors. Three libraries were constructed for each template with varying levels of mutagenesis: High, Medium and Low. In addition, a library containing a randomized 50bp sequence was constructed and tested in parallel. The library sizes obtained for this experiment were as follows: CrPV Low: 87k, CrPV

Medium: 113k, CrPV High: 187k, HIPV Low: 102k, HIPV Medium: 69k, HIPV High: 72k, PSIV Low: 238k, PSIV Medium: 83k, PSIV High: 77k, SWSS Low: 15k, SWSS Medium: 25k, SWSS High: 38k, 50N: 2.7k. After screening, several variants were isolated which enabled higher YFP expression than their respective wild-type (**Figure 6-10**). These variants were then sequenced and cloned into the terminator-containing or the stemloop containing screening vectors for re-analysis (**Figure 6-11**). We observed two promising mutants (HM3 and SM7, derived from HiPV and SWSS, respectively) which enabled a higher level of YFP expression while exhibiting minimal promoter activity. However, this increase in YFP expression came at the expense of a slight decrease in mStrawberry expression. The cause of this decrease is unknown and it is unclear whether it is indicative of a false positive result. Although these results were promising, the high level of effort required to clone each IRES candidate into several screening vectors for re-characterization highlighted the need for a scheme for negative selection during screening to eliminate false positives due to promoter activity.

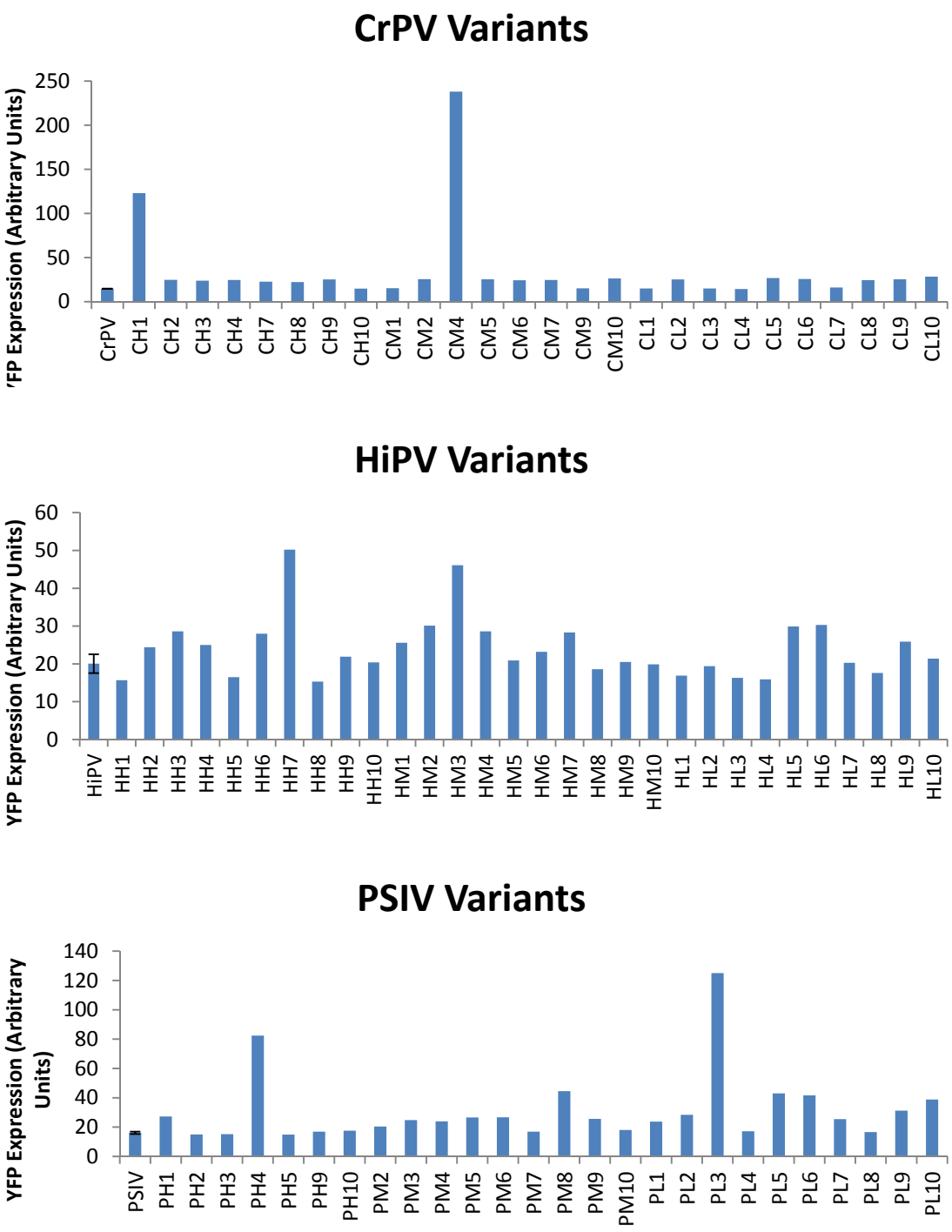


Figure 6-10: *Dicistroviridae* isolates obtained from IRES library 6

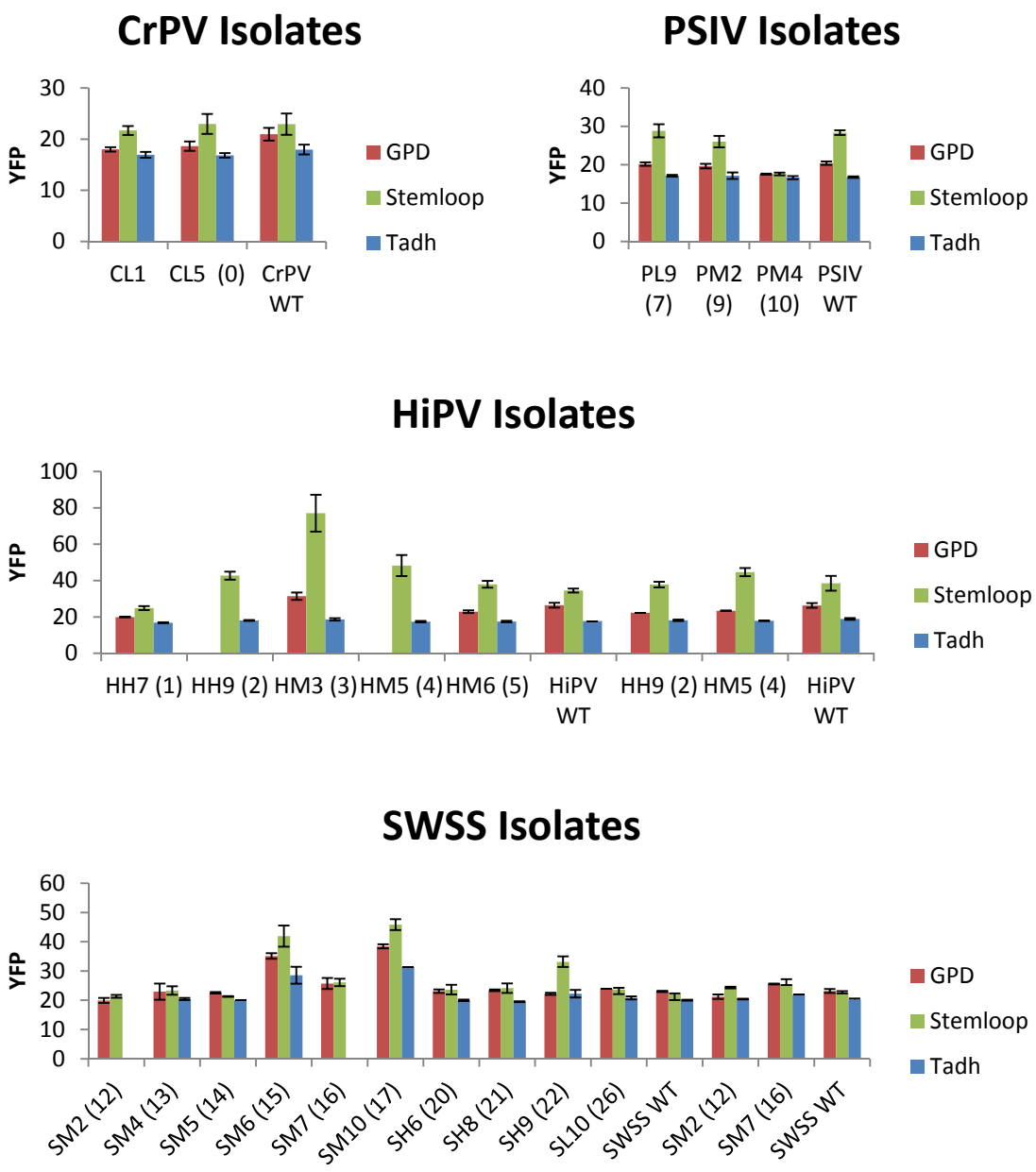


Figure 6-11: Re-characterization of isolates obtained from IRES library 6

6.2.4 IRES Screening with Inducible Promoter

In order to reduce the incidence of false positives, we replaced the GPD promoter of our screening vectors with the inducible *GAL1* promoter. In this way, IRESs could be screened for low promoter activity in a facile manner by selecting for cells which exhibit

low YFP expression upon growth in glucose. Several rounds of positive and negative selection may then be easily undertaken to enrich for sequences which enable high YFP expression in the absence of promoter activity. HIPV, PSIV, CrPV, SWSS, HM3, and SM7 were subjected to random mutagenesis and screening in the new screening vector. Three libraries were constructed for each template with varying levels of mutagenesis: High, Medium and Low. After screening, several variants were isolated which enabled higher YFP expression than their respective wild-type with low promoter activity (**Figure 6-12**). These variants were then retransformed into yeast to confirm activity (**Figure 6-13**). Unfortunately, it was observed all of the hits either exhibited negligible IRES activity or slightly increased promoter activity. After this disappointing result and in light of previous failed attempts to generate an IRES, we hypothesized that random mutagenesis may be a suboptimal strategy for identifying improved IRESs as IRES functionality is thought to result from the effect of RNA secondary structure, and random mutagenesis may have a high propensity to disrupt this structure. Therefore, we investigated alternative mutagenesis strategies in later work.

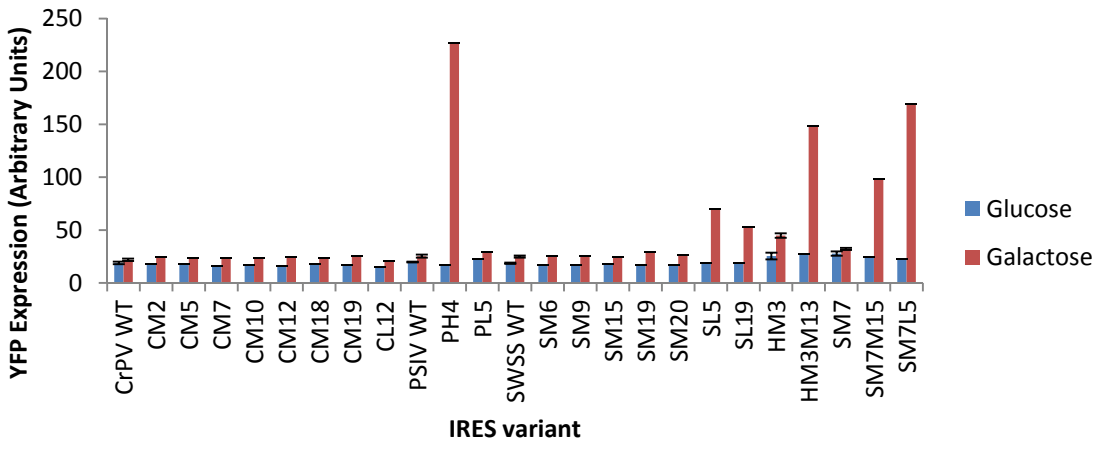


Figure 6-12: *Dicistroviridae* isolates obtained from IRES library 7

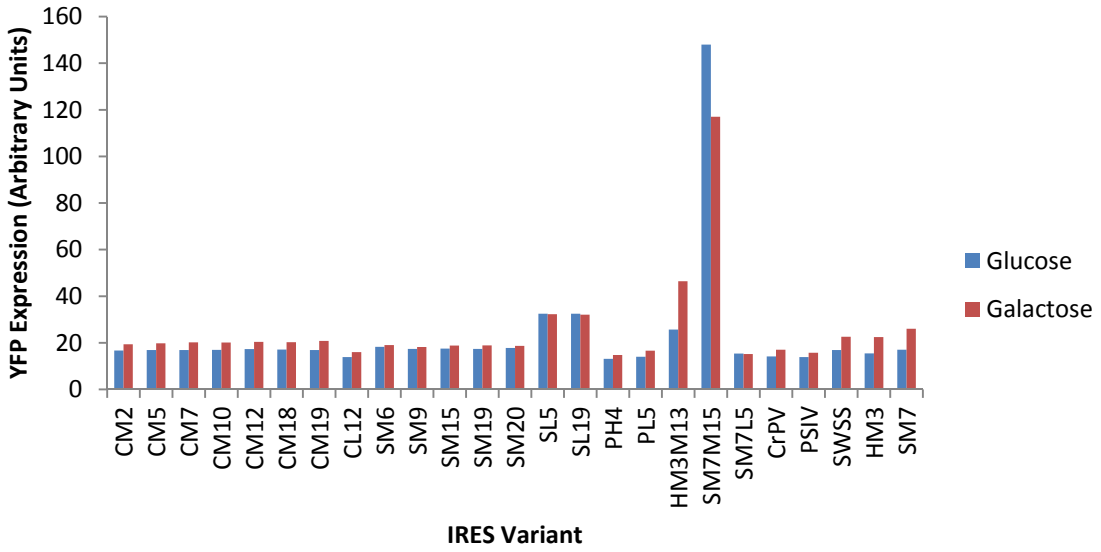


Figure 6-13: Re-characterization of isolates obtained from IRES Library 7

In addition, we characterized the activity of several additional putative IRESs using this screening system. It has been proposed that the *URE2* 5'UTR contains an internal ribosome entry site, and therefore we wished to confirm this finding (199) (**Figure 6-14**). We were also approached by the Jewett laboratory to confirm the *in vivo* efficacy of some yeast IRESs identified through an *in vitro* approach. Therefore, we cloned these sequences into our galactose screening vector and IRES activity was measured for these sequences (**Figure 6-15**). It can be seen that *URE2* appears to have slight IRES activity, while most of the IRESs from the Jewett lab function mainly as promoters, thus precluding attempts to characterize IRES activity. However, one variant, G38 appears to activate upon exposure to galactose, indicating either that this variant is a promising starting point for further development of IRES activity or that it is a galactose-responsive promoter. However, the fact that this construct enables a high amount of YFP expression in the presence of glucose indicates a high level of background promoter activity which may be problematic for future engineering efforts.

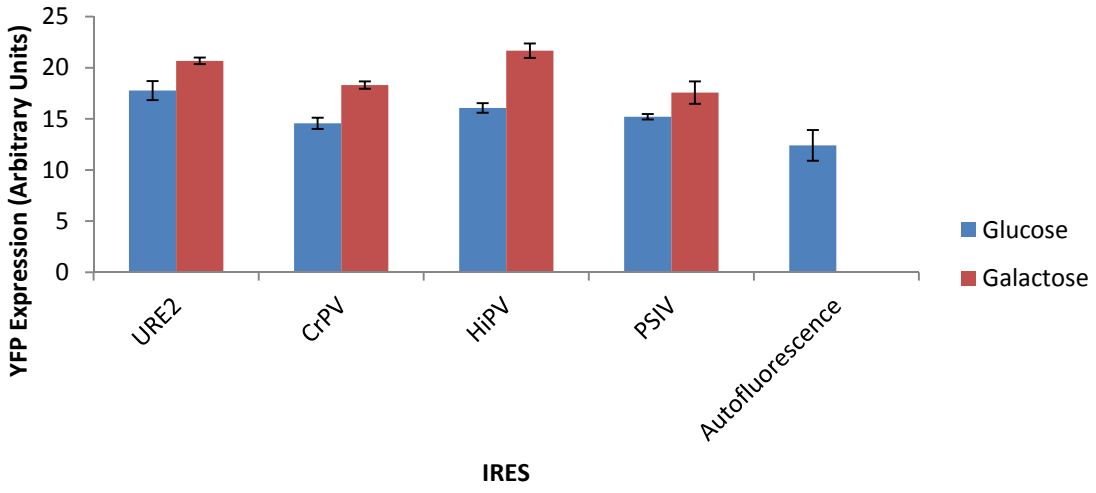


Figure 6-14: Characterization of the URE2 5'UTR using galactose screening vector

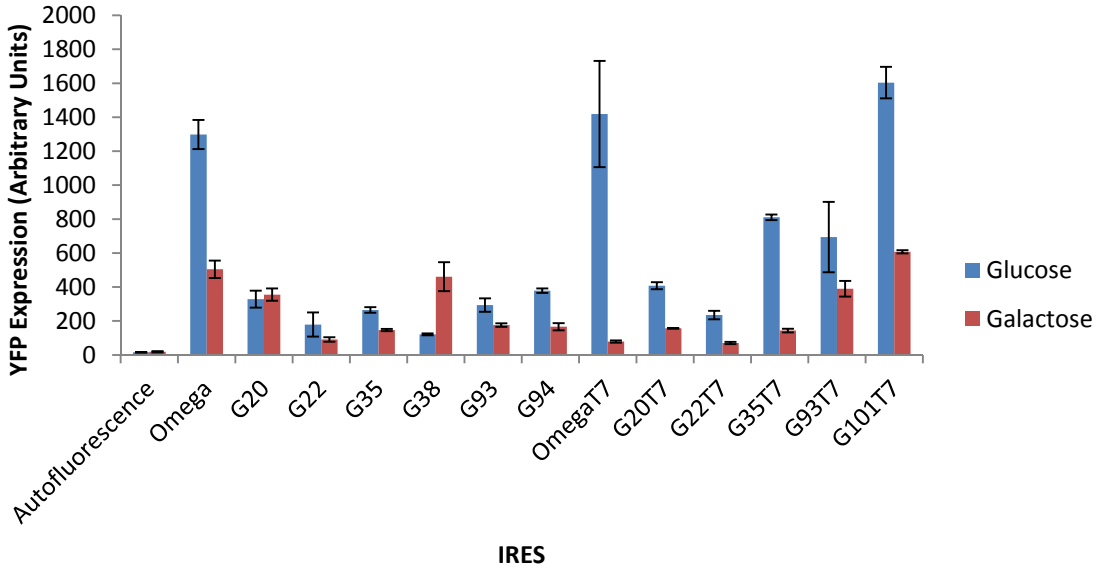


Figure 6-15: Characterization of IRES candidates from the Jewett lab using the galactose screening vector

6.2.5 Site-Directed Mutagenesis of IRESs

HIPV, PSIV and CrPV were subjected to site-directed mutagenesis as each of these IRESs have well-defined secondary structures (50). Each region of the IRES which

was predicted to be unhybridized at equilibrium was annotated and targeted for randomization. However, because of the large sequence space possible by randomizing each region in the context of each other, these regions were clustered into 5 groups of approximately equal number of randomized base pairs, thus forming 5 separate site-directed libraries for each template (**Figure 6-16**). The library sizes obtained for this experiment were as follows: CrPV #1: 480k, CrPV #2: 1056k, CrPV #3: 804k, CrPV #4: 888k, CrPV #5: 441k, HIPV #1: 639k, HIPV #2: 393k, HIVP #3: 462k, HIPV #4: 765k, HIPV #5: 396k, PSIV #1: 357k, PSIV #2: 384k, PSIV #3: 312k, PSIV #4: 261k, PSIV #5: 444k). After screening, several variants were isolated which enabled higher YFP expression than their respective wild-type (**Figure 6-17**). These variants were then sequenced, and it was found that all hits contained sequences which enabled translational fusions. This disappointing result spurred us to look at the problem of IRES functionality in yeast at a more fundamental level by probing the structure of the yeast translational machinery.

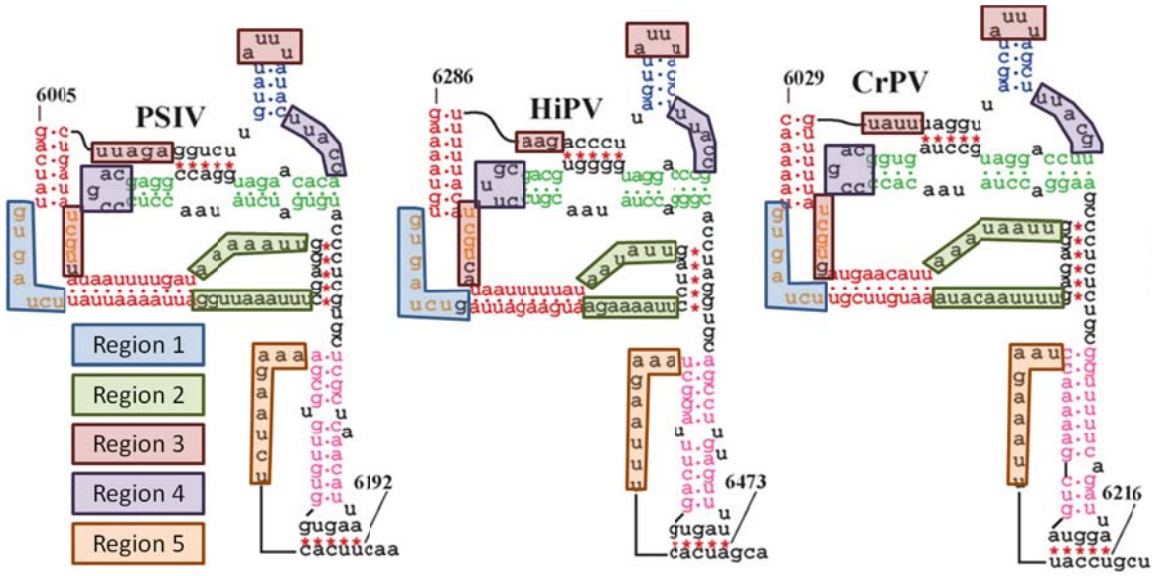


Figure 6-16: Schematic of regions targeted for site-directed mutagenesis for the *Dicistroviridae* IRESs.

Regions of similar color were randomized together to form 5 libraries for each enzyme. Figure adapted from (50).

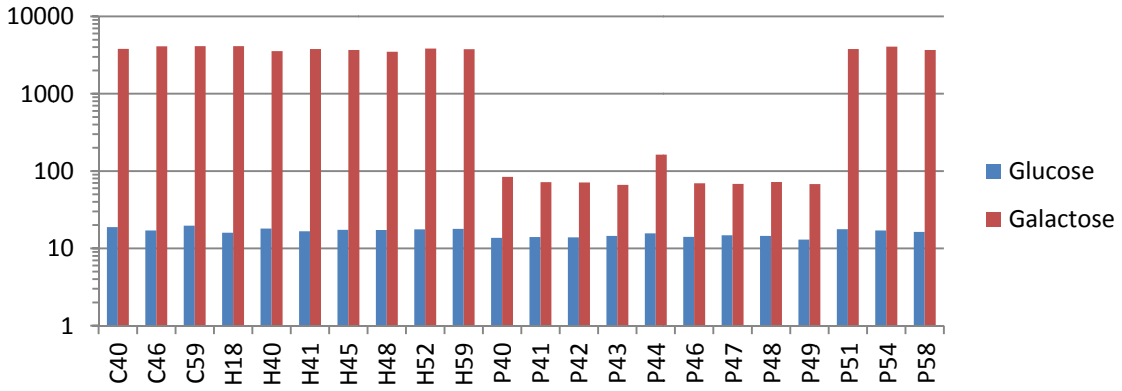


Figure 6-17: *Dicistroviridae* isolates obtained from IRES Library 8

6.2.10 Ribosomal Determinants of IRES Activity in Yeast

It has been reported that knockout of eukaryotic initiation factor 2A (*eif2a*) from yeast can greatly enhance the activity of cellular IRESs (200). It is thought that this phenomenon is due to the requirement that an AUG start codon be present in order for

EIF2Ap to bind met-tRNA and the 40S ribosome during the process of intiation complex formation. In the absence of this protein, eukaryotic initiation factor 2 (*EIF2p*) takes over and binds met-tRNA and the 40S ribosome in a GTP-dependent manner (201). Therefore, it is thought that by knocking out *EIF2A*, competitive inhibition of this protein during the translation of non-AUG transcripts (such as those containing an IRES) may be removed. In light of this hypothesis, we decided to investigate the effect of this knockout on the activity of the *dicistroviridae* IRESs. We did indeed observe a significant increase in YFP expression enabled by IRESs expressed in the Δ*EIF2A* background. Encouraged by this result, we endeavored to identify other components of the yeast translational machinery which may interfere with IRES activity. Therefore, we tested IRES functionality in knockouts of yeast translation machinery-related proteins which were available in the yeast knockout database. In addition, we investigated the effect of overexpression of each subunit of *EIF2* (*SUI2*, *SUI3*, and *GCD1*) in order to enhance the rate of AUG-independent protein synthesis (**Figure 6-18**). We observed several strain backgrounds which enabled significantly higher YFP expression than our wild-type strain, with one knockout (*RMD9*) enabling up to 9-fold increases in YFP expression. However, for each strain, we also observed that increases to YFP came concurrently with increases to mStrawberry expression, indicating that these knockouts did not increase IRES activity specifically, but rather the amount of total expressed protein present in each cell (**Figure 6-19**). Indeed, several of the high-performing knockouts (*rmd9* and *arc1*) have been annotated to result in slow cell growth, indicating that improvements observed to YFP expression may simply be the result of increased protein accumulation due to a reduced cell dilution rate. In addition, the *EIF2* overexpression constructs and blank plasmid controls exhibited uniformly increased YFP and mStrawberry expression, consistent with the reduced growth observed during maintenance of extra plasmids.

Taken together, these results indicate that *EIF2A* knockout (and other knockouts to yeast's translational machinery) may serve to simply arrest cell growth and prevent protein dilution rather than increasing IRES expression specifically. It is interesting to note that none of the studies claiming that *EIF2A* knockout increases IRES activity performed an appropriate control to measure the changes to the expression of a non-IRES-mediated transcript (199,200). However, the increase to protein expression on a per-cell basis conferred by these knockouts may serve to increase the sensitivity of a screen for variants of slightly improved IRES activity.

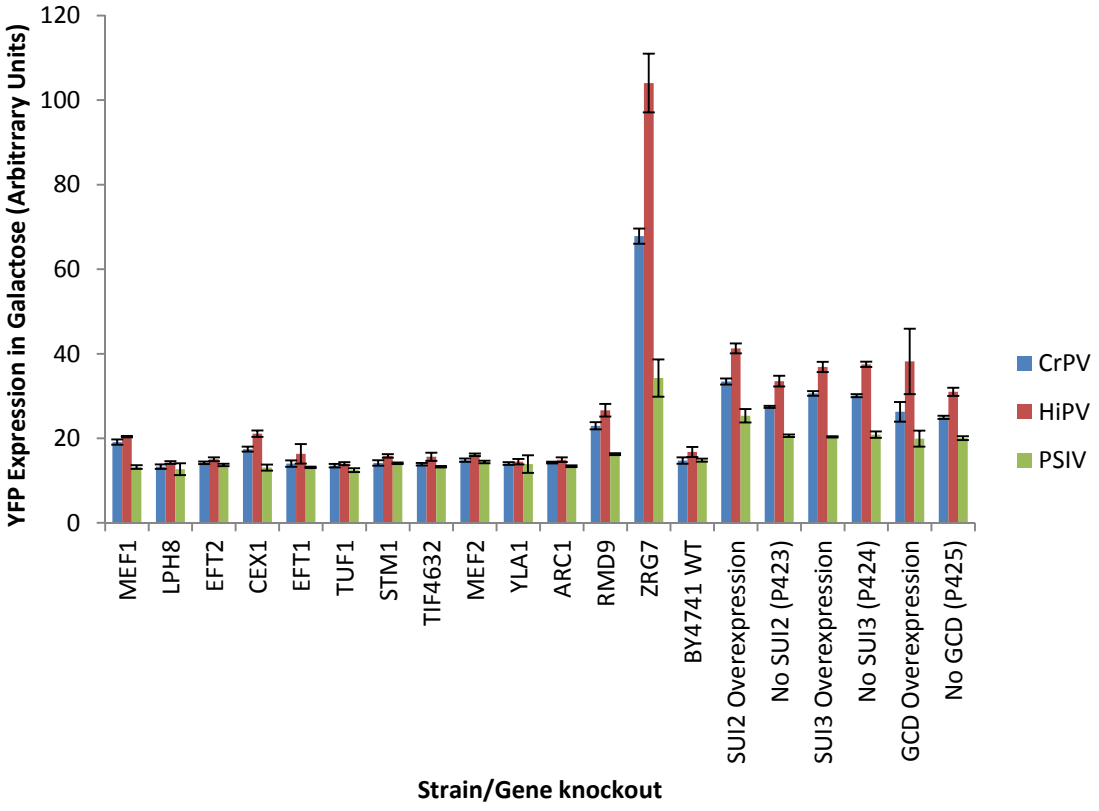


Figure 6-18: Performance of *Dicistroviridae* IRESs in strains containing altered translation machinery.

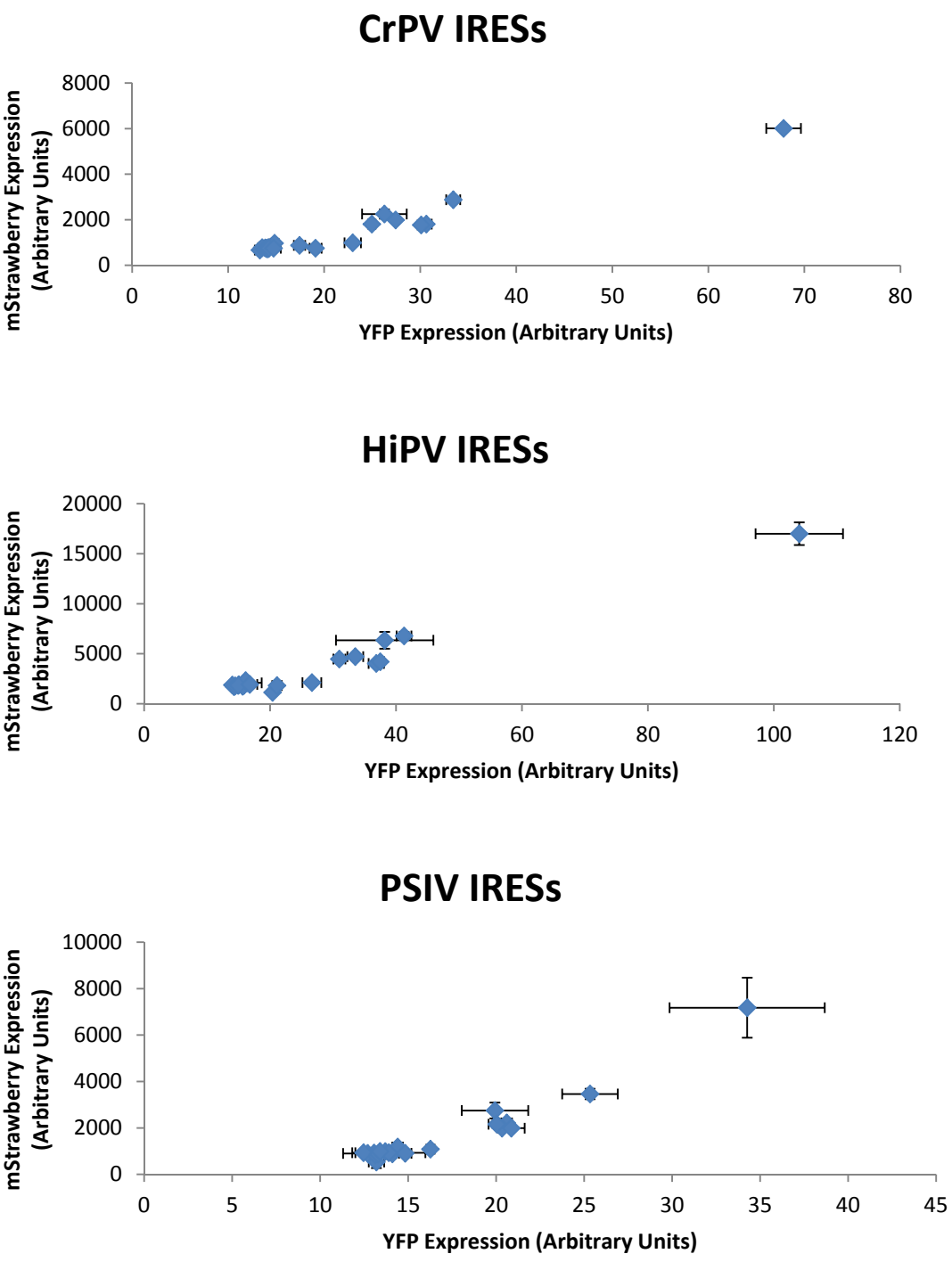


Figure 6-19: Correlation between mStrawberry and YFP expression during growth on galactose in various knockout strains.

6.3 DISCUSSION

During this study, extensive optimization of a sensitive screen to detect IRES activity was undertaken with the aim of minimizing the occurrence of false positives. Unfortunately, it became readily apparent that bicistronic reporter constructs are prone to the detection of false positives resulting from promoter activity or translational fusions. Even the use of an inducible promoter to enable facile positive and negative screening is prone to the development of inducible promoter activity in the IRES. Therefore, a novel methodology for the high-throughput detection of IRES activity is needed which is not prone to these failure modes. This work also emphasized the difficulty in the use of IRESs in multiple contexts, as many of the IRESs which have been previously reported to be functional did not show detectable activity in the screening system implemented here. For IRESs to be generally applicable as a methodology for polycistronic gene expression, they must be able to function in a variety of contexts. Interestingly, it has been shown that the efficiency with which an IRES can enable the translation of a downstream open reading frame is dependent upon the particular gene being expressed (202,203). It is reasonable that IRES functionality would be highly dependent upon context, since activity is thought to be a secondary structure-based effect. Therefore, any IRESs developed for use in a biotechnological context must be designed to be extremely robust to the disruptive effects of a wide variety of sequence contexts. Finally, this work characterized the effects of several gene deletions or overexpressions on IRES activity, and no modification to yeast's translational machinery was shown to improve IRES efficiency in a specific manner. Collectively, this work emphasizes that more work is needed to confirm the presence of IRESs in yeast and may inform future efforts directed at improving the functionality of IRES elements in yeast and other organisms.

Chapter 7: Rapid Evolution of Parts and Pathways through an *in vivo* Continuous Evolution Approach

7.1 INTRODUCTION

For classical *in vitro* directed evolution studies, the main bottleneck to realizing high library sizes (and thus high success rates of an evolutionary approach) is the transformation efficiency of the host of interest, which for yeast typically falls around 10^6 per microgram of DNA. This limits sequence coverage, especially for long sequences containing multiple genes. In addition, iterating multiple rounds of directed evolution is a laborious process, requiring several hands-on DNA manipulation steps. Although there have been systems developed which enable the continuous generation of sequence diversity in prokaryotes, these systems are limited to the development of transcriptional activators and require continuous flow bioreactors to generate and maintain selective pressure. Hence, there is a strong need for a technique to generate diversity which would allow the facile construction and screening of large libraries of arbitrary sequences in *in vivo* without the requirement for specialized and expensive equipment.

Retrotransposons are mobile genetic elements found in nearly all eukaryotic genomes (204). These sequences have high homology to viral genomes and replicate through a similar mechanism. There are five classes of retrotransposons in yeast, of which Ty1 is the most well-studied (205). Ty1 encodes proteins responsible for assembly of virus-like particles (VLPs), noninfectious virus-sized elements in which retrotransposon mRNA is converted to cDNA through an encoded reverse transcriptase. This cDNA can be integrated into the genome through either an element-encoded integrase or via homologous recombination (**Figure 7-1**). To overcome the limitations of classical directed evolution as mentioned above, a new approach termed *in vivo*

continuous evolution (ICE) was achieved through the use of a modified Ty1 retrotransposon system as shown below.

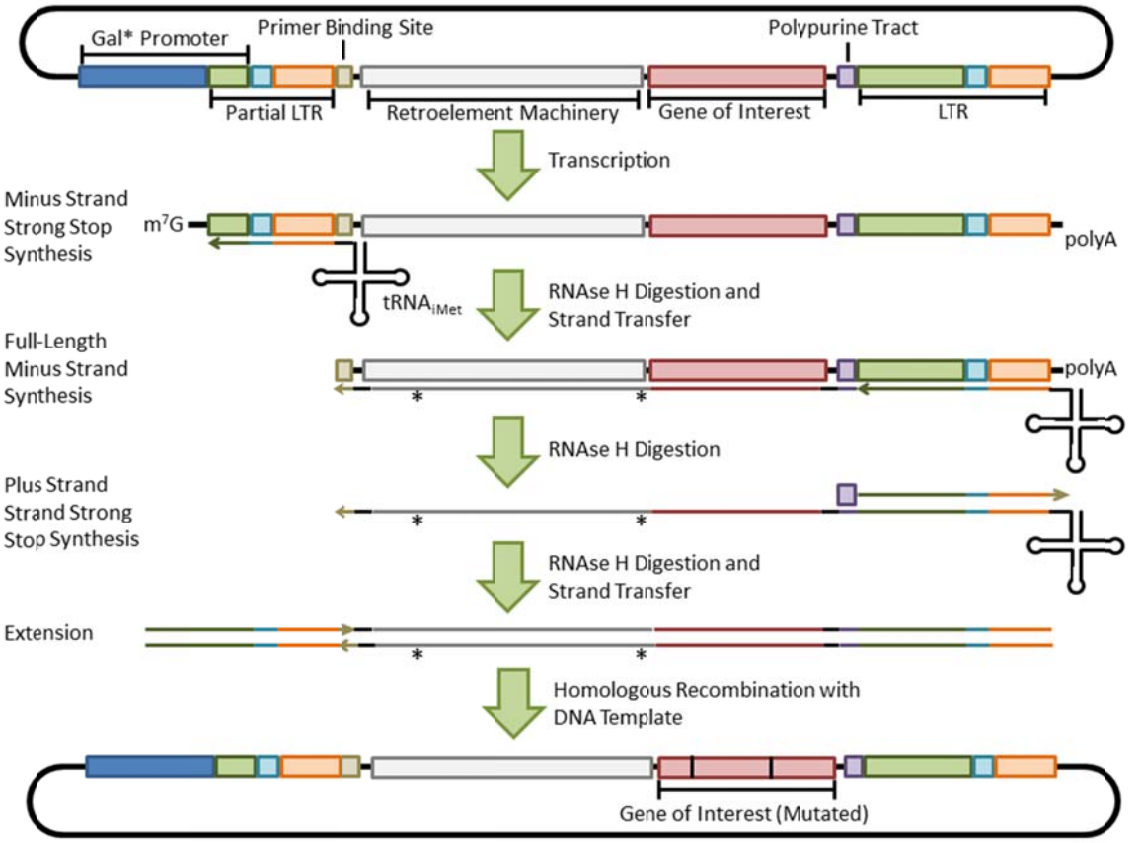


Figure 7-1: Mechanistic overview of synthetic Ty1 transposition.

First, Ty1 is transcribed from a truncated *GAL1* promoter in the presence of galactose. Then, the tRNA corresponding to initiator methionine binds to the primer binding site and serves as a primer for minus strand strong stop synthesis. After a strand transfer event, the rest of the retroelement is reverse transcribed in an error-prone manner. The RNAse H functionality of the Ty1 reverse transcriptase degrades the template RNA, with the exception of the polypurine tract, which is resistant to degradation. This polypurine tract serves as a primer for plus strand strong stop synthesis, after which a second strand transfer event completes the synthesis of double-stranded cDNA. This DNA is then integrated into a stably expressing form through the action of an encoded integrase or through homologous recombination.

The first step of *in vivo* continuous evolution is transcription of the modified Ty1 retroelement. It has been shown that this element can be inducibly expressed using galactose through fusion of the *GAL1* promoter to the 5' Long Terminal Repeat (206).

This promoter replaces the native low strength promoter found in Ty1, yet is truncated at its 3' end to prevent any *GAL1* promoter sequences from appearing in the Ty1 transcript. Inducibility ensures that successful isolates will not mutate the gene of interest after screening has taken place and before the new sequence of the gene of interest can be determined.

Reverse transcription is the critical step for mutagenesis with ICE. It has been shown that the 5' and 3' ends of Ty1 contain sequences which must be present in *cis* to ensure efficient reverse transcription (207). These sequences enable primer ($tRNA_i^{Met}$) binding, strand transfer, and critical secondary structure formation. Thus, any mutagenesis cassette must contain these sequences in order to be retrotranscribed. The mutation rate induced by the reverse transcriptase is also essential to achieving high library sizes. Native Ty1 reverse transcriptase has been shown to introduce mutations at rates of approximately 0.18 per kbp (208), which is the level of error required for mutagenesis of 5.5kb gene fragments or pathways. This indicates that Ty1 may be a promising starting point for the development of a system which generates large library sizes of pathway and gene-sized lengths of DNA.

Once a mutated cDNA has been generated, site-specific reintegration into the original locus on a plasmid or the host genome completes the *in vivo* continuous evolution cycle. Integration into nonhomologous locations is performed by the Ty1 integrase and integration into homologous locations occurs via homologous recombination. It has been shown that both processes occur in native Ty1 elements (209), but that in Ty1 elements utilizing HIV reverse transcriptase, integration occurs solely via homologous recombination (210). In this way, Ty1 retroelements can enable the generation of directed sequence diversity without the necessity for researcher intervention. Furthermore, as this diversity is generated, screening for improved

phenotypes can take place, thus enabling *continuous* generation and selection for beneficial variants. This approach thus bypasses limitations due to transformation efficiency, enables mutant generation in a facile manner, and does not require specialized equipment. Finally, because this mutagenesis process occurs independently in every cell under induction, and due to the high cell densities achievable with yeast cultures, the number of variants which can be generated with this method scales with the size of the culture, which is the fundamental upper limit to any evolutionary process. In this work, we developed and optimized this approach with the aim of enabling the creation of library sizes several orders of magnitude larger than can be achieved with the current state-of-the-art.

7.2 RESULTS

7.2.1 Construction and performance of Inducible, Marked Retrotransposon (pGALmTy1-HIV)

The yeast retrotransposon Ty1 was chosen as a scaffold for *in vivo* continuous evolution (ICE) because this element has been well-studied and has been shown to be highly amenable to engineering efforts. In particular, Boeke, et al have shown that transcription of this element may be placed under the control of a GAL promoter (206). Furthermore, it has been shown that transposition of this element may be monitored through the use of an intron-containing auxotrophic marker (211). In addition, the reverse transcriptase of this element may be replaced by the reverse transcriptase native to HIV, indicating Ty1 may be highly modular (210). As the mutation rate in the ICE technique is dictated by the reverse transcriptase, and as HIV reverse transcriptase (HIVRT) has a much higher error rate than that of Ty1 (Ty1RT) (212), we were most interested in a marked retroelement under the control of the GAL promoter containing the

HIV reverse transcriptase. Thus, the engineered retrotransposon pGALmTy1-HIV detailed in **Figure 7-2** was created on a high copy number plasmid.

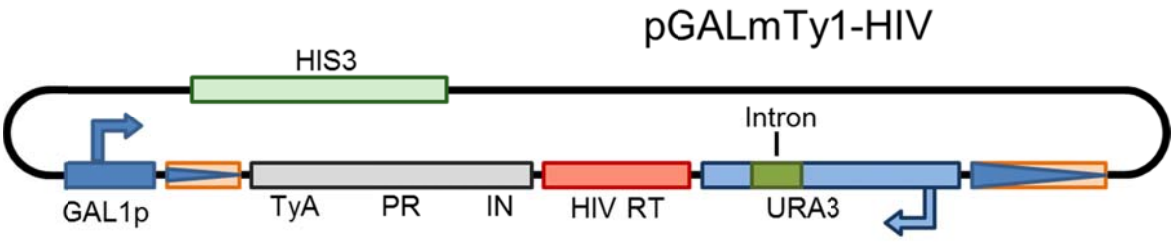


Figure 7-2: Schematic of pGALmTy1-HIV

The performance of this element was initially determined by transforming pGALmTy1-HIV into *S. cerevisiae* BY4741 and measuring transposition rate according to the Low OD induction method (see Materials and Methods section). Through this technique and a model for mutation accumulation in continuous culture, it was determined that a library size of up to 7*10^4 may be obtained using this construct in BY4741 after a week of induction in a continuous culture containing 10^10 yeast cells. Success of this proof-of-concept experiment was very exciting. To determine the limiting factor in transposition efficiency, levels of retroelement cDNA and RNA were measured through qPCR (**Figure 7-3**). Though levels of retroelement RNA increased significantly during growth on galactose, levels of retroelement cDNA showed no such increase. This suggested that reverse transcription was the rate-limiting step in transposition of this retroelement. To test this hypothesis, Ty1 reverse transcriptase was restored to pGALmTy1-HIV, generating pGALmTy1-Ty1 and transposition efficiency was measured as above (**Figure 7-4**). It is clear that Ty1 reverse transcriptase enables much higher transposition efficiencies than HIV reverse transcriptase, enabling library sizes upwards of 6*10^7 to be achieved in 10^10 yeast cells in a week.

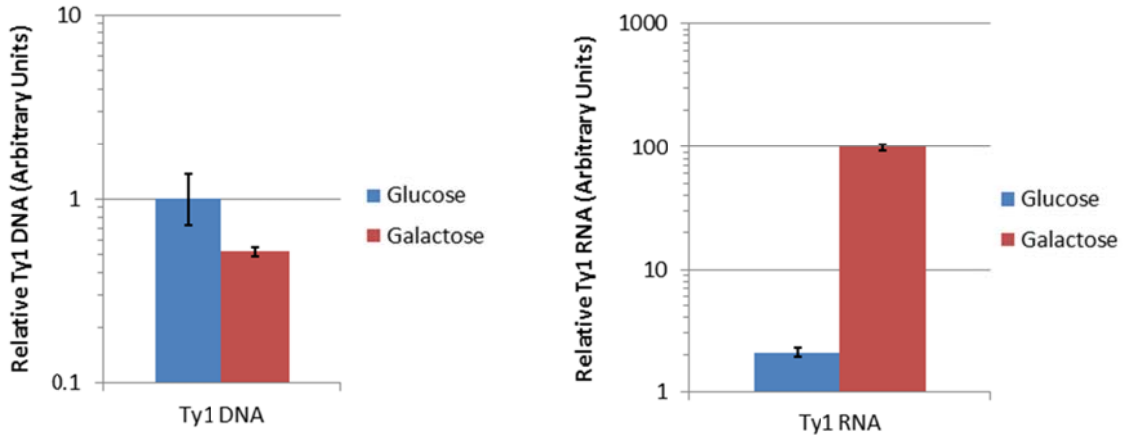


Figure 7-3: Transcript and cDNA generation by pGALmTy1-HIV.

Although transcript levels significantly increase under inducing conditions, cDNA levels show no such increase.

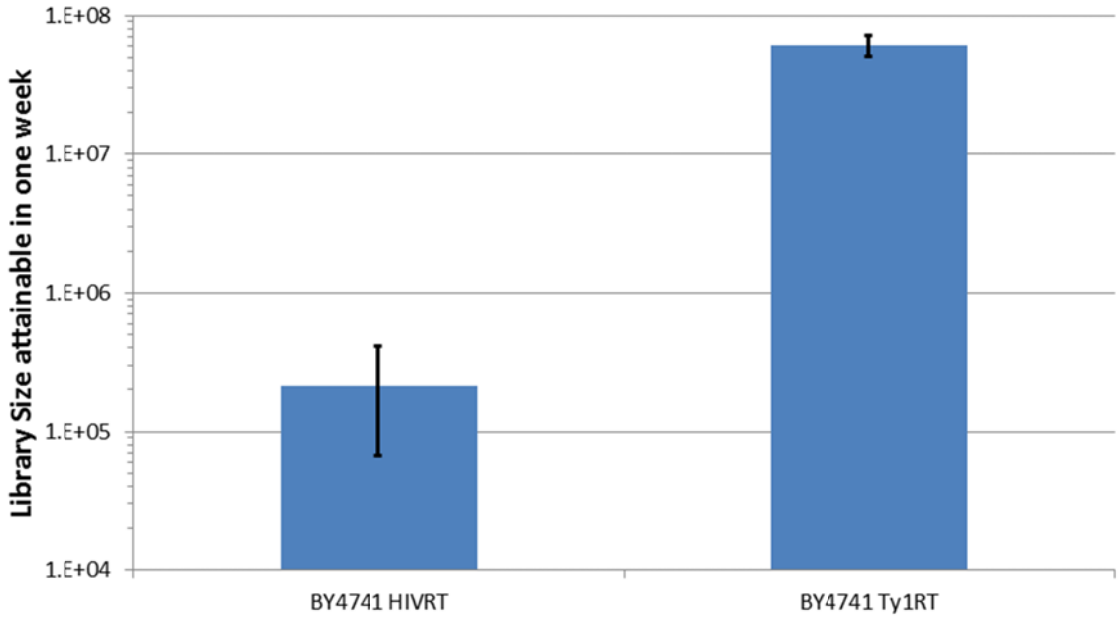


Figure 7-4: Transposition rates enabled by HIVRT and Ty1RT.

Ty1RT enables over 2 orders of magnitude more transpositions than HIVRT.

7.2.2 Strain optimization

Numerous studies have indicated gene knockouts which increase Ty1 transposition (213-218). 80 of these knockouts from the yeast systematic gene deletion project were transformed with pGALmTy1-HIV and transposition efficiency was measured according to the Low OD induction method. BY4741 Δ*mre11*, BY4741 Δ*apl2*, and BY4741 Δ*hir3* showed significantly higher transposition efficiency than BY4741 (**Figure 7-5**). The most proficient of these strains enabled library sizes of up to 1.1*10^6 to be generated in one week in a 1L culture containing 10^10 yeast cells. Interestingly, a highly beneficial gene knockout for the HIV-containing retroelement did not yield similar increases in the context of the Ty1 reverse transcriptase, indicating that performance enhancements enabled by changes to the strain background are highly context-dependent (**Figure 7-6**). Previous research has also shown that BY4741 Δ*hir3* Δ*cac3* and BY4741 Δ*hir3* Δ*cac2* could also activate Ty1 transposition to a large extent (16). This suggested that combinations of knockouts may further improve the transposition rate of pGALmTy1-HIV or pGALmTy1-Ty1. Thus, a set of single and double knockout strains were constructed in BY4741, including BY4741 Δ*cac2*, BY4741 Δ*cac3*, BY4741 Δ*hir3*/Δ*apl2*, BY4741 Δ*hir3*/Δ*mre11*, BY4741 Δ*apl2*/Δ*mre11*, BY4741 Δ*hir3*/Δ*cac2*, and BY4741 Δ*hir3*/Δ*cac3*. The full series of double and single knockouts was tested in BY4741 with either Ty1 or HIV reverse transcriptase (**Figure 7-7** and **7-8**). The best knockout with Ty1 and HIV reverse transcriptase were Δ*rrm3* and Δ*hir3*/Δ*cac3*, generating a library size of 1.07×10^7 and 1.18×10^6, which were ~1.76- and ~31.5-fold higher than the wild type, respectively. Surprisingly, very few knockouts were beneficial to Ty1-containing retroelements. Oppositely, most knockout strains for HIV-containing retroelements enabled a range from ~1.87- to ~31.5-fold higher transposition rate than the

wild type strain. The top knockout strains Δcac3, Δ*rrm3*, Δ*hir3*/Δ*cac2*, and Δ*hir3*/Δ*cac3* were used to test several strategies to further improve transposition rate.

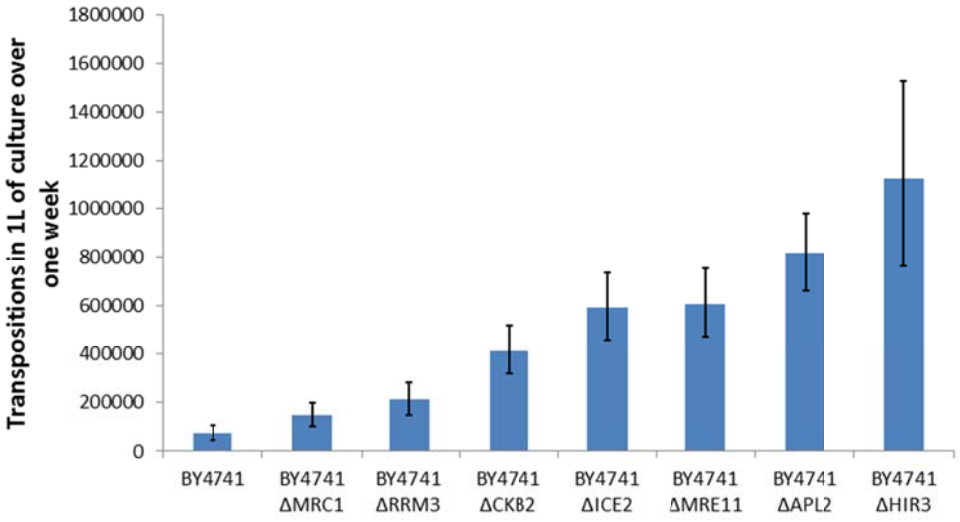


Figure 7-5: Single knockouts conferring increased transposition rates to HIVRT-expressing retroelements.

MRE11, APL2, and HIR3 knockouts conferred the highest growth rates of the knockouts tested.

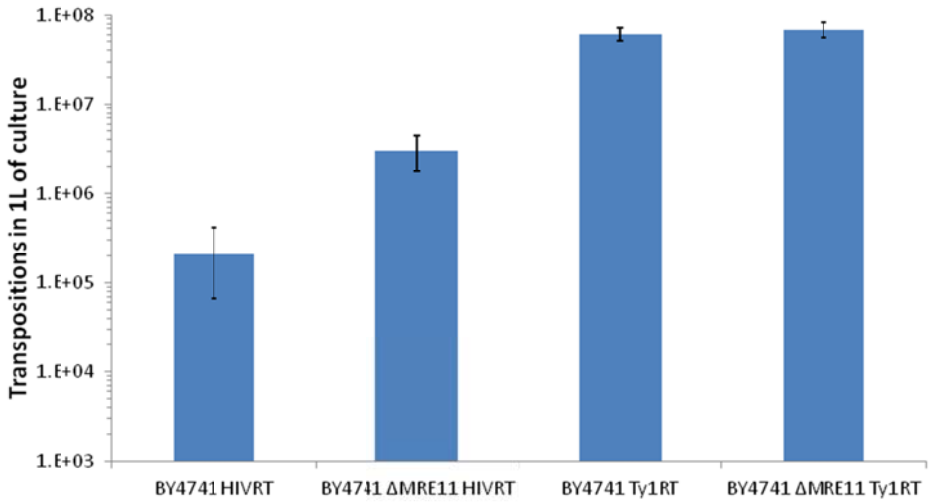


Figure 7-6: Comparison of the effects of the MRE11 knockout in retroelements expressing HIVRT and Ty1RT.

It can be seen that the MRE11 knockout significantly increases expression for the HIVRT-expressing retroelements, but that this knockout has no effect for Ty1RT-expressing retroelements.

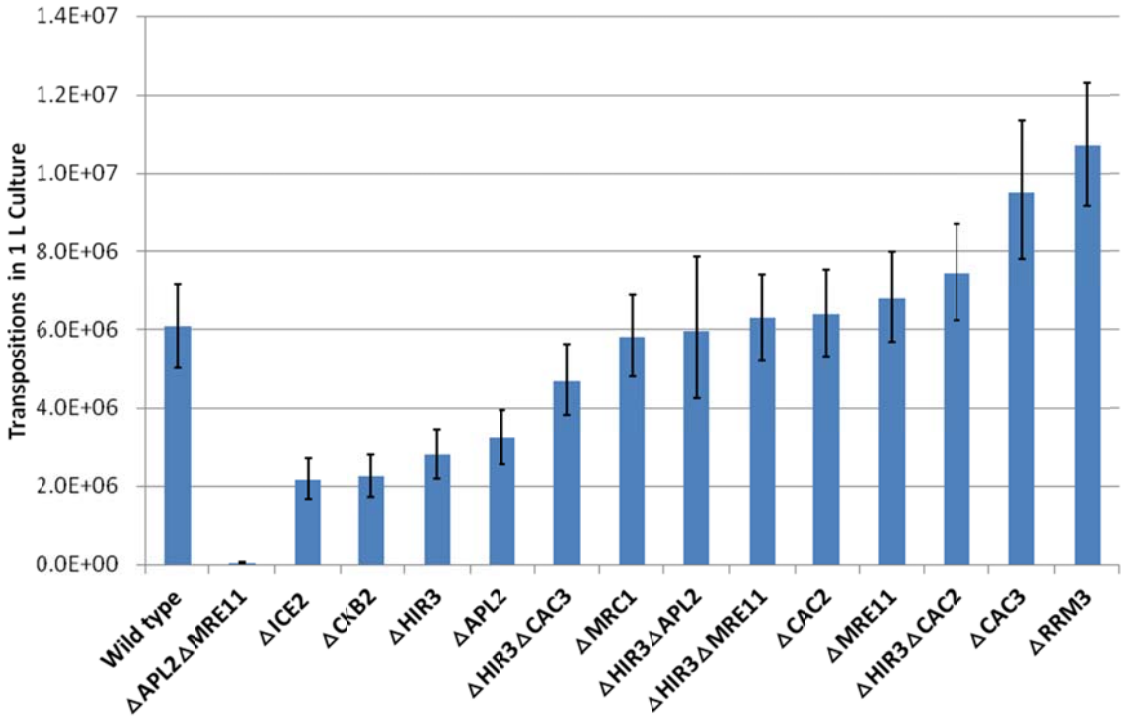


Figure 7-7: Transposition rates among various knockout strains for Ty1RT-containing retroelements.

Several knockouts, most notably Δ*rrm3*, enable significantly higher transposition rates for these retroelements.

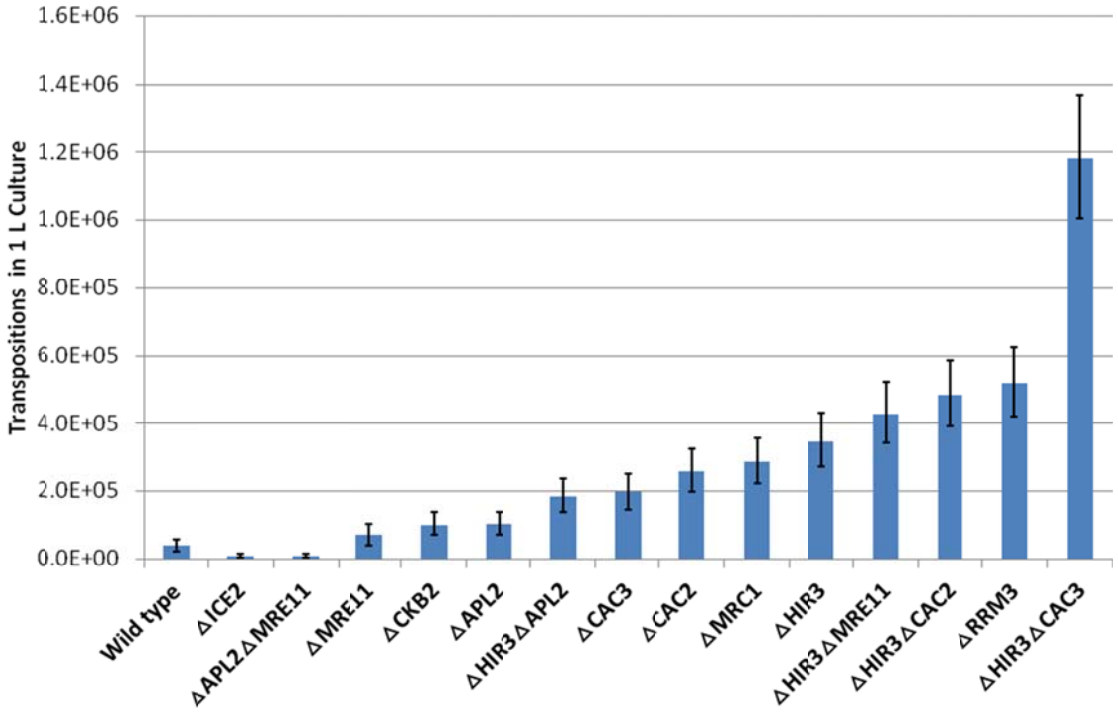


Figure 7-8: Transposition rates among various knockout strains for HIVRT-containing retroelements.

Many knockouts, most notably Δ*hir3*Δ*cac3*, enable significantly higher transposition rates for these retroelements.

7.2.3 Chimeragenesis of Ty1 and HIV Reverse Transcriptases

The outstanding performance of retroelements containing the Ty1 reverse trancriptase compared to the performance of those containing the HIV reverse transcriptase suggested that the HIV reverse transcriptase would need to be adapted to the Ty1 retroelement in order to obtain further improvements in transposition efficiency. However, the high error rate of the HIV reverse transcriptase would need to be maintained. Several possibilities presented themselves as potential reasons for the inefficiency of HIV reverse transcriptase: the inability of the reverse transcriptase to recognize the primer binding site (PBS) and polypurine tract (PPT) of Ty1 and/or incorrect posttranslational processing. Solutions to both issues were informed by

structural knowledge of HIV reverse transcriptase. HIV Reverse transcriptase is commonly understood to exist as a heterodimer, one monomer of which contains the polymerase, connection, and RNase domains, while the other monomer consists of only the polymerase and connection domain but in a divergent configuration (219). This dimer is formed through the action of an HIV protease which cleaves the connection and RNase domains (220). However, the Ty1 protease is not known to act at this junction in the Ty1 reverse transcriptase, suggesting that improper dimer formation may be the cause of HIVRT's inefficiency (221). Furthermore, it has been suggested that the HIVRT connection domain is responsible for primer tRNA binding, suggesting that swapping this domain may enable this enzyme to recognize the divergent primer tRNA of Ty1 (222). Thus, in order to combine the properties of the Ty1 and HIV reverse transcriptases, chimeras of each enzyme were generated by swapping the connection and RNAseH domains of each enzyme, as informed by a sequence alignment of the two enzymes (222). For each chimera, the HIV polymerase domain was used as it is thought that this domain is responsible for fidelity in this enzyme. These chimeras (denoted HHH, HHT, HTH, and HTT) were cloned into the pGALmTy1 vector. As HIV replication is primed by a different tRNA than Ty1 (205,219), we also generated a variant of pGALmTy1 containing the primer binding sites found in HIV (pGALmTy1H). Each chimera was cloned into this vector as well. In addition, high-copy vectors expressing truncated chimeras lacking an RNaseH domain (p425-GPD-tHH and p425-GPD-tHT) were generated.

Each chimera, primer binding site, and truncated reverse transcriptase were systematically combined in BY4741 Δ*mre11* and transposition rate was measured as above. Unfortunately, no variant was able to outperform strains containing pGALmTy1-HHH alone, indicating that chimeragenesis with Ty1 is not an appropriate strategy for

improvement of the activity of HIV reverse transcriptase (**Figure 7-9**). This also indicates that the primer binding sites present in Ty1 are not limiting the transposition rate of pGALmTy1-HIV. Finally, the poor performance of constructs co-expressing a truncated chimera may indicate that a simple co-expression may not be sufficient to generate heterodimers in the cell, or that post-translational processing of HIV reverse transcriptase is not limiting the transposition rate of pGALmTy1-HIV.

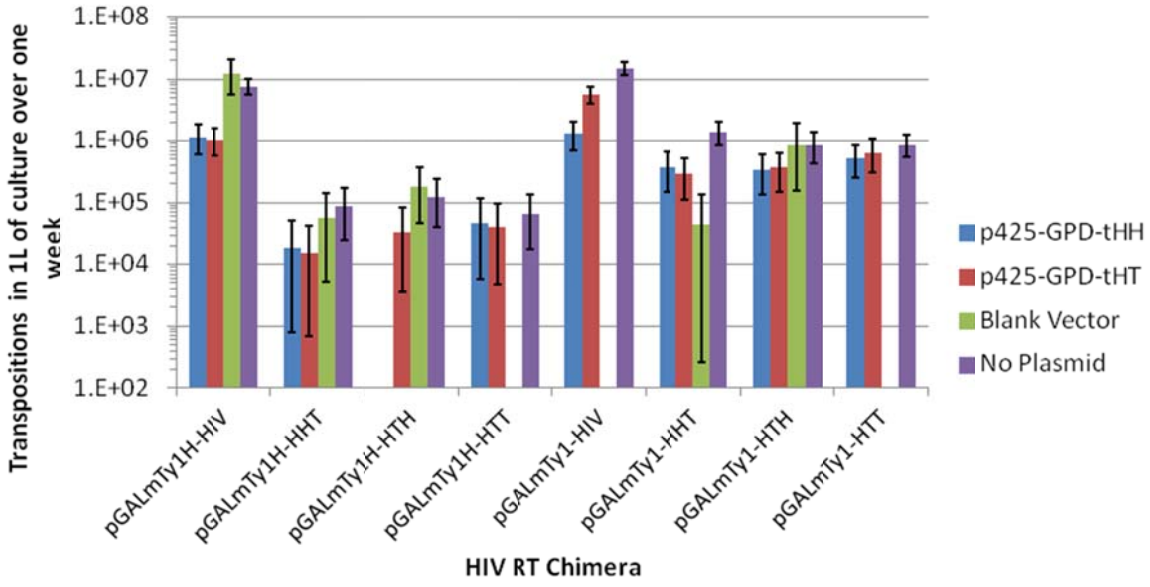


Figure 7-9: Transposition rates attained by reverse transcriptase chimeras.

Several chimeras of HIVRT and Ty1RT were constructed in order to combine the properties of the two enzymes. These chimeras were co-expressed with truncated versions of wild-type HIVRT and a chimera in order to determine whether posttranslational processing was the limiting factor for HIVRT activity. Unfortunately, it was observed that no chimera/truncation combination attained the level of transposition achieved by wild-type HIVRT alone.

7.2.4 Overexpression of Ty1 Transpositional Activators

Several genes are known to activate Ty1 transposition upon overexpression (216,223-225), so it was desired to investigate the effect of these genes on the transposition rate of pGALmTy1-Ty1 and pGALmTy1-HIV. These genes were cloned

into high copy vectors and co-expressed with either pGALmTy1-Ty1 or pGALmTy1-HIV in BY4741 or BY4741 Δ*mre11*. It was found that overexpression of *HSX1*, a gene encoding arginine tRNA, can improve transposition rate by as much as 5-fold in retroelements expressed in BY4741 Δ*mre11*, and by approximately 40% when expressed in BY4741 (**Figure 7-10**). This is particularly interesting because the rarity of $tRNA_{Arg}$ causes the ribosome to "pause" at a particular location when translating Ty1, and in a process known as frameshifting, the ribosome is able to "slip" to a codon one nucleotide downstream and continue translation (226). Given this mechanism, it is unclear how increasing the concentration of $tRNA_{Arg}$ improves transposition rate, as the protease, integrase, and reverse transcriptase of Ty1 are products of ribosomal frameshifting. However, because nucleocapsid protein is formed from the un-frameshifted polypeptide, it may be that the concentration of this species is limiting to our engineered system. In addition, increasing the concentration of a rare tRNA may simply make translation of the downstream genes more efficient, offsetting the effect of a decrease in frameshifting. It is interesting to note that the GPD promoter used for overexpression of *HSX1* is not driven by polIII, and so the 5'UTR provided by this promoter may make the resulting tRNA significantly different than the native yeast tRNA. In this way, it is possible that this nonfunctional tRNA actually interferes with the native tRNA through competitive inhibition to increase the overall rate of frameshifting.

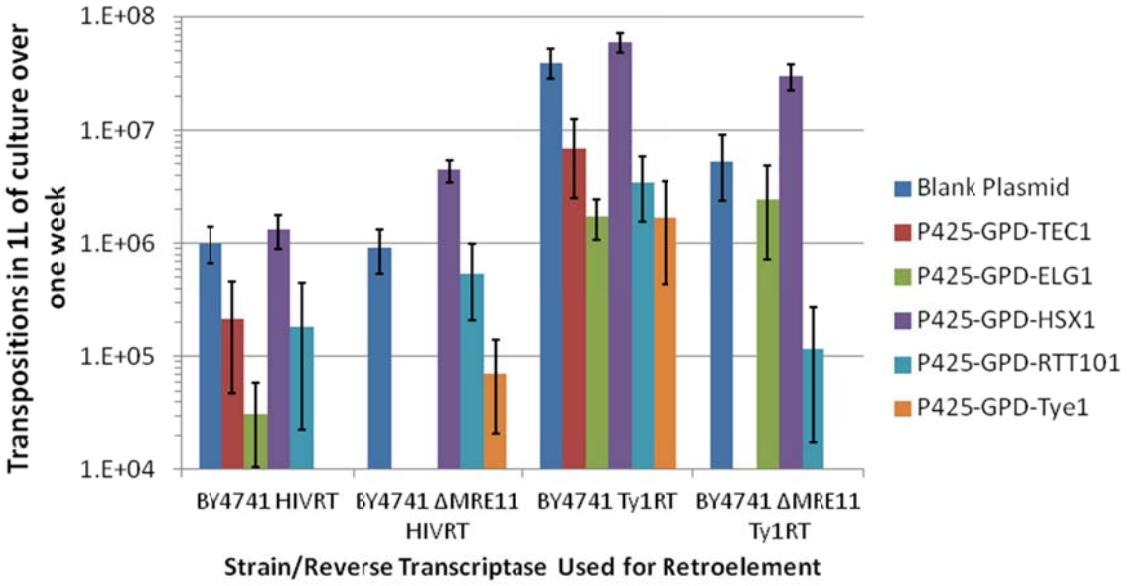


Figure 7-10: Improving transposition rate though overexpression of Ty1 transpositional activators.

Several genes reported to increase the rate of Ty1 transposition were overexpressed in BY4741 or BY4741 Δ*mre11*. *HSX1* was observed to significantly increase transposition rate in a variety of contexts.

To test the effect of this gene overexpression in other yeast strains, the top knockout strains for Ty1: BY4741 Δ*hir3*Δ*cac2* and BY4741 Δ*rrm3*, were co-transformed with pGALmTy1-Ty1-TEF1 and either p425-GPD-HSX1 or p425-GPD (control). In addition, the top knockout strains for HIV: BY4741 Δ*hir3*Δ*cac2*, BY4741 Δ*rrm3*, and BY4741 Δ*hir3*Δ*cac3*, were co-transformed with pGALmTy1-HIV-TEF1 and either p425-GPD-HSX1 or p425-GPD (control). It was found that overexpression of *HSX1* can slightly improve transposition rate of Ty1RT-containing retroelements expressed in BY4741 Δ*hir3*Δ*cac2* and BY4741 Δ*rrm3* (**Figure 7-11**). However, the overexpression of *HSX1* significantly decreased the transposition rate of HIVRT-containing retroelements expressed in BY4741 Δ*hir3*Δ*cac2*, BY4741 Δ*rrm3*, and BY4741 Δ*hir3*Δ*cac3*. We hypothesize that the added metabolic burden of overexpressing this

tRNA overwhelms a slight increase in transposition rate for these systems. Therefore, this *HSX1* overexpression strategy will not be further applied in these strains.

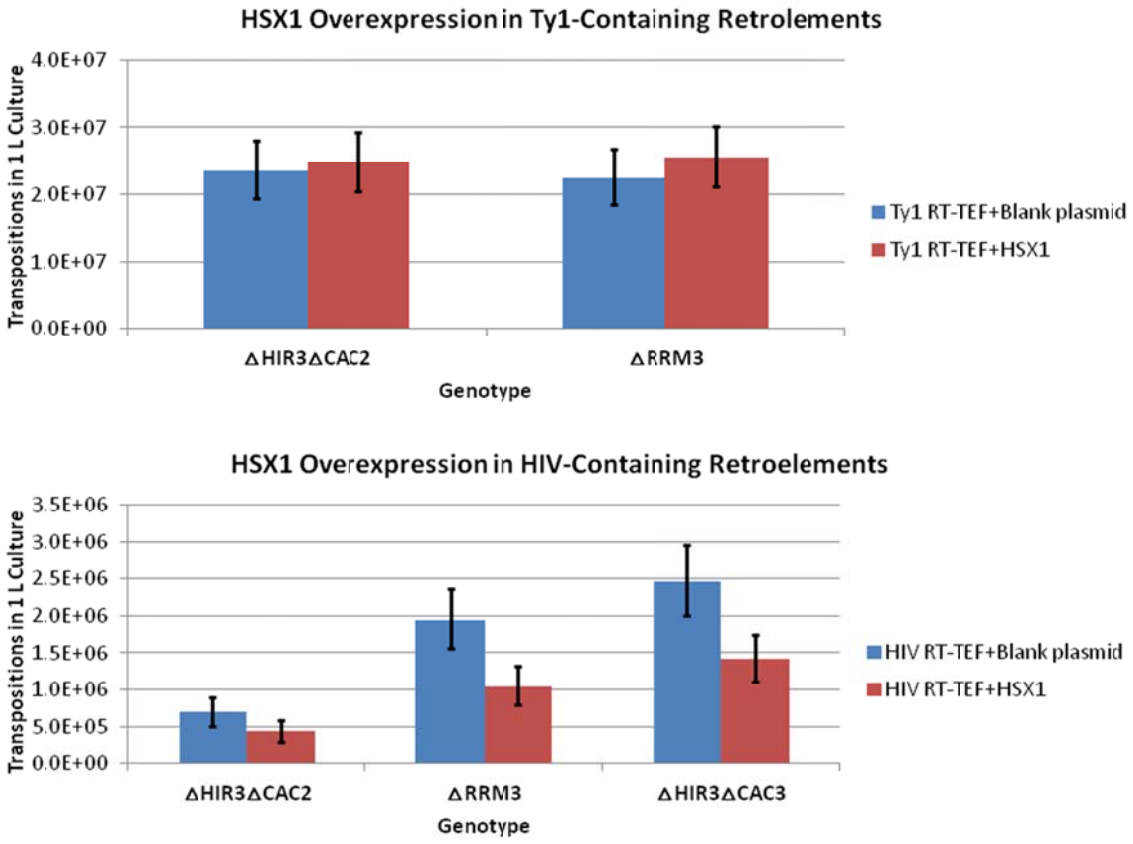


Figure 7-11: HSX1 overexpression in top-performing strains.

HSX1 overexpression was shown to slightly increase the rate of transposition for the top-performing strains expressing Ty1RT, yet significantly decrease the rate of transposition for the top-performing strains expressing HIVRT.

7.2.5 Comparison of Transposition Rates Enabled by BY4741 and CEN.PK

It has been suggested that *S. cerevisiae* CEN.PK enables higher rates of homologous recombination than BY4741, and thus may represent an attractive strain background for *in vivo* continuous evolution (ICE). Therefore, we transformed our synthetic retroelements (pGALmTy1-HIV and pGALmTy1-Ty1) into CEN.PK2-1D and

measured transposition rates. It was observed (**Figure 7-12**) that CEN.PK enables 10-fold higher transposition rates than BY4741 for retroelements utilizing HIV reverse transcriptase, and 2-fold higher transposition rates than BY4741 for those utilizing Ty1 reverse transcriptase. This promising result suggested that future strain engineering efforts should focus on CEN.PK rather than BY4741. However, because CEN.PK does not have a systematic deletion collection, it was infeasible to screen CEN.PK deletion strains for improved transposition rate at a large scale as was performed in BY4741. Therefore, in addition to recapitulating known beneficial knockouts identified in BY4741, we also overexpressed *HSX1* in CEN.PK, which has been shown to enable increased transposition rate in BY4741. We observed that the effects of *HSX1* overexpression are similar between BY4741 and CEN.PK, thus indicating this strain's potential advantage over BY4741 for use with the ICE system.

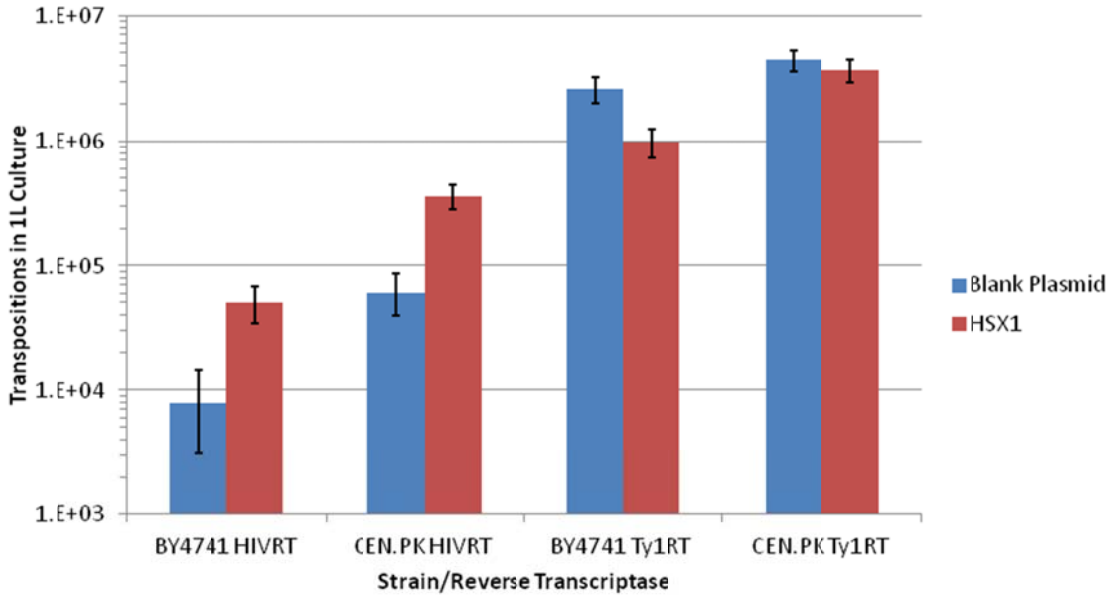


Figure 7-12: Determination of transposition rate in CEN.PK. Ty1RT and HIVRT-expressing retroelements were introduced into CEN.PK.

It can be seen that CEN.PK enables a much higher rate of transposition than BY4741, and that this increase is magnified during *HSX1* co-expression.

Then, a set of single knockout strains were constructed in CEN.PK2, including CEN.PK2-ΔICE2, Δ*cac2*, Δ*cac3*, Δ*rrm3*, Δ*apl2*, Δ*hir3*, Δ*mre11*, Δ*mrc1*, and Δ*ckb2*. Although BY4741 Δ*rrm3*, Δ*cac2*, and Δ*cac3* showed the highest transposition rate of the BY4741 strains when using the Ty1 reverse transcriptase (**Figure 7-13A**), these same knockouts in CEN.PK showed no substantial benefit. A similar result was also observed with HIV reverse transcriptase, in which case BY4741 Δ*rrm3* had ~5-fold higher transposition rate than CEN.PK Δ*rrm3* (**Figure 7-13B**). Only in the case of Δ*ice2* did CEN.PK2 have a higher transposition rate than BY4741. Although further engineering of the CEN.PK strain background could have resulted in higher transposition rates, the immediate availability of knockouts which made the BY4741 strain background superior indicated that BY4741 was more suitable for further demonstrations of the ICE system.

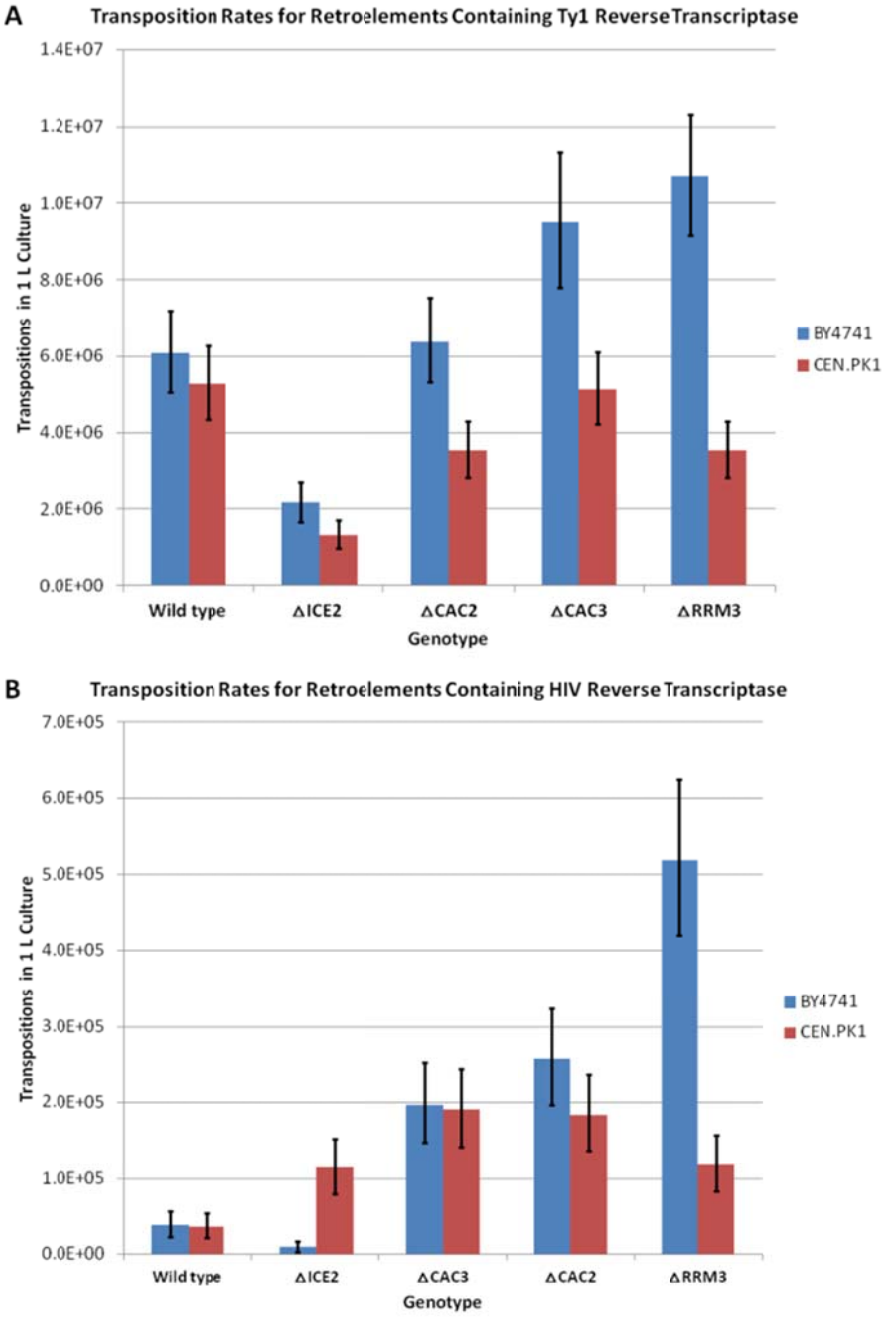


Figure 7-13: Transposition rates enabled by CEN.PK knockout strains.

It can be seen that most knockouts which were determined to be beneficial to BY4741 were not beneficial to CEN.PK

7.2.6 Increasing Expression Level of *URA3* Increases the Transposition Rate of Ty1-Containing Retroelements

We wished to investigate the extent to which the expression level of the gene of interest affected transposition rate. To test this, we cloned 3 promoters of varying strengths (pCYC1, pTEF1, or pGPD) in place of the HIS3 promoter (which normally drives *URA3* expression in pGALmTy1) and tested transposition rate. It was found (**Figure 7-14**) that while substitution of alternative promoters generally decreased the transposition rate of HIVRT-containing retroelements, substitution of the strong promoter TEF was found to increase the transposition rate of Ty1-containing retroelements by approximately 33%. We hypothesize that at low expression levels (such as that conferred by pCYC, for example), the expression level of *URA3* is insufficient to allow some transposants to grow in uracil dropout media. At very high expression levels (such as that conferred by pGPD), we hypothesize that substantial presence of transcriptional machinery interferes with the progress of RNA polymerases initiated at the upstream *GAL1* promoter and thus decrease the rate of transposition. Nevertheless, these experiments demonstrate that our engineered retroelement is suitable to evolve high-expression level pathways which are particularly relevant to metabolic engineering applications. Using this optimized retroelement, the top knockout strains achieved improved transposition rates (**Figure 7-15**). The Ty1 transposition rate in BY4741 Δ*hir3*Δ*cac2* and BY4741 Δ*rrm3* improved by 1.84- and 2.03-fold respectively, generating a maximum library size of 1.97×10^7. Meanwhile, the HIV transposition rate in BY4741 Δ*rrm3* improved by 2.72-fold, generating a library size of 1.41×10^6. This retroelement with *TEF1* promoter was used in future experiments.

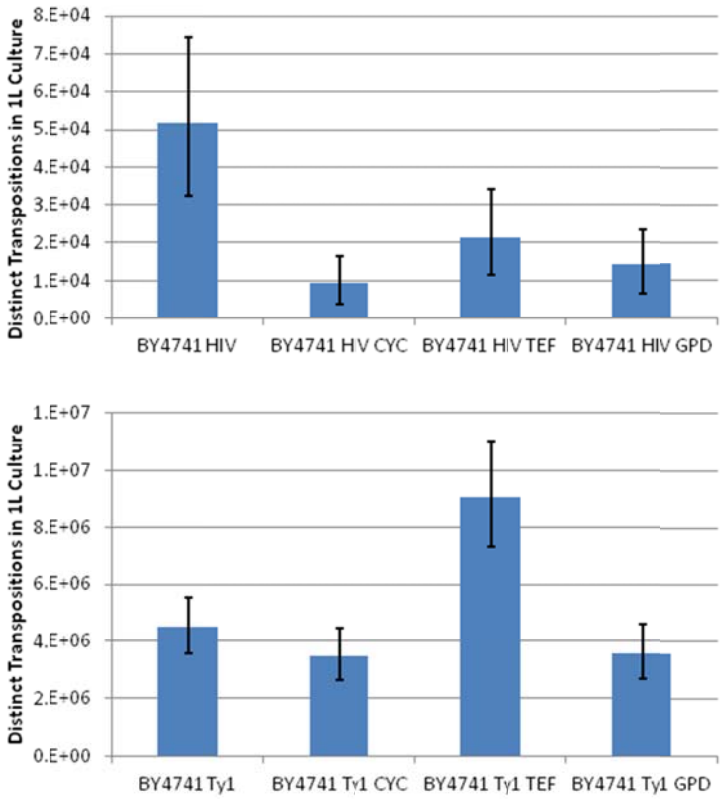


Figure 7-14: Substitution of alternative promoters in the retroelement.

It was observed that increasing expression of *URA3* increased the rate at which transpositions were detected in the synthetic retroelement.

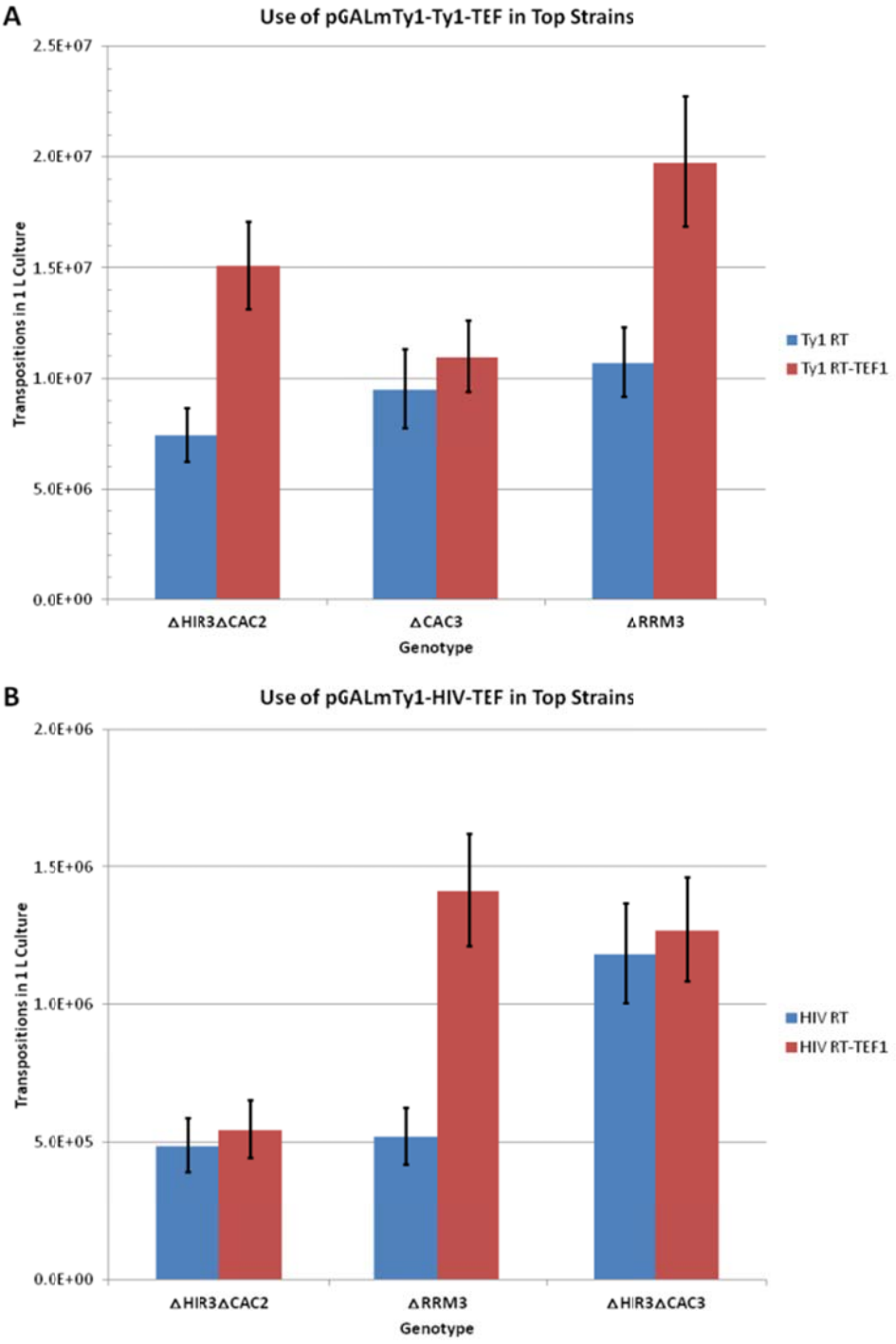


Figure 7-15: Use of the *TEF* promoter to drive *URA3* expression in top strains.

7.2.7 Measurements of Transposition Rates at High Culture Volumes and for Extended Periods of Time

Previous estimates of transposition rates have been based on experiments in which yeast expressing our synthetic retroelements have been grown in 1mL of liquid medium and allowed to transpose for 3 days. However, as ICE is most effective for large (~1L) culture volumes and must be able to apply a sustained rate of mutation over the timescale of several weeks, we performed a transposition rate test in 50mL cultures over the period of one week to determine what effects, if any, culture volume and time had on transposition rate (**Figure 7-16**). It can be seen that the number of transposants steadily increases with time for strains expressing pGALmTy1-HIV or pGALmTy1-Ty1 even though cells had reached stationary phase. Furthermore, the transposition rate observed in 50mL of culture is in agreement with that observed at the 1mL scale. This justified the economical use of 1mL volumes during large-scale experiments to identify strains with high transposition rates and also indicates that synthetic retroelements remain active in the absence of cell growth, confirming that ICE may be implemented for large culture volumes over long periods of time.

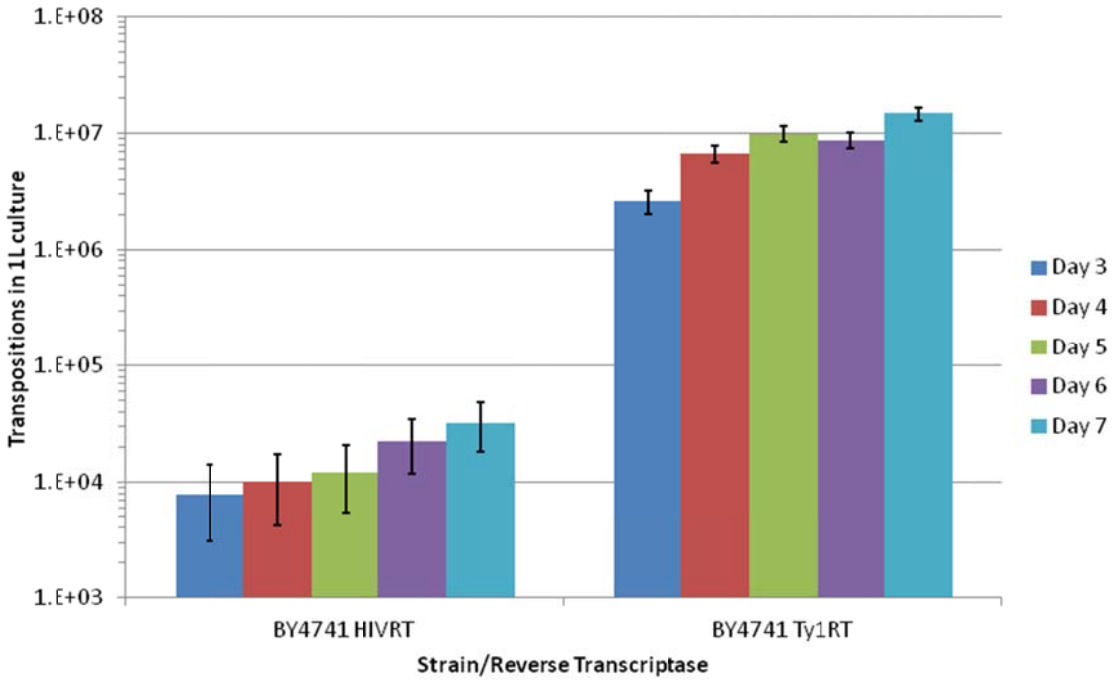


Figure 7-16: Measurement of transposition rates in cultures grown for extended periods of time.

7.2.8 Measurements of Transposition Rates for Non-growing Cultures

The ability of our synthetic retroelements to remain active after cessation of cell growth indicated that sustained induction of mutagenesis may be achieved in a high-density cell culture. High cell densities are desirable for ICE because they enable a large number of cells to be mutated simultaneously and, importantly, prevent cell growth from interfering with measurements of mutation rate. As a result, the number of transpositions in an experiment may be directly inferred from simple averages without the need for fluctuation analysis. In order to determine whether we could achieve increased library sizes with this technique, 50mL cultures were inoculated with yeast expressing pGALmTy1-HIV or pGALmTy1-Ty1 to an optical density of 1 and transposition was induced for up to 3 days. It was observed (**Figure 7-17**) that the ability to induce mutations in a large number of cells greatly increased transposition rate relative to

experiments in which mutations were induced as cells grew from low to high optical densities. This technique may be beneficial for applications in which mutagenesis and screening cannot be performed simultaneously and for cases in which a large number of mutants must be generated in a short time. This technique was then applied to the best knockout strains identified so far: BY4741 Δ*rrm3* with pGALmTy1-Ty1, and BY4741 Δ*hir3*Δ*cac3* with pGALmTy1-HIV. Measurement of transposition rate indicated large library sizes of 10^9 and 10^7 for Ty1 and HIV reverse transcriptase, respectively, which were the highest transposition rates obtained so far (**Figure 7-18**). Strains BY4741 Δ*rrm3*, Δ*cac3*, and Δ*hir3*Δ*cac2* were then transformed with pGALmTy1-Ty1-TEF1 and transposition was measured in the same way. Excitingly, we observed very large library sizes with the three strains, which obtained rates of ~3.5×10^9, 2.0×10^9, and 1.0×10^9 per liter, respectively (**Figure 7-19**). Such large library sizes were also maintained over one week, which supports the use of a long period of evolution in further work.

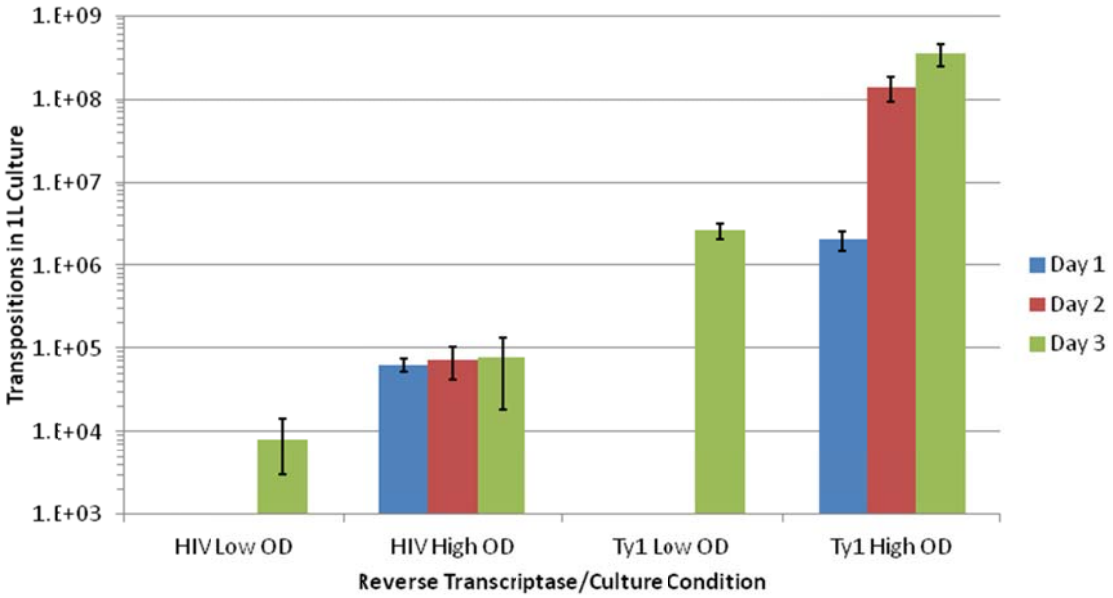


Figure 7-17: Eliminating growth increases transposition.

Transposition was induced for several days after cells had reached stationary phase. It was observed that high optical densities significantly increased the rate of transposition.

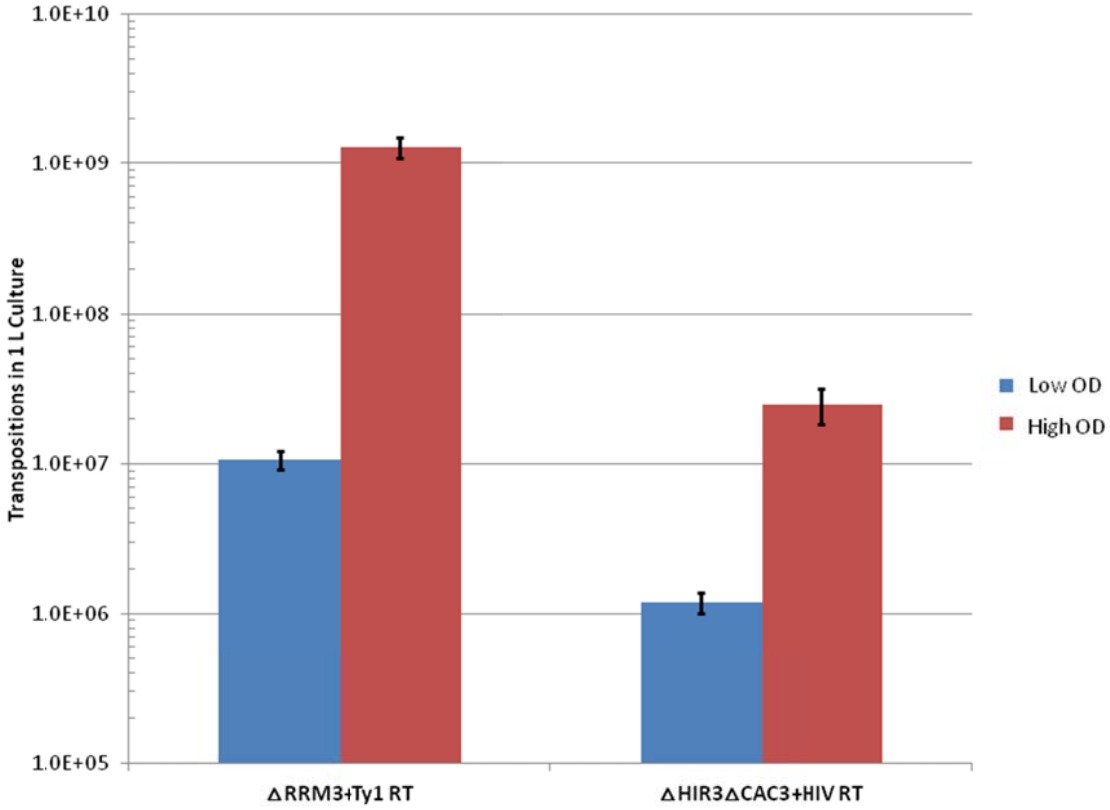


Figure 7-18: Transposition rate of top strains in high cell density cultures using retroelements expressing either Ty1RT or HIVRT.

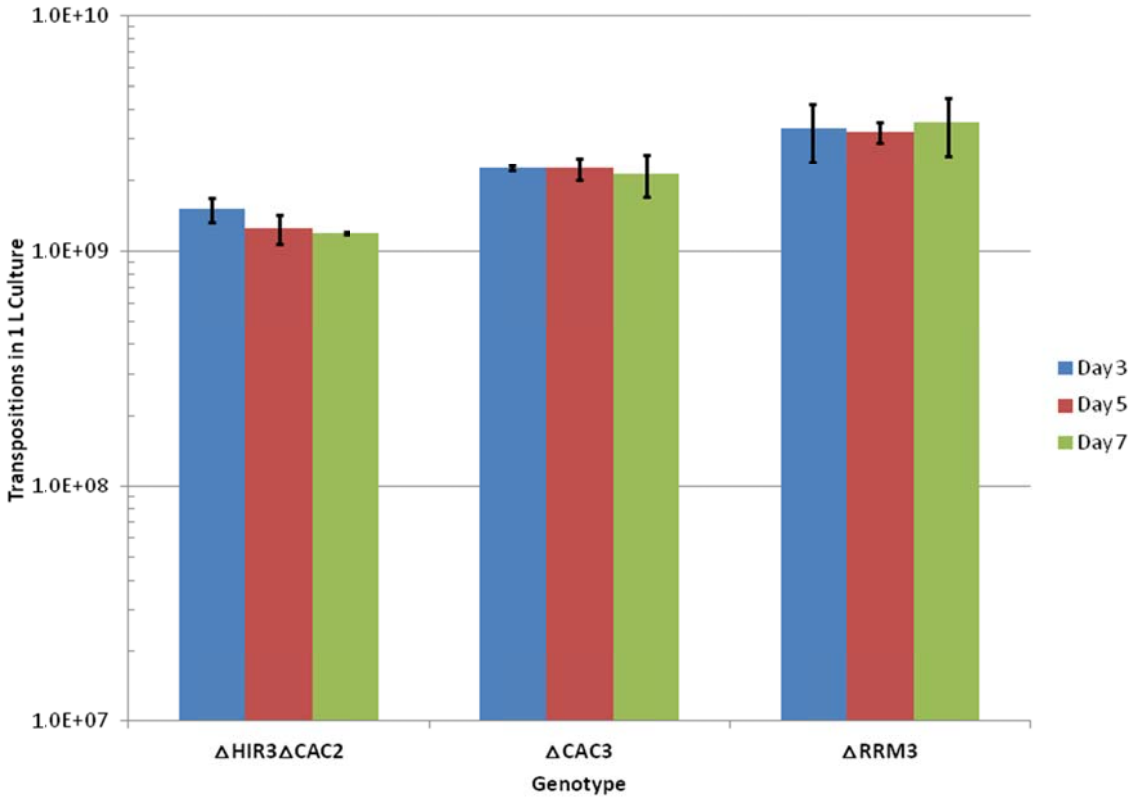


Figure 7-19: Transposition rate of top Ty1 strains expressing pGALmTy1-Ty1-TEF after induction in high cell density conditions.

7.2.9 Detection of Mutations Conferred by ICE through Next-Generation Sequencing

In order to measure the mutagenic capacity of ICE, several strains enabling the highest transposition rate were used in a High OD transposition test with either Ty1RT or HIVRT. Total DNA was extracted after galactose induction and *URA3* sequences resulting from transposition events were amplified via PCR. PCR products were then pooled and next-generation sequencing was performed on the mixture. It was found that Ty1RT enabled mutation rates $1.3*10^{-4}$ per base pair higher than that of random genetic drift, which is in line with previous estimates for this enzyme (208). Of these mutations, 55.0% were transversions, 41.4% were transitions, and 4.0% were indels. Combining this mutation rate and the transposition rate for the most active Ty1 strains, we estimate

that we could achieve upwards of 5.6*10^6 distinct variants of a 1kb gene after 1 week of galactose induction of 1L of culture (227), which is substantially higher than what can be achieved with traditional *in vitro* techniques and is also substantially easier to achieve. However, it was also observed that HIVRT enabled mutation rates at a significantly lower value than Ty1RT, which was puzzling given that HIVRT is known to be highly error-prone, especially for the variant we were testing, which had been reported to be 2-3 times more error-prone than wild-type HIVRT (228). We hypothesized that this HIVRT variant was only minimally active in yeast, such that endogenous Ty1RT provides much of the reverse transcriptase activity in this system. Nevertheless, detection of mutations in our system was highly encouraging and laid the foundation for future optimization of the engineered retroelement system as well as the undertaking of preliminary evolution experiments.

Additional analysis of the next-generation sequencing data was conducted to better characterize the mutations introduced during the ICE process. In particular, each type of mutation was quantified and normalized to a control, resulting in bias indicators that can be used to compare the mutational spectrum introduced through ICE to that introduced through the use of error-prone DNA polymerases. These results are summarized below:

	Ty1	Mutazyme II	Taq
Bias Indicators			
Ts/Tv	0.75	0.9	0.8
AT->GC/GC->AT	2.1	0.6	1.9
A,T->N (%)	76.8	50.7	75.9
G,C->N (%)	19.6	43.8	19.6
Insertions and Deletions			
Insertions (%)	0.17	0.7	0.3
Deletions (%)	3.8	4.8	4.2
Mutation Frequency			
Mutations per kb	0.13	3 - 16	4.9

Table 7-1: Mutational spectrum of Ty1 reverse transcriptase

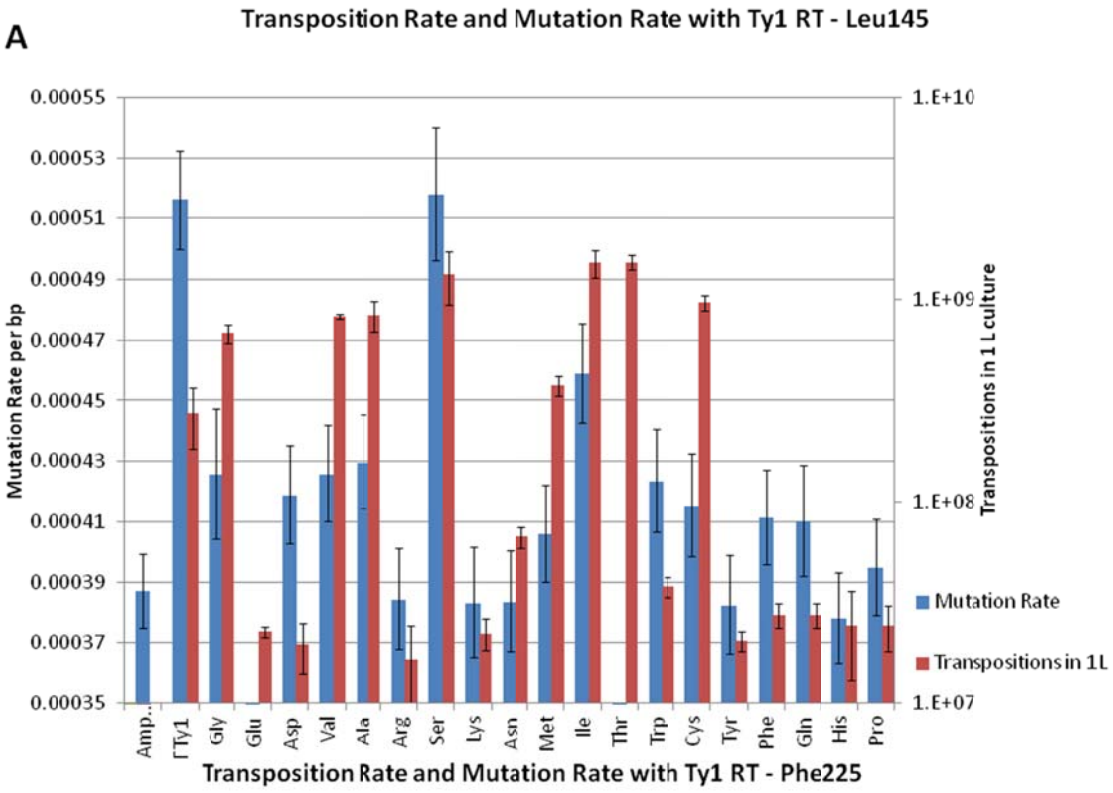
Although the overall mutation rate in the Ty1 ICE system is lower than that of these error-prone polymerases, the distribution of mutations it introduces is fairly comparable to Taq polymerase, which is commonly used for directed evolution experiments.

7.2.10 Next-Generation Sequencing of Saturation Mutagenesis Libraries

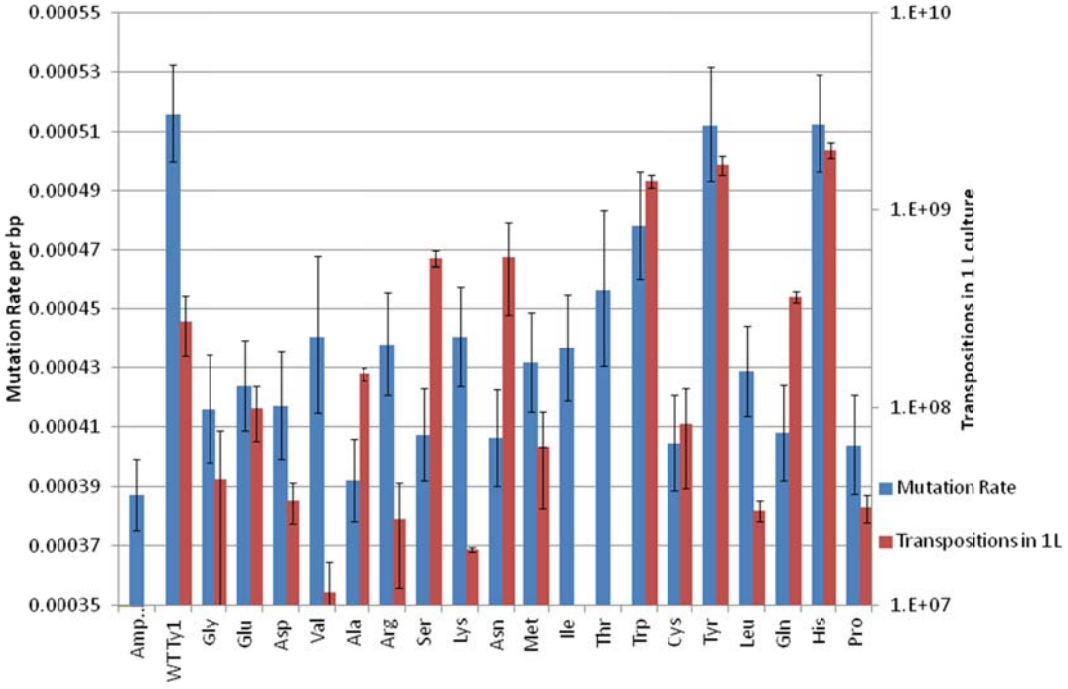
In order to increase the error rate of the native Ty1 reverse transcriptase while maintaining its relatively high transposition rate, a library of enzyme variants was created by performing site-specific saturation mutagenesis on several residues. The amino acids in positions 145, 225, and 226 are highly conserved and have been identified as playing a key role in fidelity (228-232). A library of 19 variants containing each amino acid substitution was created for each of these three sites, using pGALmTy1-Ty1 containing the *URA3* expression cassette as a template. These 57 plasmids were then transformed into the BY4741 Δ*rrm3* strain, and the transposition rate of each was measured at High OD, as described in the Methods section. Furthermore, mutational analysis was conducted by extracting total DNA after the High OD galactose induction, with *URA3* sequences resulting from transposition events amplified via PCR. PCR products were then pooled and next-generation sequencing was performed on the mixture. It was found that in many Ty1 RT mutants, transposition activity was significantly reduced, and no increase in mutation rate could be observed (**Figure 7-21**). However, an increase in both

transposition rate and mutation rate was observed in strains with three different point mutations in the Ty1 RT: L145S, F225Y and F225H. It is estimated that a system incorporating the Ty1 RT with any of these three mutations would increase the library size of unique transposants by approximately 5-fold.

A


Transposition Rate and Mutation Rate with Ty1 RT - Leu145
Mutation Rate per bp
Transpositions in 1 L culture
0.00055
0.00053
0.00051
0.00049
0.00047
0.00045
0.00043
0.00041
0.00039
0.00037
0.00035
1.E+10
1.E+09
1.E+08
1.E+07
Mutation Rate
Transpositions in 1L
Amp..
WTTy1
Gly
Glu
Asp
Val
Ala
Arg
Ser
Lys
Asn
Met
Ile
Thr
Trp
Cys
Tyr
Phe
Gln
His
Pro

B


Transposition Rate and Mutation Rate with Ty1 RT - Phe225
Mutation Rate per bp
Transpositions in 1 L culture
0.00055
0.00053
0.00051
0.00049
0.00047
0.00045
0.00043
0.00041
0.00039
0.00037
0.00035
1.E+10
1.E+09
1.E+08
1.E+07
Mutation Rate
Transpositions in 1L
Amp..
WTTy1
Gly
Glu
Asp
Val
Ala
Arg
Ser
Lys
Asn
Met
Ile
Thr
Trp
Cys
Tyr
Leu
Gln
His
Pro

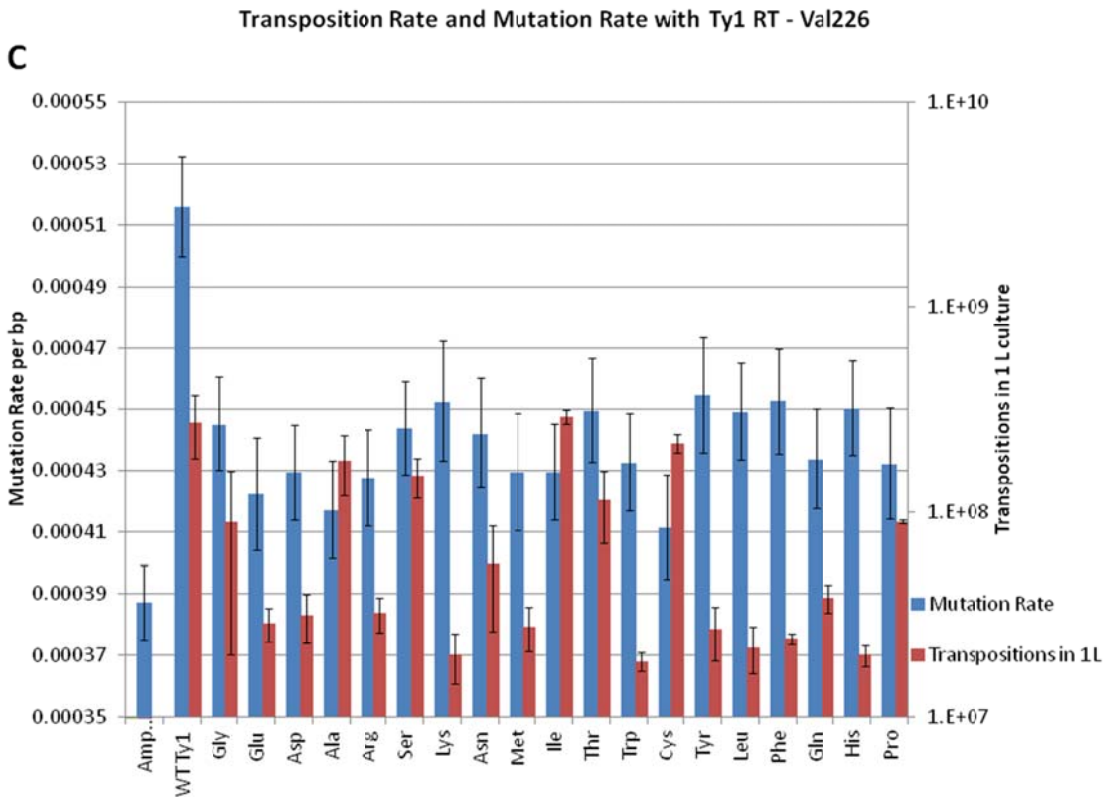


Figure 7-20: Mutation Rates enabled by Ty1RT saturation mutagenesis libraries.

Several additional mutations to Ty1 RT were constructed in an effort to further improve retrotransposition activity and/or decrease retrotransposition fidelity. First, the three mutations previously identified as beneficial (L145S, F225Y and F225H) were combined to determine the effect of double mutants (L145S/F225Y and L145/F225H). In addition, three more Ty1 RT mutants were made (Y151A, K93R, and R94K) based on previous studies of HIV RT (233-235); each analogous mutation in HIV RT has been shown to decrease fidelity without significantly impacting activity. Each of these five Ty1 RT mutants was constructed using pGALmTy1-Ty1 as a template, then transformed into BY4741 and grown under galactose induction to measure relative transposition rate.

Initial low-OD experiments showed that these mutants, excluding Y151A, either maintain or slightly reduce the retrotransposition rate (**Figure 7-22**).

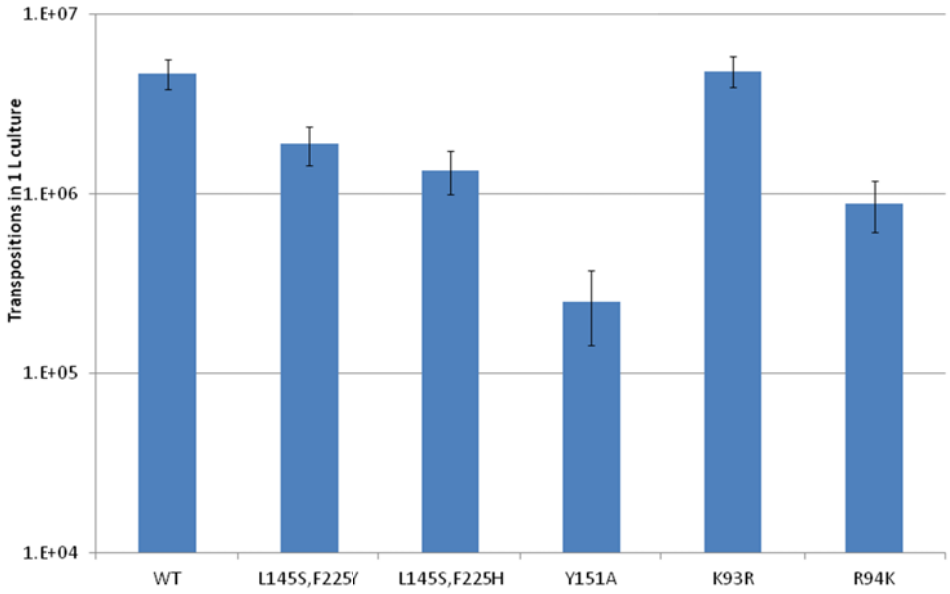


Figure 7-21: Transposition rates enabled by additional Ty1RT mutants.

7.2.11 Establishing Baseline Transposition Activity Without Reverse Transcriptase Overexpression

In order to better characterize transposition activity in both Ty1 and HIV systems, a new construct was made by removing the Ty1 reverse transcriptase gene from pGALmTy1-Ty1. The new construct still contained the integrase gene, with a new stop codon added after the integrase/reverse transcriptase protease cleavage site. This new construct, called pGALmTy1-ART, was inserted into a fresh BY4741 strain, and the transposition rate under galactose induction was measured. It was found that the retrotransposition rate in strains containing this construct was higher than in those strains containing pGALmTy1-HIV, indicating that the "activity" of this retroelement is primarily due to the activity of natively expressed Ty1 RT (**Figure 7-23**). These results implied that the HIV RT was not being successfully expressed in an active form in strains

with pGALmTy1-HIV. Furthermore, it was possible that the inclusion of the HIV RT gene in the construct actually interfered with the retrotransposition process catalyzed by native Ty1 RT background activity, further lowering the overall rate. One hypothesis was that the HIV RT was not being successfully cleaved from the integrase-reverse transcriptase polypeptide precursor, leading to an inactivity of both enzymes. It was also possible that the HIV RT is being cleaved but is not able to catalyze one or more of the reverse transcription reactions in the Ty1 VLP environment. This interference would explain why the observed retrotransposition rate in these strains is even lower than in those with no RT overexpression. These data indicated that more work was needed to optimize HIV RT expression and activity before it could be used in the ICE system.

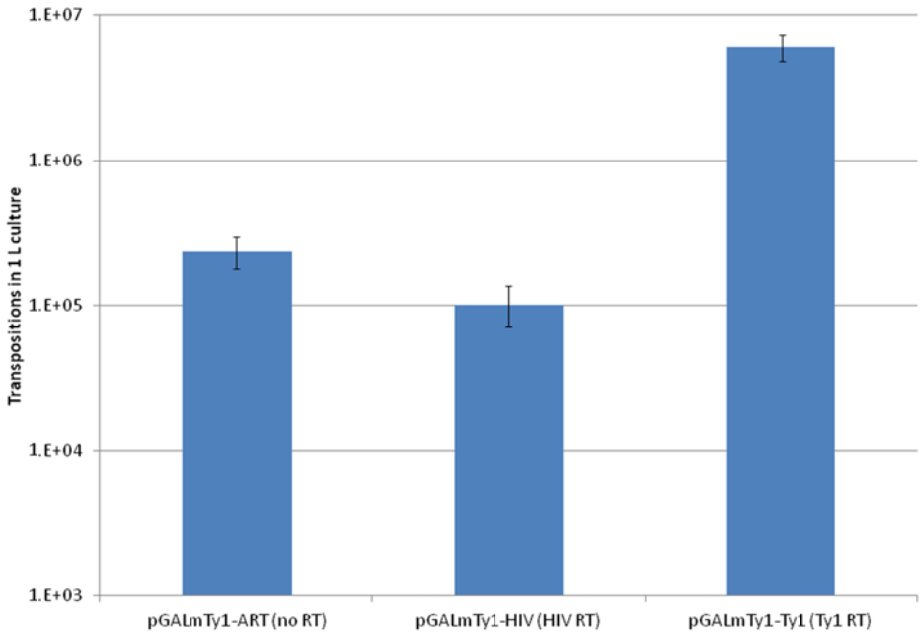


Figure 7-22: Transposition rate in the absence of reverse transcriptase expression.

The reverse transcriptase was removed from pGALmTy1 and the rate of transposition was measured. It was observed that the intronless retroelement enabled a higher transposition rate than the retroelement containing HIVRT.

7.2.12 Engineering HIV Reverse Transcriptase to Improve Expression and Activity

In an effort to create a retroelement system incorporating HIV RT with a higher retrotransposition activity, several variants of the pGALmTy1-HIV construct were made. First, two HIV RT mutations introduced previously were reverted to create a wild-type HIV RT variant. Next, the retroelement RT primer binding sites were changed from those native to Ty1 into those native to HIV. Finally, the sequence between the integrase gene and the reverse transcriptase gene, coding for the protease cleavage site wherein the polypeptide is cleaved to form the two mature enzymes, was altered to code for the native Ty1 cleavage site with either 0, 3, or 6 additional amino acids from the Ty1 RT gene downstream. It was hypothesized that if the HIV RT was inactive due to improper protease cleavage, including more native Ty1 RT amino acids at this site would facilitate more effective cleavage and thus more active enzyme. Sixteen constructs were made through combinations of each of these factors. Each was then transformed into a BY4741 strain and the transposition rate under galactose induction was measured. No clear pattern was found in the retrotransposition activity of these 16 pGALmTy1-HIV constructs (**Figure 7-24**). There was some variation between strains, but none displayed a higher retrotransposition rate than strains containing pGALmTy1-ART, the construct with no reverse transcriptase. This data seemed to indicate that the low activity previously observed with pGALmTy1-HIV was probably not due to improper protease cleavage, but more likely due to HIV RT interfering with reverse transcription catalyzed by natively present Ty1 RT. The variable activity observed in the different variants may be only due to the extent to which each mutant interferes with this process.

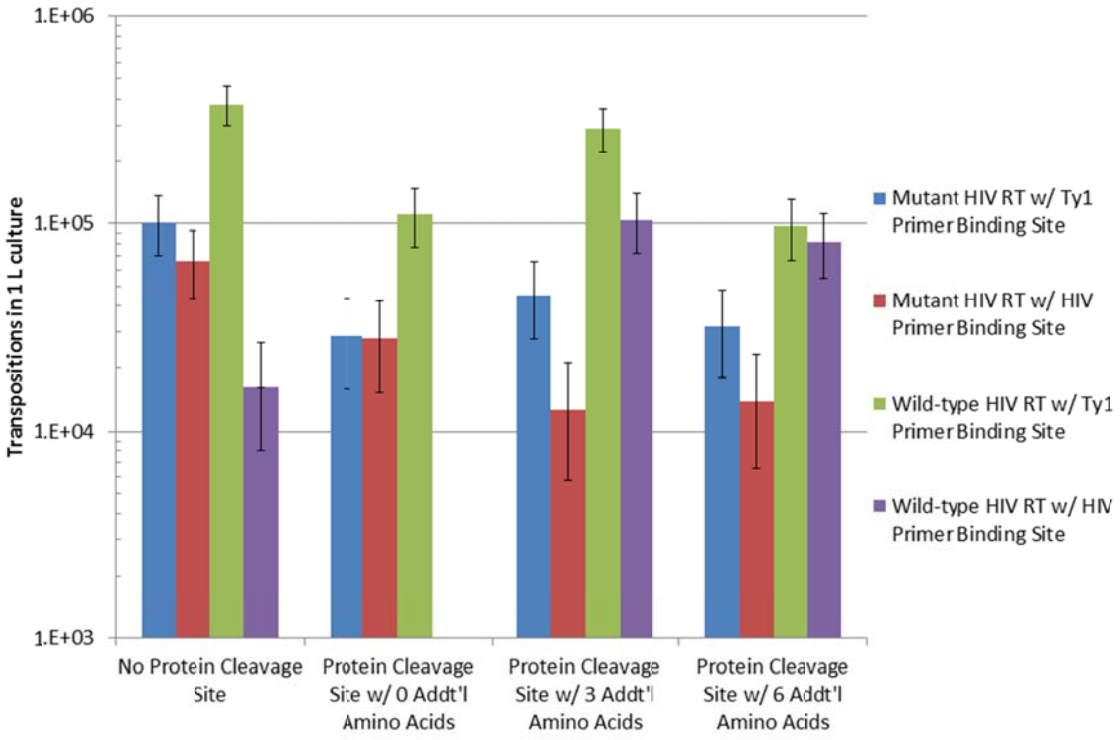


Figure 7-23: Transposition rate of HIV reverse transcriptases containing protease cleavage sites.

Various protease cleavage sites were introduced into wild-type or a mutagenic HIVRT when expressed in retroelements containing either the Ty1 or HIV primer binding sites.

7.2.13 Ty1 and HIV Reverse Transcriptase Fluorescent Fusion Proteins

In order to quantify and characterize the expression of both Ty1 and HIV reverse transcriptase as part of the ICE system, fluorescent fusion proteins were made by fusing the YFP gene downstream of the reverse transcriptase gene in the optimized retroelement (either pGALmTy1-Ty1 or pGALmTy1-HIV), with a linker region between them to allow proper folding and fluorescence. These constructs were transformed into BY4741, and fluorescence was measured by flow cytometry after growth in either glucose or galactose. The results (**Figure 7-25**) indicated that both Ty1 and HIV RT are being expressed only on galactose induction, as expected. Unfortunately, this implied that the

inability of our retroelement to utilize HIV RT is not due solely to inadequate expression, but that the activity of HIV RT is suboptimal in the yeast cell environment.

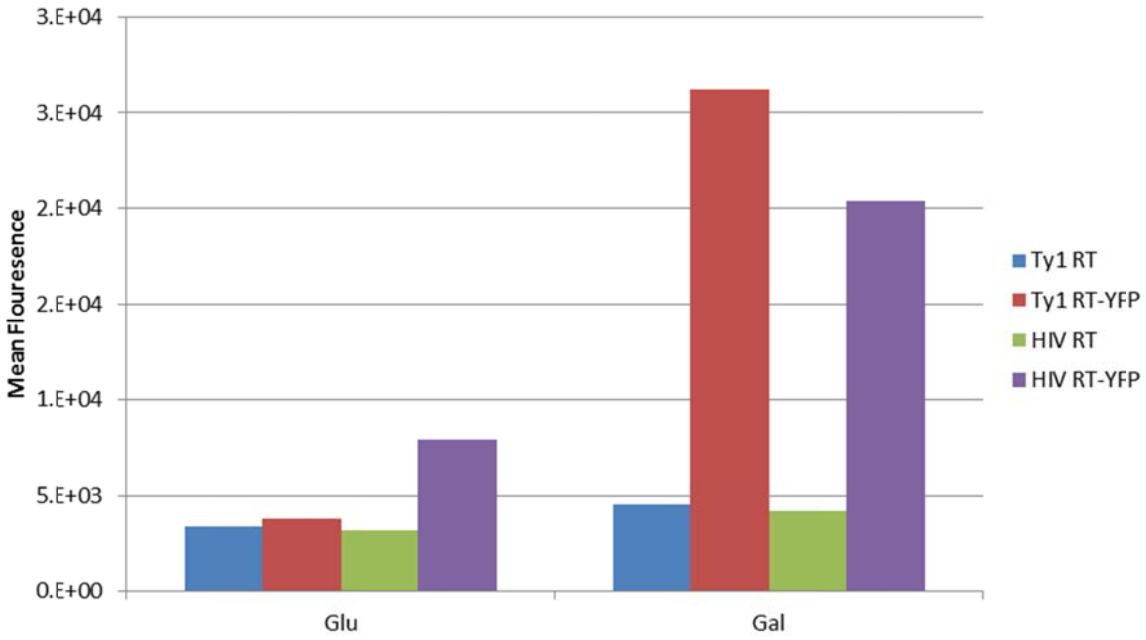


Figure 7-24: Fluorescence exhibited by RT-YFP fusion proteins.

Reverse transcriptases were fused to YFP to measure RT expression level. It can be seen that the expression of both proteins increases during growth on galactose, as expected.

7.2.14 Measurement of mRNA and cDNA Generation of Synthetic Retrotransposons

In order to verify the functionality of our synthetic retrotransposons on a deeper level, we measured transcript abundance and cDNA levels generated by several synthetic retrotransposons during induction of transposition. We measured transcript and cDNA abundance for retroelements containing Ty1RT in both wild-type BY4741 as well as BY4741 Δ*rrm3*, for retroelements containing HIVRT in both wild-type BY4741 as well as BY4741 Δ*hir3*Δ*cac*3, and finally for retroelements which do not contain a reverse transcriptase. It was observed (**Figure 7-26**) that although each synthetic retroelement

was able to generate high levels of mRNA upon induction, only elements containing Ty1RT were able to generate significant levels of cDNA. These results suggested that HIVRT is nonfunctional in our retroelement, a conclusion that is supported by its low transposition rate and the questionably low levels of mutations we observed for HIVRT-containing retroelements as measured by next-generation sequencing. Based upon previous experiments which failed to generate functional chimeric retroelements and chimeric reverse transcriptases, we were doubtful that HIVRT can be easily modified to be functional in our Ty1-based retroelement.

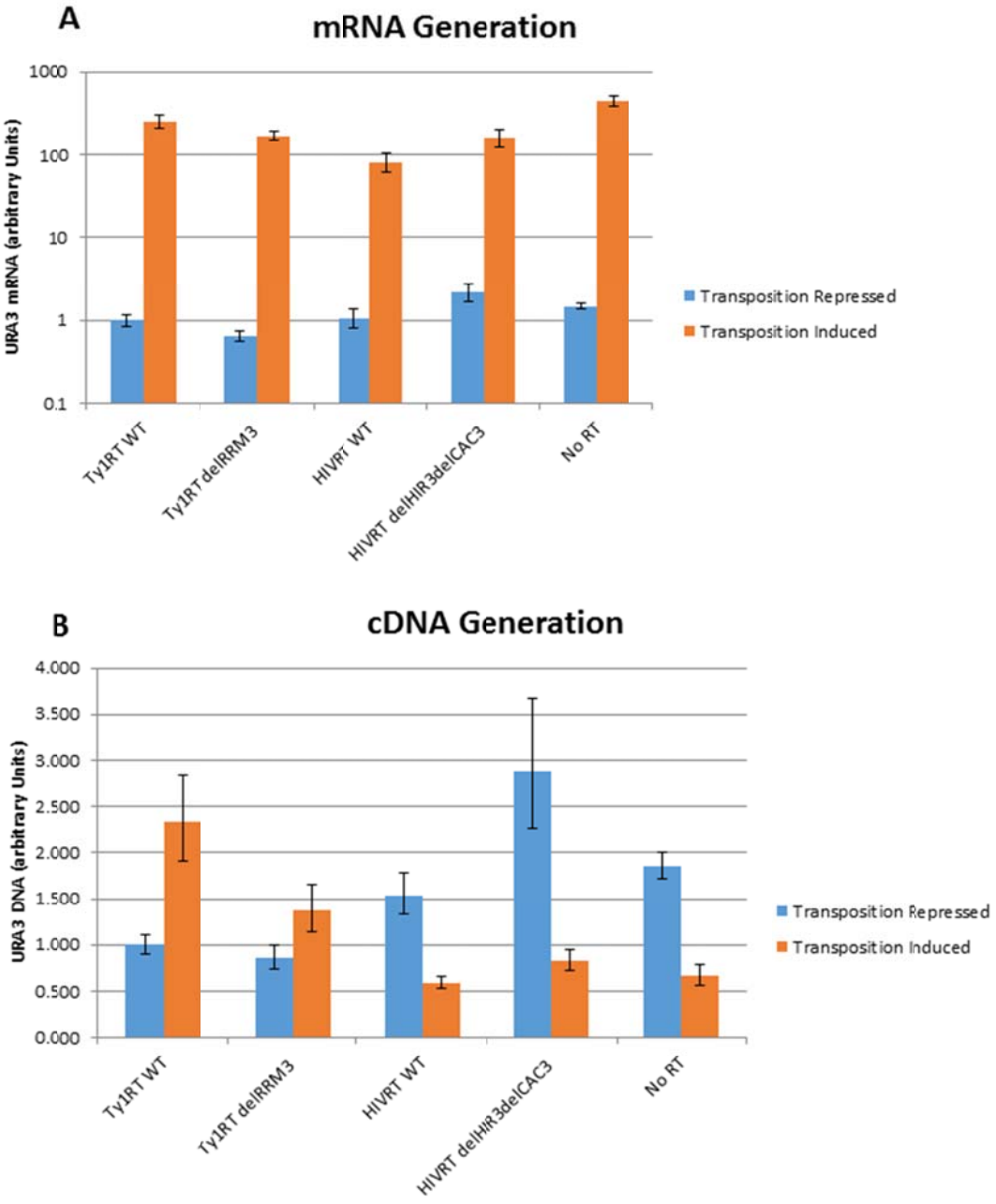


Figure 7-25: Measurement of transcript and cDNA levels produced by HIVRT and Ty1RT.

A) mRNA generation. B) cDNA generation.

7.2.15 Decreasing Proposed Genomic Integration of Tranposants through Integrase Engineering

We hypothesized that the pGALmTy1-Ty1 plasmid generates cDNAs which are preferentially integrated into the genome as opposed to plasmids. Because such genomic integration may generate effects which are not related to the function of the enzyme being mutated, it is desirable to minimize the extent of this phenomenon. However, it is known that in Ty1, the integrase and reverse transcriptase are processed as a polyprotein which is subsequently cleaved (236). Therefore, it is possible that stop codon mutagenesis of the integrase will also remove reverse transcriptase functionality, which is undesirable. To generate a mutant version of Ty1 in which the functionality of Ty1 integrase is eliminated while still maintaining reverse transcriptase activity, two series of Ty1 integrase mutants were constructed. In the first series, 5 variants were constructed in which stop codons were inserted at one of 5 positions spaced 100bp apart. In the second series, 6 variants were constructed in which 100bp sections of the integrase were deleted between the 5 positions used above for stop codon insertion. The 5' and 3' boundaries of the integrase also served to define these variants. The transposition rate of these variants was tested, and the extent of genomic integration was tested by counting the size of the resulting colonies on media lacking uracil. Small colonies were considered to be the result of genomic integration, as expression levels of *URA3* (and hence growth rate in uracil-deficient media) are likely to be much lower as a single genomic integration than as a part of a high-copy plasmid, and large colonies were considered to be the result of plasmid integration. It can be seen (**Figure 7-27**) that one integrase mutant in particular: Ty1intdel4, maintained a similar level of plasmid-based insertion while reducing the level of genomic insertion by approximately half. This mutant was considered to be possibly beneficial to use during evolution as the likelihood of off-target effects resulting from

genomic integration could be reduced. In a further attempt to reduce genomic integration without compromising reverse transcription activity, a Ty1 integrase mutation previously found to inhibit integrase activity was constructed (237). The construct was then retransformed into a BY4741 strain and the Low-OD transposition rate test was carried out. However, it was found that total transposition activity was significantly reduced when this integrase mutation was present, rendering this mutant unsuitable for the ICE system.

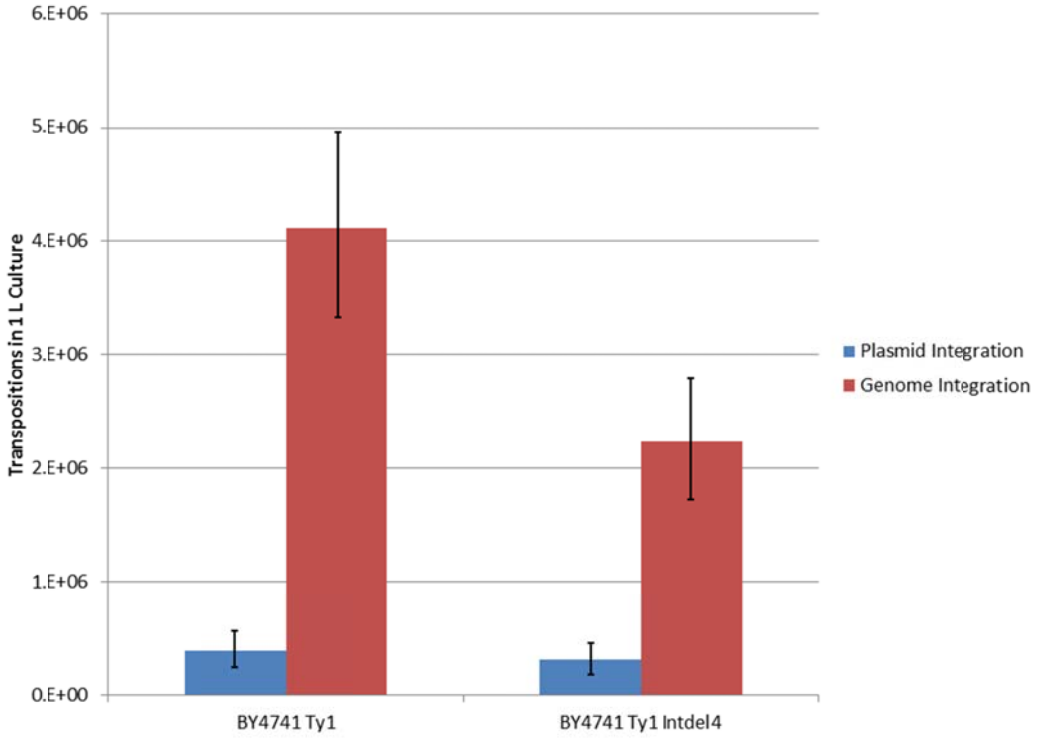


Figure 7-26: Deletion in integrase reduces proposed genomic integration.

A deletion was introduced into the integrase of Ty1 which reduced the proportion of small colonies produced during transposition rate tests. As these small colonies were thought to be the result of genomic integration, it was concluded that this integrase modification reduced the prevalence of genomic integration in favor of plasmid integration.

7.2.16 Integration of Ty1 cDNA

In order to characterize the extent of genomic integration of the synthetic retroelement more thoroughly, a retroelement was introduced into a *URA3*-marked vector which contains a *HIS3* cargo. In this way, the same retrotransposition rate test could be carried out in this system with the opposite selection (histidine dropout plates instead of uracil). Initial tests with this vector demonstrated that the overall rate of retrotransposition was approximately equivalent to the *URA3*-marked construct. This construct was then used to characterize the extent of genomic integration of reverse transcribed cDNA according to the scheme described below.

Cultures of BY4741 containing this new construct were first grown and plated on both YPD plates and YPD plates containing 5-FOA, which is toxic to cells expressing *URA3*. Colonies were observed on both plates, indicating that cells can spontaneously lose a *URA3*-containing plasmid upon plating if 5-FOA is present. In addition, a separate aliquot of cells were plated on uracil dropout plates. The resulting colonies were then picked and spread directly on YPD+5-FOA plates and growth was observed, further indicating that cells expressing a *URA3*-marked plasmid can be cured in the presence of 5-FOA. Next, retrotransposition was induced at high OD using galactose media, followed by plating on histidine- and uracil-dropout media as well as histidine-dropout media with 5-FOA present. Since 5-FOA precludes growth of cells containing *URA3*, but histidine-dropout media requires a functional HIS3 (and thus retrotransposition and re-integration of cDNA), the only cells that could grow on these plates were those which re-integrated cDNA directly into the genome, then lost the original plasmid. However, while colony counts of the histidine and uracil dropout plates confirmed the expected retrotransposition rate, *no* colonies grew on 5-FOA-containing histidine-dropout plates. This test then implied that all cDNA reverse transcribed from the retroelement-containing

plasmid is reintegrated into the plasmid, and no copies are integrated into the yeast genome.

7.2.17 Effects of Transcript Length on Ty1 Retrotransposition

During the ICE retrotransposition process, the entire synthetic retroelement is transcribed and reverse transcribed. However, all experiments exploring the rate of retrotransposition had been done using one cassette, containing the *URA3* gene with an artificial intron. It is possible that inserting a longer DNA sequence, and thus requiring a longer RNA transcript to be reverse transcribed, could affect this rate. To determine the effect of transcript length on retrotransposition rate, several constructs were created by inserting truncated genes without promoters (denoted "cargo" DNA) into the retroelement between the *URA3* gene and the reverse transcriptase gene. These constructs were then tested using high-OD galactose induction in BY4741 Δ*rrm3* after varying lengths of time to explore the effect of lengthening induction time on the rate of retrotransposition. These experiments revealed a clear relationship between length of "cargo" sequence and a reduced rate of retrotransposition; exogenous sequences up to ~5000 bp reduces transposition by an order of magnitude, and up to ~6000 bp reduces this rate by a further order of magnitude. However, lengthening the induction time at high OD from 3 to 7 days can slightly increase the number of retrotransposition events, especially for constructs containing the longest "cargo" sequences (**Figure 7-28**).

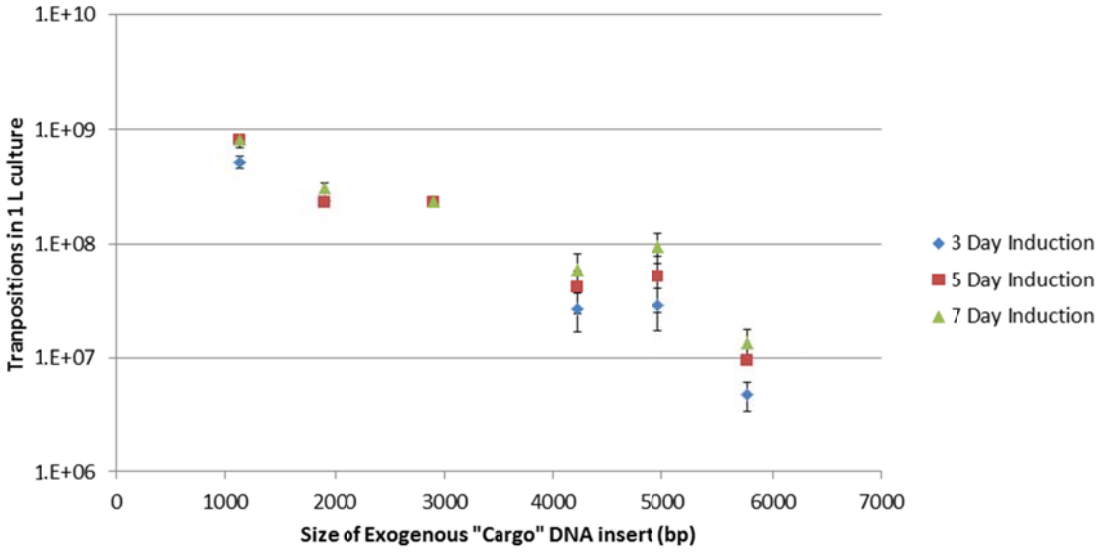


Figure 7-27: Effect of cargo size on transposition rate.

Several lengths of cargo DNA were cloned into pGALmTy1-Ty1 and transposition rate was measured. It can be seen that increasing cargo size to 6kb reduces transposition rate by approximately two orders of magnitude.

7.2.18 Construction of the *SPT15*-containing Retroelement System

Previously, the *SPT15* gene which encodes the TATA-binding protein in *S. cerevisiae* was employed to significantly increase the tolerance of yeast grown in the presence of ethanol (78). To demonstrate the application of ICE through the engineering of transcription factors, the genes *SPT15* and *SPT15*-300 were cloned into the optimized retroelement pGALmTy1-Ty1-TEF1, yielding pGALmTy1-Ty1-Spt15-TEF1 and PGALmTy1-Ty1-SPT15-TEF1. The best knockout strain for Ty1 reverse transcriptase, BY4741 Δ*rrm3*, was then transformed with either of these plasmids.

7.2.19 Characterizing Induction of GAL promoter by Growth in Xylose

During ICE-enabled evolution, the transcription and reverse transcription of the retroelement is induced using the inducible *GAL1* promoter. This promoter is highly repressed by growth in glucose medium, allowing alternating rounds of thus

mutagenesis and selection/screening. However, during the evolution of genes involved in xylose catabolism, a xylose growth selection is employed, and it is not currently understood how the *GAL1* promoter functions in the presence of xylose. To characterize this effect, a plasmid with a fluorescent reporter gene under the control of pGAL1 was transformed into a xylose-consuming strain of *S. cerevisiae* (4), and the fluorescence was measured under different growth conditions. It was observed that growth in glucose followed by xylose resulted in low levels of expression, and that growth in galactose followed by xylose resulted in much higher levels of expression, although less than that observed during growth on galactose (**Figure 7-29**). This seems to indicate that growth in xylose is permissive for pGAL1 activity, although after a time transcription rate is reduced.

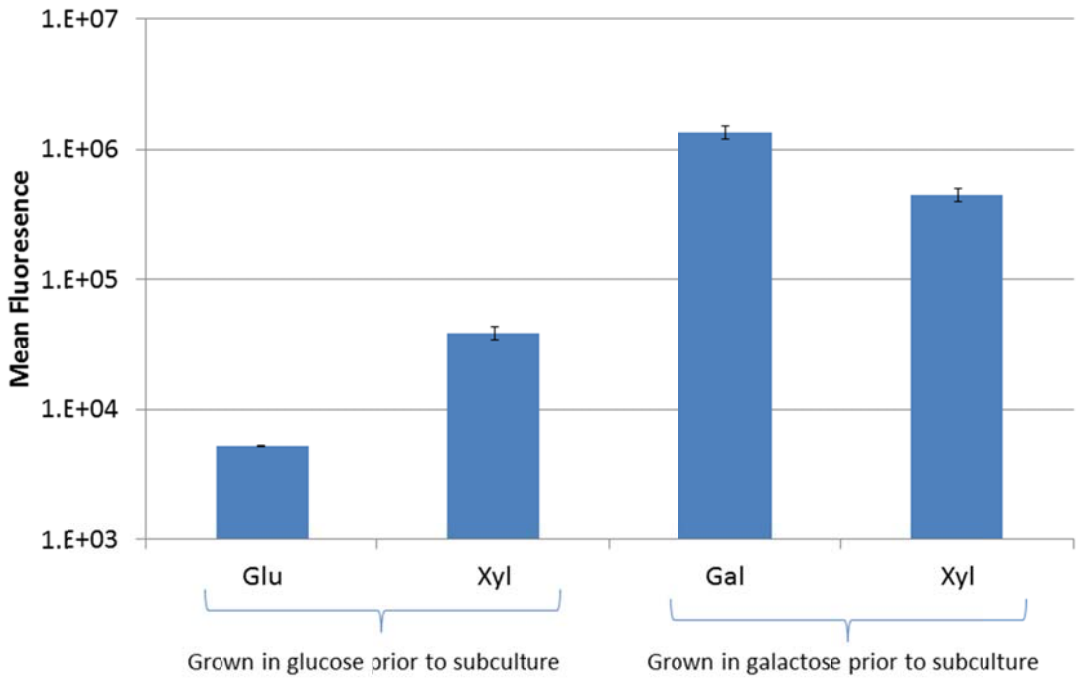


Figure 7-28: Induction of pGal1 by various carbon sources.

A fluorescent reporter was placed under the control of the Gal1 promoter, and fluorescence was measured in the presence of various carbon sources. It can be seen that galactose greatly increases the expression level enabled by the Gal1 promoter, as expected, and also that xylose seems to be permissive for pGal1 expression as well.

7.2.20 Construction of *XylA*-containing Retroelement System

An efficient xylose isomerase enzyme in yeast has the potential for significant improvements to xylose utilization, since this bacterial pathway bypasses cofactor requirements found in yeast's native oxidoreductase pathways. Previously, classical directed evolution of the *Piromyces* sp. xylose isomerase (encoded by *xylA*) has led to a beneficial mutant (designated as *xylA3*) with a 77% increase in enzymatic activity (238). The same genes *xylA* and *xylA3* were evolved using the ICE system by inserting *xylA* and *xylA3* into the optimized retroelement pGALmTy1-Ty1-TEF1, yielding pGALmTy1-Ty1-MCS-XylA-TEF1 and pGALmTy1-Ty1-XylA3-TEF1. Based on strain BY4741 Δ*rrm3*, two additional Δ*gre3* strains with or without an extra copy of xylulokinase (*XKS1*) integrated into the genome were constructed and transformed with the *XylA*(3)-containing retroelement system. GRE3, which encodes an aldose reductase, was knocked out in order to reduce endogenous xylose utilization and allow any potential improvements in *XylA* to confer a greater phenotypic advantage, thus increasing the sensitivity of a growth-based screen.

Three strains: BY4741 Δ*rrm3*, BY4741 Δ*rrm3*/Δ*gre3*, and BY4741 Δ*rrm3*/Δ*gre3*/*XKS1* were transformed with plasmid pGALmTy1-Ty1-MCS-XylA-TEF1 or pGALmTy1-Ty1-XylA3-TEF1, yielding six strains BY4741Δ*rrm3*-*XylA*, BY4741Δ*rrm3*-*XylA3*, BY4741Δ*rrm3*/Δ*gre3*-*XylA*, BY4741Δ*rrm3*/Δ*gre3*-*XylA3*, BY4741 Δ*rrm3*/Δ*gre3*/*XKS1*-*XylA*, and BY4741 Δ*rrm3*/Δ*gre3*/*XKS1*-*XylA3*. All six strains were confirmed to grow on xylose-containing agar plates.

7.2.21 Effects of Vector Copy Number on Ty1 Retrotransposition

All experiments using optimized ICE retroelements up until this point were based on a high-copy plasmid vector containing the 2μ replication site. In order to establish the effect of plasmid copy number on the retrotransposition process, a low-copy version of pGALmTy1-Ty1 was constructed by replacing the 2μ sequence with the CEN6/ARSH replication site. The two vectors were then separately transformed into BY4741 Δ*rrm3*, and the retrotransposition rate under galactose induction was measured (low-OD method). Results indicated that using a low copy vector did not affect the overall number of transpositions, implying that the amount of template DNA was not the limiting step in the retrotransposition process (**Figure 7-30**). These results suggested that using a low-copy vector should not negatively impact the evolution of a gene or pathway using ICE.

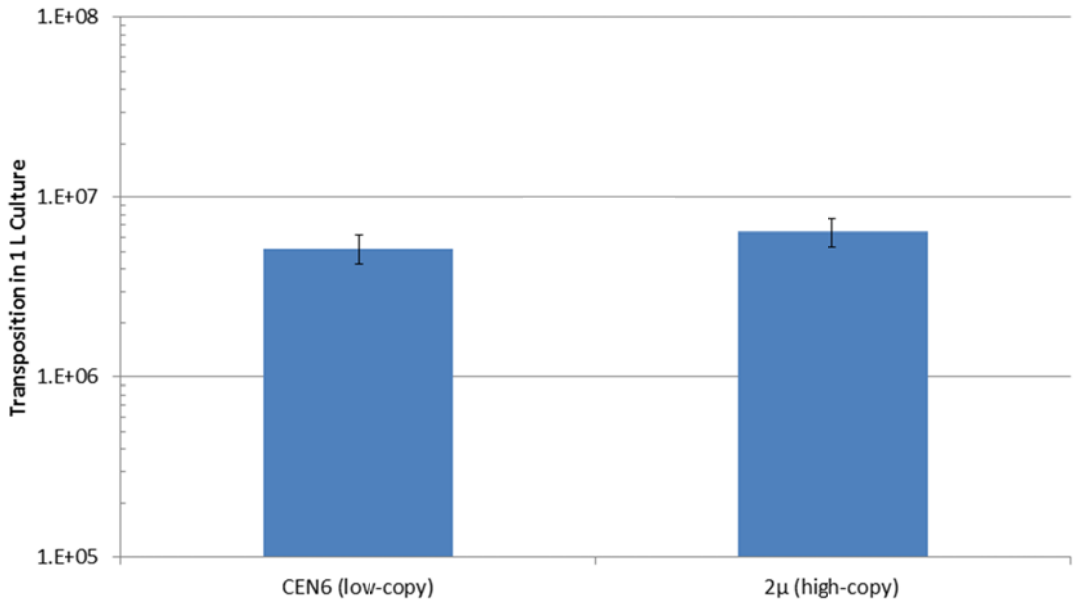


Figure 7-29: Retrotransposition rate of low-copy retroelements.

The synthetic retroelement was placed on a low copy vector and transposition was induced. It was observed that copy number of the retroelement does not limit transposition in this context.

7.2.22 Inefficient Plasmid Segregation Limited the Success of ICE

Inefficient plasmid segregation could severely limit the success of ICE. If a small fraction of a mother's plasmid population is transferred to her daughter cell, then a mother will have to bud multiple times before she will be likely to pass on a mutated plasmid, even if the mutated plasmid is beneficial. The progenitor to the P4xx series of plasmids used in this study, the pRS series, has been reported to be lost through mitotic segregation at rates of approximately 4.4% of progeny per doubling. This loss rate would imply (see materials and methods) that, on average, a mother cell containing 30 copies (239) will only transfer 3 of them to her daughter cells. Given that only one out of 30 plasmids is likely to contain a mutation in the first place, this implies that the transfer rate of mutated plasmids can be extremely low, even if the mutated plasmid confers a growth advantage. On the other hand, if a mother cell contains 3 plasmid copies (as is the case for low-copy centromeric vectors), a rate of loss of 4.4% would correspond to 2 plasmid copies being transferred to a daughter cell and thus a much higher probability that a daughter cell would contain a mutated plasmid. Therefore, we synthesized low-copy versions of our retroelement in order to reduce the effect of segregation efficiency and improve the ability of ICE to select for improved mutants.

7.2.23 Construction of Low-copy Vectors for Evolution Experiments

Since experiments have shown that including the optimized retroelement on a low-copy vector does not significantly affect the rate of transposition, low-copy versions of the retroelement constructs containing *XylA*, *Spt15*, *XylA*-3, and *Spt15*-300 were made. In a cell expressing a small number of plasmids, any putatively beneficial mutation obtained through the ICE retrotransposition cycle will be more likely to be selected for than in a cell with a high number of plasmids all expressing the wild-type gene.

7.2.24 Reduction of Wild-type Background through the Inclusion of Introns in Synthetic Retrotransposon

Because we were unable to observe any mutant generation during initial evolution experiments of *XylA*, *XylA*3, *SPT15*, and *SPT15*-300, we also wished to include a feature to our retroelement which would enable facile recovery of constructs which have undergone retrotransposition (and so would be more likely to contain mutations). Therefore, we generated constructs which contain an intron interrupting the coding sequence of *XylA*, *XylA*3, *SPT15*, and *SPT15*-300. The purpose of this intron was twofold. Firstly, the presence of an intron would prohibit the expression of enzymes contained within retroelements which have not undergone transposition, ensuring that mutants would not be produced within a cell containing a high background of wild-type enzyme, thus potentially amplifying the effect of the mutant. Secondly, isolation of mutants would be facilitated through the inclusion of a restriction site within the intron, allowing the researcher to enrich for transposed retroelements during plasmid isolation through a simple restriction digest before transformation into *E. coli*. This strategy, coupled with the use of low-copy vectors to propagate our retroelement, was expected to result in higher isolation efficiencies of mutated retroelements.

7.2.25 Development of Nonevolving Controls for Evolution Experiments

A major challenge of ICE is the propensity for strain adaptation to take place as directed evolution is occurring. As this process can potentially confound the identification of beneficial mutants, it is important to understand the extent to which strain adaptation is occurring in future evolution experiments. To address this need, we developed 2 control plasmids for each target which each lack the Ty1 reverse transcriptase, thus reducing retrotransposition (and thus mutation) by several orders of magnitude. In one of these control plasmids (pGALmTy1-(x)intron), the gene to be

optimized is interrupted by an intron, whereas the other control (pGALmTy1-(x)) maintains an intact copy of the gene. In this way, cells containing the pGALmTy1-(x)intron plasmid are left to adapt in the context of the retroelement overexpression alone, and cells containing the pGALmTy1-(x) plasmids will adapt in the context of both retroelement and gene overexpression. By comparing the growth rates of the experimental strain and the two controls, the experimenter will be able to determine if the mutagenic activity of the reverse transcriptase towards the gene of interest is conferring an additional phenotypic benefit than strain adaptation alone, thus indicating that the experimental strain contains beneficial mutants. We thus constructed control strains for each gene of interest in this study: *SPT15*, *SPT15*-300, *XylA*, and *XylA*3.

7.2.26 Evolution Study of *Spt15* and *Spt15*-300

Two versions of *Spt15*, which encodes the TATA-binding protein in *Saccharomyces cerevisiae,* were chosen for *in vivo* continuous evolution: wild-type *Spt15* and a previously identified beneficial mutant *Spt15*-300 (1). In prior work, evolution of *Spt15* was undertaken with a high-copy vector without an artificial intron, resulting in a high level of wild-type background. Therefore, the retroelement was cloned in a low-copy plasmid to eliminate inefficient plasmid segregation. Also, the coding sequences of *Spt15* and *Spt15*-300 were interrupted with an artificial intron. The six plasmids, pGALmTy1-Ty1-Spt15intron-TEF1 (low copy), pGALmTy1-Spt15intron-TEF1 (low copy), pGALmTy1-Spt15-TEF1 (low copy), pGALmTy1-Ty1-Spt15-300intron-TEF1 (low copy), pGALmTy1-Spt15-300intron-TEF1 (low copy), and pGALmTy1-Spt15-300-TEF1 (low copy) were then transformed into BY4741 Δ*rrm3*, and the resulting strains were designated as STI, SAI, SA, S3TI, S3AI, and S3A. These strains were then used for evolution in either the continuous or oscillatory mode.

In continuous mode, all six strains were pre-cultured in glucose medium and then induced in galactose medium under high OD for three days. The cultures were then transferred to a medium containing 120g/L galactose and 6% ethanol, which was designed to provide a high selective pressure while simultaneously maintaining induction of the pGAL1 promoter. As these strains grew to saturation, they were subcultured to fresh media containing the same amount of galactose but 0.5% more ethanol. Additionally, the inoculum volume was decreased (**Figure 7-31** (top)). It can be seen (**Figure 7-32A,B**) that STI attained higher growth rates than SAI or SA after subcultures 2, 4, and 5, while S3TI attained higher growth rates than S3AI or S3A after subcultures 2 and 3. We expect that STI continued to accumulate beneficial mutations throughout its evolutionary trajectory, allowing it to surpass both of the control strains. Although S3TI initially outperformed the control strains (indicating the presence of beneficial mutations), it appears that an adaptive mutation appeared in the genome of S3A after subculture 4, enabling it to grow faster than S3TI.

In the oscillatory strategy, all six strains were pre-cultured in glucose medium and then induced in galactose medium under high OD for three days. Several repeated rounds of retrotransposition induction and selection were then undertaken by serial culture in an alternating sequence of galactose and glucose plus ethanol medium. Here, galactose medium was used for induction while glucose plus ethanol medium was used for selection. A higher concentration of ethanol was added in the selective medium over each subculture in order to further enrich for beneficial mutants (**Figure 7-31** (bottom)). The growth curves of the oscillatory evolution process for *Spt15* and *Spt15*-300 are shown in **Figure 7-32C and D**, respectively. Each galactose induction period is indicated as orange dotted lines. In total, there were seven subcultures and the ethanol concentration was increased up to 8.25% (v/v). Unfortunately, the strain (STI/S3TI) with

retroelement cassette showed similar growth profile compared with the control strains, and no significant improvement was observed. It was observed that the control strains SA and S3A could also grow up in the selective medium, possibly due to the acquisition of an adaptive mutation within the genome. Surprisingly, the control strains (SAI/S3AI) which include an artificial intron in the target gene and lack the reverse transcriptase also showed comparable growth profiles in the selective medium. One possible explanation for these results is that alternating between selective and nonselective conditions might permit global modifications to the genome through adaptive evolution. These preliminary results indicated that ICE has the potential to generate strains with significantly improved phenotypes than could be attained through strain adaptation alone, indicating the generation of improved mutants. In addition, these results indicated that for ethanol tolerance, ICE is best utilized in a continuous mode.

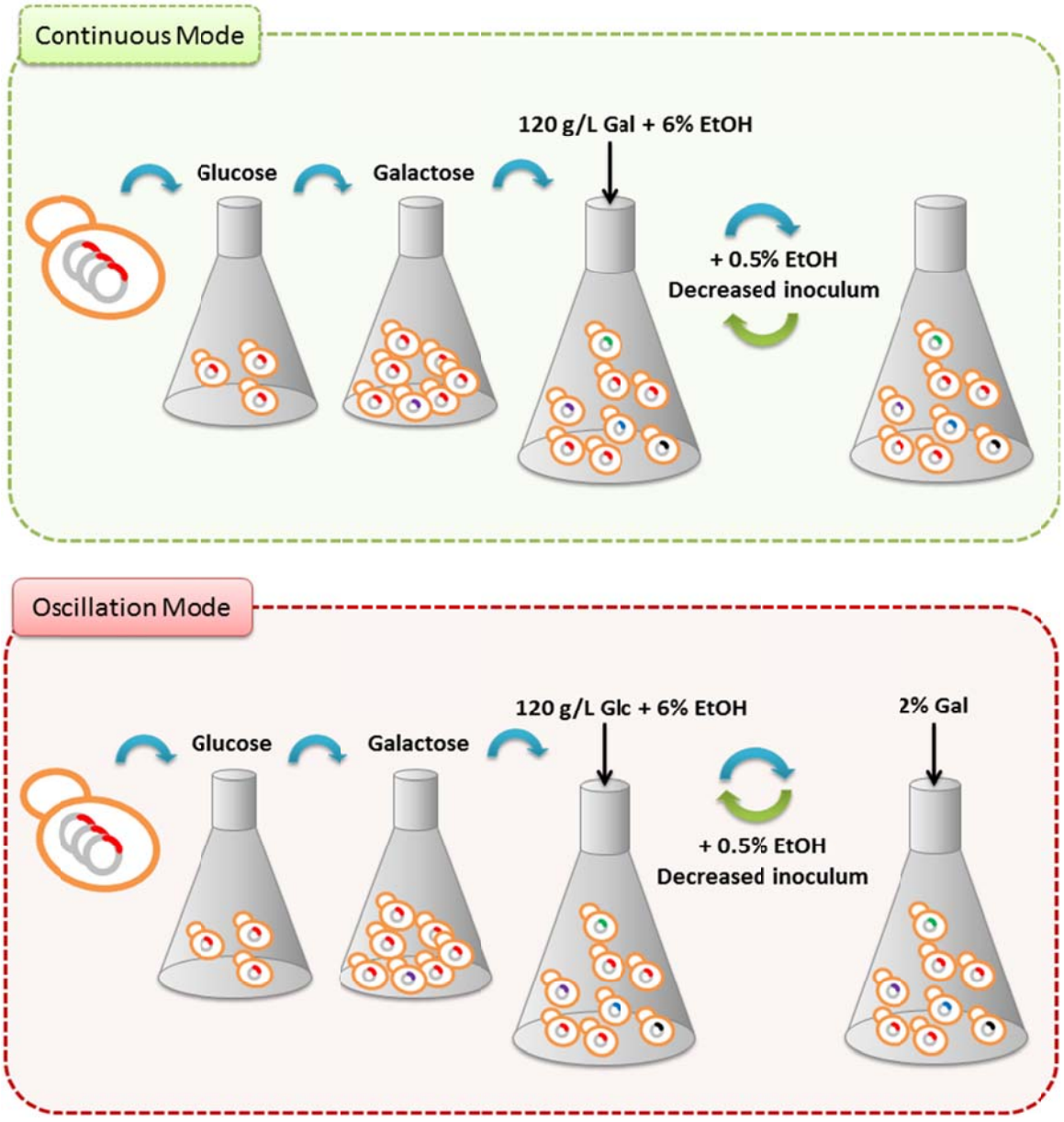


Figure 7-30: Overview of evolution strategies.

Top: Continuous mode. In the continuous mode, selection takes place concurrently with mutant generation.
Bottom: Oscillation mode. In the oscillatory mode, selection and mutant generation occur separately.

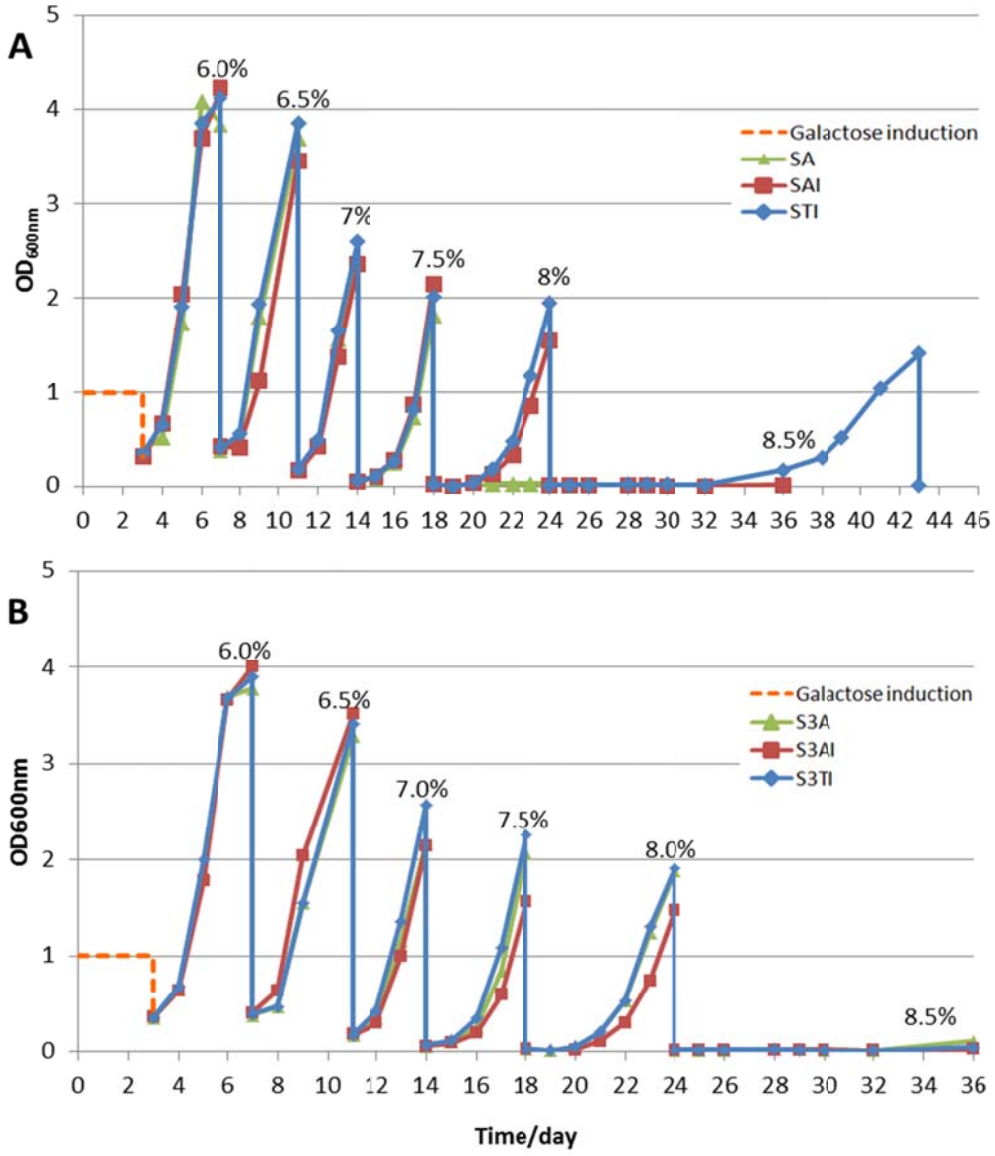


Figure 7-31: Growth of strains expressing *SPT15* or *SPT15*-300 evolution cassettes.

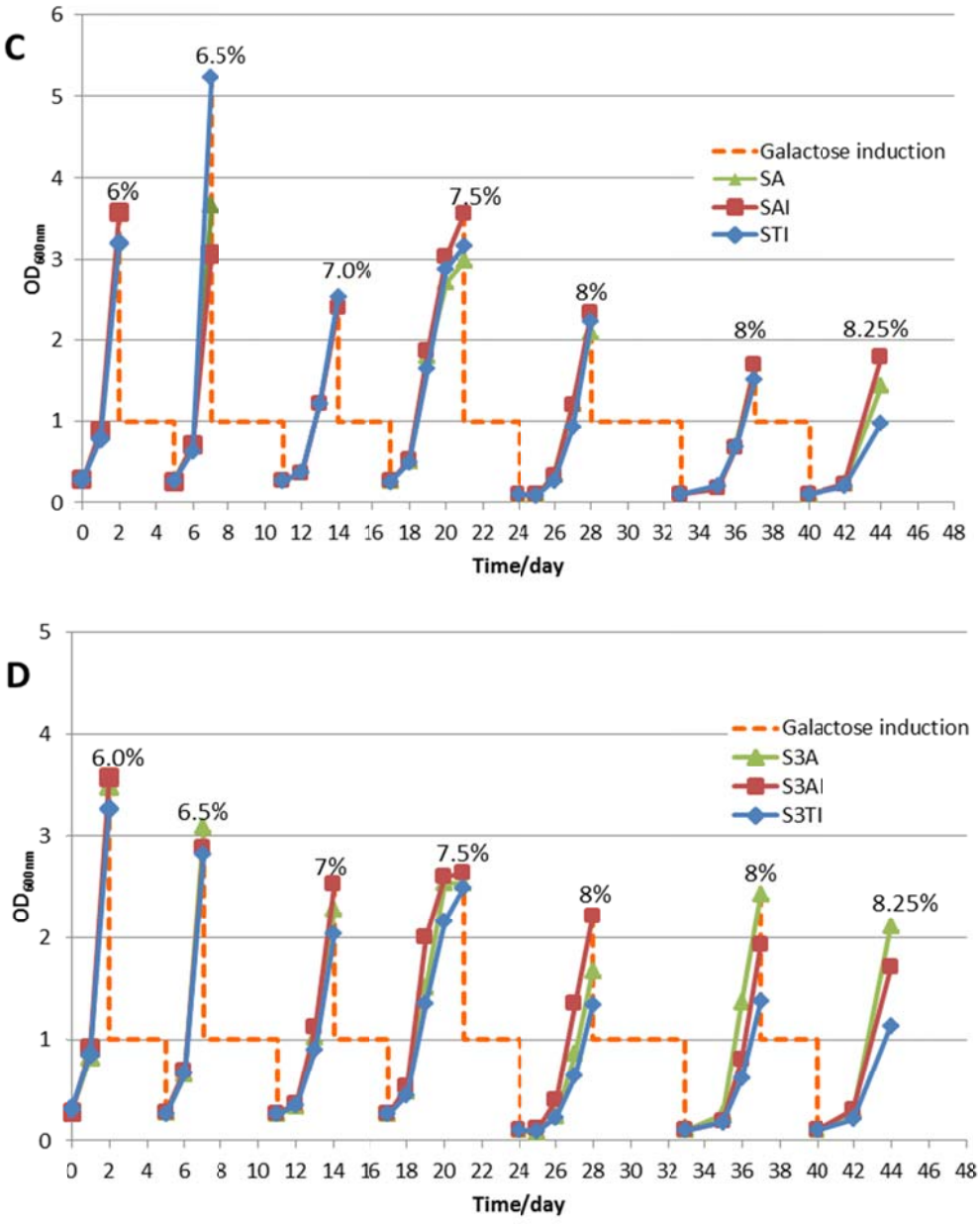


Figure 7-31 (continued): Growth of strains expressing *SPT15* or *SPT15*-300 evolution cassettes.

Strains were transformed with plasmids containing evolution cassettes for *SPT15* and *SPT15*-300 and the associated control strains. These strains were then subcultured in increasing concentrations of ethanol in either continuous or oscillatory mode. A) *SPT15* in continuous mode. B) *SPT15*-300 in continuous mode. C) *SPT15* in oscillatory mode. D) *SPT15*-300 in oscillatory mode. It can be seen that continuous mode enables the greatest separation between the experimental and control strains.

7.2.27 Mutant Recovery

Cultures from each round of evolution were processed to confirm mutant sequences using several strategies. In the first strategy, 1.5 mL culture was collected to purify yeast plasmid. This was further digested with *Asc*I (a restriction enzyme specific to the intron sequence) to eliminate any untransposed background plasmid. This digested plasmid was then transformed into *E. coli* for sequencing. Sequencing results indicated that either the wild type plasmid with intron was still present or plasmids in which the full retroelement was excised were present. In later subcultures, the fraction of plasmids in which the full retroelement was excised was increased. This phenomenon was also observed in previous experiments attempting evolution using high-copy plasmids and in experiments using *URA3*AI.

In the second strategy, genomic DNA was extracted from single colonies. Any integrated TEF-Spt15 or TEF-Spt15-300 expression cassettes were then PCR amplified and digested with *Asc*I. After a gel check of digestion product, the cassette with intron should be cleaved into two small fragments of 762 and 449 bp, or should appear as a 1.21kb band if it is uncleaved. In contrast, cassettes without intron should appear as a 1.14kb product. It was observed that all tested colonies showed two bright bands of 762 bp and 449 bp, while only one colony from subculture 4 of *Spt15* oscillation evolution (referred as SO4-1) had two additional faint bands of 1.14 kb and 1.21 kb. Both bands were gel extracted and sequenced, confirming that the intron was indeed excised from the smaller product. However, as the sequencing read showed a mixture of intron-containing and intronless sequences, no mutants could be clearly distinguished.

In the third strategy, genomic DNA was digested with *Asc*I first before amplifying the TEF-Spt15 or TEF-Spt15-300 expression cassettes. Still, only SO4-1 showed the promising 1.14 kb band. However, this strategy increased the concentration of the 1.14 kb product, as visualized by gel electrophoresis. The sequence results showed that the intron of plasmid in SO4-1 was excised, but no mutations were present.

In the fourth strategy, instead of amplifying the whole gene cassette, two pairs of intron-spanning primers were specifically designed to amplify either the intronless TEF promoter or the intronless *Spt15*(300). These primers were then used to PCR gDNA extracted from a single colony or gDNA from 3 mL from the evolution culture. Among all the TEF amplicons, 15 out of 64 showed evidence of a transposed product, while others have faint off-target products. Among all the *Spt15* or *Spt15*-300 amplicons, 30 out of 64 PCRs showed evidence of a transposed product. Unfortunately, all 15 TEF promoter products and 30 *SPT15*(300) products were wild-type.

7.2.28 Genomic Integration of Optimized Ty1 Retroelement

Until this point, all ICE experiments have used plasmids containing the optimized retroelement. However, during cDNA reintegration, it is not clear whether the cDNA integrates back into a plasmid or elsewhere in the genome. This could result in multiple copies of the retroelement, which would impair both the evolution of genes and pathways contained within and the recovery/characterization of these parts. To improve this process, a strain containing a genomic integration of the optimized retroelement was created and transposition rate was measured. Although transposition rate was significantly reduced to $4*10^5$ per liter, each colony obtained from the transposition rate analysis showed one unique transposition event which replaced its parent sequence in the genome. Furthermore, one sequence excitingly contained an amino acid substitution of

the *URA3* gene, indicating not only that the Ty1 reverse transcriptase can generate detectable levels of mutation through standard sequencing analysis, but also that *URA3* is tolerant to some level of amino acid substitution, indicating that measurement of transposition rate using this gene may not be highly sensitive to mutation rate.

7.2.29 Improvement of Transposition Rate of Genome-Encoded Retroelements

In order to further improve the transposition rate enabled by genomically encoded retroelements, we investigated the effect of inducing transposition at low temperatures, as it has been shown that 22º C greatly enhances the rate of transposition (240). Although we previously investigated this temperature for increasing the transposition rate of plasmid-encoded retroelements, no increases to transposition rate were observed. To our pleasant surprise, we observed that induction at this temperature greatly increased transposition rate to almost 10^8 per liter (**Figure 7-33**). This places the efficiency of the genome-encoded retroelement in the same regime as that of plasmid-encoded retroelements (but with a greater capacity for sequence retrieval) and so inspired us to reconstitute the evolution cassettes for *SPT15*(300), *XylA*(3) and the *XylA-XKS* pathway in a genomic context.

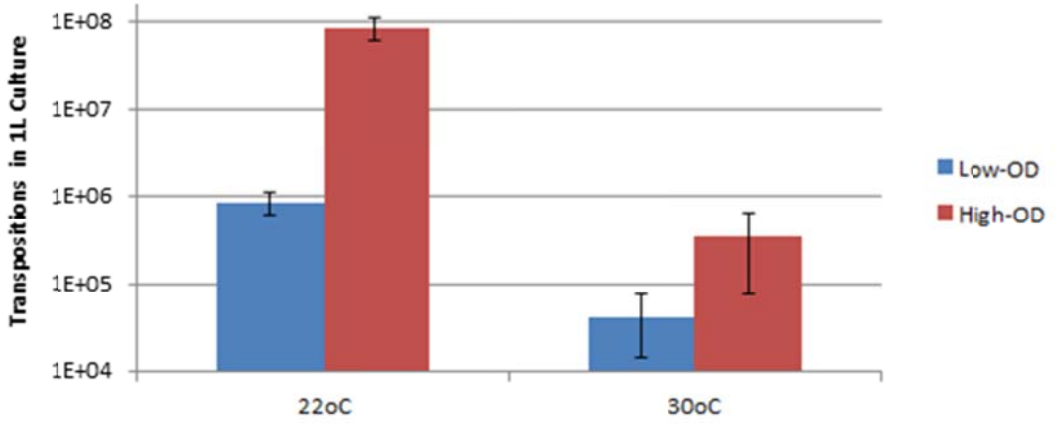


Figure 7-32: Low temperatures accelerate transposition rate of genome-encoded retroelements

7.3 DISCUSSION

Through the modification of a yeast retrotransposon, we have developed a system enabling the *continuous* evolution of any desired gene product in yeast. We have characterized the effect of a multitude of gene deletions on transposition activity and have identified gene deletions which improve this rate. We have also characterized the activity of the Ty1 reverse transcriptase in alternative strains of yeast, as well as the effect of several gene deletions in these strains. Furthermore, we have identified gene overexpressions which greatly increase the rate of Ty1 transposition. In addition, we have identified mutations in the Ty1 reverse transcriptase which improve transposition rate as well as mutation rate, thus contributing to our growing knowledge of this enzyme. Importantly, we have conclusively shown that the HIV reverse transcriptase cannot functionally replace the Ty1 reverse transcriptase in our synthetic retroelement, as shown by measurements of cDNA levels generated by this heterologous enzyme. This deficiency is not due to incorrect proteolytic processing, but rather may result from a lack of human tRNAs in yeast or some other unknown limitation of this enzyme's functionality in this organism. Instead, we expect that latent Ty1 activity in the yeast

genome contributed to the reported activity of this enzyme (210). Finally, we have optimized the Ty1 retrotransposon for use as a tool for directed evolution, both by modifications which increase the activity of the retroelement itself, but also by developing strategies to recover mutants after the directed evolution process is complete. The library size for a 1kb gene attainable through the approach as it is currently implemented is $1.1*10^6$, which is equal to that which can be obtained through the current state-of-the-art, yet has the potential to be generated *continuously* in yeast. We have shown the utility of this approach for improving the tolerance of yeast to high concentrations of ethanol and galactose. The failure to isolate causal mutations for our initial demonstration of ICE illustrates a major limitation of undertaking within-cell mutagenesis in a multi-copy plasmid system. It is expected that integration of evolution cassettes into the yeast genome, coupled with optimization of transposition rate as undertaken for plasmid-based systems, will enable successful implementation of ICE for future evolution experiments.

To overcome limitations to library size due to mutation rate, it will be necessary to develop highly mutagenic reverse transcriptases. To this end, we devised a two-color fluorescence assay to enable simultaneous measurement of both transposition rate and mutation rate through flow cytometry. In this assay, the cargo of the retroelement is a translational fusion of mStrawberry and YFP. Before transposition, neither protein may be expressed due to the presence of an intron interrupting the coding sequence of mStrawberry. After transposition has occurred, both genes will be expressed. However, if any one of the genes has been mutated by the reverse transcriptase, it may not be functional. This scheme will enable the deduction of both transposition rate and mutation rate from the fraction of cells expressing YFP only, RFP only, or both fluorescent proteins. In practice, a culture expressing a single reverse transcriptase variant will be

induced and allowed to transpose over the course of several days, after which a large number of cells will be queried using flow cytometry. Using this technique, several hundred reverse transcriptase variants may be analyzed per day, and the variant enabling high library sizes will be selected for use in downstream applications and for further optimization of mutation rate.

In the future, we also plan to implement ICE for the evolution of entire pathways. First, the effect of terminators on Ty1 retrotransposition must be ascertained. Incorporating at least one terminator in the retroelement is important for the evolution of multiple gene pathways in a single construct, but it is possible that any bi-directionality in terminator activity could also interfere with the transcription or reverse transcription process. Once this effect has been characterized, and suitable mono-directional terminators have been found, evolution will commence on two model pathways: xylose and arabinose utilization.

In order to initially demonstrate that ICE can simultaneously co-evolve a collection of synthetic parts to improve the performance of an entire pathway, we will construct a Ty1 retroelement system for the evolution of xylose utilization through a pathway composed of xylose isomerase (*XylA* or *XylA*3) and xylulokinase (*XKS1*). The expression of xylose isomerase and xylulokinase will be controlled under the TEF1 and GPD promoters, respectively. Two short synthetic terminators, Tkc1 and Tkc6, will be investigated for termination of transcription (unpublished work). Upon integration of this construct into BY4741 Δ*rrm3* and BY4741 Δ*rrm3*/Δ*gre3*, we plan to use *in vivo* continuous evolution to improve the activity of the xylose catabolism pathway.

As a second proof-of-concept for the evolution of pathways in yeast, ICE will be implemented for the evolution of arabinose catabolism. Arabinose is the second most-abundant pentose sugar in lignocellulosic biomass, yet yeast does not possess the ability

to effectively utilize this carbon source. Our lab has recently isolated a strain of yeast (*U. bevomyces*) which displays the remarkable ability to grow on arabinose as the sole carbon source in minimal media. Genome sequencing indicated a 5-gene pathway which may be responsible for this phenotype, and it has been shown that the introduction of this pathway is also sufficient to confer this phenotype to *S. cerevisiae*. In spite of this, the ability of *S. cerevisiae* to utilize this carbon source (as measured by cell growth rate) remains very poor. Therefore, we intend to improve the ability of yeast to utilize arabinose by subjecting the entire 5-gene pathway to *in vivo* continuous evolution. In addition, we will also be constructing shortened versions of this pathway because preliminary experiments indicate that all five genes may not be necessary for arabinose utilization. Because this pathway is so long (up to 8.6kb), several challenges must be addressed, including the effect of cargo size on retrotransposition as well as the effect of terminators on retrotransposition. Once these effects have been characterized (and mitigated if necessary), this pathway will be evolved using ICE. These demonstrations will unequivocally show the utility of ICE as a tool to accelerate the development of improved strains through the generation of large library sizes in a facile manner.

Chapter 8: Optimization of a Yeast RNA Interference System for Controlling Gene Expression and Enabling Rapid Metabolic Engineering

8.1 INTRODUCTION

In this chapter, we demonstrate the utility of a synthetic RNAi system in yeast for gene expression control. First, we elucidate key design principles for the construction of hairpin RNA expression cassettes in yeast as well as optimize expression of key components of the RNAi pathway. We then use these parameters to demonstrate the controlled regulation of a synthetic fluorescent protein. Finally, we demonstrate that this heterologous RNAi pathway can enable rapid strain prototyping by examining three industrially relevant strains of yeast (BY4741, CEN.PK2-a, and Sigma 10560-4A) to quickly identify routes for the improvement in titer of itaconic acid (a top value-added chemical from biomass (241)), thus demonstrating that this synthetic approach can speed the design-build-test cycle in yeast.

As a further extension of this method, we develop an optimized RNAi system enabling the use of cDNA fragments as guide RNAs. This work enables us to apply RNA interference to detect beneficial knockdowns on a transcriptome-wide scale for the improvement of 1-butanol, isobutanol, and lactic acid tolerance in yeast. This work represents the first high-throughput search for knockdown targets on a genome-wide scale in yeast. This work thus accelerates high-throughput strain modification towards improvement of relevant phenotypes. This approach may be further extended through the use of RNAseq to elucidate the transcriptome-wide responses to gene knockdown, thus enabling directed learning about the phenotype of interest and informing future strain engineering efforts as shown in **Figure 8-1**.

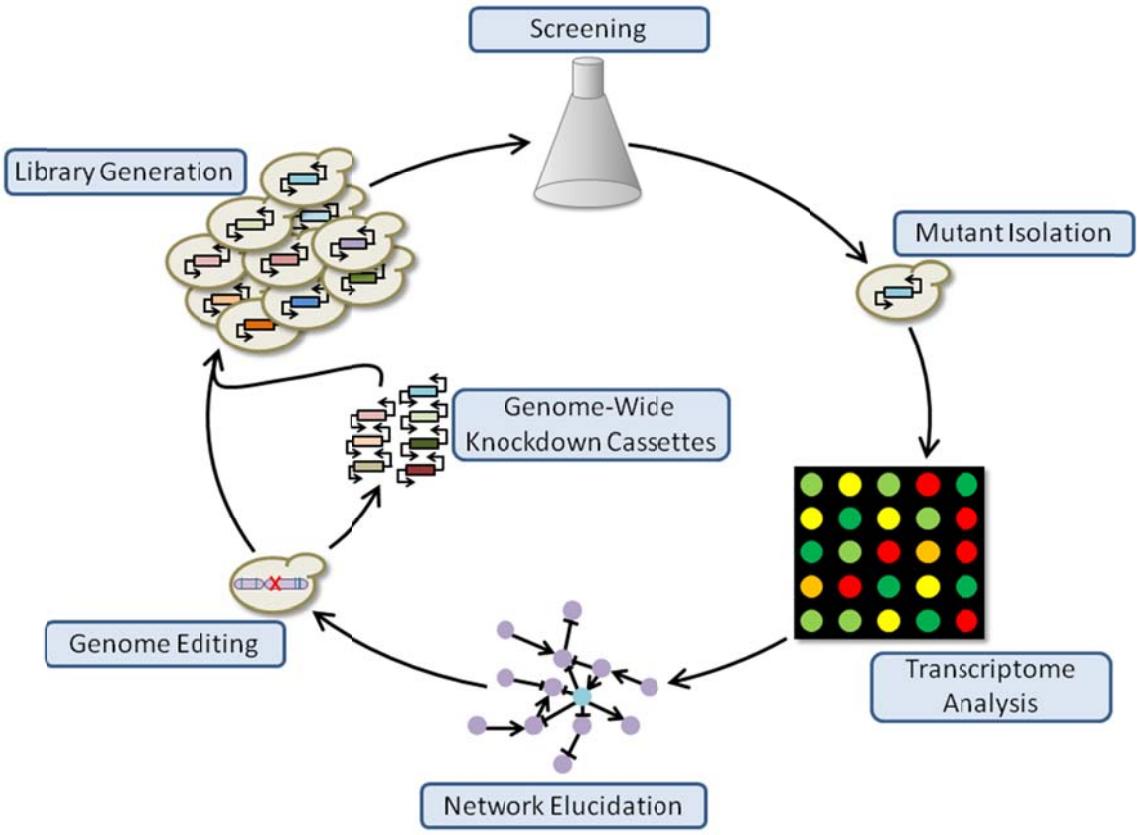


Figure 8-1: Implementation of RNAi for rapid strain engineering on the genome-scale.

Knockdown cassettes may be generated from total cDNA and transformed into the strain of interest to generate an *in vivo* knockdown library. This library may be screened for the phenotype of interest and the causal knockout cassette identified. Then, the transcriptomic changes enabled by this knockdown will enable metabolic engineers to learn about the gene network underlying the phenotype of interest, thus informing future strain engineering efforts in an iterative fashion.

8.2 RESULTS

In order to allow for rapid prototyping of yeast strains, we sought to develop a synthetic, portable version of the RNAi pathway. Previous reports have demonstrated that the RNAi machinery from *S. castelli* can be introduced into *S. cerevisiae* through genomic integration of *S. castelli* Argonaute and Dicer. This prior work demonstrated effective gene silencing during co-expression of a gene-specific hairpin (141). Here, we sought to utilize RNAi as a portable engineering tool and required Argonaute, Dicer, and

a hairpin construct to be expressed on portable plasmids (**Figure 1-1**). The effectiveness of the downregulation pathway was initially characterized through the knockdown of yellow fluorescent protein (YFP) expression. This setup enabled the elucidation of design principles influencing the extent of downregulation in yeast. Our initial design cycle (**Table 8-1**, Design Cycle 0) consisted of YFP, Argonaute, and Dicer each expressed on low copy (centromeric) plasmids, and a 100 bp hairpin RNA complementary to YFP expressed on a high copy (2μm) plasmid. This scheme enabled us to easily change design variables by swapping out plasmids; more streamlined constructs containing Argonaute, Dicer, and a hairpin on the same plasmid may be beneficial for further downstream applications. Following from this design, we investigated the influence of hairpin expression level, hairpin length, the copy number of the target gene, and the copy number of the hairpin plasmid on the efficiency of downregulation. These parameters were then optimized and used to demonstrate the effectiveness of rapid prototyping for metabolic engineering in multiple base strains of yeast.

Design Cycle	Hairpin Expression	Hairpin Length	Target Copy Number	Hairpin Copy Number
0	Low (pCYC1)	Short (100bp)	Low (plasmid)	High (2μm)
1	High (pTDH3)	Short (100bp)	Low (plasmid)	High (2μm)
2	High (pTDH3)	Long (200bp)	Low (plasmid)	High (2μm)
3	High (pTDH3)	Long (200bp)	Single (genome)	High (2μm)
4	High (pTDH3)	Long (200bp)	Single (genome)	Low (Tryptophan)
5	High (pTDH3)	Long (200bp)	Single (genome)	Low (G418)

Table 8-1: Description of the Design Cycles used in the optimization of RNAi in yeast

8.2.1 Increased Hairpin Expression Level Improves RNAi Efficiency

First, we investigated the effect of hairpin RNA expression level on gene knockdown efficiency. We hypothesized that increasing the expression level of the

hairpin would increase the amount of dsRNA substrate available for Argonaute and Dicer, thus increasing the magnitude of downregulation. Heterologous expression of YFP was driven by either a weak (pCYC1) or strong (pTDH3) promoter. When the hairpin was expressed from a weak (pCYC1) promoter, we obtained insignificant downregulation of YFP fluorescence regardless of reporter level (**Figure 8-2**). However, we observed that increased expression of the hairpin (from a strong pTDH3 promoter, Design Cycle 1) resulted in increased knockdown capacity. Specifically, we found a 2.3-fold increase in downregulation when YFP is weakly expressed and upwards of 3-fold increase in the extent of downregulation when YFP is strongly expressed (**Figure 8-2**). In total, this construct enabled up to 80% downregulation to be obtained. We additionally confirmed that the RNAi system had an insignificant effect upon growth rate (**Figure 8-3**). These results confirm both that RNA interference is functional in yeast and also highlight that the absolute extent of downregulation may be altered by synthetically controlling the expression of the hairpin RNA. This approach represents a significant reduction in labor compared to current genomic manipulation techniques(242) and enables metabolic engineers to quickly test the effects of multiple expression levels on a phenotype of interest. This technique also enables the capacity to simultaneously alter the extent and timing of gene downregulation by coupling the expression of the hairpin RNA to an inducible promoter(243) or a logic circuit (244). For the remainder of this work, we optimized the synthetic RNAi system in the context of high hairpin expression, as this condition resulted in the strongest knockdown level.

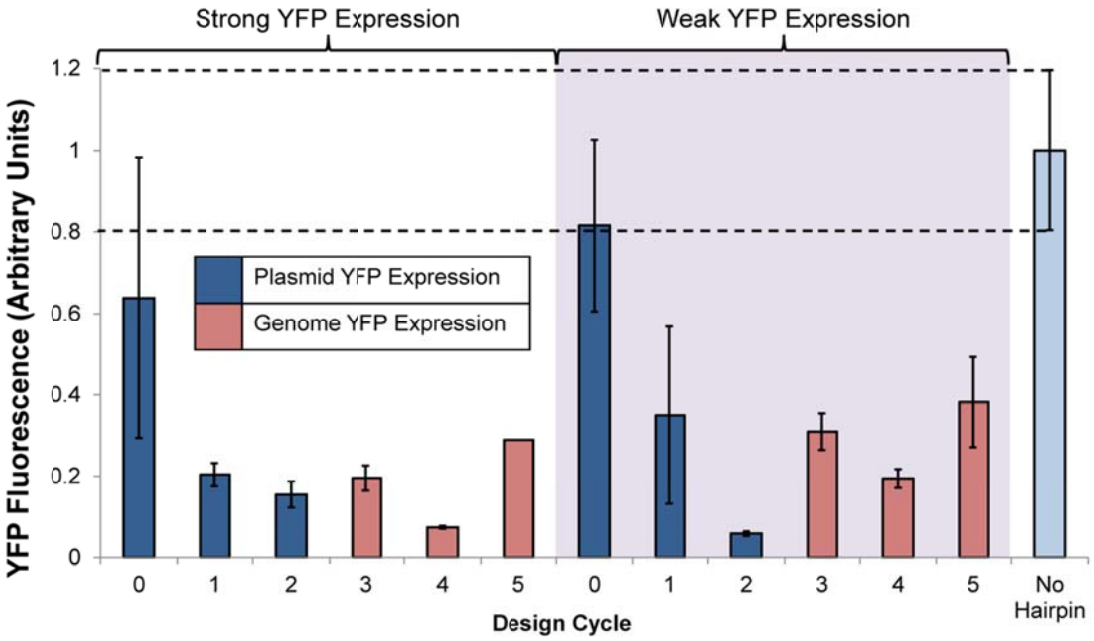


Figure 8-2: Gene knockdowns attained by each design cycle

YFP expression was downregulated through expression of Argonaute, Dicer, and a YFP-specific hairpin using the schemes listed in **Table 8-1** in order to elucidate design rules for RNAi in yeast. Red bars indicate the downregulation of plasmid-borne YFP and blue bars indicate the downregulation of YFP expressed from the genome. For each condition, the knockdown was normalized to its corresponding "no hairpin" control. Bars with a white background indicate downregulation of strongly expressed (pTDH3) YFP and a purple background refers to the downregulation of weakly expressed (pCYC1) YFP. Dashed lines denote the representative range of YFP expression levels observed in cells which do not express a hairpin. Error bars represent the standard deviation observed among three biological replicates. Through iteratively improving upon our synthetic RNAi pathway, expression of genomically-encoded proteins was downregulated by up to 93%.

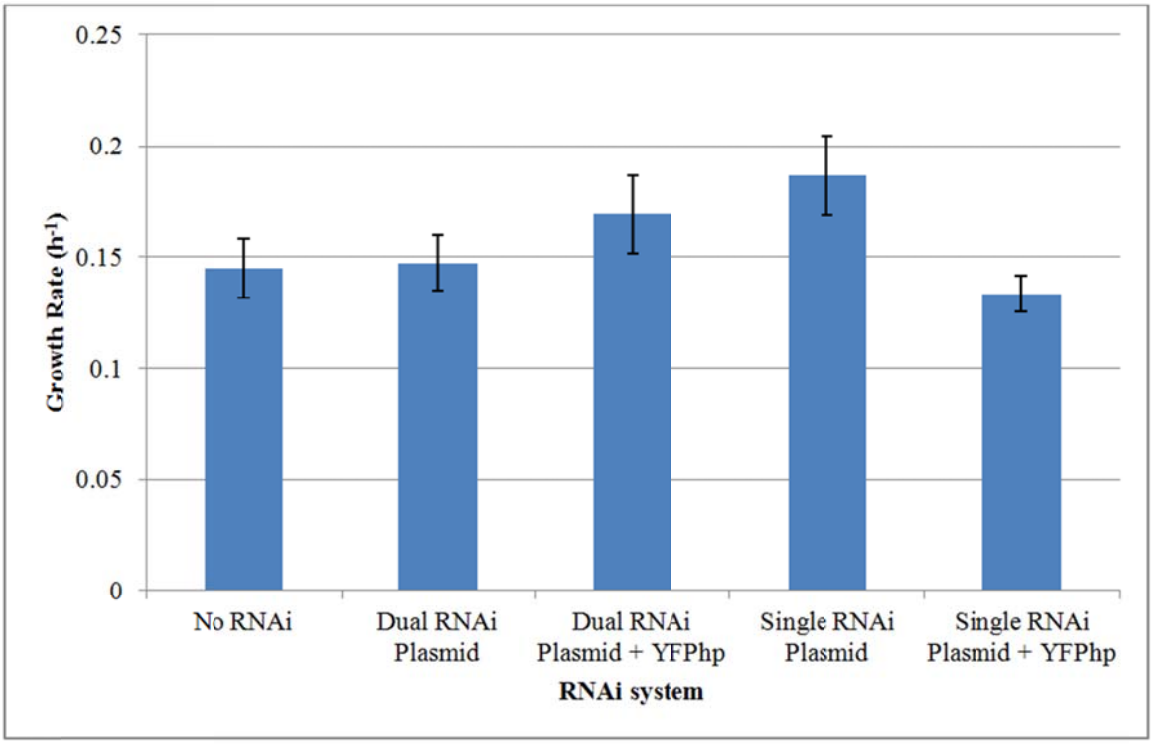


Figure 8-3: Growth Rate of Yeast Expressing the RNAi system.

No significant relationships were found between the strains as measured by a 1-tailed Student's t-test.

8.2.2 Increased Hairpin Length Improves RNAi Efficiency

To further improve downregulation, we next investigated the influence of hairpin RNA length. In endogenous RNAi systems, small interfering RNA efficiency is highly sequence-dependent, with some hairpins resulting in nearly complete reduction in expression whereas others have little effect (245). This disparate impact of RNAi is thought to result from stable secondary structures of the target mRNA occluding recognition and degradation by Argonaute (246). While long double-stranded RNAs are known to induce the interferon response in mammalian cells (247), no such defense mechanism exists in yeast. As a result, we hypothesized that increased hairpin length would improve downregulation efficiency by providing a greater diversity of guide

RNAs and thus a greater probability that Argonaute will recognize and cleave an unstructured part of the corresponding mRNA substrate (248). To test this hypothesis, we increased hairpin length from 100 bp to 200 bp (Design Cycle 2) and observed an improvement in downregulation efficiency by 30% when YFP is strongly expressed, and by nearly 6-fold when YFP is weakly expressed (**Figure 8-2**). Through this second design round, we were able to obtain inhibition levels of up to 94%, a significant improvement upon the 80% described above. It should be noted that the construction and propagation of inverted repeats of this increased length in *E. coli* were difficult, potentially due to interference with DNA replication machinery (249), thus necessitating the use of an intron-containing spacer region to ensure plasmid stability (250).

8.2.3 Decreasing Hairpin-Containing Plasmid Copy Number Improves RNAi Efficiency

The ability of RNAi to confer a useful phenotype is dependent upon the cell-to-cell variability in the extent of downregulation. Interestingly, flow cytometry analysis revealed a bimodal distribution of downregulation, with some cells almost completely downregulating YFP expression, and others exhibiting little downregulation (**Figure 8-4A**). To investigate the cause of this phenomenon, we explored the effectiveness of expressing the hairpin RNA on a centromeric (low-copy) plasmid containing either an auxotrophic (*TRP1*) or an antibiotic resistance marker (*KanMX*). For this design cycle (which was performed in the context of genomic YFP expression), we observed that a low-copy auxotrophic vector (Design Cycle 4) enabled up to 93% downregulation in the fluorescence of strongly-expressed YFP and 80% downregulation of weakly-expressed YFP, an improvement of 2.6-fold and 1.6-fold, respectively, over Design Cycle 3. Furthermore, the population of weakly downregulated cells was significantly reduced from previous design cycles (**Figure 8-4A**). While no improvements to efficiency were

seen when using a vector expressing antibiotic resistance, even when exposed to a saturating, 10-fold excess of antibiotic (Design Cycle 5), these results highlight the potential to use such a construct in heterotrophic strains (**Figure 8-2**). Interestingly, the extent of knockdown did not correlate the coefficient of variation in expression levels enabled by these plasmid constructs (**Figure 8-4B**), indicating that a mechanism other than copy number control is responsible for the improved knockdown observed when using low copy auxotrophic vectors. Nevertheless, these promising results inspired us to use this low-copy vector for hairpin expression in future experiments. Collectively, these design cycles were able to develop a synthetic RNAi system in yeast capable of efficient gene knockdown for metabolic engineering applications.

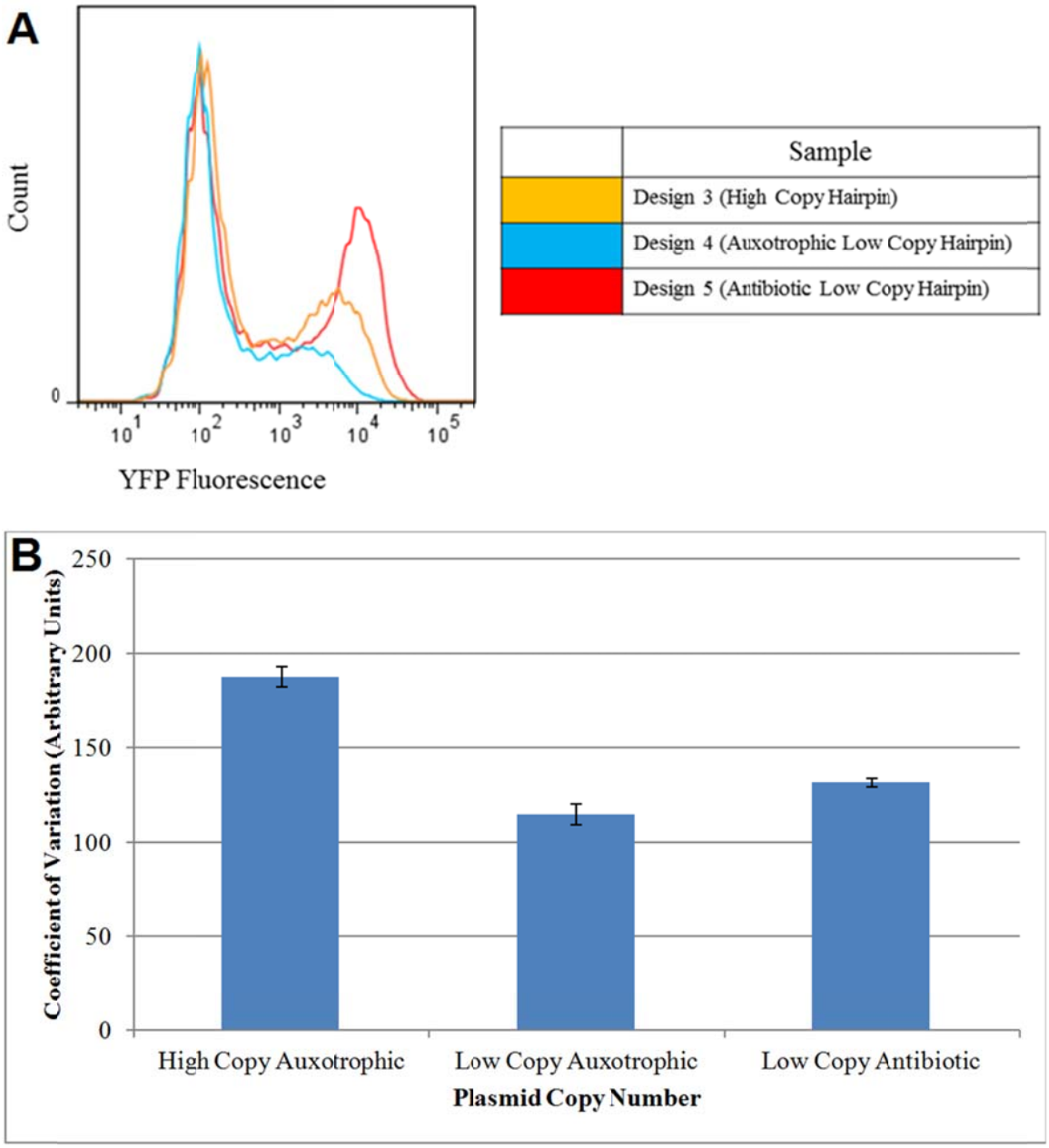


Figure 8-4: Cell-to-cell variation in strains expressing the RNAi system.

(A) Distribution of Knockdown Level in Strains Expressing Hairpins from High and Low Copy Plasmids.
(B) Variance in the Copy Number of High and Low Copy Plasmids

8.2.4 Implementation of RNAi in Alternate Yeast Strains

To demonstrate the generality and portability of this approach, we wished to implement RNAi in two additional commonly-used strains: CEN.PK and Sigma. We expected that the portability of this system would enable rapid prototyping in multiple strains simultaneously. In order to conserve auxotrophic markers in our system and further decrease expression noise, we condensed vector design by co-expressing Argonaute and Dicer from the same low-copy plasmid. In addition, we re-designed our YFP-specific hairpin to target regions of YFP mRNA with increased variability in secondary structure, as it has been indicated that these regions are ideal targets for RNAi (246). As a result, the length of the hairpin was increased to 240 bp. Finally, since genome modification techniques are rather inefficient for Sigma, we were unable to generate YFP-integrated versions of this strain and so tested downregulation of plasmid-borne YFP in all three strains. Across these strains, we achieved between 85% and 77% downregulation of YFP fluorescence, and between 90% and 97% downregulation of YFP mRNA (**Figure 8-5**). CEN.PK showed the highest overall downregulation competency, whereas Sigma showed the lowest. These strain-specific differences could be due to variations in the translation efficiencies of Argonaute and Dicer. In addition, although decreases in YFP fluorescence were well-correlated with decreases in YFP mRNA, knockdowns in fluorescence intensity were consistently lower than for mRNA levels. Regardless of these slight differences, these results demonstrate that our synthetic RNA interference is portable and efficient in a wide variety of strains, thus enabling reduction-of-function experiments and rapid prototyping to be easily performed in many strain backgrounds at a small marginal cost.

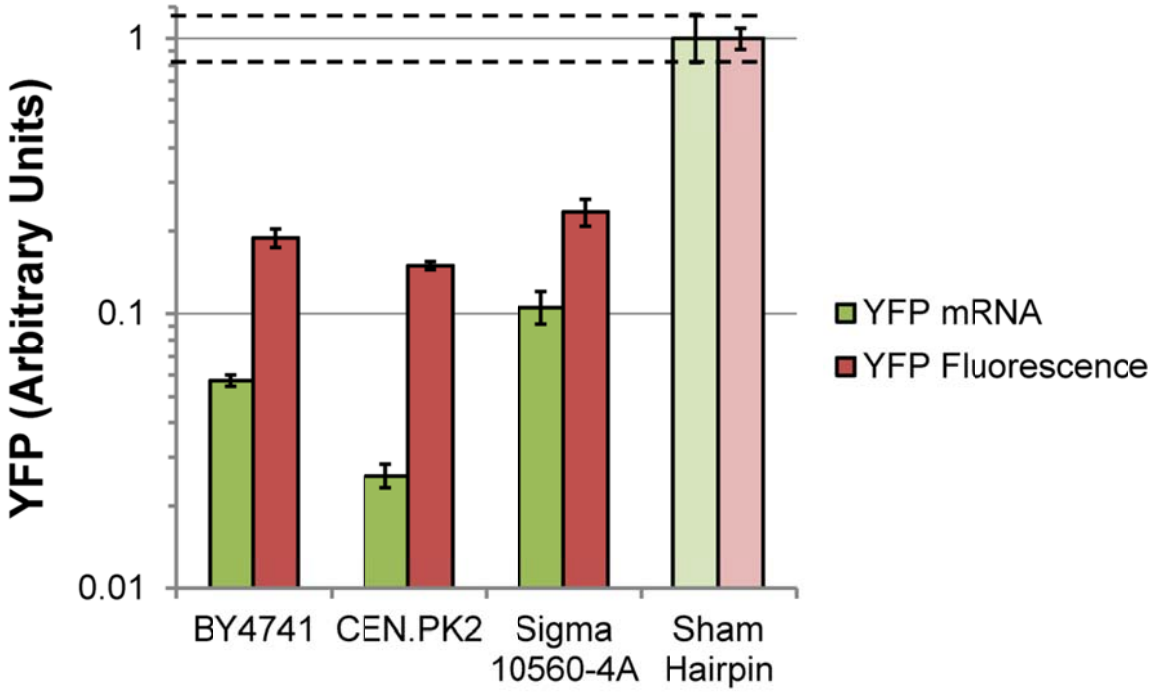


Figure 8-5: Gene knockdown in alternate strains of yeast

YFP expression was downregulated in BY4741, CEN.PK2-a, and Sigma 10560-4A using our synthetic RNA interference pathway. Red bars indicate YFP mRNA levels and blue bars indicate YFP fluorescence. Error bars represent the standard deviation observed among three biological replicates. Dashed lines represent the range of transcript and expression levels characteristic of strains expressing a sham (*ADE3*-specific) hairpin. We observed strong downregulation of YFP fluorescence and mRNA in each strain, indicating that RNAi is suitable for rapid prototyping in multiple genomic contexts.

8.2.5 Rapid Prototyping of Itaconic Acid Production in Yeasts through RNA Interference

As a final demonstration for the synthetic RNAi system in yeast, we sought to enable rapid prototyping to identify promising routes for metabolic engineering. Specifically, we undertook a combinatorial experiment whereby knockdown of a desirable gene target was simultaneously combined with various base strains. Genome scale metabolic modeling has determined that *ade3* deletions in yeast can improve the heterologous production of itaconic acid (IA) from cis-aconitate during cis-aconitate decarboxylase (CAD1) overexpression (251). In order to rapidly determine the effect of

this gene knockdown in other genomic contexts, we expressed a long hairpin specific to *ADE3* under the control of three yeast promoters (pCYC1, pTEF1, and pTDH3) which collectively span a wide range of expression. We also streamlined the RNAi system by integrating Argonaute and Dicer on the same *LEU2*-marked low-copy plasmid. This change was made because the *HIS3*-marked plasmid previously used for expression of Dicer was unavailable due to the possibility that the *ADE3* knockdown would confer histidine auxotrophy. The hairpin was maintained on a separate plasmid for modularity and ease of cloning. Next, itaconic acid production was measured upon co-expression of Dicer, Argonaute, and *CAD1* in three separate strains of yeast: BY4741, CEN.PK2-a and Sigma 10560-4A. We observed significant increases in IA production for at least one expression level of hairpin RNA in each of the three strains we tested, as indicated by a Student's t-test (**Figure 8-6**). As a result, these experiments indicate that a gene knockdown is an adequate, quick surrogate test for genotype-phenotype linkages. Expression of a sham hairpin specific for YFP did not elicit significant improvements to IA production, indicating that the observed improvements to IA titer were not simply due to the presence of dsRNA in the cell. These results also indicate that of the tested strains, *S. cerevisiae* Sigma 10560-4A is the most advantageous for IA production, and that *ade*3 knockout is a promising strategy for improvement of titer in this strain.

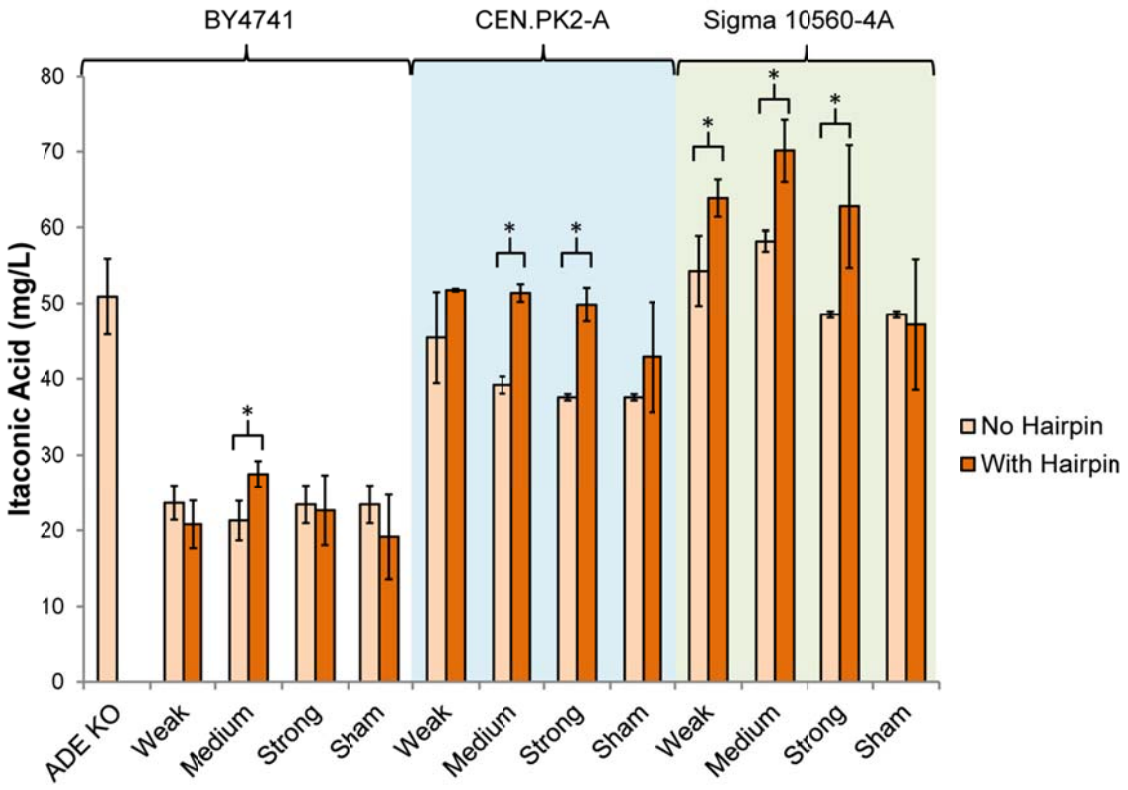


Figure 8-6: Rapid Prototyping of gene knockdowns conferring increased itaconic acid (IA) production in multiple yeast strains.

Argonaute, Dicer, cis-aconitate decarboxylase, and an ADE3-specific hairpin were co-expressed in BY4741, CEN.PK2-a and Sigma 10560-4A to rapidly identify promising engineering targets for IA production. Blue bars indicate IA production in strains without hairpin expression, and red bars indicate IA production in strains expressing a hairpin. Error bars represent the standard deviation observed among three biological replicates. Asterisks represent a statistically significant ($p<0.05$) difference in IA production as calculated by a one-tailed Student's t-test. In this experiment, sham hairpins were specific to YFP. Significant increases to IA production were observed for at least one hairpin expression level in each strain we tested, indicating the potential of the ade3 gene knockout, identifying Sigma 10560-4A as the most promising base strain for IA production, and confirming that RNAi enables rapid prototyping of engineering strategies in yeast.

We further investigated the strain-to-strain variation in the downregulation of *ade3* in each of these strains. By measuring *ADE3* mRNA levels, we found that CEN.PK2-a downregulated *ADE3* mRNA levels to the greatest extent, followed by Sigma 10560-4A and then BY4741 (**Figure 8-7**), indicating that the relatively low

increases to IA production observed in BY4741 may be due to a low *ADE3* downregulation efficiency. However, it is also expected that the relationship between *ade3* knockdown and IA induction may be different for each of these three strains.

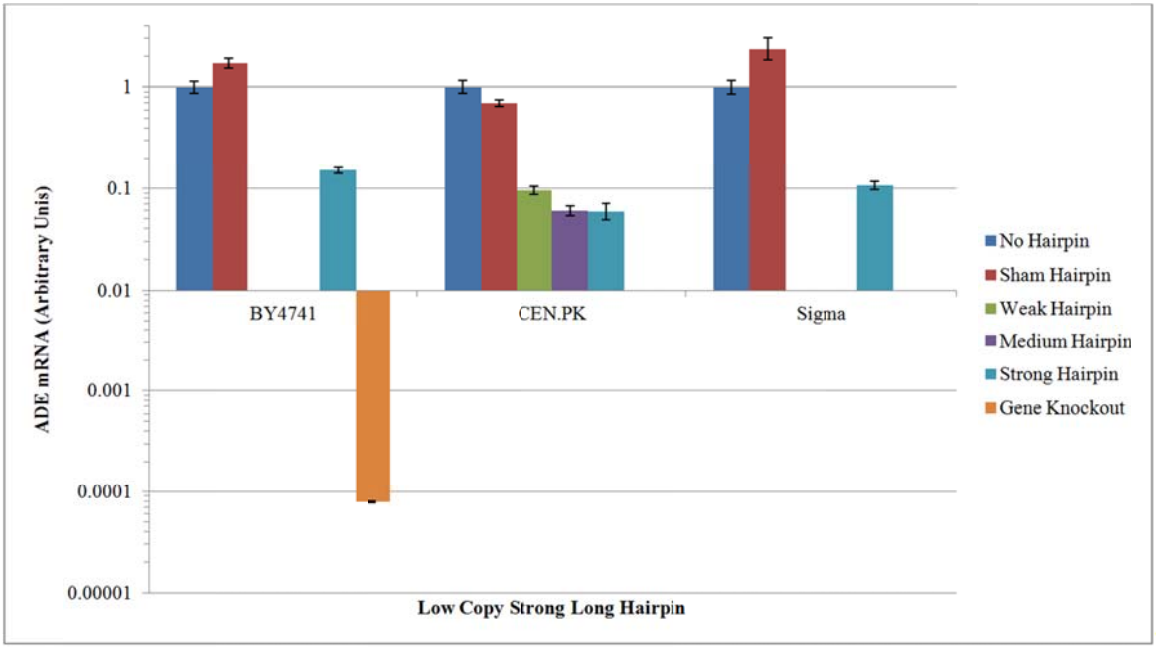


Figure 8-7: Downregulation of *ADE3* mRNA.

ADE3 mRNA was measured in selected itaconic acid-producing strains.

In this example, we demonstrated that rapid strain prototyping is possible in yeast through the use of a synthetic RNAi system. Unlike genomic knockouts, hairpins are generated in a facile manner, thus enabling RNA interference to be a potent tool for the rapid screening of knockdown strategies in multiple organisms. Moreover, for more difficult to use strains (such as Sigma), gene expression tools are more mature than genome editing tools. As a result, RNAi systems enable a wide range of expression control with more ease than genome modifications. These experiments also suggest that Sigma 10560-4A would be the best strain for itaconic acid production out of the yeasts

we tested. In this regard, this work demonstrates the potential of RNAi to significantly expedite the design-build-test cycle.

8.2.6 Characterization of RNAi in yeast using unstructured RNA

We have also developed a scheme for efficient gene knockdown which avoids the use of hairpin constructs, because generation of hairpin constructs on a library scale is quite difficult. On the other hand, a scheme which would enable significant levels of gene knockdown from non-inverted-repeat constructs would enable existing techniques for cDNA or gDNA library generation to be used for RNAi in yeast. Our initial design cycle (**Table 8-2**, Design Cycle 0) consisted of a dual promoter construct, in which one promoter (pTDH3) drives the expression of the sense strand of a 400bp YFP fragment and another promoter (pTEF1) drives the expression of the antisense strand. This dual promoter construct was expressed on a high-copy vector. In addition, RNAi machinery (argonaute and dicer) was expressed on separate low-copy vectors driven by strong promoters. In order to determine the effects of target gene expression, YFP was driven by either a strong or a weak promoter on a low-copy plasmid. In this scheme, we achieved 50% downregulation of strongly-expressed YFP and insignificant downregulation of weakly-expressed YFP (**Figure 8-8)**.

Design Cycle	Hairpin Expression	Intron	Target Copy Number	Hairpin Copy Number
0	High (pTDH3)	None	Low (plasmid)	High (2µm)
1	High (pTDH3)	None	Single (genome)	High (2µm)
2	High (pTDH3)	Rad9	Single (genome)	High (2µm)
3 (GPD)	High (pTDH3)	Rad9	Single (genome)	High (2µm)
3 (CYC)	Low (pCYC1)	Rad9	Single (genome)	Low (Tryptophan)

Table 8-2: Description of the Design Cycles used in the Optimization of Unstructured RNAi in Yeast

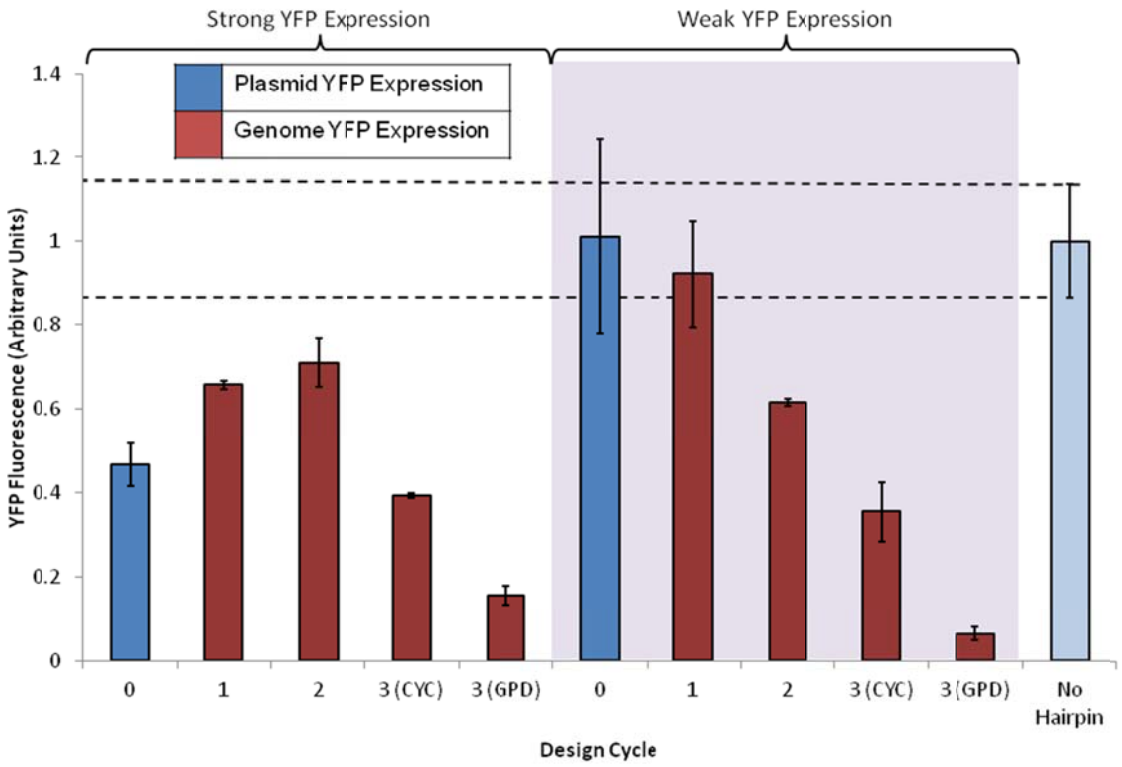


Figure 8-8: Gene knockdowns attained by each design cycle for optimization of unstructured RNAi

YFP expression was downregulated through expression of Argonaute, Dicer, and a YFP-specific hairpin using the schemes listed in **Table 8-2** in order to elucidate design rules for RNAi in yeast. Red bars indicate the downregulation of plasmid-borne YFP and blue bars indicate the downregulation of YFP expressed from the genome. For each condition, the knockdown was normalized to its corresponding "no hairpin" control. Bars with a white background indicate downregulation of strongly expressed (pTDH3) YFP and a purple background refers to the downregulation of weakly expressed (pCYC1) YFP. Dashed lines denote the representative range of YFP expression levels observed in cells which do not express a hairpin. Error bars represent the standard deviation observed among three biological replicates. Through iteratively improving upon our synthetic RNAi pathway, expression of genomically-encoded proteins was downregulated by up to 93%.

We next wished to investigate the ability of our dual-promoter construct to downregulate chromosomal gene expression. Therefore, we integrated YFP under the control of a strong or a weak promoter into the yeast genome. Under this scheme (Design Cycle 1), downregulation of strong YFP expression was decreased to 30%, whereas we still observed insignificant downregulation of weakly-expressed YFP.

In order to further increase the extent of downregulation, we integrated introns into our dual-promoter construct, as it has been shown that transcripts which are subject to RNA splicing are more effective at downregulation. These introns were placed immediately downstream of each promoter in the downregulation cassette. Using this system (Design Cycle 2), downregulation of strongly-expressed YFP remained similar to the extent of downregulation without introns, whereas downregulation of weakly-expressed YFP increased to 40%.

We also wished to express our downregulation cassette on a low-copy plasmid, as we have found that this approach significantly increases the downregulation efficacy of hairpin constructs in previous work. By expressing our downregulation cassette in this way (Design Cycle 3), we observed significantly increased downregulation: up to 85% for strongly expressed genes and 94% for weakly expressed genes. We also found that exchanging pTDH3 for pCYC1 in our downregulation cassette enabled tunable downregulation, such that strongly expressed genes were downregulated by 60% and weakly expressed genes by 65%. Taken as a whole, these results show that non-hairpin-

based downregulation cassettes can achieve significant levels of downregulation, and that the extent of this downregulation may be tuned through a simple cloning step. In addition, these results indicate that cDNA is a promising substrate to guide RNAi to downregulate transcription, enabling genome-wide searches for knockdown candidates in order to rapidly generate improved strains for a particular application.

8.2.7 Improving Isobutanol, 1-Butanol, and Lactic Acid Tolerance through a Genome-Wide Knockdown Search

Because we showed that RNAi can efficiently downregulate genomically-encoded genes using a linear guide RNAs, we wished to expand the power of RNAi from an approach for targeted strain engineering towards a method for knockout identification on a genome-wide scale. Using this method, it would be possible to transform a genomic library of knockdown cassettes and screen for beneficial phenotypes, thus identifying candidates for strain modification in a high-throughput manner. Importantly, this approach could elucidate knockdown targets for strains which are not sequenced or do not have a curated metabolic model. In addition, knockdown targets could be identified for the improvement of complex phenotypes which cannot be modeled using current approaches. In order to investigate the ability of RNAi to identify strain engineering targets on a genome-wide scale, we used this approach to improve the tolerance of yeast to 1-butanol, isobutanol, and lactic acid.

Although bioethanol is the most commonly produced liquid fuel from biomass, ethanol suffers from several issues limiting its widespread use, including high hygroscopicity and low octane rating. Butanol, on the other hand, has a higher octane rating, does not readily absorb water, and can serve as a drop-in substitute for gasoline. However, the high toxicity of butanol limits its production in a microbial setting. It has been observed in our lab that butanol concentrations of greater than 10 g/L severely limit

the growth of *S. cerevisiae*, indicating that substantial strain engineering is necessary to make biological production of butanol feasible in this organism. Furthermore, lactic acid is commonly used in fermentation processes to inhibit bacterial contamination, and so development of a method to quickly confer lactic acid tolerance to yeast would be highly desirable. It has been observed in our lab that lactic acid concentrations of greater than 9 g/L prohibit the growth of *S. cerevisiae.* Because alcohol and acid tolerance are complex phenotypes and thus are difficult to improve using rational approaches, genome-wide knockdown searches (such as that afforded by RNAi) are ideal to ensure a high likelihood of success.

We thus demonstrated the utility of RNAi knockdown libraries in improving chemical tolerance through the use of the optimized antisense RNAi substrates developed above. A cDNA library of the parent strain (*S. cerevisiae* BY4741) was generated through established procedures (252), sheared and cloned into vectors containing converging promoters to generate a library of antisense RNAi constructs. This procedure thus generated 4 libraries: either strong (RNAi cassette driven by pGPD) or weak (RNAi cassette driven by pCYC1) downregulation using 200bp or 400bp transcript fragments. Each of these libraries contained over 10^5 distinct members, thus enabling good coverage of the yeast transcriptome. These antisense constructs were then transformed into the parent strain along with the RNAi machinery and selection for high butanol or lactic acid tolerance was undertaken through serial subculture in inhibitory concentrations of these compounds, as shown in **Figure 8-9**. These screening libraries were then allowed to grow until they reached an optical density of greater than 1, at which point they were subcultured at a 1:100 ratio to a fresh culture with an increased concentration of the inhibitory compound. For lactic acid and isobutanol, a second screening was undertaken which started at a lower concentration of the compound of interest. A sample of cells

was also taken at this time in order to extract an enriched collection of downregulation cassettes. This process was repeated a total of 3 times in order to generate 4 enriched collections of downregulation cassettes for each library/compound combination. Finally, isolated members of these enriched collections were sequenced at random in order to identify the knockdown cassettes responsible for improved tolerance. These knockdown cassettes were then retransformed into their parent strain to confirm the causal nature of phenotype improvement. Those knockdown cassettes which are promising to improve the growth rate of BY4741 in inhibitory concentrations of 1-butanol and isobutanol are shown in **Table 8-3** and **Appendix Table A7-10**. Of these targets, the *ADH1* cassettes showed the greatest ability to improve the growth of yeast on 1-butanol. Interestingly, this gene catalyzes the conversion of acetaldehyde to ethanol, and is often knocked out of butanol-producing strains to improve product yield (253). Cassettes encoding fragments of *RPL28*, *SOD1*, and *SSB1* showed the greatest ability to improve the growth of yeast on isobutanol. It is interesting to note that both *RPL28* and *SSB1* are associated with translation (254,255), indicating that their deletion may impact the expression of a large number of genes. *SOD1* is an interesting target that was identified during both the 1-butanol and isobutanol selections, and is involved with the detoxification of reactive oxygen species (256). The downregulation of this gene is therefore counterintuitive, but it may indicate an inappropriate cellular response to butanol toxicity. Current work is focused on investigating the effect of total gene knockout on tolerance for these promising targets identified through a genome-wide knockdown approach.

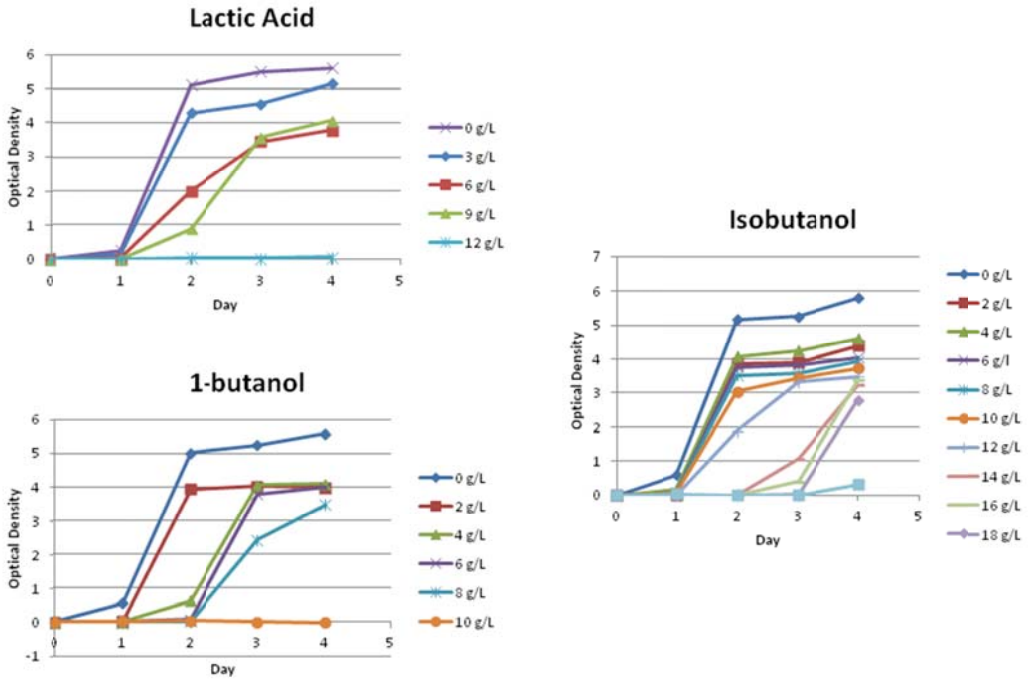


Figure 8-9: Growth rate of BY4741 in lactic acid, 1-butanol, and isobutanol

Yeast transformed with the RNAi machinery along with a blank downregulation cassette was cultured in varying concentrations of lactic acid, 1-butanol, and isobutanol in order to determine the appropriate selective condition for screening.

Compound/Concentration (g/L)	Cassette Expression Level	Genomic Target	Name
1-butanol (8 g/L)	Low (pCYC1)	*SOD1*	1B-5-1
1-butanol (8 g/L)	Low (pCYC1)	*RPL28*	1B-6-1
1-butanol (8 g/L)	Low (pCYC1)	*ADH1*	1B-8-1
1-butanol (8 g/L)	Low (pCYC1)	*SSB1*	1B-12-1
1-butanol (8 g/L)	Low (pCYC1)	*ADH1*	1B-13-1
1-butanol (8 g/L)	Low (pCYC1)	*RPL15A*	1B-14-1
1-butanol (8 g/L)	Low (pCYC1)	unknown	1B-17-1
1-butanol (8 g/L)	Low (pCYC1)	Ty1 gag-pol	1B-19-1
1-butanol (8 g/L)	Low (pCYC1)	unknown	1B-21-1
1-butanol (8 g/L)	Low (pCYC1)	YDR524C-B	1B-22-1
1-butanol (8 g/L)	Low (pCYC1)	*MHF1* or *ADH1*	1B-23-1
1-butanol (8 g/L)	High (pGPD)	Ty1 gag-pol	1B-24-1
1-butanol (8 g/L)	High (pGPD)	*ADE5,7*	1B-28-1
1-butanol (8 g/L)	High (pGPD)	unknown	1B-29-1
1-butanol (8 g/L)	High (pGPD)	*TPI1*	1B-30-1
1-butanol (8 g/L)	High (pGPD)	*ADH1*	1B-32-1
1-butanol (8 g/L)	High (pGPD)	*RP6B*	1B-33-1
Isobutanol (16g/L)	Low (pCYC1)	*RPL28*	IB-1-1
Isobutanol (16g/L)	Low (pCYC1)	unknown	IB-6-1
Isobutanol (16g/L)	Low (pCYC1)	*GRX3*	IB-7-1
Isobutanol (16g/L)	Low (pCYC1)	*GCV3*	IB-9-1
Isobutanol (16g/L)	High (pGPD)	*SCY1*	IB-12-1
Isobutanol (16g/L)	Low (pCYC1)	*SED1+TPO1*	IB-16-1
Isobutanol (16g/L)	Low (pCYC1)	*RPL36B*	IB-19-1
Isobutanol (16g/L)	Low (pCYC1)	unknown	IB-20-1
Isobutanol (16g/L)	Low (pCYC1)	*CCW12*	IB-21-1
Isobutanol (16g/L)	Low (pCYC1)	*RPS18A*	IB-24-1
Isobutanol (16g/L)	Low (pCYC1)	Ribosome or *TAR1*	IB-25-1
Isobutanol (16g/L)	Low (pCYC1)	*SOD1*	IB-26-1
Isobutanol (16g/L)	High (pGPD)	*RPL26B*	IB-28-1
Isobutanol (16g/L)	High (pGPD)	*SSB1/2*	IB-32-1
Isobutanol (16g/L)	High (pGPD)	*SSB1/2*	IB-33-1
Isobutanol (16g/L)	High (pGPD)	*RPL41B*	IB-35-1

Table 8-3: Knockdown cassettes identified for improving 1-butanol and isobutanol tolerance

8.3 DISCUSSION

In this work, we have demonstrated that RNA interference is an effective tool for expediting the design-build-test cycle and enabling rapid prototyping of engineered yeast strains. We have uncovered several important design principles influencing knockdown level, and have used an optimized scheme to demonstrate the effectiveness of RNAi through testing a putative genetic target for improved itaconic acid production. We have additionally used this approach to enable the creation of genome-wide knockdown libraries and have applied this approach to successfully identify knockdown targets for the improvement of 1-butanol and isobutanol tolerance. The portable nature of this approach (only requiring heterologous expression of Argonaute and Dicer) can enable rapid prototyping of both previously engineered and unsequenced industrial strains (esp. where polyploidy may be a substantial hurdle to genome engineering). Due to the linkage between downregulation capacity and hairpin RNA expression, it is possible to develop more advanced control of this system through the use of inducible promoters (243), sophisticated logic circuits (244), or oscillators (257). Finally, this work has the potential to be multiplexed (i.e. co-expressing many hairpin cassettes simultaneously) to investigate the impact of multiple gene knockdowns or streamlined by integrating all components necessary for RNAi (Dicer, Argonaute, and the hairpin) on the same vector. It is important to note that this system may be employed for rapid strain engineering in organisms which have not been sequenced or annotated and therefore may be highly beneficial for the rapid improvement of unsequenced industrial strains or environmental isolates. Thus, this work opens the door for metabolic engineering in yeast using RNA interference, which enables wider exploration of knockout targets, more finely tuned control of knockdown level, and greater flexibility in strain evaluation, resulting in an expedited design-build-test cycle.

Chapter 9: Conclusions and Future work

Taken together, the engineering strategies developed in this work enable a powerful approach to strain development, in which every control point for strain productivity has associated methods for either predictive design or comprehensive high-throughput perturbation methods. At the transcriptional level, we have shown that nucleosome occupancy is an important limiting factor to the activity of both native and synthetic promoters in yeast. We have developed a method that, for the first time, allows researchers to increase native promoter strength in a single design-build-test cycle through computationally informed changes to promoter sequence. Furthermore, this method enables the creation of fully synthetic yeast promoters (bearing no homology to any native sequence) which enable expression levels on par with the top 6% of highly expressed genes in yeast. Not only does this method enable the creation of stronger (or weaker) promoters, but it also provides an intriguing platform for dissecting and optimizing synthetic promoter architecture in yeast. By assembling transcription factor binding sites in a context unconfounded by nucleosome occupancy, researchers will gain clearer insight into the effects of transcriptional machinery position and orientation on promoter strength, thus uncovering the design rules behind native and synthetic yeast promoters.

At the translational level, we have shown the significant impact of mRNA secondary structure on gene expression. We showed that this effect is especially pronounced for codon-optimized genes, and that common sequences used for DNA assembly may be particularly inhibitory. It is important to note that this phenomenon is not restricted to the context of multicloning sites upon which this study was based. Rather, any sequence appearing in the 5'UTR of a transcript, synthetic or otherwise, has

the potential to inhibit gene expression. With this in mind, applications sensitive to gene expression levels (e.g. chemical production, biological part characterization, synthetic control of gene regulation, etc) must consider the effect of RNA context in the design of genetic constructs. Further application of the techniques generated in this work may be useful for the design of optimized 5'UTRs for the creation of synthetic promoters.

To accelerate the pace of protein engineering in yeast, we have developed a tool which enables mutagenesis and selection of large protein libraries in a continuous manner *in vivo*. This powerful approach takes advantage of the mutagenic capabilities of reverse transcriptases which, coupled with the replicative machinery encoded in the Ty1 reverse transcriptase, enables the generation of mutant libraries of similar magnitude to those obtained using current state-of-the-art methods using significantly less effort. We have shown that this approach is suitable to improve the tolerance of yeast to high concentrations of ethanol during osmotic stress through the mutagenesis of the global transcriptional regulator *SPT15*. We expect to improve the success rate and reliability of ICE by optimizing our synthetic retroelement to function in a genomic context by improving transposition rates through optimizing culture conditions and expression of specific endonucleases. In addition, we aim to improve the mutation rate of the reverse transcriptase through a combination of saturation mutagenesis, random mutagenesis, or expression of heterologous reverse transcriptases followed by screening for high mutation rates using several robust methods. ICE, in its present form, enables the creation of libraries of the same size as which can be obtained through current methods, yet requires much less effort. It is expected that these strategies mentioned above will in the near term enable ICE to far surpass state-of-the-art directed evolution techniques both in terms of success rate and labor intensity. In addition to enabling the continuous directed evolution of expression cassettes and pathways, ICE may in the future be used in

conjunction with high-throughput sequencing to study evolutionary processes for a broad range of enzyme classes. Excitingly, this synergy between techniques opens the door for the study of sequence-function relationships with much higher throughput than can be obtained with existing computational or experimental techniques. Taken together, this work lays the foundation for a new platform technique in the engineering and study of yeast proteins.

We have also developed a technique for rapid strain prototyping which makes use of RNA interference to tune gene expression in a facile manner. We have elucidated the design rules for the construction of tunable downregulation cassettes, and we have shown that this approach may be implemented in a variety of yeast strains to controllably reduce gene expression without the need for time-consuming genomic modifications. This approach was used for the rapid prototyping of the *ade3* gene knockout for the improvement of itaconic acid production in three strains of yeast. It was found that knockdown of *ADE3* is beneficial to itaconic acid production in multiple strain contexts, thus demonstrating the utility of RNAi to screen potential strain modifications in a rapid manner before expending a significant amount of effort during genome editing. This approach was then expanded to enable the construction of genome-wide libraries of downregulation cassettes, and the design rules for the construction of these vectors were elucidated in a similar fashion. Then, this approach was used for the identification of knockdown targets which confer increased tolerance to 1-butanol, isobutanol, and lactic acid. Upon verification of these knockdown targets through genome editing, we plan to iterate this approach in the context of an improved strain background to identify further targets. In addition, by combining this approach with RNA-seq, genome-wide changes to gene expression enabled by our knockdown cassettes may be elucidated, thus uncovering

the genetic network related to our phenotype of interest and providing highly relevant knowledge for further strain engineering efforts.

In addition to the strategies which have been successfully developed in this work, we can also gain insight by considering the approaches which did not go as planned. During the development of weak promoters, we discovered that the level of background transcription in standard expression vectors was quite large, especially for the low levels of expression we were interested in. As a result, we became interested in developing a method that would enable accurate characterizations of promoter activity in a noise-free context. Surprisingly, further work showed that we were unable to eliminate the context dependence of promoter activity by putting a terminator in front of each promoter. It is possible that long-range context interactions, such as nucleosome occupancy, may have played a role in this result. Therefore, in order to realize our original goal of developing robust, well-characterized weak promoters which enable a consistent level of gene expression regardless of context, what is needed is a synthetic insulator. Such a part would enable gene circuits to behave reproducibly regardless of genetic context, and therefore would be highly useful in a broad range of applications to address growing concerns about the generality of complex synthetic constructs.

Although we were unable to generate an IRES in yeast, our work illustrated the challenge inherent to proving IRES functionality. In many cases, sequences which had been shown to exhibit IRES functionality in one context failed to show the same function in our screening vector, and instead showed promoter activity. This certainly indicates that IRES functionality is highly context dependent, and may indicate that certain commonly used screening vectors are prone to false positives. This conjecture has also been made by others (258). Nevertheless, this work showed that a yeast IRES, if any, will be located at a substantial sequence and structural distance away from the

Dicistroviridae IRESs and that any screen for IRES functionality must not generate a positive signal if the IRES functions as a promoter. Indeed, the ideal screen for future development IRES functionality would probably avoid the use of bicistronic reporter assays altogether and rather involve translation of an uncapped mRNA transfected directly into yeast, as this would better capture the innate features of an IRES and may be less prone to false positives. It may be promising to use this assay as an orthogonal measure of IRES activity for the wild-type viral and cellular IRESs characterized in this work as well as for hits HM3 and SM7.

In a broader sense, this work also illuminates some key characteristics of engineering biological systems. Firstly, context is very important. The DNA surrounding a synthetic construct can determine whether a design functions as expected, or whether it does not work at all. However, by understanding the mechanism by which DNA context can modulate the activity of surrounding DNA parts, context effects can themselves be exploited to generate a more highly functional strain. The multicloning site and nucleosome occupancy optimization studies both demonstrated that by designing biological parts with context in mind, great increases to construct functionality can be obtained. As a second point, viruses are systems which have been evolutionarily optimized for doing genetic engineering; therefore, these systems may be a fertile source to fill many deficiencies in the metabolic engineer's toolbox. This work has shown that IRES elements, 2A peptides, retrotransposons, and RNA interference are all extremely useful to the metabolic engineer. What is also remarkable is that each of these systems was originally developed by, or in response to, the action of viruses. Although the idea of using viral parts for genetic engineering is by no means new, it is encouraging to note that as we endeavor to engineer a wider variety of organisms, we will always be able to take lessons from those systems which figured out how to deliver genetic material to our

favorite organism millions of years before we did. Lastly, this work has emphasized that large populations of living systems are extremely clever. If it is possible for a mutant strain to pass a selective pressure through a loophole, it will show up during a screen, often to the exclusion of mutants which passed the selective pressure in the way it was intended. Although this phenomenon was an annoyance during the strain engineering conducted in this work, it copels us as the people who design living systems to respect the fact that the products we sell are prone to adaptation and selection throughout their life cycle. Dr. Frances Arnold famously quipped: "You get what you screen for" because, as Dr. Ian Malcolm warned us: "Life finds a way."

Chapter 10: Materials and Methods

10.1 General Methods

10.1.1 Strains and Media

Yeast expression vectors were propagated in *Escherichia coli* DH10β. *E. coli* strains were routinely cultivated in LB medium (259) (Teknova) at 37°C with 225 RPM orbital shaking. LB was supplemented with 100 µg/mL ampicillin (Sigma) when needed for plasmid maintenance and propagation. Yeast strains were cultivated on a yeast synthetic complete (YSC) medium containing 6.7 g of Yeast Nitrogen Base (Difco)/liter, 20 g glucose/liter and a mixture of appropriate nucleotides and amino acids (CSM, MP Biomedicals, Solon, OH). All medium was supplemented with 1.5% agar for solid media.

For *E. coli* transformations, 25 µL of electrocompetent *E. coli* DH10β(259) were mixed with 30 ng of ligated DNA and electroporated (2 mm Electrporation Cuvettes (Bioexpress) with Biorad Genepulser Xcell) at 2.5 kV. Transformants were rescued for one hour at 37 °C in 1 mL SOC Buffer (Cellgro) plated on LB agar and incubated overnight. Single clones were amplified in 5 mL LB medium and incubated overnight at 37 °C. Plasmids were isolated (QIAprep Spin Miniprep Kit, Qiagen) and confirmed by sequencing.

For yeast transformations, 50 µL of chemically competent S. cerevisiae BY4741 were transformed with 1 µg of each appropriate purified plasmid according to established protocols (242), plated on the appropriate medium, and incubated for three days at 30 °C. Single colonies were picked into 1mL of the appropriate medium and incubated at 30 °C.

10.1.2 Ligation Cloning Procedures

PCR reactions were performed with Q5 Hot-Start DNA Polymerase (NEB) according to manufacturer specifications. Digestions were performed according to manufacturer's (NEB) instructions, with digestions close to the end of a linearized strand running overnight and digestions of circular strands running for 1 hour at 37 °C. PCR products and digestions were cleaned with a QIAquick PCR Purification Kit (Qiagen). Phosphatase reactions were performed with Antarctic Phosphatase (NEB) according to manufacturer's instructions and heat-inactivated for 15 min at 65 °C. Ligations (T4 DNA Ligase, Fermentas) were performed for 6 h at 22 °C followed by heat inactivation at 65 °C for 15 min.

10.1.3 Flow Cytometry Analysis

Yeast colonies were picked in triplicate from glycerol stock, grown in the appropriate medium to mid-log phase, and analyzed (LSRFortessa Flow Cytometer, BD Biosciences. Excitation wavelength: 488 nm, Detection wavelength: 530 nm). Day-to-day variability was mitigated by analyzing all comparable transformants on the same day. An average fluorescence and standard deviation was calculated from the mean values for the biological replicates. Flow cytometry data was analyzed using FlowJo software.

10.2 METHODS FOR CHAPTER 2

10.2.1 Strains and media

Saccharomyces cerevisiae strains BY4741 (*MAT a; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0*) and BY4741 ΔP_{CYC1} (*MAT a; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0;* P_{CYC1}*::ura3*) were used in this study. BY4741 ΔP_{CYC1} was generated using the "delete and repeat" knockout method (242) with the *K. lactis URA3* gene from plasmid PUG72 as the selectable marker. Primers for the generation of the knockout cassette are in **Appendix**

Table A1-2. Integration of the *CYC1* promoter variants and *yECitrine* cassettes was completed by cloning the *K. lactis URA3* gene upstream of each *CYC1* promoter variant cassette (see below for plasmid construction) and then using the "delete and repeat" method to integrate both genes into the *TRP1* locus. See **Appendix Table A1-2** for primers.

10.2.2 Plasmid construction

All plasmids used in this study were based on the p413 vectors described previously (181). These plasmids contain the *HIS3* gene as the auxotrophic marker. The *TEF1* and *CYC1* promoters were available in the parent plasmid set. The *TEF1 mutant* series of promoters and the *yECitrine* and *LacZ* genes were cloned via PCR from plasmids described previously (26,27,37,242). The *HXT7* and *HIS5* promoters were cloned via PCR from extracted BY4741 gDNA obtained using the Wizard Genomic DNA Extraction Kit from Promega (Madison, WI). Re-designed and synthetic promoters were ordered as gBlock fragments from Integrated DNA Technologies, Inc. (Coralville, IA) and then cloned via PCR (see **Appendix Table A1-1** for promoter sequences and **Appendix Table A1-2** for all primer sequences).

10.2.3 Beta-galactosidase assay

Strains expressing the *LacZ* gene were evaluated for beta-galactosidase activity through the chemiluminescent Gal-Screen system (Applied Biosystems). Yeast cultures were grown for 16 hours to mid-log phase from a starting OD_{600}=0.005. Prior to the assay, cultures were diluted with fresh media to approximately OD_{600}=0.01 to 0.07. OD_{600} was measured, and then cultures were treated with Gal-Screen Reaction Buffer according to the manufacturer's instructions. Luminescence was quantified using a Mithras LB 940 luminometer (Berthold Technologies). Day to day variation was avoided

by measuring all samples on the same day. The average luminescence across biological replicates was calculated.

10.2.4 Quantitative PCR

To measure mRNA levels resulting from re-designed promoters, quantitative PCR was performed. Yeast cultures were grown for 16 hours to mid-log phase from a starting OD_{600}=0.005, and RNA was extracted using Zymolyase digestion of the yeast cell wall followed by the Quick-RNA MiniPrep kit according to manufacturer's instructions (Zymo Research Corp.). cDNA was generated from the purified RNA via the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). Primers for qPCR were designed using the PrimerQuest® tool and obtained from Integrated DNA Technologies (see **Appendix Table A1-2** for primers). Quantitative PCR was performed on a ViiA7 qPCR system (Life Technologies) using SYBR Green Master Mix from Roche (Penzberg, Germany), following the manufacturer's instructions with an annealing temperature of 58°C and 0.25 μL of cDNA product per 20 μL reaction. The *ALG9* gene was used as a housekeeping gene, and the relative *yECitrine* transcript level was obtained by calculating the average values between three technical replicates for each sample.

10.2.5 Nucleosome mapping

Nucleosome position and density was mapped in the *CYC1* and *CYC1v3* promoters. The BY4741 ΔP_{CYC1} strain was used for this part of the study in order to prevent contaminating genomic sequence from confounding the results. Plasmids p413-*CYC1-yECitrine* and p413-*CYC1v3-yECitrine* were independently transformed into the strain as described above. Mono-nucleosome sized genomic DNA fragments were then isolated from each strain using a method described previously (260). Briefly, 200 mL of culture was grown to approximately OD_{600}=0.8. Cells were treated with 1%

formaldehyde for 30 minutes at 30°C. The reaction was stopped by adding glycine to a final concentration of 125mM and cells were centrifuged at 3000g and washed twice in 20 mL of PBS. Cells were then resuspended in in 20 mL Zymolyase buffer (1 M sorbitol, 50 mM Tris pH 7.4, 10 mM 2-mercaptoethanol), then spheroplasted with 50 U Zymolyase (Zymo Research Corp.) for 40 min at 30°C. Cells were then washed once with 10 mL Zymolyase buffer and resuspended in 2 mL NP Buffer (1 M sorbitol, 50 mM NaCl, 10 mM Tris pH 7.4, 5 mM $MgCl_2$, 0.075% NP 40, 1 mM 2-mercaptoethanol, 500 μM spermidine). Aliquots of 500 μL were split between four tubes for each sample, and $CaCl_2$ was added to a final concentration of 3 mM. Micrococcal nuclease (New England Biolabs) digestions were performed at concentrations ranging from 100 to 600 U/mL for 10 min at 37°C. Reactions were stopped by adding 100 μL stop buffer (5% SDS, 500 mM EDTA). Proteinase K (New England Biolabs) was added to each tube at a final concentration of 100 mg/mL and incubated at 65°C for approximately 8 hours. DNA was purified using phenol-chloroform-isoamyl alcohol (25:24:1) extraction and ethanol precipitation. Resuspended DNA was treated with DNase-free RNase (Promega) for 30 min at 37°C, then re-extracted using phenol-chloroform-isoamyl alcohol and ethanol precipitation. DNA was resuspended in 50 μL water and run in a 2% agarose gel. The dilution with the most apparent mono-nucleosome sized band (approximately 150 bp) was extracted using the Invitrogen Pure-Link gel extraction kit (Life Technologies, Carlsbad, CA).

A tiling array of primer sets was designed for each promoter as described previously (23) to perform quantitative PCR. Primers were designed using the PrimerQuest® tool and obtained from Integrated DNA Technologies (see **Appendix Table A1-3** for primers). Quantitative PCR was performed as described above using 0.5 μL of mono-nucleosome DNA extract (at 10 ng/μL) per 10 μL reaction. A section of the

ampicillin gene on each plasmid was used as a control to account for any variation in total plasmid copy number between the two samples. Standard curves were created for each primer set using a serial dilution of the corresponding whole plasmid with concentration varying from $5x10^7$ to $5x10^3$ copies per μL. The relative copy number for each primer set in the promoter was calculated using these standard curves and comparing to the ampicillin primer set.

10.2.6 Computational methods

Nucleosome occupancy of native yeast promoters was optimized through the use of a computational algorithm. First, transcription factor binding sites present in the wild-type sequence were manually identified through the use of the Yeast Promoter Atlas[16]. Then, nucleotides outside these sites were systematically perturbed using a custom MATLAB script, which utilized a FORTRAN implementation of the Nucleosome Positioning Prediction (NuPoP) engine (146) to predict nucleosome affinity. Minor modifications to NuPoP were made to enable the acceptance of command-line inputs. The cumulative sum of nucleosome affinities over each mutant promoter was then computed and the nucleotide substitution resulting in the largest decrease in total nucleosome affinity was saved. This single nucleotide variant was then systematically perturbed as above so that successive increases in promoter strength were achieved in an iterative fashion. This MATLAB script additionally avoided the creation of new transcription factor binding sites (148) and also restricted promoter designs to those which could be synthesized as gblocks by Integrated DNA Technologies, Inc. (Coralville, Iowa) which was the vendor chosen to provide the synthetic DNA in this project.

The identity and placement of transcription factor binding sites in the synthetic promoter scaffolds were determined using a bioinformatics analysis of glycolytic

promoters as a guide. The occurrence and relative positions of common transcription factor binding sites were catalogued and the average spacing values were calculated (See **Table 1-1**). In addition to a consensus TATA box, four transcription factor binding sites were included in the upstream activating sequence area of the synthetic promoter: a Reb1p binding site, a Rap1p binding site, and two Gcr1p binding sites. Consensus binding site sequences were used (148). *Psynth1* was designed using the average lengths between binding sites and *Psynth2* was identical, except that the minimum length of the two longest regions (between the GCR1p binding site and the TATA box and between the TATA box and the transcription start site) was used instead of the average length in an attempt to make a shorter promoter. The *TDH3* transcription start site and 5' UTR was used for both synthetic promoters in order to prevent any confounding issues from having different 5' UTR structures between promoters. Once the binding sites and relative positions were chosen, this information was then used as input to a custom MATLAB script to generate the *Psynth* series of vectors. First, the undetermined nucleotides between each transcription factor binding site were randomly seeded at a GC content of 35%. Once any inadvertent transcription factor binding sites generated in these regions were removed, nucleosome affinity was reduced in an iterative fashion as above. As before, the creation of new transcription factor binding sites or sequences which could not be synthesized was avoided. All computations were performed on an Intel Core 2 Duo processor running Windows 7.

10.3 METHODS FOR CHAPTER 3

10.3.1 Plasmid Construction

Plasmids constructed in this study were constructed through restriction digestion followed by ligation. The schemes for construction of these plasmids are detailed in **Appendix Tables A2-2,3,4**.

10.3.2 Growth Rate Analysis

Strains of interest were precultured for 3 days in the appropriate selective medium, and 1 uL of this precultured was used as an inoculum for a 250 uL culture in selective medium containing 5-FOA and reduced concentrations of uracil. Growth rate measurements were then obtained using a Bioscreen C (Growth Curves USA).

10.4 METHODS FOR CHAPTER 4

10.4.1 Plasmid Construction

10.4.1.1 Plasmid Construction: yECitrine Insert Series

Oligos 5-15 (**Appendix Table A3-3**) were annealed by combining 750 pmol of each complementary oligo in 1X T4 DNA Ligase Buffer (NEB) and incubating at 95 °C for 150 sec. The mixture was then steadily cooled from 75 °C to 25 °C over 24 min. The annealed product was cleaned with a MERmaid Spin Kit (Qbiogene), digested with XbaI, and ligated to the phosphatased XbaI fragment of a p416 (181) vector expressing yECitrine with either a mutant TEF promoter (TEFpmut5 (184)), GPD, or CYC1. Vector and insert digestions were performed for 3 hours at 37 °C and cleaned with a QIAquick PCR Purification Kit (Qiagen) and MERmaid Spin Kit, respectively. Ligations were performed at room temperature for 30 min, followed by heat inactivation. Plasmids from distinct *E. coli* colonies were isolated, sequenced, and transformed into yeast. The yECitrine Insert Series is detailed in **Appendix Table A3-1**.

10.4.1.2 yECitrine pBLUESCRIPT SK Multicloning Site Series

yECitrine was cloned from pT5Y (**Appendix Table A3-1**) using PCR. Primers matching 29 base pairs of yECitrine were used to add restriction sites to both ends of the gene, for a total of 8 different yECitrine PCR products (forward primers: 16-23, reverse primer: 25). After digestion, these yECitrine fragments were each ligated separately into the multi-cloning sites of p416-TEF, p416-GPD, and p416-CYC. The $pCYC_0$xYFP series used oligo 26 as reverse primer because the XhoI site is not unique in p416-CYC. $pGPD_0$6YFP, $pTEF_0$6YFP, $pCYC_0$6YFP, and $pCYC_0$8YFP were made with assembly PCR (see Designed Multicloning Site Series. TEFp, GPDp, or CYC1p, CYC1 terminator, and assembly oligos ($pGPD_0$6YFP, $pTEF_0$6YFP, and $pCYC_0$6YFP: 28-29, $pCYC_0$8YFP: 28 & 30) comprised the first reaction. Full-length product was amplified, digested, and ligated as for the designed MCS series). $pCYC_0$9YFP was constructed by swapping CYC1 for GPD in construct $pGPD_0$9YFP through SacI-XbaI fragmentation. This resulted in 27 distinct plasmids, detailed in **Appendix Table A3-4**.

10.4.1.3 yECitrine Designed Multicloning Site Series

Novel MCSs were generated with assembly PCR. PCR products of TEFp (primers 31-32), GPDp (primers 33-34), or CYC1p (primers 35-36) were combined with CYC1 terminator (primers 37-38) and assembly oligos (39-42, 43-45, 46-48, 49-52, or 53-56) at 30 nM each and amplified (94 °C for 1 min, 68 °C for 2 min, 72 °C for 3 min, 25 cycles). Full-length product was then amplified from 2.5 µL of this mixture (forward primers 31, 33, or 35; reverse primer 38), digested with SacI and KpnI, and ligated to a phosphatased SacI-KpnI fragment of p416. yECitrine was inserted at each restriction site as for the pBLUESCRIPT SK series (forward primers 16-24, reverse primers 25,26, or 57 as necessary) resulting in the constructs shown in Table 4. $pCYC1_11$1YFP was constructed with CYC1p, CYC1 terminator, and primer 58 using assembly PCR because

XhoI is not unique in this construct. The yECitrine Designed Multicloning Site Series is listed in **Appendix Table A3-2**.

10.4.1.4 LacZ pBLUESCRIPT SK Multicloning Site Series

LacZ was isolated from whole-genome extract of *E. coli* K12-MG1665 (Wizard Genomic DNA Purification Kit, Promega) with PCR (primers 59-60), fragmented with XbaI and ClaI, and ligated to p416-GPD. LacZ was inserted at XbaI as for the pBLUESCRIPT SK series (primers 61-62). $pTEF_0$3LacZ, $pTEF_0$5LacZ, $pTEF_0$7LacZ, and $pTEF_0$9LacZ were constructed using assembly PCR (LacZ-CYC1term (primers 38 & 63) and assembly oligos ($pTEF_0$3LacZ: 65, $pTEF_0$5LacZ: 66-67, $pTEF_0$7LacZ: 66 & 68, $pTEF_0$9LacZ: 66, 69-70) comprised the first reaction. Full-length product was amplified in a second reaction (primers 38 & 64)). Each product was digested with XbaI and KpnI, and ligated to p416-TEF. The resulting LacZ pBLUESCRIPT KS Multicloning Site Series is detailed in **Appendix Table A3-5**.

10.4.1.5 GFP pBLUESCRIPT SK Multicloning Site Series

GFP was isolated from pZE-GFP (185) using PCR (forward primers 71-75, reverse primer 76), fragmented, and ligated to p416-TEF at XbaI, BamHI, EcoRI, ClaI, and XhoI as for the pBLUESCRIPT SK series. The resulting GFP pBLUESCRIPT SK Multicloning Site Series is detailed in **Appendix Table A3-6**.

10.4.2 RT-PCR Assay

For each tested variant, the replicate yielding the most typical fluorescence measurement was grown to an optical density of 0.5 and its RNA was extracted (Ribopure Yeast Kit, Ambion). 100 ng RNA was reverse-transcribed and quantified in triplicate using an iScript One-Step RT-PCR Kit with SYBR Green (Biorad) immediately

after RNA extraction. yECitrine transcript levels were measured relative to that of ALG9 (Primers 1-4) on a 7900HT Real Time PCR Instrument (Applied Biosystems).

10.4.3 β-Galactosidase Assay

Yeast colonies were picked in triplicate, grown in YSC Ura^- to an optical density of 0.5, and prepared according to manufacturer's instructions (Novabright β-Galactosidase Enzyme Reporter Gene Chemiluminescent Detection Kit for Yeast Cells, Invitrogen). Luminescence was quantified with a SpectraMax M3 Multi-Mode Microplate Reader (Molecular Devices). Day-to-day variability was accounted for by analyzing all comparable transformants on the same day.

10.4.4 Computational Studies and Modeling Efforts

Nupack2.1.2 (176) was used to perform all RNA folding calculations. Folding conditions of 30 °C, 1 M Na^+, and 0 M Mg^{2+} were utilized. All reported energies are the free energies of the ensemble of potential structures, as opposed to the minimum free energy structure. Pseudoknots were not considered due to computational limitations. 1st and 2nd round computations were run on an intel Xeon processor running MATLAB. 3rd round computations were run on all cores of an intel core i7 processor running MATLAB. Most optimizations were run over 24 hours.

10.4.4.1 1st round of optimization

The first set of MCSs (pTEF1xYFP and pCYC11xYFP) were designed with the goal of maximizing the ensemble free energy of the complete 5'UTR (261-264). Design proceeded using a hill-climbing algorithm in a two-step process, using the free energy of the longest possible 5'UTR (i.e. cloning into the last possible restriction site in the MCS) as its score. The restriction sites were first reordered to maximize free energy, followed

by the addition of up to 5 bp between each restriction site to further increase free energy **(Figure 10-1B,C)**.

10.4.4.2 2nd round of modeling and optimization

To address the limitations of the first model of structure-based translation inhibition, a model framework was developed incorporating two (or more) regions whose free energy of folding correlates with protein production. These free energy barriers can occur as the complex is scanning along the 5'UTR or as the complex is binding to the 5' cap structure. If N_i is the number of complexes in state *i* and N_{i+1} is the number of complexes in the next state, then we have:

$$N_{i+1} = N_i exp(-\beta * \Delta G)$$

where ΔG is the magnitude of the free energy barrier and β represents the Boltzmann constant of the system (i.e. how energetic each complex is and thus how likely it is to traverse energetic barriers). Such results from statistical mechanics are valid due to the large number of yeast cells measured. If there are N complexes in the first (unbound) state, we have:

$$N_i = N * \prod_i exp(-\beta_i * \Delta G_i)$$

where β_i are the Boltzmann constants at each state, ΔG_i are the free energies of each barrier between them. We can rewrite the product to yield

$$N_i = N * exp(\sum_i -\beta_i * \Delta G_i)$$

Assuming there are *i* states and the rate of translation initiation (hence protein production) is proportional to the number of initiation complexes in the last state (the state closest to the start codon), we have

$$f = C * exp(\sum_i -\beta_i * \Delta G_i)$$

where f is the fluorescence value and C is a proportionality constant (since the data have been normalized to the fluorescence of a particular construct). If we take the logarithm of both sides, we can correlate the logarithm of the fluorescence to barrier free energies by fitting the Boltzmann constants and the proportionality constant, C:

$$\log(\hat{f}) = \sum_i -\hat{\beta}_i * \Delta G_i + \hat{C}$$

where the hat denotes the estimator of a variable. This framework was used to develop models for the 2nd and 3rd rounds of modeling.

Models and novel MCSs were evaluated using the ensemble free energies of two disjoint segments of RNA as predictors. The boundaries for each segment were measured relative to the start codon. Although possibly between the boundaries of each segment, nucleotides which were not between the start of the 5'UTR and 30 bp after the start codon were not included in folding calculations.

In addition to the pBLUESCRIPT SK MCS data, the yECitrine expression resulting from a number of other post-promoter "inserts" (see "yECitrine insert series") were also used to train the predictive model for each promoter. A hill-climbing algorithm was implemented to search for the two segments whose free energies best correlated with the data for all the available constructs according to the framework above (**Figure 10-1A**). The correlation coefficient was used to score each potential model.

Hill-climbing algorithms were similarly used to search for the best possible MCS in a two-step process similar to the first round of optimization (**Figure 10-1B,C**). For each potential MCS, a score was calculated using the model developed above. A positive value was given to those positions which, when yECitrine was inserted at that site, resulted in a higher predicted fluorescence than had been predicted at the same position (e.g. the 3rd site from the end of the promoter) in other MCSs. A negative score was

similarly given to underperforming positions. The total score for each potential MCS was the sum of these positive and negative values, and the MCSs with the greatest scores were selected.

10.4.4.3 3rd round of modeling

Due to increased computational resources, the third round of modeling used an exhaustive search of pairs of disjoint predictive regions in all available data instead of a hill-climbing algorithm (**Figure 10-1D**). The predicted residual sum of squares (PRESS) was used to score each pair, as computed by the hat matrix.

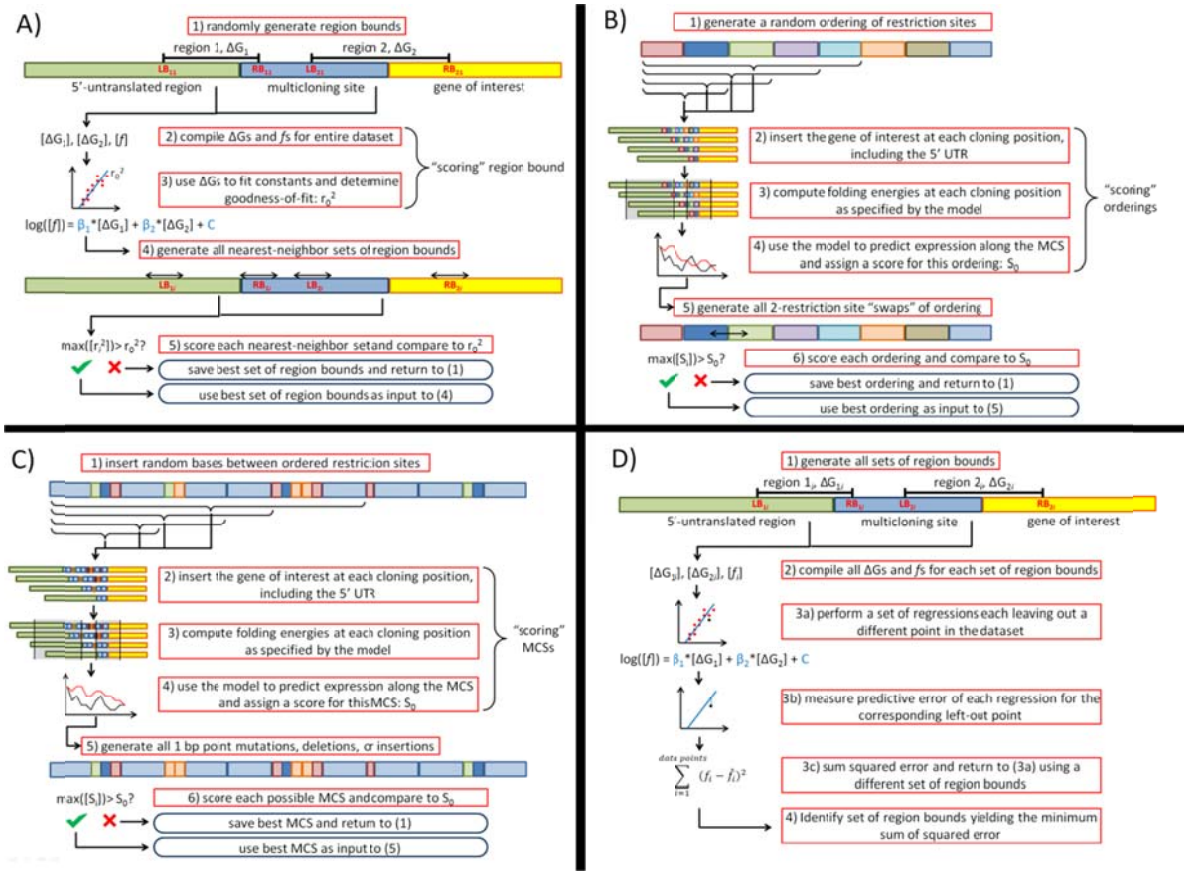


Figure 10-1: Model Construction and Multicloning Site Design Methodology.

(**A**) The first round of modeling implemented a hill-climbing algorithm to search for regions of the mRNA whose free energy of folding correlated strongly with fluorescence. MCSs were then designed via a two-

step process: (**B**) a hill-climbing algorithm to find the optimal ordering of restriction sites followed by (**C**) a hill-climbing algorithm to further decrease the likelihood of secondary structure formation. (**D**) The second round of modeling undertook an exhaustive search of all possible pairs of regions to find the set which showed the greatest predictive ability.

10.5 METHODS FOR CHAPTER 5

10.5.1 Plasmid Construction

Plasmids in this study were constructed by first PCRing the appropriate template with the indicated primers, as detailed in **Appendix Tables A4-1,2,3,** and **4**. These PCR products were then digested with DpnI (NEB), gel-purified (GeneJet Gel Extraction Kit, Thermo Scientific), phosphorylated with T4 polynucleotide kinase (NEB), and ligated with T4 DNA ligase (NEB).

10.5.2 Western Blotting

10.5.2.1 Characterization of a Panel of 2A Sites

50mL of exponential-phase cells expressing the desired bicistronic reporter construct in the appropriate growth medium were pelleted and protein was extracted (Y-PER Yeast Protein Extraction Reagent, Thermo Scientific). This extract was then denatured (4% SDS, 10% 2-mercaptoethanol, 20% glycerol, 0.004% bromophenol blue, 0.125 M Tris-HCl, pH 6.8), applied to an SDS-PAGE gel (12% stacking, 6% separating) and proteins were separated through electrophoresis (mini-PROTEAN system, BioRAD). This gel was then incubated in transfer buffer (25 mM Tris, 190 mM Glycine, 20% methanol, pH 8.3) and proteins were transferred to a PVDF membrane through electrophoresis at 150 mA for 1.4 h. This membrane was then blocked for 1h with TBST+milk (20mM Tris pH 7.5, 150 mM NaCl, 0.1% Tween 20, 5% milk), then exposed to an HRP-conjugated anti-his antibody for 1 h and washed 3 times with TBST. Conjugated HRP was then visualized with Supersignal West Pico Chemiluminescent Substrate (Thermo).

10.5.2.2 Characterization of 2A Variants

50mL of exponential-phase cells expressing the desired bicistronic reporter construct in the appropriate growth medium were pelleted and protein was extracted (Y-PER Yeast Protein Extraction Reagent, Thermo Scientific). This extract was then denatured (4% SDS, 10% 2-mercaptoethanol, 20% glycerol, 0.004% bromophenol blue, 0.125 M Tris-HCl, pH 6.8), applied to an SDS-PAGE gel (12% stacking, 6% separating) and proteins were separated through electrophoresis (mini-PROTEAN system, BioRAD). This gel was then incubated in transfer buffer (25 mM Tris, 190 mM Glycine, 20% methanol, pH 8.3) and proteins were transferred to a nitrocellulose membrane. This membrane was then blocked overnight with TBST+milk (20mM Tris pH 7.5, 150 mM NaCl, 0.1% Tween 20, 5% milk), washed 5 times then TBST, then exposed to an HRP-conjugated anti-his antibody for 30 mins and washed again 5 times with TBST. Conjugated HRP was then visualized with Immun-star HRP substrate (Bio-rad).

10.6 METHODS FOR CHAPTER 6

10.6.1 Plasmid Construction

Plasmids for this study were constructed according to the schemes detailed in Appendix A5. The procedure for recombination cloning was identical to that used in chapter 6, and the procedure for phosphorylation ligation was identical to that used in chapter 4.

10.7 METHODS FOR CHAPTER 7

10.7.1 Recombination Cloning in Yeast

1 ug of each PCR fragment was digested with DpnI and cotransformed into *S. cerevisiae* BY4741 according to the procedure described in (242). This transformation mixture was then plated on the appropriate dropout medium and allowed to grow for 3

days at 30 C. Yeast colonies from this plate were scraped and plasmids were extracted (Zymoprep Yeast Miniprep Kit, Zymo Research). This plasmid mixture was then transformed into *E. coli* DH10β and plated. Individual colonies were then amplified in liquid culture and plasmids were extracted. Correctly assembled plasmids were confirmed through restriction digestion and sequencing.

10.7.2 Analysis of Transposition Efficiency

10.7.2.1 Plate-based induction

Three to five biological replicates of a yeast strain carrying the engineered retrotransposon of interest were used to inoculate 1 mL liquid cultures lacking histidine and containing galactose, thus inducing retroelement transcription. After 3 days of growth at 30 C, cultures were plated on agar containing glucose and either lacking histidine or lacking both histidine and uracil and allowed to grow for 3 days at 30 C. Colonies were counted manually or through automated software (265) and counts were used as inputs to the Fluctuation Analysis Calculator (266) implementing the Ma, Sandri, and Sarkar Maximum Likelihood Estimation method (267). Calculated mutation rates per cell were divided by the time spent in galactose medium to determine the transposition rate per cell per generation (r as defined above) as well as 95% confidence intervals. This value was then used to estimate a library size as described above.

10.7.2.2 Low OD induction

Three biological replicates of a yeast strain carrying the engineered retrotransposon of interest were used to inoculate 50 mL liquid cultures lacking histidine and containing galactose, thus inducing retroelement transcription. After 3 days of growth at 30 C, cultures were plated on agar containing glucose and either lacking histidine or lacking both histidine and uracil and allowed to grow for 3 days at 30 C.

Colonies were counted manually or through automated software (265) and counts were used as inputs to the Fluctuation Analysis Calculator (266) implementing the Ma, Sandri, and Sarkar Maximum Likelihood Estimation method (267). Calculated mutation rates per cell were divided by the time spent in galactose medium to determine the transposition rate per cell per generation as well as 95% confidence intervals. This value was then used to estimate a library size.

10.7.2.3 High OD induction

For high OD tests, cells were first cultivated in 50 mL liquid cultures lacking histidine and containing glucose and then resuspended in 50 mL liquid cultures lacking histidine and containing galactose to an initial OD of 1. After 3 days of growth at 30 C, cultures were plated on agar containing glucose and either lacking histidine or lacking both histidine and uracil and allowed to grow for 3 days at 30 C. Colonies were counted manually or through automated software (20) and counts were averaged. This average was then used as an estimate for the number of transpositions which occurred during the experiment.

10.7.3 qPCR Analysis

Yeast strains carrying pGALmTy1-HIV were grown to mid-log phase (OD=0.5) in 5 mL YSC containing either glucose or galactose. Total RNA was extracted (Ribopure Yeast Kit, Life Technologies) from half of each culture and converted to cDNA (High Capacity cDNA Reverse Transcription Kit, Life Technologies). Total DNA was extracted (Wizard Genomic DNA Purification Kit, Promega) from the other half of the culture. qPCR was conducted using 10ng of either cDNA or total DNA (FastStart SYBR Green Master, Roche) using primers specific for an intronless *URA3* (*URA3*RTPCRF and *URA3*RTPCRR) and with Alg9 as an internal standard (Alg9F and Alg9R).

10.7.4 Models

10.7.4.1 Model for Mutation Accumulation in Continuous Culture

If a collection of yeast cells is held at a constant cell density and transposition is induced, mutants will begin to accumulate. If we assume that wild-type and mutant cells grow at the same rate, then the fraction of unmutated cells will decrease solely due to mutation and not competition with faster-growing cells. If f_0 represents this fraction and r represents the mutation rate per cell per unit of time, then we have

$$\frac{df_0}{dt} = -rf_0$$

with $f_0(0)=1$. The fraction of singly mutated cells (f_1) similarly decreases in proportion to its size but is replenished by the decrease in f_0. For this population and indeed for all cells with containing n transpositions we have

$$\frac{df_n}{dt} = rf_{n-1} - rf_n$$

for n>0 and with $f_n(0)=0$. These equations may be solved to yield

$$f_n(t) = \frac{1}{n!}(rt)^n * \exp(-rt)$$

Thus, computation of the mutation rate per cell per time enables a highly detailed picture of the culture to be ascertained. In this work, mutation rates will be calculated per cell per 1.5 hours (one doubling time) and library sizes will be calculated assuming 10^10 cells growing for one week in continuous culture by computing (1-f_0(100))*10^10.

10.7.4.2 Computational framework for deducing transposition rate and mutation rate from the two-color assay

Assume the following:

f_t = transpositions / cell (~0.1 for Ty1)

r_m = mutation rate / bp (~0.0001 for Ty1)

L_1 = Length of gene 1 (YFP = 717bp)

L_2 = Length of gene 2 (RFP = 711bp)

f_{I_1}= fraction of mutations which inactivate gene 1 (assumed to be approximately 0.3)

f_{I_2}= fraction of mutations which inactivate gene 2 (assumed to be approximately 0.3)

p_{nm} = probability of getting n mutations in gene 1 and m mutations in gene 2, given transposition

I_{00} = prob of no inactivation given n mutations in gene 1 and m mutations in gene 2

I_1 = prob of inactivating only gene 1 given n mutations in gene 1 and m mutations in gene 2

I_2 = prob of inactivating only gene 2 given n mutations in gene 1 and m mutations in gene 2

I_{12} = prob of inactivating both genes given n mutations in gene 1 and m mutations in gene 2

Probabilities of:

No transposition:

$$1 - f_t$$

Transpose and inactivate 1:

$$f_t * \sum_{n=1}^{L_1} \sum_{m=0}^{L_2} (p_{nm} I_1)$$

Transpose and inactivate 2:

$$f_t * \sum_{n=0}^{L_1} \sum_{m=1}^{L_2} (p_{nm} I_2)$$

Transpose and inactivate both:

$$f_t * \sum_{n=1}^{L_1} \sum_{m=1}^{L_2} (p_{nm} I_{12})$$

Transpose and inactivate none:

$$f_t * \sum_{n=0}^{L_1} \sum_{m=0}^{L_2} (p_{nm} I_{00})$$

What is p_{nm}?

$$p_{nm} = \binom{L_1}{n} {r_m}^n (1 - r_m)^{L_1 - n} \binom{L_2}{m} {r_m}^n (1 - r_m)^{L_2 - m}$$

What is I_{00}?

$$I_{00} = \binom{n}{0} {f_{I_1}}^0 (1 - f_{I_1})^n \binom{m}{0} {f_{I_2}}^0 (1 - f_{I_2})^m$$

$$I_{00} = (1 - f_{I_1})^n (1 - f_{I_2})^m$$

What is I_1?

$$I_1 = \sum_{i=1}^{n} \binom{n}{i} {f_{I_1}}^i (1 - f_{I_1})^{n-i} \binom{m}{0} {f_{I_2}}^0 (1 - f_{I_2})^m$$

$$I_1 = (1 - (1 - f_{I_1})^n)(1 - f_{I_2})^m$$

What is I_2?

$$I_2 = \binom{n}{0} {f_{I_1}}^0 (1 - f_{I_1})^n \sum_{i=1}^{m} \binom{m}{i} {f_{I_2}}^i (1 - f_{I_2})^{m-i}$$

$$I_2 = (1 - f_{I_1})^n (1 - (1 - f_{I_2})^m)$$

What is I_{12}?

$$I_{12} = \sum_{i=1}^{n} \binom{n}{i} {f_{I_1}}^i (1 - f_{I_1})^{n-i} \sum_{i=1}^{m} \binom{m}{i} {f_{I_2}}^i (1 - f_{I_2})^{m-i}$$

$$I_{12} = (1 - (1 - f_{I_1})^n)(1 - (1 - f_{I_2})^m)$$

Therefore we have the fraction of cells not fluorescent:

$$1 - f_t + f_t * \sum_{n=1}^{L_1} \sum_{m=1}^{L_2} \binom{L_1}{n} {r_m}^n (1 - r_m)^{L_1 - n} \binom{L_2}{m} {r_m}^n (1 - r_m)^{L_2 - m} (1 - (1 - f_{I_1})^n)(1 - (1 - f_{I_2})^m)$$

$$1 - f_t + f_t * \sum_{n=1}^{L_1} \binom{L_1}{n} {r_m}^n (1 - r_m)^{L_1 - n} (1 - (1 - f_{I_1})^n) \sum_{m=1}^{L_2} \binom{L_2}{m} {r_m}^n (1 - r_m)^{L_2 - m} (1 - (1 - f_{I_2})^m)$$

$$1 - f_t + f_t * \left(1 - \left(1 + (1 - f_{I_1}) r_m - r_m\right)^{L_1}\right) \left(1 - \left(1 + (1 - f_{I_2}) r_m - r_m\right)^{L_2}\right)$$

And similarly for the fraction of cells in region exhibiting dual fluorescence:

$$f_t * \sum_{n=0}^{L_1} \sum_{m=0}^{L_2} \binom{L_1}{n} {r_m}^n (1-r_m)^{L_1-n} \binom{L_2}{m} {r_m}^n (1-r_m)^{L_2-m} \left(1-f_{I_1}\right)^n \left(1-f_{I_2}\right)^m$$

$$f_t * \sum_{n=0}^{L_1} \binom{L_1}{n} {r_m}^n (1-r_m)^{L_1-n} \left(1-f_{I_1}\right)^n \sum_{m=0}^{L_2} \binom{L_2}{m} {r_m}^n (1-r_m)^{L_2-m} \left(1-f_{I_2}\right)^m$$

$$f_t * \left(1+\left(1-f_{I_1}\right)r_m - r_m\right)^{L_1} \left(1+\left(1-f_{I_2}\right)r_m - r_m\right)^{L_2}$$

RFP only:

$$f_t * \sum_{n=1}^{L_1} \sum_{m=0}^{L_2} \binom{L_1}{n} {r_m}^n (1-r_m)^{L_1-n} \binom{L_2}{m} {r_m}^n (1-r_m)^{L_2-m} \left(1-\left(1-f_{I_1}\right)^n\right)\left(1-f_{I_2}\right)^m$$

$$f_t * \sum_{n=1}^{L_1} \binom{L_1}{n} {r_m}^n (1-r_m)^{L_1-n} \left(1-\left(1-f_{I_1}\right)^n\right) \sum_{m=0}^{L_2} \binom{L_2}{m} {r_m}^n (1-r_m)^{L_2-m} \left(1-f_{I_2}\right)^m$$

$$f_t * \left(1-\left(1+\left(1-f_{I_1}\right)r_m - r_m\right)^{L_1}\right) \left(1+\left(1-f_{I_2}\right)r_m - r_m\right)^{L_2}$$

YFP only:

$$f_t * \sum_{n=0}^{L_1} \sum_{m=1}^{L_2} \binom{L_1}{n} {r_m}^n (1-r_m)^{L_1-n} \binom{L_2}{m} {r_m}^n (1-r_m)^{L_2-m} \left(1-f_{I_1}\right)^n \left(1-\left(1-f_{I_2}\right)^m\right)$$

$$f_t * \sum_{n=0}^{L_1} \binom{L_1}{n} {r_m}^n (1-r_m)^{L_1-n} \left(1-f_{I_1}\right)^n \sum_{m=1}^{L_2} \binom{L_2}{m} {r_m}^n (1-r_m)^{L_2-m} \left(1-\left(1-f_{I_2}\right)^m\right)$$

$$f_t * \left(1+\left(1-f_{I_1}\right)r_m - r_m\right)^{L_1} \left(1-\left(1+\left(1-f_{I_2}\right)r_m - r_m\right)^{L_2}\right)$$

A MATLAB script has been developed to infer mutation rate and transposition rate given the proportions of cells which are nonfluorescent, exhibit dual fluorescence, or which only express a single fluorescent protein. We have shown this script to accurately infer these parameters for mutation rates up to 0.0009, a greater than 6-fold increase in mutation rate over wild-type Ty1RT, and a mutation rate which would enable significantly higher library sizes for smaller DNA sequences.

10.7.4.3 Plasmid Segregation Inefficiency Calculations

If we assume that each plasmid contained within the mother cell has an equal probability (p) of being transferred to the daughter cell (total plasmid number = n) and we observe some fraction of daughter cells which do not contain a plasmid (f), then we must have the following relationship:

$$f = (1-p)^n$$

Thus, we can infer p from measurements of n and f. Then, the average number of plasmids delivered to a daughter cell per budding is simply equal to p*n.

10.7.5 Next-Generation Sequencing

10.7.5.1 Next-Generation Sequencing Sample Preparation

Ten replicates from BY4741 Δ*rrm3* plus pGALmTy1-Ty1 or BY4741 Δ*hir3*Δ*cac3* plus pGALmTy1-HIV were cultivated in 50 mL liquid cultures lacking histidine and containing glucose. After 3 days of growth at 30 C, 1 mL culture was removed and the plasmids were extracted using Zymoprep Yeast Plasmid Miniprep Kit II (Zymo Research). The rest of the culture was then resuspended in 50 mL liquid cultures lacking histidine and containing galactose to an initial OD of 1. After 3 days of growth at 30 C, 1 mL culture was extracted to obtain plasmids, and 1 mL culture was plated on agar containing glucose and either lacking histidine or lacking both histidine and uracil and allowed to grow for 3 days at 30 C. Colonies were counted manually or through automated software (268) and counts were averaged. This average was then used as an estimate for the number of transpositions which occurred during the experiment. Two sequencing primer pairs with different barcodes were used to amplify the ampicillin sequence region from fresh pGALmTy1-Ty1 plasmid and pGALmTy1-Ty1 plasmid extracted from glucose medium, and 20 primer pairs amplified the *URA3* sequence region from the 20 minipreps of galactose cultures. The PCR products were purified and the concentrations were determined by nanodrop. A final concentration of 50 ng/μL sample was prepared by combining 22 PCR purified products, with a 5:2 molar basis of ampicillin amplicon to *URA3* amplicon. This mixture was then sequenced using an

Illumina Miseq in 2x250bp paired-end mode. All PCR fragments and their corresponding primers are listed below.

10.7.5.2 Analysis of Next Gen Sequencing Data

Paired-end reads were matched up and error-corrected using pandaseq (269) using stringent quality filtering (threshold=0.9). Matched pairs were then divided up based upon barcode sequence using sabre, allowing for single nucleotide mutations (since each barcode was at least 2bp away from one another) and barcodes were removed with the trimmingreads.pl script of the NGS QC toolkit (270). After combining reads originating from the same culture into the same file, alignment to the unmutated amplicon was performed using ssaha2 (271). Custom shell scripts were then used to extract the total number of mutations identified (**Appendix B.2**) and 95% confidence intervals for mutation counts were computed using the method of the Clopper-Pearson Interval (272)

10.7.6 Vector Construction

10.7.6.1 Construction of Vectors with Homologous Recombination in Yeast

All vectors which were constructed using yeast homologous recombination were assembled according to the schemes listed in **Tables A6-1,2, and 3**

10.7.6.2 Generation of Transpositional Activator Expression Plasmids

TEC1, ELG1, RTT101, *HSX1*, and TYE1 were amplified from the genome of *S. cerevisiae* using the primers listed in Table1. These PCR fragments were digested with XmaI and XhoI and ligated to p425-GPD treated with XmaI, XhoI, and Antarctic phosphatase (NEB) using T4 DNA ligase according to the manufacturer's instructions, generating p425-GPD-TEC1, p425-GPD-ELG1, p425-GPD-RTT101, p425-GPD-HSX1,

and p425-GPD-TYE1. These plasmids were co-transformed with the appropriate retroelement in both BY4741 and BY4741 Δ*mre11* and transposition rate was measured.

10.7.6.3 Generation of Truncated Reverse Transcriptase Expression Plasmids

Truncated reverse transcriptases containing the HIV polymerase domain and either the HIV or Ty1 connection domain (tHT and tHH) were amplified from pGALmTy1-HIV and pGALmTy1-HTT with the primers listed in Table 3. These PCR fragments were digested with SpeI and XhoI and ligated to p425-GPD (181) treated with SpeI, XhoI, and Antarctic phosphatase (NEB) using T4 DNA ligase according to the manufacturer's instructions, generating p425-GPD-tHH andp425-GPD-tHT. These plasmids were co-transformed with the appropriate retroelement in BY4741 and transposition rate was measured.

10.7.6.4 Saturation Mutagenesis of Ty1 Reverse Transcriptase

The pGALmTy1-Ty1 containing the *CAN1* cassette was mutated at each of the three Ty1 Reverse Transcriptase target sites using a QuikChange Multi Site-directed Mutagenesis Kit (Agilent Cat#200514), using Ty145QCF, Ty225QCF, and Ty226QCF according to the manufacturer's direction. The resulting library was transformed into electrocompetent *E. coli* as described above and plated. Single clones were amplified in 5 mL LB medium and incubated overnight at 37 °C. Plasmids were isolated using Zyppy Plasmid Mini-Prep Kit (Zyppy Cat#D4037) and the mutations were confirmed by sequencing.

To construct the double mutants L145S/F225Y and L145/F225H, the QuikChange Multi Site-directed Mutagenesis Kit (Agilent Cat#200514) was used to introduce the L145S mutation into the previously made plasmids containing the F225Y or F225H mutations; in both cases, the QMTy1RTL145Sf and QMTy1RTL145Sr primers were

used. For the other three point mutations, the pGALmTy1-Ty1 was used as a template in three PCR reactions using either primers Ty1RTL151Af and Ty1RTL151Af, Ty1RTK93Rf and Ty1RTK93R94rev, or Ty1RTR94Kf and Ty1RTK93R94rev. Each reaction amplified a linear DNA fragment ~14 kbp in length, containing the entire retroelement sequence with a new single point mutation. This PCR fragment was phosphorylated using Polynucleotide Kinase (New England Biolabs), and then subsequently ligated using T4 DNA Ligase (New England Biolabs) to form complete plasmids.

10.7.6.5 Insertion of URA3-intron system into Ty1 Saturation Mutagenesis Library

For each of the 57 variants in the three Ty1 saturation mutagenesis libraries, the *CAN1* expression cassette was replaced by the *URA3*-intron cassette to allow testing of transposition rate. First, *URA3*AI-2 was amplified through PCR using pGALmTy1-Ty1 as a template and BefpptR and His3AIgenomeflankF primers. The plasmid backbone was then digested with EcoRI and BsrGI (New England Biolabs) according to manufacturer instructions, which excised an approximately 650 bp region of the *CAN1* gene. Transformation of the digested plasmid with the PCR product swapped the entire *CAN1* gene with the *URA3*-intron system through homologous recombination.

10.7.6.6 Construction of Retroelement Without Reverse Transcriptase

The pGALmTy1-Ty1 was used as a template in PCR with primers His3AIgenomeflankF and ARTrev, resulting in a linear DNA fragment ~11 kbp in length, containing all parts of the retroelement except for the reverse transcriptase and incorporating a new stop codon after the integrase-reverse transciptase protein cleavage site. This PCR fragment was phosphorylated using Polynucleotide Kinase (New England

Biolabs), then subsequently ligated using T4 DNA Ligase (New England Biolabs) to form the plasmid pGALmTy1-ART. The plasmid sequence was verified by enzyme digestion and sequencing.

10.7.6.7 Construction of HIV Reverse Transcriptase Variants

First, two variants of pGALmTy1-HIV with both HIV and Ty1 primer binding sites were used as templates to revert two previously-made mutations in the HIV RT. This was done stepwise using the QuikChange Multi Site-directed Mutagenesis Kit (Agilent Cat#200514), first with primers HIVT1361GG1362AF and HIVT1361GG1362AR, then with primers HIVA343TQCF and HIVA343TQCR.

Next, the protein cleavage site was inserted using PCR with the reverse primer PCSinsR and either PCS0insF, PCS3insF, or PCS6insF as the forward primer. This resulted in a linear DNA fragment ~14 kbp in length, containing the entire retroelement sequence in addition to a new protein cleavage site and coding for either 0, 3, or 6 additional amino acids from the Ty1 RT. This PCR fragment was phosphorylated using Polynucleotide Kinase (New England Biolabs), and then subsequently ligated using T4 DNA Ligase (New England Biolabs) to form a complete plasmid. This process resulted in 16 pGALmTy1-HIV variants with each combination of the above factors – primer binding site, wild-type RT, and 4 variations of protein cleavage site.

10.7.6.8 Construction of Vectors with Inactivated Integrase

The pGALmTy1-Ty1 was used as a template in PCR with primers Ty2600F and Ty2600R, resulting in a linear DNA fragment ~14 kbp in length, containing the entire retroelement sequence in addition to a new 15-bp sequence in the integrase gene. This PCR fragment was phosphorylated using Polynucleotide Kinase (New England Biolabs),

and then subsequently ligated using T4 DNA Ligase (New England Biolabs) to form a complete plasmid.

10.7.6.9 Construction of "Cargo"-containing Retroelements

The pGALmTy1-Ty1 vector was used as a template in PCR to create a linear DNA fragment with a break between the *URA3* and RT gene. To create various cargo genes, PCR fragments with homology were created using LacZ, eGFP, *CAN1*, and the mStrawberry-YFP gene (used in the two-color assay) as templates. Primers were constructed such that homologous recombination added each gene after the stop codon of the RT with a frame-shift mutation in the start codon of the template gene, and none included a promoter sequence. In this way, the DNA added should have a minimal effect due to transcription or expression. The largest, pGALmTy1-Ty1-Cargo5, was made by digesting Cargo3 and the mStraw-YFP PCR fragment with EcoRI (New England Biolabs), and then subsequently ligating the two using T4 DNA Ligase (New England Biolabs) to form a complete plasmid. All plasmid sequences were verified by enzyme digestion and sequencing.

10.7.6.10 Construction of SPT15, XylA, and XylA Pathway Vectors

The plasmids pGALmTy1-Ty1-XylA, pGALmTy1-Ty1-XylA3, pGALmTy1-Ty1-Spt15 were constructed through homologous recombination cloning using the scheme outlined in the tables below. The TEF1 promoter was then inserted into these plasmids in the following way. The homologous *XylA*-TEF1 *Spt15*-TEF1 cassettes were amplified, digested with *Dpn*I, and co-transformed with the above plasmids digested with *Not*I, yielding pGALmTy1-Ty1-XylA-TEF1, pGALmTy1-Ty1-XylA3-TEF1, and pGALmTy1-Ty1-Spt15-TEF1. A multiple cloning site was then inserted into pGALmTy1-Ty1-XylA3 after the Ty1 reverse transcriptase. A MCS cassette was

amplified through overlap-extension PCR of MCS1 and MCS2, and this construct was co-transformed with PXKSmcs digested with *Dpn*I, yielding pGALmTy1-Ty1-MCS-XylA3-TEF1. The plasmid pGALmTy1-Ty1-MCS-XylA3-TEF1 was digested with *Not*I and *EcoR*I and ligated with NotI-XylA-TEF1-EcoRI cassette amplified from p415-TEF-XylA, yielding the plasmid pGALmTy1-Ty1-MCS-XylA-TEF1.

10.7.6.11 Construction of Low-copy Vectors

The low-copy version of pGALmTy1-Ty1 was constructed through homologous recombination cloning using the scheme outlined in the tables below, using a p413 vector as a template for the CEN6/ARSH replication sequence and pGALmTy1-Ty1 as a template to make a linearized fragment lacking a replication site. To construct low-copy versions of the four constructs used in current evolution experiments (XylA, *XylA*-3, *Spt15*, and *Spt15*-300), the same scheme and primers were used, and the high-copy version of each was used as a template for the homologous recombination.

10.7.6.12 Construction of synthetic retroelements with intron-containing cargos

pGALmTy1-Ty1-MCS-XylA-TEF1, pGALmTy1-Ty1-MCS-XylA3-TEF1, pGALmTy1-Ty1-Spt15-TEF1, and pGALmTy1-Ty1-Spt15-300-TEF1 were used as templates for PCR to generate products *XylA*intronnosite, *XylA*3intronnosite, *Spt15*intronnosite, and *Spt15*-300intronnosite. These PCR products were digested with DpnI (New England Biolabs), phosphorylated with T4 polynucleotide kinase (New England Biolabs), and ligated with T4 DNA ligase (Fermentas, Inc) to generate pGALmTy1-Ty1-MCS-XylAintronnosite-TEF1, pGALmTy1-Ty1-MCS-XylA3intronnosite-TEF1, pGALmTy1-Ty1-Spt15intronnosite-TEF1, and pGALmTy1-Ty1-Spt15-300intronnosite-TEF1. These vectors were then used to generate PCR products *XylA*intron, *XylA*3intron, *Spt15*intron, and *Spt15*-300intron, which were treated

as above to generate pGALmTy1-Ty1-MCS-XylAintron-TEF1, pGALmTy1-Ty1-MCS-XylA3intron-TEF1, pGALmTy1-Ty1-Spt15intron-TEF1, and pGALmTy1-Ty1-Spt15-300intron-TEF1.

10.7.6.13 Construction of Nonevolving Controls

pGALmTy1-Ty1-MCS-XylA-TEF1 (low copy), pGALmTy1-Ty1-MCS-XylA3-TEF1 (low copy), pGALmTy1-Ty1-Spt15-TEF1 (low copy), pGALmTy1-Ty1-Spt15-300-TEF1 (low copy), pGALmTy1-Ty1-MCS-XylAintron-TEF1 (low copy), pGALmTy1-Ty1-MCS-XylA3intron-TEF1 (low copy), pGALmTy1-Ty1-Spt15intron-TEF1 (low copy), and pGALmTy1-Ty1-Spt15-300intron-TEF1 (low copy) were used as templates for PCR to generate products *XylA*lcnoRT, *XylA*3lcnoRT, *SPT15*lcnoRT, *SPT15*-300lcnoRT, *XylA*lcintnoRT, *XylA*3lcintnoRT, *SPT15*lcintnoRT, and *SPT15*-300lcintnoRT, respectively. These PCR products were digested with DpnI (New England Biolabs), phosphorylated with T4 polynucleotide kinase (New England Biolabs), and ligated with T4 DNA ligase (Fermentas, Inc) to generate pGALmTy1-MCS-XylA-TEF1 (low copy), pGALmTy1-MCS-XylA3-TEF1 (low copy), pGALmTy1-Spt15-TEF1 (low copy), pGALmTy1-Spt15-300-TEF1 (low copy), pGALmTy1-MCS-XylAintron-TEF1 (low copy), pGALmTy1-MCS-XylA3intron-TEF1 (low copy), pGALmTy1-Spt15intron-TEF1 (low copy), and pGALmTy1-Spt15-300intron-TEF1 (low copy).

10.7.6.14 Construction of Ty1 and HIV Reverse Transcriptase Fluorescent Fusion Proteins

To construct both Ty1-RT and HIV-RT fusion proteins, the pGALmTy1-Ty1 and pGALmTy1-HIV vectors were first linearized by digestion with BamHI and NotI (New England Biolabs). PCR was used to create a fragment including the YFP gene with overlap such that the YFP would recombine before the stop codon of the RT (Either

YFPHARTfusF or YFPTARTfusF and YFPfusR). Homologous recombination cloning using the scheme outlined in the tables below was carried out, and the correct plasmids were confirmed by digestion and sequencing. Next, the linker was inserted using PCR of the entire vector with primers incorporating the additional sequence (either Ty1RTlinkerR or HIVRTlinkerR and YFPlinkerF). This linearized vector was then phosphorylated (using T4 PNK, New England Biolabs), and subsequently ligated using T4 DNA Ligase (New England Biolabs) to form a complete plasmid

10.7.6.15 Construction of Ty1 Two-color Fluorescent Retroelement system

To construct the two-color fluorescent retroelement system, pGALmTy1-Ty1-TEF-mStrawberry-intron-P2A-YFP, PCR was used to create a fragment including the mStrawberry-P2A-YFP gene with restriction sites for XmaI and EcoRI (using primers mStraw-YFPf and mStraw-YFPr). Both this PCR fragment and pGALmTy1-Ty1-TEF-XylA3 were digested with XmaI and EcoRI, and then subsequently ligated using T4 DNA Ligase (New England Biolabs) to form a complete plasmid, pGALmTy1-Ty1-TEF-mStrawberry-P2A-YFP. Next, the artificial intron was inserted using PCR of the entire vector with primers incorporating the additional sequence (mStrawIntPmef and mStrawIntPmer). This linearized vector incorporated a PmeI site in the artificial intron to enable more efficient ligation; the PCR product was then digested with PmeI (New England Biolabs), and subsequently ligated using T4 DNA Ligase (New England Biolabs) to form a complete plasmid. This construction scheme was carried out using a high-copy plasmid as a template, and it was then transferred to a low-copy version through homologous recombination cloning, using a p413 vector as a template for the CEN6/ARSH replication sequence and pGALmTy1-Ty1-TEF-mStrawberry-intron-P2A-YFP as a template to make a linearized fragment lacking a replication site.

10.7.6.16 Construction of Xylose Catabolism Pathway Vectors

To construct the xylose catabolism pathway vector, fragments TEF1, *XylA*, *XylA*3, Tkc1, GPD, *XKS1*, and Tkc6 were amplified. The plasmid pGALmTy1-Ty1-Tkc1-XylA-TEF1 was constructed by three sequential ligations with TEF1, *XylA*, and Tkc1 into plasmid pGALmTy1-Ty1-MCS. The plasmid pGALmTy1-Ty1-Tkc1-XylA3-TEF1 was constructed by three sequential ligations with TEF1, *XylA*3, and Tkc1 into plasmid pGALmTy1-Ty1-MCS. The plasmid pGALmTy1-Ty1-Tkc6-*XKS1*-GPD was constructed by three sequential ligations with GPD, *XKS1*, and Tkc6 into plasmid pGALmTy1-Ty1-MCS. Then the GPD-*XKS1*-Tkc6 cassette was amplified and ligated into plasmids pGALmTy1-Ty1-Tkc1-XylA-TEF1 and pGALmTy1-Ty1-Tkc1-XylA3-TEF1 respectively, yielding xylose isomerase pathway plasmids pGALmTy1-Ty1-Tkc6-*XKS1*-GPD-Tkc6-XylA-TEF1 and pGALmTy1-Ty1-Tkc6-*XKS1*-GPD-Tkc6-XylA3-TEF1.

10.7.6.17 Construction of Arabinose Pathway Vectors

The DNA parts which comprised the arabinose pathway will be amplified through PCR, and individual expression cassettes will then be constructed by combining a promoter, gene, and terminator through assembly PCR. Finally, these assembly PCR products will be transformed together with the pGALmTy1-Ty1 backbone to generate pGALmTy1-Ty1-ara3gene, pGALmTy1-Ty1-aranoLXR, pGALmTy1-Ty1-aranoXKS, and pGALmTy1-Ty1-ara5gene, as detailed below.

10.7.7 Strain Construction

10.7.7.1 Construction of gene knockouts in S. cerevisiae BY4741 and CEN.PK2

For all knockouts, a *loxP-kanMX-loxP* deletion cassette was constructed from plasmid PUG6 (24). One kilobase of homologous sequence was amplified from the

upstream region of gene of interest in the genome, and then ligated at the 5' end of the *loxP-kanMX-loxP* module. A second kilobase of homologous sequence amplified from the downstream region of the gene was then ligated at the 3' end of the *loxP-kanMX-loxP* module. The whole gene disruption cassette was amplified and transformed into *S. cerevisiae* BY4741 and CEN.PK2, using a standard lithium acetate transformation method (242) and a version optimized for CEN.PK2 (273), respectively. Cells were then plated onto YPD plus G418 plates (200 μg/mL G418). After one day of growth, the microcolonies were replicated onto new YPD plus G418 plates. The resulting colonies were amplified in 3 mL YPD+G418 and the genomic DNA was extracted using Wizard Genomic DNA Purification kit (Promega). Correct knockouts were confirmed by PCR.

Confirmed single knockout strains were transformed with the *cre* expression plasmid pSH47 (242). Cre recombinase was induced by incubating cell in YPG (galactose) medium for 24 h. The cells were subsequently streaked onto YPD and replica-plated onto YPD plus G418. The *cre* expression plasmid in G418-sensitive colonies was removed by incubating cells in YPD plus 5-FOA for 24 h, thus excising the plasmid and yielding a clean version of knockout strain with a single *loxP* site in the chromosome. Sequential gene knockouts were introduced with the same protocol using the clean strain, yielding a double-knockout strain. The constructed knockout strains are listed in **Table A6-4**.

10.7.7.2 Construction of GRE Knockout strains

The plasmid p415-TEF-XKS was first constructed through ligation of XbaI*XKS1*XhoI with p415-TEF digesed with *Xba*I and *Xho*I. The GRE3KO+XKS and GRE3KO cassettes were amplified from p415-TEF-XKS, which share 40 nt of homology to upstream and downstream of *gre3* gene. The gene disruption cassettes were then

transformed into strain BY4741 Δ*rrm3*, using a standard lithium acetate transformation method (242). Cells were plated onto a yeast synthetic complete plate lacking leucine containing glucose. 10 colonies from each plate were amplified and the genomic DNA was extracted using a Wizard Genomic DNA Purification kit (Promega). The correct knockouts were confirmed by PCR. All PCR fragments and their corresponding primers are listed below. The confirmed knockout strains were transformed with the appropriate plasmids for *in vivo* continuous evolution.

10.7.8 Oscillation Evolution Strategy

A strain containing the appropriate retroelement was first pre-cultured in a yeast synthetic complete medium lacking histidine and containing 2% glucose for 1 to 2 days. The pre-cultured cells were then resuspended into 20 mL of culture containing 2% galactose to an initial OD of 1, thus inducing retroelement transposition. For the oscillation evolution strategy, after 3 days of growth at 30 C, cells were transferred to 500-mL selective liquid culture (a yeast synthetic complete medium lacking histidine and containing 120 g/L glucose and 6% to 8.25% ethanol) to an appropriate initial OD of 0.1－0.2. The flasks were then tightly sealed with rubber stoppers and parafilm to prevent ethanol evaporation. After exponential phase, the culture underwent additional retrotransposition induction and selection by serially culturing in a sequence of galactose and selective media for multiple alternating cycles. For each round of selection, a 1 μL sample of culture was plated on glucose to isolate colonies.

10.7.9 Continuous Evolution Strategy

In the continuous evolution strategy, each culture was first grown up in glucose, then subcultured into 20 mL of 20 g/L galactose media at OD 1.0 for 3 days to induce retrotransposition. The induced culture was then centrifuged, washed with water, and

inoculated into 500 mL of 120 g/L galactose with 6.0% ethanol. The cultures were grown at 30°C until they reached stationary phase (OD > ~2.0), then subcultured again. The first subculture was 10% of the volume (50 mL), with each subsequent subculture using half the volume as the previous (5%, 2.5%, etc…). In addition, each subculture had 0.5% higher ethanol concentration (6.5%, 7.0%, etc…). The OD of each culture was measured each day (see Figures 3 and 4).

10.7.10 Mutant Isolation Method

All the PCR, digestion, gel electrophoresis, gel extraction, and sequencing followed general molecular cloning procedures. Here the Q5-hot start high-fidelity DNA polymerase (New England Biolabs) was used as the polymerase. Primers used for sequencing were listed in Table 1. The primer LacZBegSeqrev was used to sequence plasmid-based evolution cassette. The primers NotITEFF and SacIISpt15R were used to amplify and sequence genome-based evolution cassette. The intron-spanning primers NotITEFF and *Spt15*NointronR were used to amplify and sequence TEF promoter, while primers *Spt15*NointronF and SacIISpt15R were used to amplify and sequence *Spt15* or *Spt15*-300.

10.8 METHODS FOR CHAPTER 8

10.8.1 Strains and Media

Yeast strains are listed in **Appendix Table A7-2** and **A7-7**

10.8.2 Cloning Procedures

Restriction enzyme-based plasmid construction schemes are detailed in **Appendix Table A7-5**. Oligonucleotides were purchased from Integrated DNA Technologies (Coralville, IA). PCR reactions were performed with Q5 Hot-Start High-Fidelity DNA

Polymerase from New England Biolabs (Ipswich, MA) according to manufacturer specifications and the schemes listed in **Appendix Table A7-4**. Homologous recombination-based plasmid construction schemes are detailed in **Appendix Table A7-6**. All assembly reactions were performed according to standard procedures (274).

10.8.3 RT-PCR Assay

For each tested variant, the replicate yielding the most typical fluorescence measurement or itaconic acid yield was grown to an optical density of 0.5 and its RNA was extracted (Quick-RNA Miniprep, Zymo Research Corporation). 2 μg RNA was reverse-transcribed (High Capacity cDNA Reverse Transcription Kit, Applied Biosystems) and quantified in triplicate (SYBR Green PCR Master Mix, Life Technologies) immediately after RNA extraction. Transcript levels were measured relative to that of a housekeeping gene (ALG9) (Viia 7 Real Time PCR Instrument, Life Technologies). Primers used for quantification are listed in **Appendix Table A7-3**.

10.8.4 Itaconic Acid Production

Strains of interest were precultured for 3 days in the appropriate selective medium, and 30uL of this culture were used as inoculum for a 3mL culture in the same medium, which was grown for 3 days in a rotary drum incubator at 30 ºC. This culture was then pelleted down (4 min at 1600 x g), and the supernatant was filtered using a 0.22 μm syringe filter (Corning). 2.0 μL of filtrate was analyzed with a HPLC Ultimate 3000 (Dionex) using a Zorbax SB-Aq column (Agilent) in a mobile phase composed of 99.5% potassium phosphate buffer (pH=2.0) and 0.5% acetonitrile at 30 ºC. Flow rate was maintained at 1.25 mL/min and absorption was measured at 210 nm.

10.8.5 Growth Rate Analysis

Strains of interest were precultured for 3 days in the appropriate selective medium, and 1 uL of this precultured was used as an inoculum for a 250 uL culture in the same selective medium. Growth rate measurements were then obtained using a Bioscreen C (Growth Curves USA).

10.8.6 cDNA Library Generation

Total RNA was extracted from yeast using the RNA Extraction kit (Ambion) and converted to cDNA using a (High Capacity cDNA Reverse Transcription Kit, Applied Biosystems) according to manufacturer's instructions with the exception that primer oligodTEcoRIR2 substituted for the random hexamer primer provided with the kit. This cDNA was then purified using the Qiagen PCR cleanup kit and ligated to primer RNALigAd using T4 RNA Ligase (NEB) according to manufacturer's instructions. This ligation was purified using the Qiagen PCR cleanup kit and amplified using Q5 hot-start DNA polymerase and primers XmaIFlankF and EcoRIFlankR2 according to manufacturer's instructions. Amplicons ranging in size from 500bp to 5kb were gel-extracted (Genejet gel purification kit) and re-purified using the Qiagen PCR cleanup kit. This purified, double stranded, full-length cDNA was then sheared using a covaris sonicator to an average length of 200bp or 400bp, blunt-ended and phosphorylated using the End-It DNA End Repair Kit (Epicentre), and ligated to either p414-CYC-rad9-MCS-rad9'-TEF' or p414-GPD-rad9-MCS-rad9'-TEF'.

Appendices

APPENDIX A: SUPPLEMENTARY TABLES

Appendix A1

Promoter Name	Sequence
TEF1v1	ATAGCTTCAAAATGTTTCTACTCCTTTTTTACTCTTCCAGATTTTCTC GGACTCCGCGCATCGCCGTACCACTTCAAAACACCCAAGCACAGCA TACTAAATTTCCCCTCTTTCTTCCTCTAGGGTGTCGTTAATTACCCGT ACTAAAGGTTTGGAAAAGAAAAAAAAGACCGCCTCGTTTCTTTTTTT TCGTCGAAAAAGGCAATAAAAATTTTTTTCACGTTTCTTTTTCTTGA AAATATTTTTTTTGATTTTTTTCTCTTTCGATGACCTCCCATTGATAT TTAAGTTAATAAACGGTCTTCAATTTCTCAAGTTTCAGTTTCATTTTT TTTGTTCTATTACAACTTTTTTTACTTCTTGCTCATTAAAAAGAAAGC ATAGCAATCTAATCTAAGTTT
HXT7v1	CTCGTAGGAAAAATTTCGGGCCCCTGCGTTTTTTTCTGAGGTTCATT TTTTACATTTGCTTCTGCTGGATAATTTTCAGAGAAAAAAAGGAAAA ATTATATGAAAAAAAGTTTTTTTTCAAGGAAAAAACCCTATTTTTTT TCGAGATCCCCTGTAACTTATTGGCAACTGAAAGAATGAAAAGGAA AAAAATAAAAAATATACTAGAACTGAAAAAAAATTAGTATAAATA GAGACGATATATGCCAATACTTCACAATGTTCGAATCTTTTTTTTAT TTTCAGCTATTGAAAAAAAATAAAACATCAAGAACAAACAAGCTCA ACTTGTCTTTTCTAAGAACAAAGAATAAACACAAAAACAAAAAGTT TTTTTAATTTTAATCAAAAA
HIS5v1	AAATGGTTAAAAATTGTTATCATAAATAAGGTGACCGGTTATATTG AGACCTTTCCTGGACAGTAACTAATACAGAAGCCATTGGTAATGCA ATAATTTTTTTGATCATGTGACTACGATCCGGGTGAGACTATTAAAA AAAGGAGTCAAGCATTGAAATAATTAATGACTAATCCGAAGTTAAT TGTTAGGAGTCAATTGTTTTTTCCAATGAATGGAATCTGAGATGACT AAACTACCAATTTTCAATAGTTCATGGTATAGTGACGTAGTTAGTGC TTTTTTTTCTTGGATCTGTTGACTCACTTCAATTGATGTTTCTTACCC TGACATGACATACTTGATTTTTTATCTCTCACGTTATATAACTTGAA AAGGATGCACACAGTTCTGTTCAATATACCCTCCAATATGTAAAAA AAGTTTTTTCATTGATTACTCTTAATTTTTTTCCTGCTAAACCAGCAG TACGTGTGTGCCGTATATATTAAAATTACACT
CYC1v1	ATTTGGCGAGCGTTGGTTGGTGGATCAAGCCCACGCGTAGGCAATC CTCGAGCAGATCCGCCAGGCGTGTATATATAGCGTGGATGGCCAGG CAATTTTAGTGCTGACACATACAGGCATATATATATGTGTGCGACG AAAAATGATCATATGGCATGCATGTGCTCTGTATGTATATAAAACTC TTGTTTTCTTTTTTTCTCTAAATATTCTTTCCTTATACATTAGGACCTT TGCAGCATAAATTACTATACTTCTATAGACACGCAAAAACAAATAC ACACACTAA

CYC1v2	ATTTGGCGAGCGTTGGTTGGTGGATCAAGCCCACGCGTAGGCAATC CTCGAGCAGATCCGCGAGGCGTGTATATATAGCGTGGATGGCCAGG CAACTTTAGTGCTGACACATACAGGCATATATATATGTGTGCGACG ACACATGATCATATGGCATGTATGTGCTCTGTATGTATATAAAACTC TTTTTTTCTTTTTTTCTCTAAATTTTTTTTCCTTATACATTAGGACCTT TGCAGCATAAATTACTATACTTCTATAGACACGCAAACACAAATAC ACACACTAA
CYC1v3	ATTTCGCGCGCGTTGGTTAGTAAAAAAAGCCCACGCGTAGGGAATC CTCGAGCATATACGCGAGGCGCGTATATATAGCGCGTATGTTCAGG TAAATTTAGTGCTGACACATACAGGCATATATATATGTGCGCGTATA TACATGATTATATGGCATGTATGTGCTCTGTATGTATATAAAACTCT TTTTTTCTTTTTTTTCTCTAAATTTTTTTTCCTTATACATTAGGACCTTT GCAGCATAAATTACTATACTTCTATAGACACGCAAATACAAATACA CACACTAA
TDH3v1	AGTTTATCATTATCAATACTCGCCATTTCAAAGAATACGTAAATAAT TAATAGTAGTGATTTTCCTAACTTTTTTTAGTCAAAAAATTAGCCTTT TAATTCTGCTGTAACCCGTACATGCCCAAAATAGGGGGCGGGTTAC ACAGAATATATAACATCGTAGGTGTCTGGGTGAACAGTTTATTCCTG GCATCCACTAAATATAATGGAGCCCGCTTTTTAAGCTGGCATCCAG AAAAAAAATGAATCCCAGCACCAAAATATTTTTTTCTTCACCAACC ATCAGTTCATAGGTCCATTCTCTTAGCGCAACTACAGAGAACAGGG GCACAAACAGGCAAAAAACGGGCACAACCTCAATGGAGTGATGCA ACCTGCCTGGAGTAAATGATGACACAAGGCAATTTACCCGCGCATG TATCTATCTCATTTTTTTACACCTTCTATTACCTTCTGCTCTCTCTGAT TTGGAAAAAGCTGAAAAAAAAGGTTGAAACCAGTTCCCTGAAATTA TTCCCCTACTTGACTAATAAGTATATAAAGACGGTAGGTATTGATTG TAATTCTGTAAATCTATTTCTTAAACTTCTTAAATTCTACTTTTATAG TTAGTCTTTTTTTTAGTTTTAAAAAACCAGAACTTAGTTTCGACGGA T
GAL1v1	ACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGA CTCTCCTCCGCGCGTCCGCGTCTTCACCGGTCGCGTTCCTGAAACGC AGATGTGCCTCGCGCCGCACTGCTCCGAAAAATAAAGATTCTACAA TACTAGCTTTTTTGGTTATGAAGAGGAAAAATTGGCAGTAACCTGG CCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGAT GATAATGCGATTAGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAG CGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGAAAAAG CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTTTATTA CTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAATTGTTAATAT ACCTCTATACTTTAACGTCAAGGAGAAAAAAC
Psynth1v1	TCCGGGTAACGCCGACACAGTAAGTAACGAGATGTATGGGTGTCCT AACTAAAAGGCTTCCAACTCAACATTGAATCAGGTAATCCTAGATC AAGGCTTCCATACACAGGTTTATATTAATACATATACGACAACTCTC CAATTCGCTCATAATTACAACAAAGATCGAACTGAGAGAGACTTAG ACTCGTACAACTACATTTTTCGTTAACTTTTTAACATACGCGAGGGT ATTAAACTTAGCTGACGCAACTCTAGTTGTATCTCGCGATAATTTCT TTTTACTTGTCTATTTATAAAAACCAAGCTAATAACTTCATACGTCT TATTGTATTTAGACTATTTCTTTTTAACCTAACTATAGCAGAACCCG

	CGGGTAATTACTTAAAACACCAAGAACTTAGTTTCGAATAAACACA CATAAACAAACAAA
Psynth1v2	TCCGGGTAACGCCGAAAAAATAAGTAACGCGATGTATGGGTGTACT AAAAAAAAGGCTTCCAATAAAAAATTGAATCAGGTAATCCTATATC AAGGCTTCCATATATAGGTTTATATTAATACATATACGAAAAAACTC TTTTTCGCGCATAATTATAATAAAAATCGAACTGAGAGAGACTTAG ACTCGTACAACTATTTTTTTTGTTAATTTTTTTATATACGCGCGGGTA TTAAACTTAGCTGACGCGATTCTATTTGTATCTCGCGATAATTTCTTT TTTCTTCTCTATTTATAAAAACCAAGCTAATAACTTCATACGTCTTTT TGTATATAGACTTTTTCTTTTTTTCCTAACTATAGGAGAACCCGCGG GTAATTTTTTAAAACACCAAGAACTTAGTTTCGAATAAACACACAT AAACAAACAAA
Psynth1v3	TCCGGGTAACGCCGAAAAAATTATATACGCGATGTATGGGTGTATT AAAAAAAAGGCTTCCAATAAAAAAAAGAATCAGGTAATCCTTTTTC GCGGCTTCCATATATATTTTTTTATTAATACATATACGAAAAAAGTC TTTTTCGCGGGTAATTATAATAAAAATCGAACTGAGAGAGACTTTC ACACGTACTACTATTTTTTTTATTATTTTTTTTATATACGCGCGGGTA AAAAAATTAACTAACGCGATTTTTTTTCTTTCGCGCGAAAATTTCTT TTTTCTTCTCTATTTATAAAAAGGAAGGAAAAAAGTTCTTACCTCTT TTTGTATATACACTTTTTCTTTTTTTCCTTAGTAAAGGAGAACGCGCG GGTATTTTTTTAAAACACCAAGAACTTAGTTTCGAATAAACACACAT AAACAAACAAA
Psynth2v1	TCCGGGTAACTGCGGGTGACCGCAATCTTAGATGTATGGGTGTTAA CTGAGCTAGGCTTCCATGCATTTAGAGAACTTATTAACTGAATAGTT AGGCTTCCAAACGAGCTAGTTCTCGCGTGTCATCTAAAAAAATTCTA GACTGGTGATACTTATAACTATAAAAAAACTGACACTTCTCCCTAAT CGTAGTATTGTATATATTTTTTTAAAAAAAACTTGCAACCATTTAAA ACACCAAGAACTTAGTTTCGAATAAACACACATAAACAAACAAA
Psynth2v2	TCCGGGTAACCGCGGGTAACCTCAATCTTAGATGTATGGGTGTTAA CTGAGCTCGGCTTCCATGTATTAAAAAAAATTATTAACTGAAAAAA AAGGCTTCCAAACTAGTTTTTTTTCGCGTGTGATCAAAAAAAATTCT AGACGGGTAATACATATAAGTATAAAAAAACTGACACTTCTCCCTA ATCGTAGTATTGTATATATTTTTTTAAAAAAAACTTGCAACCATTTA AAACACCAAGAACTTAGTTTCGAATAAACACACATAAACAAACAAA
Psynth2v3	TCCGGGTAACCGCGGGTAACATATATATTAGATGTATGGGTGTAAA AAAAGCGCGGCTTCCATGTATTAAAAAAAATTTTTTTCTGAAAAAA AAGGCTTCCATACTAATTTTTTTTCGCGCGGGTAGAAAAAAAATACT AGTCGGGTAATACATATAAGTATAAAAAAAGAGACACTTCTCCCTA ATCGTACTATTGTATATATTTTTTTAAAAAAAAGTTCGAAGCTTTTA AAACACCAAGAACTTAGTTTCGAATAAACACACATAAACAAACAAA

Appendix Table A1-1: Sequences of re-designed and synthetic promoters.

Underlined sequences in the synthetic promoters Psynth1 and Psynth2 are designated transcription factor binding sites and 5' UTR sequences.

Primer set name	Forward Primer	Reverse Primer
yECitrine	GGCGCTACTAGTATGTCTAAAGGTGAAGAATTATTCACTGG	ACGCGTCGACTTATTTGTACAATTCATCCATACCATG
LacZ	GGCGCTTCTAGAACTAGTATGACCATGATTACGGATTCACT	ACGCGTCGACTTATTTTTGACACCAGACCAACTG
General promoter primers	TAAAGGGAACAAAAGCTGGAGCTC	CAGTGAATAATTCTTCACCTTTAGACATACTAGTTCTAGA
***HXT7* promoter**	TGACTGAGCTCCTCGTAGGAACAATTTCGGG	GGCGCTACTAGTTCTAGATTTTTGATTAAAATTAAAAAAACTTTTTGTTTT
***HIS5* promoter**	TGACTGAGCTCAAATGGTTAAAAATTGTTATCATA	GGCGCTACTAGTTCTAGAAGTGTAATTTTAATATATACGGCA
P_{CYC1} knockout cassette	TGAATCTAAAATTCCCGGGAGCAAGATCAAGATGTTTTCACAGCTGAAGCTTCGTACGC	TAGCACCTTTCTTAGCAGAACCGGCCTTGAATTCAGTCATGCATAGGCCACTAGTGGATCTG
K. lactis URA3	TGACTGAGCTCCAGCTGAAGCTTCGTACGC	GCATAGGCCACTAGTGGATCTG
***TRP1* integration cassette**	TGGAGATATTCCTTATGGCATGTCTGGCGATGATAATAAAGGGAACAAAAGCTGGAGCTC	ACACCAATAACGCCATTTAATCTAAGCGCATCACCAACGGTACCCAATTCGCCCTATAGT
***yECitrine* qPCR primers**	TTCTGTCTCCGGTGAAGGTGAA	TAAGGTTGGCCATGGAACTGGCAA
***ALG9* qPCR primers**	ATCGTGAAATTGCAGGCAGCTTGG	CATGGCAACGGCAGAAGGCAATAA

Appendix Table A1-2: Primer sequences for cloning of promoters, yECitrine and LacZ genes, knockout and integration cassettes, and primers for qPCR of yECitrine.

All re-designed and synthetic promoters were cloned using the "general promoter primers" with the exception of HIS5v1, which was cloned using the HIS5 promoter primer set

Primer set	Mid-amplicon location relative to start codon	Forward primer	Reverse primer
CYC1_1	-313.5	CGCGCAATTAACCCTCACTAA	AACCAACGCTCGCCAAAT
CYC1_2	-271.5	ATTTGGCGAGCGTTGGT	CGGATCTGCTCGAGGATTG
CYC1_3	-222.5	GCAATCCTCGAGCA	GCCTGTATGTGTCA

		GATCC	GCACTAA
CYC1_4	-189	ATGGCCAGGCAACTTTAG	GTGTCGTCGCACACATA
CYC1_5	-170.5	TAGTGCTGACACATACAGG	CACATGCATGCCATATGAT
CYC1_6	-120.5	GTGCGACGACACATGAT	GGTCCTAATGTATAAGGAAAGAATATTTAG
CYC1v3_1	-315.5	CGCGCAATTAACCCTCACTAAA	AACGCGCGCGAAATGAG
CYC1v3_2	-270.5	GCGCGCGTTGGTTAGTAAA	ATATGCTCGAGGATTCCCTACG
CYC1v3_3	-233.5	CCCACGCGTAGGGAATC	TCAGCACTAAATTTACCTGAACATAC
CYC1v3_4	-208.5	GTATATATAGCGCGTATGTTCAGGTA	GCCTGTATGTGTCAGCACTAA
CYC1v3_5	-179.5	TAGCGCGTATGTTCAGGTAAA	CAGAGCACATACATGCCATATAATC
CYC1v3_6	-167	GTGCTGACACATACAGGCATA	CAGAGCACATACATGCCATATAATC
CYC_7*	-52	TTTCCTTATACATTAGGACCTTTGCAG	AGTGTGTGTATTTGTATTTGCGTGT
CYC_8*	-11	GACACGCAAATACAAATACACACA	TTGGGACAACACCAGTGAATAA
CYC_9*	129.5	TTCTGTCTCCGGTGAAGGTGAA	TAAGGTTGGCCATGGAACTGGCAA
Ampicillin control*	N/A	TGTAACTCGCCTTGATCGTTGGGA	TTGTTGCCATTGCTACAGGCATCG

Appendix Table A1-3: Primers for nucleosome mapping tiling array.

Primers sets marked with a * were used for both CYC1 and CYC1v3. All other sets were used for a specific promoter as noted.

Appendix A2

Plasmid Name	Source
P413-CYC1	(181)
P413-NUP57-*URA3*	Kate Curran
P413-TFC1-*URA3*	Kate Curran
P416-TEFpmut7-YFP	(27)

Appendix Table A2-1: Plasmids used in this study

Primer Name	Sequence

p416ura3BamH1fwd	CGC-GGATCC-ATGTCGAAAGCTACATATAAGGAACGT
p416ura3EcoR1rev	CCG-GAATTC-TTAGTTTTGCTGGCCGCATC
CYC1p F	CCCCCC-GAGCTC-ATTTGGCGAGCG
CYC1pmXbaSpeR	GG-ACTAGT-TCTAGA-TTAGTGTGTGTATTTG
MCS-Fwd-SpeI	G-ACTAGT-ATGTCTAAAGGTGAAGAATTATTCACTGG
MCS-Rev-2	CCCCG-CTCGAG-TTATTTGTACAATTCATCCATACCATGGG
CYC1pmut16F	CCCCC-GAGCTC-CAGCTTTTGTTCCCTTTAGTGAATCC-TCTAGA-ACTAGT-CC
CYC1pmut16R	GG-ACTAGT-TCTAGA-GGATTCACTAAAGGGAACAAAAGCTG-GAGCTC-GGGGG
CYC1pmut17R	GG-ACTAGT-TCTAGA-GGATCCTCAGCACTAAAGTTG
CYC1pmut18R	GG-ACTAGT-TCTAGA-GGATCCACTAGATTAGTGTGTGT
CYC1pmut19R	GG-ACTAGT-TCTAGA-GGATCCACAACTATATATACACGC
CYC1pmut20F	CCCCC-GAGCTC-CAGCTTTTGTTCCCTTTAGTGGATCC-TCTAGA-ACTAGT-CC
CYC1pmut20R	GG-ACTAGT-TCTAGA-GGATCCACTAAAGGGAACAAAAGCTG-GAGCTC-GGGGG
CYC1pmut21R	GG-ACTAGT-TCTAGA-TATAGACACGCAAACACAAATACA
CYC1pmut22F(2)	ccccc-gagctc-atttggcgagcgttggttggtggatcaagcc
CYC1pmut22R(2)	GG-ACTAGT-TCTAGA-attgcctacgcgtgggcttgatccaccaacc

Appendix Table A2-2: Primers used in this study (IDT)

Name	Primer 1	Primer 2	Template
URA3	p416ura3BamH1fwd	p416ura3EcoR1rev	BY4741 genome
CYC1	CYC1p F	CYC1pmXbaSpeR	P413-CYC1
CYC1libs	CYC1p F	CYC1pmXbaSpeR	P413-CYC1
YFP	MCS-Fwd-SpeI	MCS-Rev-2	P416-TEFpmut7-YFP
CYC1mut3	CYC1pF	CYC1pmXbaSpeR	P413-CYC1mut3-*URA3*
CYC1mut7	CYC1pF	CYC1pmXbaSpeR	P413-CYC1mut7-*URA3*
CYC1mut8	CYC1pF	CYC1pmXbaSpeR	P413-CYC1mut8-*URA3*
CYC1mut13	CYC1pF	CYC1pmXbaSpeR	P413-CYC1mut13-*URA3*
CYC1mut16	CYC1pmut16F	CYC1pmut16R	None
CYC1mut17	CYC1pF	CYC1pmut17R	P413-CYC1mut17-*URA3*
CYC1mut18	CYC1pF	CYC1pmut18R	P413-CYC1mut18-*URA3*
CYC1mut19	CYC1pF	CYC1pmut19R	P413-CYC1mut19-*URA3*
CYC1mut20	CYC1pmut20F	CYC1pmut20R	None
CYC1mut22	CYC1pmut22F(2)	CYC1pmut22R(2)	P413-CYC1mut22-*URA3*
CYC1mut23	CYC1pF	CYC1pmut17R	P413-CYC1-*URA3*

Appendix Table A2-3: PCR products generated in this study

Name	Backbone	Insert	Restriction Enzyme 1	Restriction Enzyme 2

P413-CYC1-*URA3*	P413-CYC1	*URA3*	BamHI	EcoRI
P413-CYC1libs-*URA3*	P413-CYC-*URA3*	CYC1libs	SpeI	SacI
P423-CYC1-YFP	P423-CYC1	YFP	SpeI	XhoI
P423-CYC1mut3-YFP	P423-CYC1-YFP	CYC1mut3	SacI	SpeI
P423-CYC1mut7-YFP	P423-CYC1-YFP	CYC1mut7	SacI	SpeI
P423-CYC1mut8-YFP	P423-CYC1-YFP	CYC1mut8	SacI	SpeI
P423-CYC1mut13-YFP	P423-CYC1-YFP	CYC1mut13	SacI	SpeI
P423-CYC1mut16-YFP	P423-CYC1-YFP	CYC1mut16	SacI	SpeI
P423-CYC1mut17-YFP	P423-CYC1-YFP	CYC1mut17	SacI	SpeI
P423-CYC1mut18-YFP	P423-CYC1-YFP	CYC1mut18	SacI	SpeI
P423-CYC1mut19-YFP	P423-CYC1-YFP	CYC1mut19	SacI	SpeI
P423-CYC1mut20-YFP	P423-CYC1-YFP	CYC1mut20	SacI	SpeI
P423-CYC1mut22-YFP	P423-CYC1-YFP	CYC1mut22	SacI	SpeI
P423-CYC1mut23-YFP	P423-CYC1-YFP	CYC1mut23	SacI	SpeI

Appendix Table A2-4: Plasmids generated through restriction ligation

Name	Sequence
CYC1mut3	atttggcgagcgttggttggtggatcaagcccacgcgtaggcaatcctcgagcagatccgccaggcgtgtatatata gcgtggatggcctggcaactttagtgctgacacatacaggcatatatatatgtgtgcgacgacacatgatcatatggc

	atgcatgtgctctgtatgtatataaaactcttgttttcttcttttctctaaatattctttccttatacattaggacctttgcagcat aaattactatacttctatagacacgcaaacacaaatacacacactaa
CYC1mut7	atttggcgagcgttggttggtggatcaagcccacgcgtaggcaatcctcgagcagatccgccagtcgtgtatatata gcgtggatggccaggcaactttagtgctgacacatacaggcatatatatatgtgtgcgacgacacatgatcatatggc atacatgtgctctgtatgtatataaaactcttgttttcttcttttctctaaatattctttccttatacattaggacctttgcagcat aaattactatacttctatagccacgcaaacacaaatacacacactaa
CYC1mut8	atttggcgagcgttggttggaggatcaagcccacgcgtaggcaatcctcgagcagatccgccaggcgtgtatatat agcgtggatggcctggcaactttagtgctgacacatacaggcatatatatatgtgtgcgacgacacatgatcatatgg catgcatgtgctctgtatgtatataaaactcttgttttcttcttttctctaaatattctttccttatacattaggacctttgcagc ataaattactatacttctatagacacgcaaacacaaatacacacactaa
CYC1mut13	atttggcgagcgttggttggtggatcaagcccacgcgtaggcaatcctcgagcagatccgccaggcgtgtatatata gcgtggatggccaggcaactttagtgctgacacatacaggcatataaatatgcgtgcgacgacacatgatcatatgg catgcatgtgctctgtatgtatataaatctcttgttttcttcttttctctaaatattctttccttatacattaggacctttgcagca taaattactatacttctatagacacgcaaacacaaatacacacactaa
CYC1mut16	CAGCTTTTGTTCCCTTTAGTGAATCC
CYC1mut17	atttggcgagcgttggttggtggatcaaacccacgcgtaggcaatcctcgagcagatccgccaggcgtgtatatata gcgtgggtggccaggcaactttagtgctga
CYC1mut18	atttggcgagcgttggttggtggatcaagcccacgcgtaggcaatcctcgagcagatccgccaggcgtgtatatata gcatggatggccaggcaacttttgtgctgacacatacaggcatatatatatgtgtgcgacgacacatgatcatatggc atgcatgtgctctgtatgtatataaaactcttgttttcttcttttctctaaatattctttccttatacattaggaccgttgcagca taaattactatacttctatagacacgcaaacacaaatacacacactaa
CYC1mut19	atttggcgagcgttggttggtggatcaagcccacgcgtaggcaatcctcgagcagatccgccaggcgtgtatatata gt
CYC1mut20	CAGCTTTTGTTCCCTTTAGTGGATCC
CYC1mut22	atttggcgagcgttggttggtggatcaagcccacgcgtaggcaat
CYC1mut23	atttggcgagcgttggttggtggatcaagcccacgcgtaggcaatcctcgagcagatccgccaggcgtgtatatata gcgtggatggccaggcaactttagtgctga

Appendix Table A2-5: Promoter mutants generated in this study

Appendix A3

pT5Y	TEFpmut5-TCTAGA-AAA-YFP
pT21(1) F	TEFpmut5-TCTAGA-GAATTC-TCTAGA-AAA-YFP
pT21(1) FF	TEFpmut5-TCTAGA-GAATTC-TCTAGA-GAATTC-TCTAGA-AAA-YFP
pT21(1) FFFF	TEFpmut5-TCTAGA-GAATTC-TCTAGA-GAATTC-TCTAGA-GAATTC-TCTAGA-GAATTC-TCTAGA-AAA-YFP
pT21(1) FFFFFF	TEFpmut5-TCTAGA-GAATTC-TCTAGA-GAATTC-TCTAGA-GAATTC-TCTAGA-GAATTC-TCTAGA-GAATTC-TCTAGA-GAATTC-TCTAGA-AAA-YFP
pT21(2) F	TEFpmut5-TCTAGA-GGTTGG-TCTAGA-AAA-YFP
pT21(2) RR	TEFpmut5-TCTAGA-CCAACC-TCTAGA-CCAACC-TCTAGA-AAA-YFP
pT21(2) RRR	TEFpmut5-TCTAGA-CCAACC-TCTAGA-CCAACC-TCTAGA-CCAACC-TCTAGA-AAA-YFP

pT21(3) F	TEFpmut5-TCTAGA-GGGCCC-TCTAGA-AAA-YFP
pT21(3) FF	TEFpmut5-TCTAGA-GGGCCC-TCTAGA-GGGCCC-TCTAGA-AAA-YFP
pT21(4) F	TEFpmut5-TCTAGA-AAATTT-TCTAGA-AAA-YFP
pT21(4) FF	TEFpmut5-TCTAGA-AAATTT-TCTAGA-AAATTT-TCTAGA-AAA-YFP
pT26 F	TEFpmut5-TCTAGA-GA-GAATTC-AGG-TCTAGA-AAA-YFP
pT26 RR	TEFpmut5-TCTAGA-CCT-GAATTC-TC-TCTAGA-CCT-GAATTC-TC-TCTAGA-AAA-YFP
pT36 F	TEFpmut5-TCTAGA-AGTAGCC-GAATTC-TGTCAGTT-TCTAGA-AAA-YFP
pT46 F	TEFpmut5-TCTAGA-AGGGACAAACTT-GAATTC-ATTTAGGCGTAGT-TCTAGA-AAA-YFP
pGY	GPD-TCTAGA-AAA-YFP
pG21(1) F	GPD-TCTAGA-GAATTC-TCTAGA-AAA-YFP
pG21(2) RR	GPD-TCTAGA-CCAACC-TCTAGA-CCAACC-TCTAGA-AAA-YFP
pG21(2) R	GPD-TCTAGA-CCAACC-TCTAGA-AAA-YFP
pG21(2) F	GPD-TCTAGA-GGTTGG-TCTAGA-AAA-YFP
pG21(3)	GPD-TCTAGA-GGGCCC-TCTAGA-AAA-YFP
pG21(4)	GPD-TCTAGA-AAATTT-TCTAGA-AAA-YFP
pG26 FRRR	GPD-TCTAGA-GA-GAATTC-AGG-TCTAGA-CCT-GAATTC-TC-TCTAGA-CCT-GAATTC-TC-TCTAGA-CCT-GAATTC-TC-TCTAGA-AAA-YFP
pG26 FRFF	GPD-TCTAGA-GA-GAATTC-AGG-TCTAGA-CCT-GAATTC-TC-TCTAGA-GA-GAATTC-AGG-TCTAGA-GA-GAATTC-AGG-TCTAGA-AAA-YFP
pG26 FFFR	GPD-TCTAGA-GA-GAATTC-AGG-TCTAGA-GA-GAATTC-AGG-TCTAGA-GA-GAATTC-AGG-TCTAGA-CCT-GAATTC-TC-TCTAGA-AAA-YFP
pG26 R	GPD-TCTAGA-CCT-GAATTC-TC-TCTAGA-AAA-YFP
pG36 FF	GPD-TCTAGA-AGTAGCC-GAATTC-TGTCAGTT-TCTAGA-AGTAGCC-GAATTC-TGTCAGTT-TCTAGA-AAA-YFP
pG36 RRF -1	GPD-TCTAGA-AACTGACA-GAATTC-GGCTACT-TCTAGA-AACTGACA-GAATTC-GGCTACT-TCTAGA-AGTACC-GAATTC-TGTCAGTT-TCTAGA-AAA-YFP
pG36 RFF	GPD-TCTAGA-AACTGACA-GAATTC-GGCTACT-TCTAGA-AGTAGCC-GAATTC-TGTCAGTT-TCTAGA-AGTAGCC-GAATTC-TGTCAGTT-TCTAGA-AAA-YFP
pG46 RR	GPD-TCTAGA-ACTACGCCTAAAT-GAATTC-AAGTTTGTCCCT-TCTAGA-ACTACGCCTAAAT-GAATTC-AAGTTTGTCCCT-TCTAGA-AAA-YFP
pG46 R	GPD-TCTAGA-ACTACGCCTAAAT-GAATTC-AAGTTTGTCCCT-TCTAGA-AAA-YFP
pG46 FF -1	GPD-TCTAGA-AGGACAAACTT-GAATTC-ATTTAGGCGTAGT-TCTAGA-AGGGACAAACTT-GAATTC-ATTTAGGCGTAGT-TCTAGA-AAA-YFP
pG46 F	GPD-TCTAGA-AGGGACAAACTT-GAATTC-ATTTAGGCGTAGT-TCTAGA-AAA-YFP
pCY	CYC1-TCTAGA-AAA-YFP

pC21(1) F	CYC1-TCTAGA-GAATTC-TCTAGA-AAA-YFP
pC21(2) R	CYC1-TCTAGA-CCAACC-TCTAGA-AAA-YFP
pC21(2) RF	CYC1-TCTAGA-CCAACC-TCTAGA-GGTTGG-TCTAGA-AAA-YFP
pC21(2) RFF	CYC1-TCTAGA-CCAACC-TCTAGA-GGTTGG-TCTAGA-GGTTGG-TCTAGA-AAA-YFP
pC21(2) FFF	CYC1-TCTAGA-GGTTGG-TCTAGA-GGTTGG-TCTAGA-GGTTGG-TCTAGA-AAA-YFP
pC21(3) F	CYC1-TCTAGA-GGGCCC-TCTAGA-AAA-YFP
pC21(3) FF	CYC1-TCTAGA-GGGCCC-TCTAGA-GGGCCC-TCTAGA-GGGCCC-TCTAGA-AAA-YFP
pC21(4) F	CYC1-TCTAGA-AAATTT-TCTAGA-AAA-YFP
pC26 R	CYC1-TCTAGA-CCT-GAATTC-TC-TCTAGA-AAA-YFP
pC26 RFRF	CYC1-TCTAGA-CCT-GAATTC-TC-TCTAGA-GA-GAATTC-AGG-TCTAGA-CCT-GAATTC-TC-TCTAGA-GA-GAATTC-AGG-TCTAGA-AAA-YFP
pC26 RFRR	CYC1-TCTAGA-CCT-GAATTC-TC-TCTAGA-GA-GAATTC-AGG-TCTAGA-CCT-GAATTC-TC-TCTAGA-CCT-GAATTC-TC-TCTAGA-AAA-YFP
pC36 F	CYC1-TCTAGA-AGTAGCC-GAATTC-TGTCAGTT-TCTAGA-AAA-YFP
pC36 RF	CYC1-TCTAGA-AACTGACA-GAATTC-GGCTACT-TCTAGA-AGTAGCC-GAATTC-TGTCAGTT-TCTAGA-AAA-YFP
pC36 RRF	CYC1-TCTAGA-AACTGACA-GAATTC-GGCTACT-TCTAGA-AACTGACA-GAATTC-GGCTACT-TCTAGA-AGTAGCC-GAATTC-TGTCAGTT-TCTAGA-AAA-YFP
pC36 RRRF	CYC1-TCTAGA-AACTGACA-GAATTC-GGCTACT-TCTAGA-AACTGACA-GAATTC-GGCTACT-TCTAGA-AACTGACA-GAATTC-GGCTACT-TCTAGA-AGTAGCC-GAATTC-TGTCAGTT-TCTAGA-AAA-YFP
pC46 F	CYC1-TCTAGA-AGGGACAAACTT-GAATTC-ATTTAGGCGTAGT-TCTAGA-AAA-YFP
pC46 RR	CYC1-TCTAGA-ACTACGCCTAAAT-GAATTC-AAGTTTGTCCCT-TCTAGA-ACTACGCCTAAAT-GAATTC-AAGTTTGTCCCT-TCTAGA-AAA-YFP

Appendix Table A3-1: yECitrine Insert Series

pTEF$_1$1YFP	TEF-ATCGAT-YFP
pTEF$_1$2YFP	TEF-ATCGAT-AATAC-GGATCC-YFP
pTEF$_1$3YFP	TEF-ATCGAT-AATAC-GGATCC-CC-ACTAGT-YFP
pTEF$_1$4YFP	TEF-ATCGAT-AATAC-GGATCC-CC-ACTAGT-ACAAA-GAATTC-YFP
pTEF$_1$5YFP	TEF-ATCGAT-AATAC-GGATCC-CC-ACTAGT-ACAAA-GAATTC-CCCCC-CTGCAG-CCC-CTCGAG-YFP
pTEF$_1$6YFP	TEF-ATCGAT-AATAC-GGATCC-CC-ACTAGT-ACAAA-GAATTC-CCCCC-CTGCAG-CCC-CTCGAG-CCC-TCTAGA-YFP
pTEF$_1$7YFP	TEF-ATCGAT-AATAC-GGATCC-CC-ACTAGT-ACAAA-GAATTC-

	CCCCC-CTGCAG-CCC-CTCGAG-CCC-TCTAGA-AAAAA-AAGCTT-YFP
pTEF$_1$8YFP	TEF-ATCGAT-AATAC-GGATCC-CC-ACTAGT-ACAAA-GAATTC-CCCCC-CTGCAG-CCC-CTCGAG-CCC-TCTAGA-AAAAA-AAGCTT-ACAAC-GTCGAC-YFP
pTEF$_1$9YFP	TEF-ATCGAT-AATAC-GGATCC-CC-ACTAGT-ACAAA-GAATTC-CCCCC-CTGCAG-CCC-CTCGAG-CCC-TCTAGA-AAAAA-AAGCTT-ACAAC-GTCGAC-ACAAA-GATATC-AAAAA-CCCGGG-YFP
pTEF$_2$1YFP	TEF-ACCCC-ATCGAT-YFP
pTEF$_2$2YFP	TEF-ACCCC-ATCGAT-CACCC-ACTAGT-YFP
pTEF$_2$3YFP	TEF-ACCCC-ATCGAT-CACCC-ACTAGT-TCTAGA-YFP
pTEF$_2$4YFP	TEF-ACCCC-ATCGAT-CACCC-ACTAGT-TCTAGA-CCCCC-GAATTC-YFP
pTEF$_2$5YFP	TEF-ACCCC-ATCGAT-CACCC-ACTAGT-TCTAGA-CCCCC-GAATTC-CCCCC-AAGCTT-YFP
pTEF$_2$6YFP	TEF-ACCCC-ATCGAT-CACCC-ACTAGT-TCTAGA-CCCCC-GAATTC-CCCCC-AAGCTT-TAAA-GTCGAC-YFP
pTEF$_2$7YFP	TEF-ACCCC-ATCGAT-CACCC-ACTAGT-TCTAGA-CCCCC-GAATTC-CCCCC-AAGCTT-TAAA-GTCGAC-CCCCC-CTCGAG-YFP
pTEF$_2$8YFP	TEF-ACCCC-ATCGAT-CACCC-ACTAGT-TCTAGA-CCCCC-GAATTC-CCCCC-AAGCTT-TAAA-GTCGAC-CCCCC-CTCGAG-ACCCC-GGATCC-YFP
pTEF$_2$9YFP	TEF-ACCCC-ATCGAT-CACCC-ACTAGT-TCTAGA-CCCCC-GAATTC-CCCCC-AAGCTT-TAAA-GTCGAC-CCCCC-CTCGAG-ACCCC-GGATCC-CCCGGG-YFP
pGPD$_2$1YFP	GPD-A-AAGCTT-YFP
pGPD$_2$2YFP	GPD-A-AAGCTT-ACTAGT-YFP
pGPD$_2$3YFP	GPD-A-AAGCTT-ACTAGT-ATCGAT-YFP
pGPD$_2$4YFP	GPD-A-AAGCTT-ACTAGT-ATCGAT-AAA-GAATTC-YFP
pGPD$_2$5YFP	GPD-A-AAGCTT-ACTAGT-ATCGAT-AAA-GAATTC-TCTAGA-YFP
pGPD$_2$6YFP	GPD-A-AAGCTT-ACTAGT-ATCGAT-AAA-GAATTC-TCTAGA-GGATCC-YFP
pGPD$_2$7YFP	GPD-A-AAGCTT-ACTAGT-ATCGAT-AAA-GAATTC-TCTAGA-GGATCC-CTGCAG-CTCGAG-YFP
pGPD$_2$8YFP	GPD-A-AAGCTT-ACTAGT-ATCGAT-AAA-GAATTC-TCTAGA-GGATCC-CTGCAG-CTCGAG-GTCGAC-YFP
pGPD$_2$9YFP	GPD-A-AAGCTT-ACTAGT-ATCGAT-AAA-GAATTC-TCTAGA-GGATCC-CTGCAG-CTCGAG-GTCGAC-CCCGGG-YFP
pCYC1$_1$1YFP	CYC1-CTCGAG-YFP
pCYC1$_1$2YFP	CYC1-CTCGAG-AAAAA-GTCGAC-YFP
pCYC1$_1$3YFP	CYC1-CTCGAG-AAAAA-GTCGAC-AAC-GGATCC-YFP
pCYC1$_1$4YFP	CYC1-CTCGAG-AAAAA-GTCGAC-AAC-GGATCC-AAAAA-AAGCTT-YFP
pCYC1$_1$5YFP	CYC1-CTCGAG-AAAAA-GTCGAC-AAC-GGATCC-AAAAA-AAGCTT-

	AAAAA-ACTAGT-YFP
pCYC1$_1$6YFP	CYC1-CTCGAG-AAAAA-GTCGAC-AAC-GGATCC-AAAAA-AAGCTT-AAAAA-ACTAGT-ACAC-ATCGAT-YFP
pCYC1$_1$7YFP	CYC1-CTCGAG-AAAAA-GTCGAC-AAC-GGATCC-AAAAA-AAGCTT-AAAAA-ACTAGT-ACAC-ATCGAT-AAAAA-GAATTC-YFP
pCYC1$_1$8YFP	CYC1-CTCGAG-AAAAA-GTCGAC-AAC-GGATCC-AAAAA-AAGCTT-AAAAA-ACTAGT-ACAC-ATCGAT-AAAAA-GAATTC-CCAA-CTGCAG-CC-TCTAGA-YFP
pCYC1$_1$9YFP	CYC1-CTCGAG-AAAAA-GTCGAC-AAC-GGATCC-AAAAA-AAGCTT-AAAAA-ACTAGT-ACAC-ATCGAT-AAAAA-GAATTC-CCAA-CTGCAG-CC-TCTAGA-ACAAT-GATATC-AAAAA-CCCGGG-YFP
pCYC1$_2$1YFP	CYC1-AACCC-GGATCC-YFP
pCYC1$_2$2YFP	CYC1-AACCC-GGATCC-AAGCTT-YFP
pCYC1$_2$3YFP	CYC1-AACCC-GGATCC-AAGCTT-AAAAC-TCTAGA-YFP
pCYC1$_2$4YFP	CYC1-AACCC-GGATCC-AAGCTT-AAAAC-TCTAGA-AACCA-ACTAGT-YFP
pCYC1$_2$5YFP	CYC1-AACCC-GGATCC-AAGCTT-AAAAC-TCTAGA-AACCA-ACTAGT-ACCCA-ATCGAT-YFP
pCYC1$_2$6YFP	CYC1-AACCC-GGATCC-AAGCTT-AAAAC-TCTAGA-AACCA-ACTAGT-ACCCA-ATCGAT-CCTAA-GAATTC-YFP
pCYC1$_2$7YFP	CYC1-AACCC-GGATCC-AAGCTT-AAAAC-TCTAGA-AACCA-ACTAGT-ACCCA-ATCGAT-CCTAA-GAATTC-AAAA-CCCGGG-YFP
pCYC1$_2$8YFP	CYC1-AACCC-GGATCC-AAGCTT-AAAAC-TCTAGA-AACCA-ACTAGT-ACCCA-ATCGAT-CCTAA-GAATTC-AAAA-CCCGGG-AGAAG-CTGCAG-TAAAA-GTCGAC-YFP

Appendix Table A3-2: pTEF$_1$xYFP, pTEF$_2$xYFP, pGPD$_2$xYFP, pCYC1$_1$xYFP, and pCYC1$_2$xYFP.

1	RT YFP Fwd	TTCTGTCTCCGGTGAAGGTGAA
2	RT YFP Rev	TAAGGTTGGCCATGGAACTGGCAA
3	RT ALG9 Fwd	ATCGTGAAATTGCAGGCAGCTTGG
4	RT ALG9 Rev	CATGGCAACGGCAGAAGGCAATAA
5	21(1)	GC-TCTAGA-GAATTC-TCTAGA-GC
6	21(2)F	GC-TCTAGA-GGTTGG-TCTAGA-GC
7	21(2)R	GC-TCTAGA-CCAACC-TCTAGA-GC
8	21(3)	GC-TCTAGA-GGGCCC-TCTAGA-GC
9	21(4)	GC-TCTAGA-AAATTT-TCTAGA-GC
10	26F	GC-TCTAGA-GA-GAATTC-AGG-TCTAGA-GC
11	26R	GC-TCTAGA-CCT-GAATTC-TC-TCTAGA-GC
12	36F	GC-TCTAGA-AGTAGCC-GAATTC-TGTCAGTT-TCTAGA-GC
13	36R	GC-TCTAGA-AACTGACA-GAATTC-GGCTACT-TCTAGA-GC
14	46F	GC-TCTAGA-AGGGACAAACTT-GAATTC-ATTTAGGCGTAGT-TCTAGA-GC

15	46R	GC-TCTAGA-ACTACGCCTAAAT-GAATTC-AAGTTTGTCCCT-TCTAGA-GC
16	YFPXbaIF	GC-TCTAGA-ATGTCTAAAGGTGAAGAATTATTCACTGG
17	YFPSpeIF	G-ACTAGT-ATGTCTAAAGGTGAAGAATTATTCACTGG
18	YFPBamHIF	CG-GGATCC-ATGTCTAAAGGTGAAGAATTATTCACTGG
19	YFPXmaIF	TCC-CCCGGG-ATGTCTAAAGGTGAAGAATTATTCACTGG
20	YFPEcoRIF	G-GAATTC-ATGTCTAAAGGTGAAGAATTATTCACTGG
21	YFPClaIF	CC-ATCGAT-ATGTCTAAAGGTGAAGAATTATTCACTGG
22	YFPSalIF	TAACGC-GTCGAC-ATGTCTAAAGGTGAAGAATTATTCACTGG
23	YFPXhoIF	CCCCG-CTCGAG-ATGTCTAAAGGTGAAGAATTATTCACTGG
24	YFPHindIIIF	CCCCCC-AAGCTT-ATGTCTAAAGGTGAAGAATTATTCACTGG
25	YFPXhoIR	CCCCG-CTCGAG-TTATTTGTACAATTCATCCATACCATGGG
26	YFPSalIR	TAACGC-GTCGAC-TTATTTGTACAATTCATCCATACCATGGG
27	YFP fwd	ATGTCTAAAGGTGAAGAATTATTCACTGGTG
28	6&8constfwd	TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT
29	6constrev	AGTGAATAATTCTTCACCTTTAGACAT-AAGCTT-GATATC-GAATTC-CTGC
30	8constrev	GAATAATTCTTCACCTTTAGACAT-GTCGAC-GGT-ATCGAT-AAGCTT-GATATC-GAATTC-CTG
31	TEFpF	CCCCCC-GAGCTC-ATAGCTTCAAAATGTTTCTAC
32	TEFpR	AAACTTAGATTAGATTCGTATGCTTTCTTTC
33	GPDpF	CCCCCC-GAGCTC-AGTTTATCATTATCAATACTCGCCA
34	GPDpR	ATCCGTCGAAACTAAGTTCTGG
35	CYC1pF	CCCCCC-GAGCTC-ATTTGGCGAGCG
36	CYC1pR	TTAGTGTGTGTATTTGTGTTTGCG
37	CYCtermF	ATTAGTTATGTCACGCTTACATTCACG
38	CYCtermR	GG-GGTACC-GGCCGCAAAT
39	TEFdesMCS1-1	TTGT-ACTAGT-GG-GGATCC-GTATT-ATCGAT-AAACTTAGATTAGATTCGTATGCTTTCTTTC
40	TEFdesMCS1-2	C-GGATCC-CC-ACTAGT-ACAAA-GAATTC-CCCCC-CTGCAG-CCC-CTCGAG-CCC-TCTAGA-AAAAA
41	TEFdesMCS1-3	CCCGGG-TTTTT-GATATC-TTTGT-GTCGAC-GTTGT-AAGCTT-TTTTT-TCTAGA-GGG-CTCGAG-G
42	TEFdesMCS1-4	ACAAA-GATATC-AAAAA-CCCGGG-TCATGTAATTAGTTATGTCACGCTTACATTCA
43	TEFdesMCS2-1	GAAAGCATAGCAATCTAATCTAAGTTT-ATCGAT-GTCGAC-GGATCC-ACTAGT-GAATTC-TCT
44	TEFdesMCS2-2	CTGCAG-GATATC-CTCGAG-CCCGGG-AAGCTT-TCTAGA-GAATTC-ACTAGT-GGATCC-G
45	TEFdesMCS2-3	TGAATGTAAGCGTGACATAACTAAT-CTGCAG-GATATC-CTCGAG-C
46	GPDdesMCS2-1	CCAGAACTTAGTTTCGACGGAT-A-AAGCTT-ACTAGT-ATCGAT-AAA-GAATTC-TCTAGA-GGAT

47	GPDdesMCS2-2	GATATC-CCCGGG-GTCGAC-CTCGAG-CTGCAG-GGATCC-TCTAGA-GAATTC-TTT-ATCGAT-ACT
48	GPDdesMCS2-3	CGAC-CCCGGG-GATATC-ATTAGTTATGTCACGCTTACATTCA
49	CYCdesMCS1-1	AAGCTT-TTTTT-GGATCC-GTT-GTCGAC-TTTTT-CTCGAG-TTAGTGTGTGTATTTGTGTTTGC
50	CYCdesMCS1-2	CAAC-GGATCC-AAAAA-AAGCTT-AAAAA-ACTAGT-ACAC-ATCGAT-AAAAA-GAATTC-CCAA-CTG
51	CYCdesMCS1-3	CCCGGG-TTTTT-GATATC-ATTGT-TCTAGA-GG-CTGCAG-TTGG-GAATTC-TTTTT-ATCGAT-G
52	CYCdesMCS1-4	TGAATGTAAGCGTGACATAACTAAT-CCCGGG-TTTTT-GATATC-ATTGT
53	CYCdesMCS2-1	GCAAACACAAATACACACACTAA-AACCC-GGATCC-AAGCTT-AAAAC-TCTAGA-AACCA-ACTA
54	CYCdesMCS2-2	T-CCCGGG-TTTT-GAATTC-TTAGG-ATCGAT-TGGGT-ACTAGT-TGGTT-TCTAGA-GTTTT-AAGCT
55	CYCdesMCS2-3	CTAA-GAATTC-AAAA-CCCGGG-AGAAG-CTGCAG-TAAAA-GTCGAC-TAAAC-CTCGAG-GATATC
56	CYCdesMCS2-4	TGAATGTAAGCGTGACATAACTAAT-GATATC-CTCGAG-GTTTA-GTCGA
57	YFPXmaIR	TCC-CCCGGG-TTATTTGTACAATTCATCCATACCATGGG
58	CD1-1R	AGTGAATAATTCTTCACCTTTAGACAT-CTCGAG-TTAGTGTGTGTATTTGTGTTTGC
59	LacZExtF	GC-TCTAGAAAA-ATGACCATGATTACGGATTCACTGG
60	LacZExtR	ACGCGTCGACGGTATCGAT-TTATTTTTGACACCAGACCAACTGG
61	LacZXbaIF	GC-TCTAGA-ATGACCATGATTACGGATTCACTGG
62	LacZXhoIR	CCCCG-CTCGAG-TTATTTTTGACACCAGACCAACTGGT
63	LacZF	ATGACCATGATTACGGATTCACTG
64	Mummcsfwd	GC-TCTAGA-ACTAGT-GGATCC-CCC
65	TL3F	TCTAGA-ACTAGT-GGATCC-ATGACCATGATTACGGATTCAC
66	TL57F	TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC
67	TL5R	GTGAATCCGTAATCATGGTCAT-GAATTC-CTGCAG-CCCG
68	TL7R	GTGAATCCGTAATCATGGTCAT-ATCGAT-AAGCTT-GATATC-GAATTC-CTGCAG-CCCG
69	TL9R	CTCGAG-GTCGAC-GGT-ATCGAT-AAGCTT-GATATC-GAATTC-CTGCAG-CCCG
70	TL9F	GATACC-GTCGAC-CTCGAG-ATGACCATGATTACGGATTCAC
71	GFPXbaIF	GC-TCTAGA-ATGCGTAAAGGAGAAGAACTTTT
72	GFPBamHIF	CG-GGATCC-ATGCGTAAAGGAGAAGAACTTTT
73	GFPEcoRIF	G-GAATTC-ATGCGTAAAGGAGAAGAACTTTT
74	GFPClaIF	CC-ATCGAT-ATGCGTAAAGGAGAAGAACTTTT
75	GFPXhoIF	CCCCG-CTCGAG-ATGCGTAAAGGAGAAGAACTTTT
76	GFPXhoIR	CCCCG-CTCGAG-TTAAACTGCTGCAGCGTAG

Appendix Table A3-3: Oligos (IDT)

pTEF$_0$1YFP	TEF-TCTAGA-YFP
pTEF$_0$2YFP	TEF-TCTAGA-ACTAGT-YFP
pTEF$_0$3YFP	TEF-TCTAGA-ACTAGT-GGATCC-YFP
pTEF$_0$4YFP	TEF-TCTAGA-ACTAGT-GGATCC-CCCGGG-YFP
pTEF$_0$5YFP	TEF-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-YFP
pTEF$_0$6YFP	TEF-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-YFP
pTEF$_0$7YFP	TEF-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-ATCGAT-YFP
pTEF$_0$8YFP	TEF-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-ATCGAT-ACC-GTCGAC-YFP
pTEF$_0$9YFP	TEF-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-ATCGAT-ACC-GTCGAC-CTCGAG-YFP
pGPD$_0$1YFP	GPD-TCTAGA-YFP
pGPD$_0$2YFP	GPD-TCTAGA-ACTAGT-YFP
pGPD$_0$3YFP	GPD-TCTAGA-ACTAGT-GGATCC-YFP
pGPD$_0$4YFP	GPD-TCTAGA-ACTAGT-GGATCC-CCCGGG-YFP
pGPD$_0$5YFP	GPD-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-YFP
pGPD$_0$6YFP	GPD-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-YFP
pGPD$_0$7YFP	GPD-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-ATCGAT-YFP
pGPD$_0$8YFP	GPD-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-ATCGAT-ACC-GTCGAC-YFP
pGPD$_0$9YFP	GPD-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-ATCGAT-ACC-GTCGAC-CTCGAG-YFP
pCYC1$_0$1YFP	CYC1-TCTAGA-YFP
pCYC1$_0$2YFP	CYC1-TCTAGA-ACTAGT-YFP
pCYC1$_0$3YFP	CYC1-TCTAGA-ACTAGT-GGATCC-YFP
pCYC1$_0$4YFP	CYC1-TCTAGA-ACTAGT-GGATCC-CCCGGG-YFP
pCYC1$_0$5YFP	CYC1-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-YFP
pCYC1$_0$6YFP	CYC1-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-YFP
pCYC1$_0$7YFP	CYC1-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-ATCGAT-YFP
pCYC1$_0$8YFP	CYC1-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-ATCGAT-ACC-GTCGAC-YFP
pCYC1$_0$9YFP	CYC1-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-ATCGAT-ACC-GTCGAC-CTCGAG-YFP

Appendix Table A3-4: pTEF$_0$xYFP, pGPD$_0$xYFP and pCYC1$_0$xYFP.

pTEF$_0$1LacZ	TEF-TCTAGA-LacZ
pTEF$_0$3LacZ	TEF-TCTAGA-ACTAGT-GGATCC- LacZ
pTEF$_0$5LacZ	TEF-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC- LacZ
pTEF$_0$7LacZ	TEF-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-ATCGAT- LacZ
pTEF$_0$9LacZ	TEF-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-ATCGAT-ACC-GTCGAC-CTCGAG- LacZ

Appendix Table A3-5: pTEF$_0$xLacZ

pTEF$_0$1GFP	TEF-TCTAGA-GFP
pTEF$_0$3GFP	TEF-TCTAGA-ACTAGT-GGATCC- GFP
pTEF$_0$5GFP	TEF-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC- GFP
pTEF$_0$7GFP	TEF-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-ATCGAT- GFP
pTEF$_0$9GFP	TEF-TCTAGA-ACTAGT-GGATCC-CCCGGG-CTGCAG-GAATTC-GATATC-AAGCTT-ATCGAT-ACC-GTCGAC-CTCGAG- GFP

Appendix Table A3-6: pTEF$_0$xGFP

Appendix A4

Plasmid Name	Source
P426-GAL-mStrawberry- PSIV -YFP	Chapter 5

Appendix Table A4-1: Plasmids used in this study

Primer Name	Sequence
P2AYFPGlcF	ACAAGATGGTGATGTTGAAGAAAATCCAGGACCA-tctaaaggtgaagaattattcactgg
P2AstrawGlcR	TTTAATAAAGAAAAGTTGGTAGCACCAGAACC-CTTGTACAGCTCGTCCATGC
T2AYFPGlcF	AACTTGTGGTGATGTTGAAGAAAATCCAGGACCA-tctaaaggtgaagaattattcactgg
T2AstrawGlcR	AATAAAGAGCCACGACCTTCACCAGAACC-CTTGTACAGCTCGTCCATGC
E2AYFPGlcF	ATTAGCCGGTGATGTTGAAAGTAACCCTGGTCCT-tctaaaggtgaagaattattcactgg
E2AstrawGlcR	TTTAATAAAGCATAGTTGGTACATTGACCAGAACC-CTTGTACAGCTCGTCCATGC
HistagYFPF	CATCATCACCATCATCAC-taactcgagtcatgtaattagttatgtca
HistagYFPR	tttgtacaattcatccataccatgg
P2AdYFPF	TTT-tctaaaggtgaagaattattcactggt

P2AdR	TCCTGGATTTTCTTCAACATCAC
P2Av3YFPF	CAGGACGGAGACGTCGAGGAGAACCCTGGTCCTtctaaaggtgaaga attattcactggt
P2Av3mStrawberryR	CTTCAACAATGAGAAATTAGTTGCTCCTGATCCCTTGTACAG CTCGTCCATGC

Appendix Table A4-2: Primers used in this study (IDT)

Name	Primer 1	Primer 2	Template
P2A	P2AYFPGlcF	P2AstrawGlcR	P426-GAL-mStrawberry-PSIV-YFP
T2A	T2AYFPGlcF	T2AstrawGlcR	P426-GAL-mStrawberry- PSIV -YFP
E2A	E2AYFPGlcF	E2AstrawGlcR	P426-GAL-mStrawberry- PSIV -YFP
P2Ahis	HistagYFPF	HistagYFPR	P426-GAL-mStrawberry-P2A-YFP
E2Ahis	HistagYFPF	HistagYFPR	P426-GAL-mStrawberry-E2A-YFP
P2Av2	P2Av3YFPF	P2Av3mStrawberryR	P426-GAL-mStrawberry-P2A-YFPhisx6
P2Ad	P2AdYFPF	P2AdR	P426-GAL-mStrawberry-P2A-YFPhisx6

Appendix Table A4-3: PCR products generated in this study

Name	Construction Fragment	Figs used in
P426-GAL-mStrawberry-P2A-YFP	P2A	4-1
P426-GAL-mStrawberry-T2A-YFP	T2A	4-1
P426-GAL-mStrawberry-E2A-YFP	E2A	4-1
P426-GAL-mStrawberry-P2A-YFPhisx6	P2Ahis	4-2,3
P426-GAL-mStrawberry-E2A-YFPhisx6	E2Ahis	4-2
P426-GAL-mStrawberry-P2Av2-YFPhisx6	P2Av2	4-3
P426-GAL-mStrawberry-P2Ad-YFPhisx6	P2Ad	4-3

Appendix Table A4-4: Plasmids generated in this study

Appendix A5

Plasmid Name	Source
P416-GPDMCSrev2	(37)
pIRES-hrGFP	(275)
P413-GPD	(181)
P416-TEFmut7-YFP	(27)
pGPD$_0$1YFP	(37)
P426-GPD	(181)
pKT102	(182)
pKT127	(182)
pKT149	(182)

pmStrawberry	Amanda
pKT103	(182)
pKT120	(182)
CrPV Vector	Johnny
HIPV Vector	Johnny
PSIV Vector	Johnny
SWSS Plasmid	Ellington Lab
CrTMV Plasmid	Ellington Lab
Gypsy Plasmid	Ellington Lab
Omega plasmid	Jewett Lab
G20 plasmid	Jewett Lab
G22 plasmid	Jewett Lab
G35 plasmid	Jewett Lab
G38 plasmid	Jewett Lab
G93 plasmid	Jewett Lab
G94 plasmid	Jewett Lab
OmegaT7 plasmid	Jewett Lab
G20T7 plasmid	Jewett Lab
G22T7 plasmid	Jewett Lab
G35T7 plasmid	Jewett Lab
G93T7 plasmid	Jewett Lab
G101T7 plasmid	Jewett Lab
P423-GPD	(181)
P424-GPD	(181)
P425-GPD	(181)

Appendix Table A5-1: Plasmids used in this study

Primer Name	Sequence
HIS3 Fwd	gg-actagt-atgacagagcagaaagcc
HIS3 Rev	cc-atcgat-tctacataagaacactttggtgg
MCS-Fwd-XbaI	GC-TCTAGA-ATGTCTAAAGGTGAAGAATTATTCACTGG
YFP-BamHI-REV	CG-GGATCC-TTATTTGTACAATTCATCCATACCATGGG
Ires Fwd	g-gaattc-CGACTGCATAGGGTTACC
Ires Rev2	gc-tctaga-GGTTGTGGCCATTATCATCG
SacI IRES fwd	cccccc-gagctc-cgactgcatagggttacc
50N 2	CGAACTGGTTGTGG-gaattcNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN-tctaga-CGAGATACACTGCC
50N Fwd	CGAACTGGTTGTGGgaattc
50N Rev	GGCAGTGTATCTCG-tctaga
50NFwd3	gttccctccaccaaaggtgttcttatgtaggaattc
50NRev3	accagtgaataattcttcacctttagacattctaga

50N3	caaaggtgttcttatgtag-gaattc-NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN-tctaga-atgtctaaaggtgaagaat
IRESBBFwd	tctagaatgtctaaaggtgaagaattattc
IRESBBRev	gaattctttatcgatcgattctacataaga
IREShomFwd	tcttatgtagaatcgatcgataaagaattc
IREShomRev	gaataattcttcacctttagacattctaga
GPD fwd	CCCCCC-GAGCTC-AGTTTATCATTATCAATACTCGCCA
MCS-Rev-2	CCCCG-CTCGAG-TTATTTGTACAATTCATCCATACCATGGG
SacITadhF	CCCCCC-GAGCTC-GCGAATTTCTTATGATTTATGATTTTTATTATTAAAT
TadhHindIIISpeIR	gg-actagtaagcttt-CATGCCGGTAGAGGTGT
SpeImStrawberryF	CC-ACTAGT-ATGGTGAGCAAGGGCG
EcoRImStrawberryR	g-gaattc-tttatcgatcgatt-CTACTTGTACAGCTCGTCCAT
SpeIyECitrineF	CC-ACTAGT-ATGTCTAAAGGTGAAGAATTATTCACTGG
EcoRIyECitrineR	g-gaattc-tttatcgatcgatt-ttatttgtacaattcatccataccatgg
GPDsacF	CCC-GAGCTC-AGTTTATCATTATCAATACTCGCCATT
GPDR	GC-ACTAGT-TCTAGAATCCGTCGAAACTAAGT
StemloopInsert	GG-ACTAGT-GGATCC-ACTAGT-CC
CrPVshorthomF	aggtgttcttatgtagaatcgatcgataaagaattc-AGCAAAAATGTGATCTTGCTTGTA
HiPVshorthomF	ttcttatgtagaatcgatcgataaagaattc-AAAATGTGTGATCTGATTAGAAGTAAGAA
PSIVshorthomF	tcttatgtagaatcgatcgataaagaattc-GCTGACTATGTGATCTTATTAAAATTAGGT
CrPVshorthomR	acaccagtgaataattcttcacctttagatctaga-agcaagtgagattcttttcgcacaa
HiPVshorthomR	agtgaataattcttcacctttagatctaga-tgctagtgaaattctttagcctaaagtcca
PSIVshorthomR	cagtgaataattcttcacctttagatctaga-agcaggtaaatttcttaggtttttcgact
IRESBBFwd2	tctaaaggtgaagaattattcactgg
IRESBBRev2	gaattctttatcgatcgattCTACTTG
IRESmutF	CCGCCACTCCACCGGCGGCATGGACGAGCTGTACAAGTAGaatcgatcgataaagaattc
IRESmutR	atcaccatctaattcaaccaaaattgggacaacaccagtgaataattcttcacctttaga
p150F	GCATGGACGAGCTGTACAAGTAGaatcgatcgataaagaattc-CCCAGTTCGATCCTGGG
p150R	gtgaataattcttcacctttagacattctaga-TATTGTAATAGGTAATTACAGTTGTCCT
YAP1F	GGACGAGCTGTACAAGTAGaatcgatcgataaagaattc-TACCGATTAAGCACAGTACCT
YAP1R	gaataattcttcacctttagacattctaga-TTTAAGAAACAACTTTTCCTTCTTTAAACG

gypsyF	AGCTGTACAAGTAGaatcgatcgataaagaattc-gttcaaatcttgtgtcgaaataaacc
gypsyR	acaccagtgaataattcttcacctttagacattctaga-ttagattggtgggtcagattgt
SWSSF	GAGCTGTACAAGTAGaatcgatcgataaagaattc-ctaagcgatactttaattggtcact
SWSSR	accagtgaataattcttcacctttagacattctaga-cgaattgttgaagaacactgtaag
CrTMVF	CGGCATGGACGAGCTGTACAAGTAGaatcgatcgataaa-gaattcgtcgattcggttgca
CrTMVR	gtgaataattcttcacctttagacattctaga-tttcttctttcaaattaaacgaatcagg
SacIGalF	CCCCCC-GAGCTC-ACGGATTAGAAGCCGCC
GALHindIIIR	GG-actagt-aagcttt-ggtttttctccttgacgttaaagt
T7promF	TGGACGAGCTGTACAAGTAGaatcgatcgataaagaattc-TAATACGACTCACTATAGGG
OmegaWTF	TGTACAAGTAGaatcgatcgataaagaattc-AGTATTTTTACAACAATTACCAACAACAA
OmegaWTR	aattcttcacctttagacattctaga-ATGTAATTGTAAATAGTAATTGTAATGTTGTTTG
OmegaG20F	CTGTACAAGTAGaatcgatcgataaagaattc-AGTATTTTTACAACAATTACCAGTGAGA
OmegaG20R	gacaacaccagtgaataattcttcacctttagacattctaga-CCTCCCTCCTCTCTCACT
OmegaG22F	GTACAAGTAGaatcgatcgataaagaattc-AGTATTTTTACAACAATTACCAGTGTTATG
OmegaG22R	caacaccagtgaataattcttcacctttagacattctaga-CTCTCACCCCATAACACTGG
OmegaG35F	TACAAGTAGaatcgatcgataaagaattc-AGTATTTTTACAACAATTACCAGTTAAAGAA
OmegaG35R	ttcttcacctttagacattctaga-AAATTATATTCTTTAACTGGTAATTGTTGTAAAAAT
OmegaG38F	ACAAGTAGaatcgatcgataaagaattc-AGTATTTTTACAACAATTACCAGTTTAGTTAA
OmegaG38R	gaataattcttcacctttagacattctaga-CTTAACTTTAACTAAACTGGTAATTGTTGT
OmegaG93F	TGTACAAGTAGaatcgatcgataaagaattc-AGTATTTTTACAACAATTACCAGTTTGTG
OmegaG93R	gaataattcttcacctttagacattctaga-TTAATATTTTCACAAACTGGTAATTGTTGT
OmegaG94F	TGTACAAGTAGaatcgatcgataaagaattc-AGTATTTTTACAACAATTACCAGTAAGGT
OmegaG94R	cagtgaataattcttcacctttagacattctaga-ATACTCACAAACCTTACTGGTAATTG
OmegaG101F	TACAAGTAGaatcgatcgataaagaattc-AGTATTTTTACAACAATTACCAGTTAAAGAA
OmegaG101R	gaataattcttcacctttagacattctaga-CTTAACTTTAACTAAACTGGTAATTGTTGT
URE2F	GGACGAGCTGTACAAGTAGaatcgatcgataaagaattc-AATAACAACGGCAACCAAGTG
URE2R	ggacaacaccagtgaataattcttcacctttagacattctaga-TTCACCACGCAATGCCT
CrPV-Loop1F	NNNNNNN-tgcttgtaaatacaattttgagaggtt

CrPV-Loop1R	atttttgctttttagaagtcgtaaacct
CrPV-PSF	NNNNNNNNNN-gagagg-NNNNNNNN-ttacaagtagtgctatttttgtatttaggtt
CrPV-PSR	ttacaagcaagatcacatttttgctt
CrPV-connhp1F	NNNNN-atttttg-NNNNN-aggttagct-NNNN-agctttacgttccaggatgc
CrPV-connhp1R	tacttgtaatttattaacctctcaaaattgtatttac
CrPV-connhp2F	NNNNN-ttccaggatgcctagtgg-NNNNN-ccacaatatccaggaagccc
CrPV-connhp2R	agctaaatagctaacctaaatacaaaaatagc
CrPV-conn3F	NNNNNNNNN-tacctgctacatttcaagattctagat
CrPV-conn3R	ggtttttcgactacctaatctgaaaaa
HiPV-Loop1F	NNNNNNNN-attagaagtaagaaaattcctagttataatatttttaatactg
HiPV-Loop1R	acattttcgttgagcacaaagc
HiPV-PSF	NNNNNNNN-cctag-NNNNNN-tatttttaatactgctacatttttaagaccct
HiPV-PSR	tacttctaatcagatcacacattttcgt
HiPV-connhp1F	NNNNNN-acattttt-NNN-acccttagtt-NNNN-agctttaccgcccagga
HiPV-connhp1R	attaaaaatattataactaggaattttcttacttctaatcag
HiPV-connhp2F	NNNNN-gcccaggatggggtgcag-NNNNN-ctgcaatatccagggcacc
HiPV-connhp2R	agctaaataactaagggtcttaaaaatgtag
HiPV-conn3F	NNNNNNNNN-cactagcaaataataataattctagatctaaaggt
HiPV-conn3R	agcctaaagtccactaaaactacaa
PSIV-Loop1F	NNNNNNN-tattaaaattaggttaaatttcgaggttaaaaatagt
PSIV-Loop1R	atagtcagctttcttcttcaaagaagt
PSIV-PSF	NNNNNNNNNN-cgagg-NNNNNNN-tagttttaatattgctatagtcttagaggtct
PSIV-PSR	taattttaataagatcacatagtcagctttctt
PSIV-connhp1F	NNNNN-atagtc-NNNNN-ggtcttgtat-NNNN-atacttaccacacaagatggacc
PSIV-connhp1R	tattaaaactatttttaacctcgaaatttaacctaat
PSIV-connhp2F	NNNNN-acacaagatggaccggag-NNNNN-ctccaatatctagtgtaccctcgt
PSIV-connhp2R	gtataaatatacaagacctctaagactatagca
PSIV-conn3F	NNNNNNNNN-cacttcaagaaaaagaattttctagatctaaag
PSIV-conn3R	tcgcacaacaccacttaatgt
SUI2SpeIF	GG-actagt-ATGTCCACTTCTCATTGCAGAT
SUI2XhoIR	CCCCG-ctcgag-TTACTCGTCGTCTGACTCATCC
SUI3SpeIF	GG-actagt-ATGTCCTCCGATTTAGCTGC
SUI3XhoIR	CCCCG-ctcgag-TCACATTCTCCTTCTCTTACCAAC
GCD1SpeIF	GG-actagt-ATGTCAATTCAGGCTTTTGTCTTT
GCD1SpeIR	CCCCG-ctcgag-TTAACGCTCAAATAATCCGTCATCT

Appendix Table A5-2: Primers used in this study (IDT)

Name	Primer 1	Primer 2	Template
HIS3	HIS3 Fwd	HIS3 Rev	P413-GPD
YFP	MCS-Fwd-XbaI	YFP-BamHI-REV	P416-TEFmut7-YFP
EMCV	Ires Fwd	Ires Rev2	pIRES-hrGFP
EMCVlib	Ires Fwd	Ires Rev2	pIRES-hrGFP

IRESproms	SacI IRES fwd	Ires Rev2	Promising IRES Isolates
50N	50N Fwd	50N Rev	50N 2
M4lib	Ires Fwd	Ires Rev2	P416-GPD-HIS3-M4-YFP
IREShrlib	IREShomFwd	IREShomRev	P416-GPD-HIS3-EMCV-YFP
IRESlcBBhr	IRESBBFwd	IRESBBRev	P416-GPDMCSrev2
50Nhr	50NFwd3	50NRev3	50N 3
GPD-HIS3-EMCV-YFP	GPD fwd	MCS-Rev-2	P416-GPD-HIS3-EMCV-YFP
IREShcBBhr	IRESBBFwd	IRESBBRev	P426-GPD-HIS3-EMCV-YFP
Tadh	SacITadhF	TadhHindIIISpeIR	Yeast genome
CFP	SpeIyECitrineF	EcoRIyECitrineR	pKT102
GFP	SpeIyECitrineF	EcoRIyECitrineR	pKT127
Sapphire	SpeIyECitrineF	EcoRIyECitrineR	pKT149
mStrawberry	SpeImStrawberryF	EcoRImStrawberryR	pmStrawberry
Venus	SpeIyECitrineF	EcoRIyECitrineR	pKT103
YFPup	SpeIyECitrineF	EcoRIyECitrineR	pKT120
GPD	GPDsacF	GPDR	P426-GPD-HIS3-EMCV-YFP
stem	StemloopInsert	None	None
CrPVhr1	CrPVshorthomF	CrPVshorthomR	CrPV Vector
HiPVhr1	HiPVshorthomF	HiPVshorthomR	HIPV Vector
PSIVhr1	PSIVshorthomF	PSIVshorthomR	PSIV Vector
IREShcBBterm	IRESBBFwd	IRESBBRev	P426-Tadh-HIS3-YFP
EMCVhr	IRESmutF	IRESmutR	
CrPVhr2	IRESmutF	IRESmutR	P426-Tadh-HIS3-CrPV-YFP
HiPVhr2	IRESmutF	IRESmutR	P426-Tadh-HIS3-HIPV-YFP
PSIVhr2	IRESmutF	IRESmutR	P426-Tadh-HIS3-PSIV-YFP
50NB4hr	IRESmutF	IRESmutR	IRES isolate from Library 4
50ND7hr	IRESmutF	IRESmutR	IRES isolate from Library 4
50ND3hr	IRESmutF	IRESmutR	IRES isolate from Library 4
50NB8hr	IRESmutF	IRESmutR	IRES isolate from Library 4
IRESDFBBhrterm	IRESBBFwd2	IRESBBRev2	P426-Tadh-mStrawberry-YFP
Stem-mStrawberry	HindIII	EcoRI	P426-GPD-stem-mStrawberry-YFP

P150	p150F	p150R	Yeast genome
YAP1	YAP1F	YAP1R	Yeast genome
SWSS	SWSSF	SWSSR	SWSS Plasmid
CrTMV	CrTMVF	CrTMVR	CrTMV Plasmid
Gypsy	gypsyF	gypsyR	Gypsy Plasmid
IRESDFBBhr	IRESBBFwd2	IRESBBRev2	P426-GPD-mStrawberry-YFP
Isolatehr	IRESmutF	IRESmutR	IRES isolate
GAL	SacIGalF	GALHindIIIR	Yeast genome
Omega	OmegaWTF	OmegaWTR	Omega plasmid
G20	OmegaG20F	OmegaG20R	G20 plasmid
G22	OmegaG22F	OmegaG22R	G22 plasmid
G35	OmegaG35F	OmegaG35R	G35 plasmid
G38	OmegaG38F	OmegaG38R	G38 plasmid
G93	OmegaG93F	OmegaG93R	G93 plasmid
G94	OmegaG94F	OmegaG94R	G94 plasmid
omegaT7	T7promF	OmegaWTR	OmegaT7 plasmid
G20T7	T7promF	OmegaG20R	G20T7 plasmid
G22T7	T7promF	OmegaG22R	G22T7 plasmid
G35T7	T7promF	OmegaG35R	G35T7 plasmid
G93T7	T7promF	OmegaG93R	G93T7 plasmid
G101T7	T7promF	OmegaG101R	G101T7 plasmid
URE2	URE2F	URE2R	Yeast genome
GALBBhr	IRESBBFwd2	IRESBBRev2	P426-GAL-mStrawberry-IRES-YFP
CrPVSD1	CrPV- Loop1F	CrPV- Loop1R	P426-GAL-mStrawberry-CrPV-YFP
CrPVSD2	CrPV- PSF	CrPV- PSR	P426-GAL-mStrawberry-CrPV-YFP
CrPVSD3	CrPV- connhp1F	CrPV- connhp1R	P426-GAL-mStrawberry-CrPV-YFP
CrPVSD4	CrPV- connhp2F	CrPV- connhp2R	P426-GAL-mStrawberry-CrPV-YFP
CrPVSD5	CrPV- conn3F	CrPV- conn3R	P426-GAL-mStrawberry-CrPV-YFP
HIPVSD1	HIPV- Loop1F	HIPV- Loop1R	P426-GAL-mStrawberry-HiPV-YFP
HIPVSD2	HIPV- PSF	HIPV- PSR	P426-GAL-mStrawberry-HiPV-YFP
HIPVSD3	HIPV- connhp1F	HIPV- connhp1R	P426-GAL-mStrawberry-HiPV-YFP
HIPVSD4	HIPV- connhp2F	HIPV- connhp2R	P426-GAL-mStrawberry-HiPV-YFP
HIPVSD5	HIPV- conn3F	HIPV- conn3R	P426-GAL-

			mStrawberry-HiPV-YFP
PSIVSD1	PSIV- Loop1F	PSIV- Loop1R	P426-GAL-mStrawberry-PSIV-YFP
PSIVSD2	PSIV- PSF	PSIV- PSR	P426-GAL-mStrawberry-PSIV-YFP
PSIVSD3	PSIV- connhp1F	PSIV- connhp1R	P426-GAL-mStrawberry-PSIV-YFP
PSIVSD4	PSIV- connhp2F	PSIV- connhp2R	P426-GAL-mStrawberry-PSIV-YFP
PSIVSD5	PSIV- conn3F	PSIV- conn3R	P426-GAL-mStrawberry-PSIV-YFP
SUI2	SUI2SpeIF	SUI2XhoIR	Yeast genome
SUI3	SUI3SpeIF	SUI3XhoIR	Yeast genome
GCD1	GCD1SpeIF	GCD1SpeIR	Yeast genome

Appendix Table A5-3: PCR products generated in this study

Name	Construction Fragments	Figs Used in
P416-GPD-HIS3-EMCVlib-YFP	IRESlcBBhr, IREShrlib	5-10
P416-GPD-HIS3-50N-YFP	IRESlcBBhr, 50Nhr	5-10
P426-GPD-HIS3-EMCVlib-YFP	IREShcBBhr, IREShrlib	5-11,12
P426-GPD-HIS3-50N-YFP	IREShcBBhr, 50Nhr	5-11,12
P426-Tadh-HIS3-CrPV-YFP	CrPVhr1, IREShcBBterm	
P426-Tadh-HIS3-HIPV-YFP	HIPVhr1, IREShcBBterm	
P426-Tadh-HIS3-PSIV-YFP	PSIVhr1, IREShcBBterm	
P426-Tadh-mStrawberry-EMCV-YFP	EMCVhr, IRESDFBBhrterm	5-16
P426-Tadh-mStrawberry-CrPV-YFP	CrPVhr2, IRESDFBBhrterm	5-16
P426-Tadh-mStrawberry-HiPV-YFP	HiPVhr2, IRESDFBBhrterm	5-16
P426-Tadh-mStrawberry-PSIV-YFP	PSIVhr2, IRESDFBBhrterm	5-16
P426-Tadh-mStrawberry-50NB4-YFP	50NB4hr, IRESDFBBhrterm	5-16
P426-Tadh-mStrawberry-50ND7-YFP	50ND7hr, IRESDFBBhrterm	5-16
P426-Tadh-mStrawberry-50ND3-YFP	50ND3hr, IRESDFBBhrterm	5-16
P426-Tadh-mStrawberry-50NB8-YFP	50NB8hr, IRESDFBBhrterm	5-16
P426-Tadh-mStrawberry-P150-YFP	P150, IRESDFBBhrterm	5-17
P426-Tadh-mStrawberry-YAP1-YFP	YAP1, IRESDFBBhrterm	5-17
P426-Tadh-mStrawberry-SWSS-YFP	SWSS, IRESDFBBhrterm	5-17
P426-Tadh-mStrawberry-CrTMV-YFP	CrTMV, IRESDFBBhrterm	5-17
P426-Tadh-mStrawberry-gypsy-YFP	Gypsy, IRESDFBBhrterm	5-17
P426-GPD-mStrawberry-P150-YFP	P150, IRESDFBBhr	5-17
P426-GPD-mStrawberry-YAP1-YFP	YAP1, IRESDFBBhr	5-17
P426-GPD-mStrawberry-SWSS-YFP	SWSS, IRESDFBBhr	5-17
P426-GPD-mStrawberry-CrTMV-YFP	CrTMV, IRESDFBBhr	5-17
P426-GPD-mStrawberry-gypsy-YFP	Gypsy, IRESDFBBhr	5-17
P426-Tadh-mStrawberry-IRES-YFP	Isolatehr, IRESDFBBhrterm	5-18

P426-GPD-mStrawberry-IRES-YFP	Isolatehr, IRESDFBBhr	5-19
P426-GAL-mStrawberry-URE2-YFP	URE2, GALBBhr	5-22
P426-GAL-mStrawberry-Omega-YFP	Omega, GALBBhr	5-23
P426-GAL-mStrawberry-G20-YFP	G20, GALBBhr	5-23
P426-GAL-mStrawberry-G22-YFP	G22, GALBBhr	5-23
P426-GAL-mStrawberry-G35-YFP	G35, GALBBhr	5-23
P426-GAL-mStrawberry-G38-YFP	G38, GALBBhr	5-23
P426-GAL-mStrawberry-G93-YFP	G93, GALBBhr	5-23
P426-GAL-mStrawberry-G94-YFP	G94, GALBBhr	5-23
P426-GAL-mStrawberry-omegaT7-YFP	omegaT7, GALBBhr	5-23
P426-GAL-mStrawberry-G20T7-YFP	G20T7, GALBBhr	5-23
P426-GAL-mStrawberry-G22T7-YFP	G22T7, GALBBhr	5-23
P426-GAL-mStrawberry-G35T7-YFP	G35T7, GALBBhr	5-23
P426-GAL-mStrawberry-G93T7-YFP	G93T7, GALBBhr	5-23
P426-GAL-mStrawberry-G101T7-YFP	G101T7, GALBBhr	5-23

Appendix Table A5-4: Plasmids generated through homolgous recombination

Name	Construction Fragments	Figs used in
P426-GAL-mStrawberry-CrPVSD1-YFP	CrPVSD1	5-25
P426-GAL-mStrawberry-CrPVSD2-YFP	CrPVSD2	5-25
P426-GAL-mStrawberry-CrPVSD3-YFP	CrPVSD3	5-25
P426-GAL-mStrawberry-CrPVSD4-YFP	CrPVSD4	5-25
P426-GAL-mStrawberry-CrPVSD5-YFP	CrPVSD5	5-25
P426-GAL-mStrawberry-HIPVSD1-YFP	HIPVSD1	5-25
P426-GAL-mStrawberry-HIPVSD2-YFP	HIPVSD2	5-25
P426-GAL-mStrawberry-HIPVSD3-YFP	HIPVSD3	5-25
P426-GAL-mStrawberry-HIPVSD4-YFP	HIPVSD4	5-25
P426-GAL-mStrawberry-HIPVSD5-YFP	HIPVSD5	5-25
P426-GAL-mStrawberry-PSIVSD1-YFP	PSIVSD1	5-25
P426-GAL-mStrawberry-PSIVSD2-YFP	PSIVSD2	5-25
P426-GAL-mStrawberry-PSIVSD3-YFP	PSIVSD3	5-25
P426-GAL-mStrawberry-PSIVSD4-YFP	PSIVSD4	5-25
P426-GAL-mStrawberry-PSIVSD5-YFP	PSIVSD5	5-25

Appendix Table A5-5: Plasmids generated through phosphorylation-ligation

Name	Insert	Backbone	RE1	RE2	Figs used in
P416-GPD-HIS3	HIS3	P416-GPDMCSrev2	ClaI	SpeI	
P416-GPD-HIS3-YFP	YFP	P416-GPD-HIS3	XbaI	BamHI	
P416-GPD-HIS3-EMCV-YFP	EMCV	P416-GPD-HIS3-YFP	XbaI	EcoRI	5-1,2,4,5,6,7,8,9

P416-GPD-HIS3-EMCVlib-YFP	EMCVlib	P416-GPD-HIS3-YFP	XbaI	EcoRI	5-1,2,4,5,8,9
P416-IRES- YFP	IRESproms	$pGPD_01YFP$	SacI	XbaI	5-3
P416-GPD-HIS3-50N-YFP	50Nlib	P416-GPD-HIS3-YFP	XbaI	EcoRI	5-4,5,8,9
P416-GPD-HIS3-M4lib-YFP	M4lib	P416-GPD-HIS3-YFP	XbaI	EcoRI	5-6,7
P426-GPD-HIS3-EMCV-YFP	GPD-HIS3-EMCV-YFP	P426-GPD	SacI	XhoI	5-10
P426-Tadh-HIS3-IRES-YFP	Tadh	High Copy IRES Isolates	SacI	SpeI	5-13
P426-Tadh-HIS3-YFP	Tadh	P416-GPD-HIS3-YFP	SacI	SpeI	5-14
P426-Tadh-CFP-YFP	CFP	P426-Tadh-HIS3-YFP	SpeI	EcoRI	5-14
P426-Tadh-GFP-YFP	GFP	P426-Tadh-HIS3-YFP	SpeI	EcoRI	5-14
P426-Tadh-Sapphire-YFP	Sapphire	P426-Tadh-HIS3-YFP	SpeI	EcoRI	5-14
P426-Tadh-mStrawberry-YFP	mStrawberry	P426-Tadh-HIS3-YFP	SpeI	EcoRI	5-14
P426-Tadh-Venus-YFP	Venus	P426-Tadh-HIS3-YFP	SpeI	EcoRI	5-14
P426-Tadh-YFP-YFP	YFPup	P426-Tadh-HIS3-YFP	SpeI	EcoRI	5-14
P426-GPD-mStrawberry-YFP	GPD	P426-Tadh-mStrawberry-YFP	SpeI	SacI	
P426-GPD-stem-mStrawberry-YFP	Stem (4x)	P426-GPD-mStrawberry-YFP	SpeI	None	
P426-GPD-mStrawberry-EMCV-YFP	Tadh	P426-TADH-mStrawberry-EMCV-YFP	SacI	SpeI	5-16
P426-GPD-mStrawberry-CrPV-YFP	Tadh	P426-TADH-mStrawberry-CrPV-YFP	SacI	SpeI	5-16
P426-GPD-mStrawberry-HiPV-YFP	Tadh	P426-TADH-mStrawberry-HiPV-YFP	SacI	SpeI	5-16
P426-GPD-mStrawberry-PSIV-YFP	Tadh	P426-TADH-mStrawberry-PSIV-YFP	SacI	SpeI	5-16
P426-GPD-mStrawberry-	Tadh	P426-TADH-mStrawberry-	SacI	SpeI	5-16

50NB4-YFP		50NB4-YFP			
P426-GPD-mStrawberry-50ND7-YFP	Tadh	P426-TADH-mStrawberry-50ND7-YFP	SacI	SpeI	5-16
P426-GPD-mStrawberry-50ND3-YFP	Tadh	P426-TADH-mStrawberry-50ND3-YFP	SacI	SpeI	5-16
P426-GPD-mStrawberry-50NB8-YFP	Tadh	P426-TADH-mStrawberry-50NB8-YFP	SacI	SpeI	5-16
P426-GPD-STEM-mStrawberry-EMCV-YFP5	Stem-mStrawberry	P426-GPD-mStrawberry-EMCV-YFP	HindIII	EcoRI	5-16
P426-GPD-STEM-mStrawberry-CrPV-YFP	Stem-mStrawberry	P426-GPD-mStrawberry-CrPV-YFP	HindIII	EcoRI	5-16
P426-GPD-STEM-mStrawberry-HiPV-YFP	Stem-mStrawberry	P426-GPD-mStrawberry-HiPV-YFP	HindIII	EcoRI	5-16
P426-GPD-STEM-mStrawberry-PSIV-YFP	Stem-mStrawberry	P426-GPD-mStrawberry-PSIV-YFP	HindIII	EcoRI	5-16
P426-GPD-STEM-mStrawberry-50NB4-YFP	Stem-mStrawberry	P426-GPD-mStrawberry-50NB4-YFP	HindIII	EcoRI	5-16
P426-GPD-STEM-mStrawberry-50ND7-YFP	Stem-mStrawberry	P426-GPD-mStrawberry-50ND7-YFP	HindIII	EcoRI	5-16
P426-GPD-STEM-mStrawberry-50ND3-YFP	Stem-mStrawberry	P426-GPD-mStrawberry-50ND3-YFP	HindIII	EcoRI	5-16
P426-GPD-STEM-mStrawberry-50NB8-YFP	Stem-mStrawberry	P426-GPD-mStrawberry-50NB8-YFP	HindIII	EcoRI	5-16
P426-GPD- STEM-mStrawberry-P150-YFP	Stem-mStrawberry	P426-GPD-mStrawberry-P150-YFP	HindIII	EcoRI	5-16
P426-GPD- STEM-mStrawberry-YAP1-YFP	Stem-mStrawberry	P426-GPD-mStrawberry-YAP1-YFP	HindIII	EcoRI	5-16
P426-GPD- STEM-mStrawberry-SWSS-YFP	Stem-mStrawberry	P426-GPD-mStrawberry-SWSS-YFP	HindIII	EcoRI	5-16
P426-GPD- STEM-mStrawberry-CrTMV-YFP	Stem-mStrawberry	P426-GPD-mStrawberry-CrTMV-YFP	HindIII	EcoRI	5-16

P426-GPD- STEM-mStrawberry-gypsy-YFP	Stem-mStrawberry	P426-GPD-mStrawberry-gypsy-YFP	HindIII	EcoRI	5-16
P426-GPD- STEM-mStrawberry-IRES-YFP	Stem-mStrawberry	P426-GPD-mStrawberry-IRES-YFP	HindIII	EcoRI	5-16
P426-GAL-mStrawberry-IRES-YFP	GAL	P426-GPD-STEM-mStrawberry-IRES-YFP	SacI	SpeI	5-20
P423-GPD-SUI2	SUI2	P423-GPD	SpeI	XhoI	5-26,27
P424-GPD-SUI3	SUI3	P424-GPD	SpeI	XhoI	5-26,27
P425-GPD-GCD1	GCD1	P425-GPD	SpeI	XhoI	5-26,27

Appendix Table A5-6: Plasmids generated through restriction-ligation

Name	Sequence
M4	cgactgcatagggttaccCCCCTCTCCCTCCCCCCCCCCTAACGTTACTGGCCGA AGCCGCTTGGAATAAGGCCGGTGTGCGTTTGTCCATATGTTATTTTCC ACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTGGCCCT GTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGA ATGCAAGGTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGC TTCTTAAAGACAAACAACGTCTGTAGCGACCCTTTGCAGGCAGCGGA ACCCCCCACCTGGCGACAGGTGCCTCTGCGGCCAAAAGCCACGTGTA TAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGAG TTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCA ACAAGGGGCTGAAGGATGCCCAGAAGGTACCCCATTGTATGGGATCT GATCTGGGGCCTCGGTGCACATGCTTTACATGTGTTTAGTCGAGGTTA AAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGTTTTCCTTTGA AAAACACGATGATAATGGCCACAACC
50NB4	GAATCCACTACTGGCAATAGAATACACTGTTTAGATCGCCGCGTCAC GGA
50ND7	CGGGCCTCCCTTCGTATGTCTGCCTTTCACTGTGGGACGACTGTCACG GC
50ND3	TCTGCGAAGCAACGCACCGGTCCCGGAACAGCACGAGTTGCCTTTGC GGC
50NB8	GTGTTAATCCTAAGTGTTTGGATCCGCGGGCAAAGTGAACGATGGAC TGT
HM3	tgagagtttcaaacattgtgcgaagcttcttggctcgttccctctgttcttggacagagagtccgccagaagctttgtgc tcaacgaaaatgtgtgatctgattagaagtaggaaaattcctagttataatatttttaatactgctacatttttaagaccctt agttatttagctttaccgcccaggatggggtgcagcgttcctgcaatatccagggcacctaggtgcagccttgtagttt tagtggactttaggctaaagaatttcactagcaaataataataat
SM7	ctaagcgatactttaattggtcactctttcttgtgatccaagaattgtcatggtgagcagctcccaattgagcctccacc gaagcgatttcagcttcagtcacagcaggctcatctccgttagacgacattatCtactcctccttaaatgcagtgattg caaacaataaatcgtcccttgtatcgcctcttttatacgtttaacagcgtatgttgcagattctccgtttagtatatgggtg tgtttatatctagatttacatacaatatcaatatctttactctccccttcaacaatcacattaaagaatatgactacgcgttct

	cctgttcgattaaatgcttcaaagcttagcacgtcattaactacgtcactacgagacacaagaatcactgcatctacatc ctcatcatttatggaacgaaacaattttgccgttattttaaaatctcccaagggaatggagaaatcgtaccttgtaaattct ccatcgtcaccaccaattctcaaattgtacgtagcaacactacaccgccctacttgaccagtttcttgttgaatttttcgtc tatatctgtgcaagatattgggggccgataccaatgtacaagattttgcgttagttttcgtttagccaagtaatcagttatt atttgtctaacGtctattccttgtttttcaattgcctctttttatgtcttctaggcgagatctcacatccagttctgcatccttc atgtaaactctgtctccgtgacttacagtgttcttcaacaattcg

Appendix Table A5-7: Selected IRES mutants generated in this study

Appendix A6

Primer Name	Sequence
3'LTRF	TGTTGGAATAGAAATCAACTATCATCT
AftACE2R	CTAAGAGTCTGTTTAGATCAACAGTCT
AftCPK*CAN1*R	GAGTTCTGCCCTTGGCTTCC
AftGRE3rev	CTGTTTGACGCACTGATGGGT
AftPBSF	TGTGCTTCGGTTACTTCTAAGG
AftRRM1R	AGGATTCTCCGAATAACCTCTAGC
AInoass2F	TCTAGAGGATCCCCGGGTACCGAGCTCGAATTTTTACTAACAAATG GTATTATTTATAAC
AInoass3F	GTATGTTAATATGGACTAAAGGAGGCTTTTCTGCAGGTCGACTCTA GAGGATCCCCGGG
Alg9F	ATCGTGAAATTGCAGGCAGCTTGG
Alg9R	CATGGCAACGGCAGAAGGCAATAA
APL23'F	CCCCCC-ACTAGT-CTATAAACGTCCGTTGTAGTGAACT
APL23'R	CCCCCC-CCGCGG-CCTGACATCTTTGGACGTGG
APL25'F	CCCCCC-CGTACG-TATCCTGATGGAGCACTTCG
APL25'R	CCCCCC-GTCGAC-AGTTGAAACTGTTTTTTAAGTGCAGT
APL2BegR	CCAACTCAACGCATAAATCC
APL2EndF	TAACGATGATGTGCTATTGG
ARTrev	TTATGCAATCAGGTGAATACGTTTCTT
BefCPK*CAN1*F	GGGTTTCTGTGTGGTTTCCG
BefGRE3for	ATGGGCGCATTACTACAAGAAG
BefICE2F	ATGATTCACTGTCACTTAGTGAGC
BefPPTR	ATAATGTAATAGATCGCGGCC
BefRRM1F	CATAGAACCGAGTGTAACACCA
BefRTR	ATTCTTAGTATTCCATGTGTCTCG
BegCKB2R	AACGATCAGTGATGTACTCAGG
BegCPK*CAN1*R	TCAGCGTTCTGTACTTCTCC
BegGRE3rev	GACCTTCGGAGATGGCTTTC

BegMRC1F	CAGGTATTCTTTGCGTCTGCG
BY*CAN13*'F	CCCCCC-ACTAGT-ATATGACGTTTTATTACCTTTGATCACATT
BY*CAN13*'R	CCCCCC-CCGCGG-ACCATCGTTCTGGCTGAATATAG
BY*CAN15*'F	CCCCCC-CGTACG-ATGGAACACGGAGTAAAATATTGTGT
BY*CAN15*'R	CCCCCC-GTCGAC-TGCTATGCCTTTTTTTTTTTTTGTTTTTAC
BYHIR33'F	CCCCCC-ACTAGT-GATGACCATATTTTGGAAGAAGTGTG
BYHIR33'R	CCCCCC-CCGCGG-CAAATCTTTATCGTAATCAGATAATTTTTCCAA
BYHIR35'F	CCCCCC-CGTACG-ACTAGCAATGATTCCGTTTTACATTT
BYHIR35'R	CCCCCC-GTCGAC-AATAAGCTTTATCTAGAATCTGTGTTGAGG
CAC23'F	CCCCCC-ACTAGT- TTTTTAATATATTTAATGCGGTACATAAGAATGCC
CAC23'R	CCCCCC-CCGCGG-TCACGAGAGATGAGTCCACC
CAC25'F	CCCCCC-CGTACG-AGAAAGGTCCTCAGATTGAGC
CAC25'R	CCCCCC-GTCGAC-TGTCCTGCCCCTTTGCT
CAC2BegR	GATTCTTGCTGTGTATTTGG
CAC2EndF	TGATTTAGCATGGTCTGAGG
CAC33'F	CCCCCC-ACTAGT-CCTAAACGTTCCTTGAAGCCA
CAC33'R	CCCCCC-CCGCGG-GTTCGGCTTTGGACATTTCG
CAC35'F	CCCCCC-CGTACG-GTGGTTTGTTGTCTGTCTGG
CAC35'R	CCCCCC-GTCGAC- CTTTGAAACTAAATTTGTATATTGTTTGTCAGAA
CAC3BegR	TTGGGAAGATGTAAATGAGG
CAC3EndF	AAGAAGATGGGTTAGTCAAGC
CPK*CAN13*'F	CCCCCC-ACTAGT-ATATGACGTTTTATTACCTTTAATCACATTCC
CPKCKB23'F	CCCCCC-ACTAGT-CTGTTAAAGAAAGAAGAAAAAATTCGAAATGA
CPKCKB23'R	CCCCCC-CCGCGG-AACCGGCAATGAATAAAGTGTC
CPKCKB25'F	CCCCCC-CGTACG-CTGCGTAAGTTTATTTATGAGTTTGTGT
CPKCKB25'R	CCCCCC-GTCGAC-GATCTGCTTTTCTATCAGTTCTCTAAGT
CPKMRC13'F	CCCCCC-ACTAGT-TGGTTTTTTATCTTTTCCGAAGAAGTT
CPKMRC13'R	CCCCCC-CCGCGG-CCGTGCTCACCTAAAAAACAAC
CPKMRC15'F	CCCCCC-CGTACG-TATCATTTTGAAAGCCAGAGATTTGATC
CPKMRC15'R	CCCCCC-GTCGAC-CACTAAAATATTTGGTGATAAGTTCAAAAGC
CUP1-1(2)Conf3'Fwd 2	CCTCGACATCATCTGCCC
CUP1-1(2)Conf5'Rev2	GGATGTATGGGCTAAATG
CYCtseq1	AACTCATGTGCCCTTGGTGG
EcoRITkc6XK	GCCC-GAATTC-

Sfor	TTTGAAAAAATTTATTTCTAGACAGTTATATAAAAAAAAA-TTAGA
EcoRITkc6XKSHRfor1	TAATCGGTG-GAATTC-TTTGAAAAAATTTATTTCTAGACAGTTATATAAAAAAAAATTAGA
EcoRITkc6XKSHRfor2	AGatcgataagcttgggCTGCAGCTTTAA-TAATCGGTG-GAATTC-TTTGAAAAAATTTATT
EcoRITkc6XKSHRshort	AGatcgataagcttgggCTG
EcoRITkc6XKSshort	GCCC-GAATTC-TTTGAAAAAATTTATTTCT
ELG1F	TCCC-CCCGGG-ATGAAAAGGCACGTGTCTTTAT
ELG1R	CCCCG-CTCGAG-TTATTTGTTCTTTGAAAAGCCTGAG
EndCKB2F	GACAAGGTGGCAAGAACTACAACG
EndCPK*CAN1*F	TTCCGATAGAAGAGACATTGAGGC
EndGRE3for	AACCATCCAGGCAGTACCAC
EndMRC1R	TTACAAGTACGGCTACTGACC
Gal1pFixR2	TATACTAGAAGTTCTCCTCGAGGCGGTAGAGGAATAAGAAGTAATACAAACCGAAAATGT
GALLTRR	TGAGAATTGGGTGAATGTTGAG
GalpF1	CTCACTAAAGGGAACAAAAGCTGGAGCTCCTAGTACGGATTAGAAGCCGC
GalpF2	CAGCTATGACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAAGGGAACAAAAGCT
GalpR2	TTCCATTGTTGATAAAGGCTATAATATTAGGTATACAGAATATACTAGAAGTTCTCCTCG
GRE3KOfor	GTAATATAAATCGTAAAGGAAAATTGGAAATTTTTTAAAGatagcttcaaaatgtttcta
GRE3KOleuF	TGTAATATAAATCGTAAAGGAAAATTGGAAATTTTTTAAA-atgtctgcccctaagaagat
GRE3KOrevnew	TTGTTCATATCGTCGTTGAGTATGGATTTTACTGGCTGGA-ttaagcaaggattttcttaa
HconnTRnaseF	ATCAATTGGAAAAAGAACCTATCGTTGGTGCCGAAACTTTT-GATGCTTCGTATGGCAACC
HIR3BegR	ATCTAGCGTAGGAGAAGAATTGC
HIR3EndF	ATTTGACAGCGTTTGCTTGG
His3AIF	ATCGATAAGCTTGGGCTGC
His3AIgenomeflankF	ATCGATAAGCTTGGGCTGC
HispromCANF	ACTAAAAAATGAGCAGGCAAGATAAACGAAGGCAAAG-ATGACAAATTCAAAAGAAGACGC
HISpromF	CTTTGCCTTCGTTTATCTTGCC
HISpromR	ACCACCCATAATGTAATAGATCGCGGCCGCCTCTAGTACACTCTATATTTTTTTATGCCT

HispromSPT15F	AAAAAATGAGCAGGCAAGATAAACGAAGGCAAAG-atggccgatgaggaacgt
HispromXylAF	AAATGAGCAGGCAAGATAAACGAAGGCAAAG-ATGGCTAAAGAATATTTCCCTCAAATTC
HistermCANR	GATAAGCTTGGGCTGCAGCTTTAAATAATCGGTG-CTATGCTACAACATTCCAAAATTTGT
HistermR	CACCGATTATTTAAAGCTGCAGCCC
HistermSPT15R	aagcttgggCTGCAGCTTTAAATAATCGGTG-tcacatttttctaaattcacttagcaca
HistermXylA3R	ataagcttgggCTGCAGCTTTAAATAATCGGTG-TTATTGATACATCGCGATAATAGCCT
HistermXylAR	gataagcttgggCTGCAGCTTTAAATAATCGGTG-ttaTTGATACATCGCGACAATAGCC
HIVA343TQCF	gaagaaaaagtctgttaccgttttggatgttggtgatgctt
HIVA343TQCR	aagcatcaccaacatccaaaacggtaacagacttttttcttc
HIVnoATGF	GCTTTGAAGGCTGTTCCA
HIVRThomR	ACCGATTATTTAAAGCTGCAGCCCAAGCTTATCGATTTACAAGATCTTTCTAATACCGGC
HIVT1361GG1362AF	tatgttgatggtgctgctaacagagaaactaagttgggtaaagctg
HIVT1361GG1362AR	Cagctttacccaacttagtttctctgttagcagcaccatcaacata
HSX1F	TCCC-CCCGGG-GTTCCGTTGGCGTAATGG
HSX1R	CCCCG-CTCGAG-CGTTCCGTACGGGACT
Int1delF	AAATATCCTTATCCTTTCATTCATCGAATGTATCCATTACACGACCGTCGCGAGGACTCT
Int1delR	AGAGTCCTCGCGACGGTCGTGTAATGGATACATTCGATGAATGAAAGGATAAGGATATTT
Int1stopF	acaacaaaattccgttgggtttaaccattacacgaccgtc
Int1stopR	gacggtcgtgtaatggttaaacccaacggaattttgttgt
Int2delF	GATGAGACAACAAAATTCCGTTGGGTTTATGAATTTTCTACTATTGTGAGAAATTCACTA
Int2delR	TAGTGAATTTCTCACAATAGTAGAAAATTCATAAACCCAACGGAATTTTGTTGTCTCATC
Int2stopF	gaaccatttatggttctctgcaatctaattttctactattgtgagaaattc
Int2stopR	gaatttctcacaatagtagaaaattagattgcagagaaccataaatggttc
Int3delF	CCGAACCATTTATGGTTCTCTGCAATCGAATACGACGCACTCACTTTCGATGAAGACTTA
Int3delR	TAAGTCTTCATCGAAAGTGAGTGCGTCGTATTCGATTGCAGAGAACCATAAATGGTTCGG
Int3stopF	aggaatccagattagatcaattcaattaagacgcactcactttc
Int3stopR	gaaagtgagtgcgtcttaattgaattgatctaatctggattcct

Int4delF	AAGGAATCCAGATTAGATCAATTCAATTACAAGAAGAGATCTAGC ACCCCCCAAATTTCC
Int4delR	GGAAATTTGGGGGGTGCTAGATCTCTTCTTGTAATTGAATTGATCT AATCTGGATTCCTT
Int4stopF	aacatatctgaatctaatattcttccatcatagaagagatctagcac
Int4stopR	gtgctagatctcttctatgatggaagaatattagattcagatatgtt
Int5delF	ATATCTGAATCTAATATTCTTCCATCAAAGCCGGAAAATAATTCAT CGCACAATATTGTT
Int5delR	AACAATATTGTGCGATGAATTATTTTCCGGCTTTGATGGAAGAATA TTAGATTCAGATAT
Int5stopF	cgttcaccttcaatcgatgcttctccataggaaaataattcatcgc
Int5stopR	gcgatgaattattttcctatggagaagcatcgattgaaggtgaacg
Int6delF	CGTTCACCTTCAATCGATGCTTCTCCACCGATGCGTAGTTTAGAAC CTCCGAGATCGAAG
Int6delR	CTTCGATCTCGGAGGTTCTAAACTACGCATCGGTGGAGAAGCATC GATTGAAGGTGAACG
Leufor2	cggtagtgttagacctgaacaag
Leurev1	ggtgggttgggttcttaactag
LTRCYCF	GAATATACTAAAAAATGAGCAGGCAAGATAAACGAAGGCAAAG-atttggcgagcgttggt
LTRCYCF2	CTATTCCAACATACCACCCATAATGTAATAGATCgcggccgc-atttggcgagcgttggt
LTRF1	TGTACTAGAGGATCTATTACATTATGGGTGGTATGTTGGAATAGAA ATCAACTATCATCT
LTRflankR	AGGGTTTTCCCAGTCACG
LTRGPDF	AAATGAGCAGGCAAGATAAACGAAGGCAAAG-agtttatcattatcaatactcgccatttc
LTRGPDF2	ACCACCCATAATGTAATAGATCgcggccgc-agtttatcattatcaatactcgccatttc
LTRR1	CGCGCGTAATACGACTCACTATAGGGCGAATTGGGTACCTGAGAA ATGGGTGAATGTTGA
LTRTEFF	AAAATGAGCAGGCAAGATAAACGAAGGCAAAG-atagcttcaaaatgtttctactcctttt
LTRTEFF2	TACCACCCATAATGTAATAGATCgcggccgc-atagcttcaaaatgtttctactcctttt
M13seq1	ctacgtgaaccatcacccta
MCS1For	ATAAACTAGTCGCAATAAGTGATGCTTCGTATGGCAACCA
MCS1Rev	GCGATGTATCAATAAGGATCCGCGGAATTCGGGCACCGATTATTT AAA
MCS2For	TTTAAATAATCGGTGCCCGAATTCCGCGGATCCTTATTGATACATC GC
MCS2Rev	ATCTAGATAAAGACTTCAAAGTCAATATTGAAGTTAATCA
MRE113'F	CCCCCC-ACTAGT-TTGTACTTGATCCCTATATTATATTATATCCTATTTATAACC

MRE113'R	CCCCCC-CCGCGG-AGTTCTATTGTGTGTCCAGGC
MRE115'F	CCCCCC-CGTACG-TCTTTCCAACAAACCAAGCG
MRE115'R	CCCCCC-GTCGAC-AGTCGAGTTTTATCGGATCTGAGC
MRE11BegR	TTGGTAGAGTGACTTCTTGG
MRE11EndF	CCAACGAGCAAACCCAAACG
NewAftAPL2R	CTTGTTGATCTTTCTTCCCACC
NewAftBY*CAN1*R	TATGACATTTCGCCTGAGCC
NewAftBYHIR3R	TTGACGCAAAGGAAATGTGG
NewAftCAC2R	TTGTTGCTGTTGGTCATTGG
NewAftCAC3R	TGGAAATGTTGTAGAGTGGAGG
NewAftCPKCKB2R	TATCTCTTCCTGTGCCCTCG
NewAftCPKMRC1R	CTCCACAATCGCAATCCACC
NewAftMRE11R	TGTGTTTGAGGGCTCCTTGG
NewBefAPL2F	TTCTCAACCATCCAAGTCGG
NewBefBY*CAN1*F	AGTGGAGGGTGTTGTTGTGG
NewBefBYHIR3F	TCTACGCGGTCCATAATCTCC
NewBefCAC2F	CGTTTCTGAGAGGTAACTGAGG
NewBefCAC3F	ACAACCACTTCACCCAAACCC
NewBefCPKCKB2F	CATTATTCGGCACACCTTTCACC
NewBefCPKMRC1F	TCTCTTCCTTTCAGCGGTGC
NewBefMRE11F	ATTGATGGCTGATGACGTGG
NGSAmp1F	ACTGAT-GCCAACTTACTTCTGACAACG
NGSAmp1R	GCTACC-ccgcctccatccagtc
NGSAmp2F	CGTACG-GCCAACTTACTTCTGACAACG
NGSAmp2R	TGACAT-ccgcctccatccagtc
NGSnointron10F	AGAATC-atcgcggccgccc
NGSnointron10R	GCCTAA-agaatgggcagacattacgaatg
NGSnointron11F	CTGCAG-atcgcggccgccc
NGSnointron11	ACATCG-agaatgggcagacattacgaatg

R	
NGSnointron12F	ATCACG-atcgcggccgccc
NGSnointron12R	CACGTA-agaatgggcagacattacgaatg
NGSnointron13F	TCACAT-atcgcggccgccc
NGSnointron13R	TATAGA-agaatgggcagacattacgaatg
NGSnointron14F	TGCAAA-atcgcggccgccc
NGSnointron14R	GTGCCA-agaatgggcagacattacgaatg
NGSnointron15F	TGTTAG-atcgcggccgccc
NGSnointron15R	ATAGAA-agaatgggcagacattacgaatg
NGSnointron16F	TCGAAG-atcgcggccgccc
NGSnointron16R	GAATGA-agaatgggcagacattacgaatg
NGSnointron17F	TACAGC-atcgcggccgccc
NGSnointron17R	TCTGAG-agaatgggcagacattacgaatg
NGSnointron18F	CTATAC-atcgcggccgccc
NGSnointron18R	AGCTAG-agaatgggcagacattacgaatg
NGSnointron19F	CGGAAT-atcgcggccgccc
NGSnointron19R	GCCATG-agaatgggcagacattacgaatg
NGSnointron1F	GATACA-atcgcggccgccc
NGSnointron1R	GGAACT-agaatgggcagacattacgaatg
NGSnointron20F	CACGAT-atcgcggccgccc
NGSnointron20R	GCTCAT-agaatgggcagacattacgaatg
NGSnointron2F	AGTCAA-atcgcggccgccc
NGSnointron2R	TAACCG-agaatgggcagacattacgaatg
NGSnointron3F	AGCTTT-atcgcggccgccc

NGSnointron3R	TACAAG-agaatgggcagacattacgaatg
NGSnointron4F	GGCTAC-atcgcggccgccc
NGSnointron4R	AAGCTA-agaatgggcagacattacgaatg
NGSnointron5F	ATACGA-atcgcggccgccc
NGSnointron5R	CTGATC-agaatgggcagacattacgaatg
NGSnointron6F	TTACTG-atcgcggccgccc
NGSnointron6R	AGTTCC-agaatgggcagacattacgaatg
NGSnointron7F	ACTTGA-atcgcggccgccc
NGSnointron7R	GATCTG-agaatgggcagacattacgaatg
NGSnointron8F	ACATCT-atcgcggccgccc
NGSnointron8R	AATCGT-agaatgggcagacattacgaatg
NGSnointron9F	GCCAAT-atcgcggccgccc
NGSnointron9R	CACTGT-agaatgggcagacattacgaatg
P416F	GGTACCCAATTCGCCCT
P416R	GAGCTCCAGCTTTTGTTCC
PCS0insF	GCTGTAAAACCAATTTCTCCAATTGAAACCGT
PCS3insF	GCTGTAAAAGCAGTAAAACCAATTTCTCCAATTGAAACCGT
PCS6insF	GCTGTAAAAGCAGTAAAATCAATCAAACCAATTTCTCCAATTGAAACCGT
PCSinsR	TGCAATCAGGTGAATTCGTTTC
PPTNotIF	GCGGCCGCGATCTATTACATTATGGGTGGTATGT
pUG6KOPCRnewF	CCAGCTGAAGCTTCGTACG
pUG6KOPCRnewR	CCGGCAGATCCGCGG
PXKSmcsFor	TGATTAACTTCAATATTGACTTTGAAGTCTTTATCTAGAT
PXKSmcsRev	TGGTTGCCATACGAAGCATCACTTATTGCGACTAGTTTAT
QMTy1RTL145Sf	cttgcattagacaataactactatattacacaatctgacatatcttcggcatatttgtat
QMTy1RTL145Sr	atacaaatatgccgaagatatgtcagattgtgtaatatagtagttattgtctaatgcaag
RTmutF	TCAACAGTAAGAAAAGATCATTAGAAGA
RTmutR	GGAAGGGATGCTAAGGTAGAG
RTSpeIF	GG-ACTAGT-ATGCGTAGTTTAGAACCTCCG

RTT101F	TCCC-CCCGGG-ATGATAAATGAGAGCGTTTCCAA
RTT101R	CCCCG-CTCGAG-TTAGTACTTGTAAGTTGCTGTTGAT
SacIITkc1BamHIrev	CCCCCC-GGATCC-TATATATATATAACTGTCTAGAAATAAAGAGTATCATCTTTCAAA-CCG
SacIITkc1BamHIrevshort	CCCCCC-GGATCC-TATATATATATAACT
SacIITkc1EcoRIrev	CCCCCC-GAATTC-TATATATATATAACTGTCTAGAAATAAAGAGTATCATCTTTCAAA-CCG
SacIITkc1HRrev1	ATGTATCAATAA-TATATATATATAACTGTCTAGAAATAAAGAGTATCATCTTTCAAA-CCG
SpeIXKSTkc6	cccc-actagt-ATGTTGTGTTCAGTAATTCAGAGACA
Spt15tefR	aactcctttaaacgttcctcatcggccat-aaacttagattagattgctatgctttcttt
TconnHpolR	TAATGAGTTTTCCATACCTAATTTCATGTATTTACCTCT-ATCAGGATGCAATTCGTAACC
TconnHRnaseF	AAACCTACCGAGCCAGATAATAAACTAGTCGCAATAAGT-TATGTTGATGGTGCTGCTAAC
TEC1F	TCCC-CCCGGG-ATGAGTCTTAAAGAAGACGACTTT
TEC1R	CCCCG-CTCGAG-TTAATAAAAGTTCCCATGCGATTG
TEF F	CCCCCC-GAGCTC-ATAGCTTCAAAATGTTTCTAC
tHIVRTXhoIR	CCCCG-CTCGAGTTA-AAAAGTTTCGGCACCAACG
Tkc1HRrev2	CGAGGCTATTGTCGCGATGTATCAATAA-TATATATATATAACTGTCTAGAAATAAAGAGT
Tkc1HRrev3	CCGAAACAAACCTCAGGAAAACAAGAACTATACGAGGCTATTGTCGCGAT
Tkc1HRshort	CCGAAACAAACCTCAGGAAAAC
TRnaseHconnR	ATTTGTGATTTATAATACGGTTGGTTGCCATACGAAGCATC-AAAAGTTTCGGCACCAACG
tTy1RTXhoIR	CCCCG-CTCGAGTTA-ACTTATTGCGACTAGTTTATTATCTGG
Ty145QCF	CCTGTCACTTGCATTAGACAATAACTACTATATTACACAANNKGACATATCTTCGGCATATTTG
Ty1RTconnF	TTTTGTGGATGGGTTACGAATTGCATCCTGAT-AGAGGTAAATACATGAAATTAGGTATGG
Ty1RTconnR	ACCCAACTTAGTTTCCATGTTAGCAGCACCATC-ACTTATTGCGACTAGTTTATTATCTGG
Ty1RTHDF1	CCGTTCACCTTCAATCGATGCTTCTCCACCGGAAAATAATTCATCGCACAATATTGTTCC
Ty1RTHDF2	CGCAGATAAGTGACCAAGAGACTGAGAAAAGGATTATACACCGTTCACCTTCAATCGATG
Ty1RTHDR1	AAAGCTGCAGCCCAAGCTTATCGATCTAATGAATCCATTTGTTAGTTAATAGTTTAAATG

Ty1RTHDR2	CATATTTGAGAAGATGCGGCCAGCAAAACTAACACCGATTATTTA AAGCTGCAGCCCAAG
Ty1RTK93R94 rev	CTTGTTGAAGATAAACATTGAGT
Ty1RTK93Rf	AAAAAG-GACGGTACTCATAAAGCTAG
Ty1RTL151Af	GCTTTGTATGCAGACATCAAAGA
Ty1RTL151Ar	TGCCGAAGATATGTCTAATTG
Ty1RTR94Kf	AGACGTGACGGTACTCATAAAGC
Ty225QCF	CGTATTTAAAAACAGTCAAGTGACAATTTGTTTANNKGTAGATGAT ATGGTATTGTTTAGCAAAAATCTA
Ty226QCF	GTATTTAAAAACAGTCAAGTGACAATTTGTTTATTCNNKGATGATA TGGTATTGTTTAGCAAAAATCTAAATTCA
Ty2600F	GTGCAAGTAGTCGCTGAACGGCTAAAC
Ty2600R	GCGTGCACCATGTGCTCGGGAATCC
Tye1F	TCCC-CCCGGG-ATGAACTCTATTTTAGACAGAAATGTTAGA
Tye1R	CCCCG-CTCGAG-TTATTTTTGGTCTTGTTTCAAAGTGT
TyH3FlankR	ACTTTTGGACCATCCATACCT
TyH3GenomeF	GCGCGGGCAAAGCCCAAAAG
TyH3GenomeR	TGCGCAAGCCCGGAATCGAA
TyH3PCRF1	TATCAACAATGGAATCCCAACAATTATCTCAACATTCACCCAATTC TCATGGTAGCGCCT
TyH3PCRF2	CTTCTAGTATATTCTGTATACCTAATATTATAGCCTTTATCAACAAT GGAATCCCAACAA
TyH3PCRF3	CTCGAGGAGAACTTCTAGTATATTCTGTATACCTAATATTATAGCC
TyH3PCRR1	GAACGGTTTCAATTGGAGAAATTGGAACAGCCTTCAAAGCTGCAA TCAGGTGAATTCGTT
TyH3PCRR2	ACTTTTGGACCATCCATACCTGGTTTCAACTTAACTGGAACGGTTT CAATTGGAGAAATT
TyHIVPBSF	ATGGAATCCCAACAATTATCTCAACATTCACCCAATTCTCA-GTGGCGCCCGAACAGGGAC
TyHIVPBSR	TCTTGATTTGTGTGGACTTCCTTAGAAGTAACCGAAGCACA-GTCCCTGTTCGGGCGCCAC
TyHIVPPTF	AATATAGAGTGTACTAGAGGCGGCCGCGATCTATTACATTAT-AAAAGAAAAGGGGGGACT
TyHIVPPTR	TAAATACTAGTTAGTAGATGATAGTTGATTTCTATTCCAACA-AGTCCCCCCTTTTCTTTT
***URA3*AIF**	CTCGAATTTTTACTAACAAATGGTATTATTTATAACAGCCGCCCAT GTCTCTTTGAGCAA
***URA3*AIR**	GCAGAAAAGCCTCCTTTAGTCCATATTAACATACCCGCGATGAAG GTTACGATTGGTTGA
***URA3*CYCR**	TTGCCTTCGTTTATCTTGCCTGCTCATTTTTTAGT-ttagtgtgtgtatttgtgtttgcg

***URA3*CYCR2**	gtagcagcacgttccttatatgtagctttcgacat-ttagtgtgtgtatttgtgtttgcg
***URA3*F**	TTGTAAATCGATAAGCTTGGGCTGCAGCTTTAAATAATCGGTGTTAGTTTTGCTGGCCGC
***URA3*GPDR**	CTTTGCCTTCGTTTATCTTGCCTGCTCATTTTTTAGT-atccgtcgaaactaagttctgg
***URA3*GPDR2**	gagtagcagcacgttccttatatgtagctttcgacat-atccgtcgaaactaagttctgg
***URA3*R**	AAAAAATGAGCAGGCAAGATAAACGAAGGCAAAGATGTCGAAAGCTACATATAAGGAACG
***URA3*RTPCRF**	ATCGCGGCCGCCCATGTCT
***URA3*RTPCRR**	TTGTCATGCAAGGGCTCCCTATCT
***URA3*TEFR**	TCGTTTATCTTGCCTGCTCATTTTTTAGT-aaacttagattagattgctatgctttcttt
***URA3*TEFR2**	gcacgttccttatatgtagctttcgacat-aaacttagattagattgctatgctttcttt
XbaI*XKS1*	CCCCCC-tctaga-ATGTTGTGTTCAGTAATTCAGAGA
XhoISpt15R	ccccgctcgagtcacatttttctaaattcacttagcaca
***XKS1*XhoI**	CCCCCC-ctcgag-TTAGATGAGAGTCTTTTCCAGTTC
XKSSacIITkc1 rev	GAAATAAAGAGTATCATCTTTCAAA-CCGCGG-agtttatcattatcaatactcgccatttc
XKSTkc6Rev	AAAAATTTATTTCTAGACAGTTATATAAAAAAAAA-TTAGATGAGAGTCTTTTCCAGTTCG
XKSTkc6XhoI	CCCCC-ctcgag-TTTGAAAAAATTTATTTCTAGACAGTTATATAAAAAAAAA-TTAGATGAG
XylA3seq	GAAGATGGGAAACTGACATTGG
XylAtefR	TGAATTTGAGGGAAATATTCTTTAGCCAT-aaacttagattagattgctatgctttcttt
NotITEFF	ATAAGAAT-GCGGCCGC-ATAGCTTCAAAATGTTTCTACTCCTTT
XmaITEFR	TCCC-CCCGGG-AAACTTAGATTAGATTGCTATGCTTTCT
XmaIXylA3F	TCCC-CCCGGG-ATGGCTAAAGAATATTTCCCTCAAA
BamHIXylA3R	CG-GGATCC-TTATTGATACATCGCGATAATAGCC
BamHIXylAR	CG-GGATCC-TTATTGATACATCGCGACAATAGC
BamHITKC1SacII	CG-GGATCC-TATATATATATAACTGTCTAGAAATAAAGAGTATCATCTTTCAAA-CCGCGG-G
SacIITKC1BamHI	C-CCGCGG-TTTGAAAGATGATACTCTTTATTTCTAGACAGTTATATATATATA-GGATCC-CG
SacIIGPDF	TCC-CCGCGG-AGTTTATCATTATCAATACTCGCCATT
PacIGPDR	CC-TTAATTAA-ATCCGTCGAAACTAAGTTCTGG
PacI*XKS1*F	CC-TTAATTAA-ATGTTGTGTTCAGTAATTCAGAGAC
SbfI*XKS1*R	GG-CCTGCAGG-TTAGATGAGAGTCTTTTCCAGTTCG
SbfITKC6EcoRI	GG-CCTGCAGG-TTTTTTTTTATATAACTGTCTAGAAATAAATTTTTTCAAA-GAATTC-

	C
EcoRITKC6SbfI	G-GAATTC-TTTGAAAAAATTTATTTCTAGACAGTTATATAAAAAAAAA-CCTGCAGG-CC
CEN6f	ACGAAAGGGCCTCGTGATAC
CEN6r	AGGGCGACACGGAAATGTTG
AfterOriF	AGCACAGATGCTTCGTTCAG
BeforeOriR	GCTGTTCTATATGCTGCCAC
RT-*URA3*-BBf	ATAAGCTTGGGCTGCAGC
RT-*URA3*-BBr	CGATCTAATGAATCCATTTG
RT-eGFPf	CATTTAAACTATTAACTAACAAATGGATTCATTAG-TGCGTAAAGGAGAAGAACTTTTCAC
RT-eGFPr	CCGATTATTTAAAGCTGCAGCCCAAGCTTATCGAT-TTAAACTGCTGCAGCGTAG
RT-LacZf	CATTTAAACTATTAACTAACAAATGGATTCATTAG-TGACCATGATTACGGATTCAC
RT-LacZr	CCGATTATTTAAAGCTGCAGCCCAAGCTTATCGAT-TTATTTTTGACACCAGACCAACTGG
RT-*CAN1*f	CATTTAAACTATTAACTAACAAATGGATTCATTAG-TGACAAATTCAAAAGAAGACGC
RT-*CAN1*r	CCGATTATTTAAAGCTGCAGCCCAAGCTTATCGAT-CTATGCTACAACATTCCAAA
YFPfusR	ATTCCAACATACCACCCATAATGTAATAGATC-TTATTTGTACAATTCATCCATACCATGG
YFPHARTfusF	AAGTTGGTTTCTGCCGGTATTAGAAAGATCTTG-TCTAAAGGTGAAGAATTATTCACTGGT
YFPTARTfusF	ACATTTAAACTATTAACTAACAAATGGATTCAT-TCTAAAGGTGAAGAATTATTCACTGGT
YFPlinkerF	GGTGACGGTGCTGGTTTAATTAAC-TCTAAAGGTGAAGAATTATTCACTGGT
Ty1RTlinkerR	ATGAATCCATTTGTTAGTTAATAGTTTAAATGTTT
HIVRTlinkerR	CAAGATCTTTCTAATACCGGCAGA
SPT300Vfor	ACGTTCCTCATCGGCCATAAA
SPT300Vrev	TGTGCTAAGTGAATTTAGAAAAATGTGA
SPT300Ifor	AAATAATCGGTGTCACATTTTTCTAAATTCACTTAGCACA
SPT300Irev	CATAGCAATCTAATCTAAGTTTATGGCCGATGAGGAACGT
XylAintronR	AATTCAAAAGCCTCCTTTAGTCCATATTAACATACTGGCTAAAGAA TATTTCCCTCAAAT
XylAintronMCSF	TTTACTAACAAATGGTATTATTTATAACAGTCCCGGGGGATCCAC
XylA3intronMCSF	TTTACTAACAAATGGTATTATTTATAACAGTAAACTTAGATTAGAT TGCTATGCTTTCTT

SPT15intronR	GTTAGTAAAAATTCAAAAGCCTCCTTTAGTCCATATTAACATACTG GCCGATGAGGAACG
SPT15intronF	AAATGGTATTATTTATAACAGTAAACTTAGATTAGATTGCTATGCT TTCTT
IntronSiteF	CGCCGAATTTTTACTAACAAATGGTATTATTTATAACAGT
IntronSiteR	CGCCAAAAGCCTCCTTTAGTCCATATTAAC
LacZBegSeqrev	CGGGCCTCTTCGCTATTACG
SacIISpt15R	C-CCGCGG-TCACATTTTTCTAAATTCACTTAGCACA
Spt15Nointron R	CTCATCGGCCATAAACTTAGATTA
Spt15Nointron F	TAATCTAAGTTTATGGCCGATGAG
ARTR	TTAATTCTTAGTATTCCATGTGTCTCGT
LTRendF	ATACTAGTTAGTAGATGATAGTTGATTTCTATTCCAACATACCACC CATAATGTAATAGA
LTRHXT7F	CCAACATACCACCCATAATGTAATAGATCgcggccgc-ACTTCTCGTAGGAACAATTTCGG
XRHXT7R	ACGTACTTGTTGAGCGACAT-TTTTTGATTAAAATTAAAAAAACTTTTTGTTTTTGTG
XRF	ATGTCGCTCAACAAGTACGT
TKC8HXT7R	ct-acgcgt-TTTGAAAAAATTTATTAAAAAAAAAATATATA-TTACTGGGTCTTGACGGTGA
GPDTKC8R	tgaaatggcgagtattgataatgataaact-acgcgt-TTTGAAAAAATTTATTAAAAAAAA
TKC8GPDF	TTTTTTTTTAATAAATTTTTTCAAA-acgcgt-agtttatcattatcaatactcgccatttc
LADGPDR	GGTGGGTGCCAGCAT-atccgtcgaaactaagttctgg
LADF	ATGCTGGCACCCACC
TKC1LADR	AAAGATGATACTCTTTATTTCTAGACAGTTATATATATATA-TTACACCGACGTGTAGCCT
TEFTKC1R	tagaaacattttgaagctat-cccggg-TTTGAAAGATGATACTCTTTATTTCTAGACAGTT
TKC1TEFF	CTAGAAATAAAGAGTATCATCTTTCAAA-cccggg-atagcttcaaaatgtttctactcctt
LXRTEFR	TCTTCGATGGTCATCGACAT-aaacttagattagattcgtatgctttcttt
LXRF	ATGTCGATGACCATCGAAGA
TKC5LXRR	AGATGATACTCTTTATTTATATATATATATATATATATATA-TCAAGGGAGTGTGTAGCCT
TKC5LXRR2	gatcc-TTTGAAAGATGATACTCTTTATTTATATATATATATATATATATATA-TCAAGGGA
TKC5ENO2F	TATATATATATATATAAATAAAGAGTATCATCTTTCAAA-ggatcc-gtgtcgacgctgcgg
XDHENO2R	GCGACCTGAGCAGACAT-tattattgtatgttatagtattagttgcttggt

XDHF	ATGTCTGCTCAGGTCGC
TKC6XDHR	taa-TTTGAAAAAATTTATTTCTAGACAGTTATATAAAAAAAAA-CTAGTGCTTGCCCTCGC
PGITKC6R	TGATTTTTGTTA-ttaattaa-TTTGAAAAAATTTATTTCTAGACAGTTATATAAAAAAAAA
TKC6PGIF	ATAACTGTCTAGAAATAAATTTTTTCAAA-ttaattaa-TAACAAAAATCACGATCTGGGTG
XKPGIR	GCTTCGGTGCTTTGCAT-TTTTAGGCTGGTATCTTGATTCTAAAT
XKF	ATGCAAAGCACCGAAGC
TKC22XKR	gcagg-TTTTTAGATGATACTCTTTATTTCTAGACAGTTATATA-TCAGGCCTGCTTCTGGC
RTTKC22R	AACTAACAAATGGATTCATTAG-gaattccctgcagg-TTTTTAGATGATACTCTTTATTTC
RTMCSR	AACCTCTTCCGATAAAAACATTTAAACTATTAACTAACAAATGGATTCATTAG-gaattcc
MCSTKC6R	gaattccctgcaggttaattaa-TTTGAAAAAATTTATTTCTAGACAGTTATATAAAAAAA
RTMCSTCK6R	ATTTAAACTATTAACTAACAAATGGATTCATTAG-gaattccctgcaggttaattaa-TTTG
ENO2TKC1R	cgcagcgtcgacac-ggatcccccggg-TTTGAAAGATGATACTCTTTATTTCTAGACAGTT
TKC1ENO2F	ACTGTCTAGAAATAAAGAGTATCATCTTTCAAA-cccgggggatcc-gtgtcgacgctgcgg
Lacz-*URA3*-BBf	TAAatcgataagcttgggCTGC
Lacz-*URA3*-BBr	TTGACACCAGACCAACTGG
LacZ-eGFPf	CATTACCAGTTGGTCTGGTGTCAAAAATAA-tgcgtaaaggagaagaacttttcac
mStraw-YFPf	gatcccccggg-ATGGTGAGCAAGGGCGAG
mStraw-YFPr	actcgagGAATTC-ttatttgtacaattcatccataccatggg

Appendix Table A6-1: Oligonucleotides used in this study (IDT)

PCR Fragment Name	PCR Template	Forward Primer	Reverse Primer
GALfrag1	S. cerevisiae BY4741 genome	GalpF1	Gal1pFixR2
GALfrag2	GALfrag1	GalpF2	GalpR2
Ty1frag1	S. cerevisiae BY4741 genome	TyH3GenomeF	TyH3GenomeR
Ty1frag2	Ty1frag1	TyH3PCRF1	TyH3PCRR1
Ty1frag3	Ty1frag2	TyH3PCRF2	TyH3PCRR2
Ty1frag4	Ty1frag3	TyH3PCRF3	TyH3FlankR

HIVRTfrag1	Synthetic HIVRT (See Below)	HIVnoATGF	HIVRThomR
***URA3*AIfrag1**	S. cerevisiae BY4741 genome	*URA3*F	*URA3*AIR
***URA3*frag1**	S. cerevisiae BY4741 genome	*URA3*R	*URA3*AIF
***URA3*frag2**	*URA3*frag1	*URA3*R	AInoass2F
***URA3*frag3**	*URA3*frag2	*URA3*R	AInoass3F
LTRfrag1	Synthetic Ty1 (See Below)	LTRF1	LTRR1
LTRfrag2	LTRfrag1	PPTNotIF	LTRflankR
HIS3promfrag1	S. cerevisiae BY4741 genome	HISpromF	HISpromR
P423frag1	P423-GPD (181)	P416F	P416R
Ty1RTfrag1	Synthetic Ty1RT (See Below)	Ty1RTHDF1	Ty1RTHDR1
Ty1RTfrag2	Ty1RTfrag1	Ty1RTHDF2	Ty1RTHDR2
pGALmTy1frag1	pGALmTy1-HIV	RTmutF	RTmutR
TYE1PCR	S. cerevisiae BY4741 genome	Tye1F	Tye1R
TEC1PCR	S. cerevisiae BY4741 genome	TEC1F	TEC1R
HSX1PCR	S. cerevisiae BY4741 genome	HSX1F	HSX1R
ELG1PCR	S. cerevisiae BY4741 genome	ELG1F	ELG1R
RTT101PCR	S. cerevisiae BY4741 genome	RTT101F	RTT101R
HIVPBS	none	TyHIVPBSF	TyHIVPBSR
HIVPPT	none	TyHIVPPTF	TyHIVPPTR
MidTy1-HIV	pGALmTy1-HIV	AftPBSF	BefPPTR
pGALmBackbone	pGALmTy1-HIV	3'LTRF	GALLTRR
HH-	pGALmTy1-HIV	RTmutF	TRnaseHconnR
H--	pGALmTy1-HIV	RTmutF	TconnHpolR
-T-	pGALmTy1-Ty1	Ty1RTconnF	Ty1RTconnR
--T	pGALmTy1-Ty1	TconnHRnaseF	RTmutR
--H	pGALmTy1-HIV	HconnTRnaseF	RTmutR
-TT	pGALmTy1-Ty1	Ty1RTconnF	RTmutR
tHH	pGALmTy1-HIV	RTSpeIF	tHIVRTXhoIR
tHT	pGALmTy1-HTT	RTSpeIF	tTy1RTXhoIR
BYHIR35	S. cerevisiae BY4741 genome	BYHIR35'F	BYHIR35'R
BYHIR33	S. cerevisiae BY4741 genome	BYHIR33'F	BYHIR33'R
BY*CAN15*	S. cerevisiae BY4741 genome	BY*CAN15*'F	BY*CAN15*'R
BY*CAN13*	S. cerevisiae BY4741 genome	BY*CAN13*'F	BY*CAN13*'R
CPK*CAN13*	S. cerevisiae CEN.PK2 genome	CPK*CAN13*'F	BY*CAN13*'R
CPKMRC15	S. cerevisiae CEN.PK2 genome	CPKMRC15'F	CPKMRC15'

			R
CPKMRC13	S. cerevisiae CEN.PK2 genome	CPKMRC13'F	CPKMRC13'R
CPKCKB25	S. cerevisiae CEN.PK2 genome	CPKCKB25'F	CPKCKB25'R
CPKCKB23	S. cerevisiae CEN.PK2 genome	CPKCKB23'F	CPKCKB23'R
CAC25	S. cerevisiae BY4741 genome	CAC25'F	CAC25'R
CAC23	S. cerevisiae BY4741 genome	CAC23'F	CAC23'R
CAC35	S. cerevisiae BY4741 genome	CAC35'F	CAC35'R
CAC33	S. cerevisiae BY4741 genome	CAC33'F	CAC33'R
APL25	S. cerevisiae BY4741 genome	APL25'F	APL25'R
APL23	S. cerevisiae BY4741 genome	APL23'F	APL23'R
MRE115	S. cerevisiae BY4741 genome	MRE115'F	MRE115'R
MRE113	S. cerevisiae BY4741 genome	MRE113'F	MRE113'R
BYHIR cassette	pUG6-BYHIR35-kanMX-BYHIR33	pUG6KOPCRnewF	pUG6KOPCRnewR
BYCAN cassette	pUG6-BY*CAN1*5-kanMX-BY*CAN1*3	pUG6KOPCRnewF	pUG6KOPCRnewR
CPKCAN cassette	pUG6-BY*CAN1*5-kanMX-CPK*CAN1*3	pUG6KOPCRnewF	pUG6KOPCRnewR
CPKMRC cassette	pUG6-CPKMRC15-kanMX-CPKMRC13	pUG6KOPCRnewF	pUG6KOPCRnewR
CPKCKB cassette	pUG6-CPKCKB25-kanMX-CPKCKB23	pUG6KOPCRnewF	pUG6KOPCRnewR
CAC2 cassette	pUG6-CAC25-kanMX-CAC23	pUG6KOPCRnewF	pUG6KOPCRnewR
CAC3 cassette	pUG6-CAC35-kanMX-CAC33	pUG6KOPCRnewF	pUG6KOPCRnewR
APL2 cassette	pUG6-APL25-kanMX-APL23	pUG6KOPCRnewF	pUG6KOPCRnewR
MRE11 cassette	pUG6-MRE115-kanMX-MRE113	pUG6KOPCRnewF	pUG6KOPCRnewR
ICE2 cassette	S. cerevisiae BY4741 ΔICE2 genome	BefICE2F	AftACE2R
RRM3 cassette	S. cerevisiae BY4741 ΔRRM2 genome	BefRRM1F	AftRRM1R
pGALmTy1-Ty1 BB	pGALmTy1-Ty1	HispromF	HistermR
pGALmTy1-HIV BB	pGALmTy1-HIV	HispromF	HistermR
CAN1	S. cerevisiae BY4741 genome	HispromCANF	HistermCANR

XylA-TEF1 cassette	p416-TEF-YFP	LTRTEFF2	XylAtefR
XylA3-TEF1 cassette	p416-TEF-YFP	LTRTEFF2	XylAtefR
Spt15-TEF1 cassette	p416-TEF-YFP	LTRTEFF2	Spt15tefR
Amp amplicon 1	p423-TART	NGSAmp1F	NGSAmp1R
Amp amplicon 2	BY4741 Δ*rrm3*-1 plasmid (glucose)	NGSAmp2F	NGSAmp2R
***URA3* amplicon 1**	BY4741 Δ*rrm3*-1 plasmid (galactose)	NGSnointron1F	NGSnointron 1R
***URA3* amplicon 2**	BY4741 Δ*rrm3*-2 plasmid (galactose)	NGSnointron2F	NGSnointron 2R
***URA3* amplicon 3**	BY4741 Δ*rrm3*-3 plasmid (galactose)	NGSnointron3F	NGSnointron 3R
***URA3* amplicon 4**	BY4741 Δ*rrm3*-4 plasmid (galactose)	NGSnointron4F	NGSnointron 4R
***URA3* amplicon 5**	BY4741 Δ*rrm3*-5 plasmid (galactose)	NGSnointron5F	NGSnointron 5R
***URA3* amplicon 6**	BY4741 Δ*rrm3*-6 plasmid (galactose)	NGSnointron6F	NGSnointron 6R
***URA3* amplicon 7**	BY4741 Δ*rrm3*-7 plasmid (galactose)	NGSnointron7F	NGSnointron 7R
***URA3* amplicon 8**	BY4741 Δ*rrm3*-8 plasmid (galactose)	NGSnointron8F	NGSnointron 8R
***URA3* amplicon 9**	BY4741 Δ*rrm3*-9 plasmid (galactose)	NGSnointron9F	NGSnointron 9R
***URA3* amplicon 10**	BY4741 Δ*rrm3*-10 plasmid (galactose)	NGSnointron10 F	NGSnointron 10R
***URA3* amplicon 11**	BY4741 Δ*hir3* Δ*cac3*-1 plasmid (galactose)	NGSnointron11 F	NGSnointron 11R
***URA3* amplicon 12**	BY4741 Δ*hir3* Δ*cac3*-2 plasmid (galactose)	NGSnointron12 F	NGSnointron 12R
***URA3* amplicon 13**	BY4741 Δ*hir3* Δ*cac3*-3 plasmid (galactose)	NGSnointron13 F	NGSnointron 13R
***URA3* amplicon 14**	BY4741 Δ*hir3* Δ*cac3*-4 plasmid (galactose)	NGSnointron14 F	NGSnointron 14R
***URA3* amplicon 15**	BY4741 Δ*hir3* Δ*cac3*-5 plasmid (galactose)	NGSnointron15 F	NGSnointron 15R
***URA3* amplicon 16**	BY4741 Δ*hir3* Δ*cac3*-6 plasmid (galactose)	NGSnointron16 F	NGSnointron 16R
***URA3* amplicon 17**	BY4741 Δ*hir3* Δ*cac3*-7 plasmid (galactose)	NGSnointron17 F	NGSnointron 17R
***URA3* amplicon 18**	BY4741 Δ*hir3* Δ*cac3*-8 plasmid (galactose)	NGSnointron18 F	NGSnointron 18R
***URA3* amplicon 19**	BY4741 Δ*hir3* Δ*cac3*-9 plasmid (galactose)	NGSnointron19 F	NGSnointron 19R

***URA3* amplicon 20**	BY4741 Δ*hir3* Δ*cac3*-10 plasmid (galactose)	NGSnointron20F	NGSnointron20R
***URA3*AI**	pGALmTy1-Ty1	His3AIgenomeflankF	His3transconfR
***URA3*AI-2**	pGALmTy1-Ty1	His3AIgenomeflankF	BefPPTR
XylA	pXylA	HispromXylAF	HistermXylAR
XylA3	pXylA3	HispromXylAF	HistermXylA3R
Spt15	S. cerevisiae BY4741 genome	HispromSPT15F	HistermSPT15R
CYC	P416-CYC	LTRCYCF2	*URA3*CYCR2
TEF	P416-TEF	LTRTEFF2	*URA3*TEFR2
GPD	P416-GPD	LTRGPDF2	*URA3*GPDR2
HIV-instop1	pGALmTy1-HIV	Int1stopF	Int1stopR
HIV-instop2	pGALmTy1-HIV	Int2stopF	Int2stopR
HIV-instop3	pGALmTy1-HIV	Int3stopF	Int3stopR
HIV-instop4	pGALmTy1-HIV	Int4stopF	Int4stopR
HIV-instop5	pGALmTy1-HIV	Int5stopF	Int5stopR
HIV-intdel1	pGALmTy1-HIV	Int1del1F	Int1del1R
HIV-intdel2	pGALmTy1-HIV	Int1del2F	Int1del2R
HIV-intdel3	pGALmTy1-HIV	Int1del3F	Int1del3R
HIV-intdel4	pGALmTy1-HIV	Int1del4F	Int1del4R
HIV-intdel5	pGALmTy1-HIV	Int1del5F	Int1del5R
HIV-intdel6	pGALmTy1-HIV	Int1del6F	Int1del6R
TY1-instop1	pGALmTy1-Ty1	Int1stopF	Int1stopR
TY1-instop2	pGALmTy1-Ty1	Int2stopF	Int2stopR
TY1-instop3	pGALmTy1-Ty1	Int3stopF	Int3stopR
TY1-instop4	pGALmTy1-Ty1	Int4stopF	Int4stopR
TY1-instop5	pGALmTy1-Ty1	Int5stopF	Int5stopR
TY1-intdel1	pGALmTy1-Ty1	Int1del1F	Int1del1R
TY1-intdel2	pGALmTy1-Ty1	Int1del2F	Int1del2R
TY1-intdel3	pGALmTy1-Ty1	Int1del3F	Int1del3R
TY1-intdel4	pGALmTy1-Ty1	Int1del4F	Int1del4R
TY1-intdel5	pGALmTy1-Ty1	Int1del5F	Int1del5R
TY1-intdel6	pGALmTy1-Ty1	Int1del6F	Int1del6R
HIR3 40bp cassette	pUG6	pUG6KOPCRnewF	pUG6KOPCRnewR
HIR3 cassette	HIR3 40bp cassette	BYHIRKO80A	BYHIRKO80

		F	OR
PXKSmcs	pGALmTy1-Ty1-XylA3-TEF1	PXKSmcsFor	PXKSmcsRev
MCS1	pGALmTy1-Ty1-XylA3-TEF1	MCS1For	MCS1Rev
MCS2	pGALmTy1-Ty1-XylA3-TEF1	MCS2For	MCS2Rev
XbaI*XKS1*XhoI	p*XKS1*	XbaI*XKS1*	*XKS1*XhoI
XmaIXylAXhoI	pGALmTy1-Ty1-XylA	XmaIXylAF	XhoIXylAR
NotI-XylA-TEF1-EcoRI	p415-TEF-XylA	NotITEFF	EcoRIXylAR
GRE3KO+XKS cassette	p415-TEF-XKS	GRE3KOfor	GRE3KOrev new
GRE3KO cassette	p415-TEF-XKS	GRE3KOleuF	GRE3KOrev new
***XKS1*-Tkc6 1**	p*XKS1*	SpeIXKSTkc6	XKSTkc6Rev
***XKS1*-Tkc6 2**	*XKS1*-Tkc6 1	SpeIXKSTkc6	XKSTkc6XhoI
Tkc6-*XKS1*-GPD-Tkc1 1	p415-GPD-*XKS1*-Tkc6	EcoRITkc6XKSfor	XKSSacIITkc1rev
Tkc6-*XKS1*-GPD-Tkc1 2	Tkc6-*XKS1*-GPD-Tkc1 1	EcoRITkc6XKSshort	SacIITkc1BamHIrev
EcoRI-Tkc6-*XKS1*-GPD-Tkc1	Tkc6-*XKS1*-GPD-Tkc1 1	EcoRITkc6XKSshort	SacIITkc1EcoRIrev
Tkc6-*XKS1*-GPD-Tkc1 HR1	Tkc6-*XKS1*-GPD-Tkc1 1	EcoRITkc6XKSHRfor1	SacIITkc1HRrev1
Tkc6-*XKS1*-GPD-Tkc1 HR2	Tkc6-*XKS1*-GPD-Tkc1 HR1	EcoRITkc6XKSHRfor2	Tkc1HRrev2
Tkc6-*XKS1*-GPD-Tkc1 HR3	Tkc6-*XKS1*-GPD-Tkc1 HR2	EcoRITkc6XKSHRshort	Tkc1HRrev3
SacI-Spt15-XhoI	Spt15 plasmid or genomic DNA	TEF F	XhoISpt15R
TEF1	pGALmTy1-Ty1-XylA3-TEF1	NotITEFF	XmaITEFR
XylA	pGALmTy1-Ty1-XylA	XmaIXylA3F	BamHIXylAR
XylA3	pGALmTy1-Ty1-XylA3	XmaIXylA3F	BamHIXylA3R
Tkc1	Anneal	BamHITKC1SacII	SacIITKC1BamHI
GPD	p416-GPD	SacIIGPDF	PacIGPDR
XKS1	p413-TEF-*XKS1*	Pac*IXKS1*F	Sbfl*XKS1*R
Tkc6	Anneal	SbfITKC6EcoRI	EcoRITKC6SbfI
CEN6/ARSH	p413-GPD	CEN6f	CEN6r
NoOri	pGALmTy1-Ty1	AfterOriF	BeforeOriR

RT-*URA3*-BB	pGALmTy1-Ty1	RT-*URA3*-BBf	RT-*URA3*-BBr
Cargo-eGFP	p423-GPD-eGFP	RT-eGFPf	RT-eGFPr
Cargo-LacZ	p423-GPD-LacZ	RT-LacZf	RT-LacZr
Cargo-*CAN1*	pGALmTy1-Ty1-*CAN1*	RT-*CAN1*f	RT-*CAN1*r
NoOri-XylA	pGALmTy1-Ty1-XylA-TEF1	AfterOriF	BeforeOriR
NoOri-Spt15	pGALmTy1-Ty1-Spt15-TEF1	AfterOriF	BeforeOriR
NoOri-XylA3	pGALmTy1-Ty1-XylA3-TEF1	AfterOriF	BeforeOriR
NoOri-Spt15-300	pGALmTy1-Ty1-Spt15-300-TEF1	AfterOriF	BeforeOriR
YFP-Ty1	p423-GPD-YFP	YFPTARTfusF	YFPfusR
YFP-HIV	p423-GPD-YFP	YFPHARTfusF	YFPfusR
pGALmTy1-Ty1-TEF	pGALmTy1-Ty1-Spt15-TEF1	SPT300Vfor	SPT300Vrev
Spt15-300	p413-TEF-SPT15-300	SPT300Ifor	SPT300Irev
XylAintronnosite	pGALmTy1-Ty1-MCS-XylA-TEF1	XylAintronMCSF	XylAintronR
XylA3intronnosite	pGALmTy1-Ty1-MCS-XylA4-TEF1	XylA3intronMCSF	XylAintronR
Spt15intronnosite	pGALmTy1-Ty1-MCS-Spt15-TEF1	SPT15intronF	SPT15intronR
Spt15-300intronnosite	pGALmTy1-Ty1-Spt15-300-TEF1	SPT15intronF	SPT15intronR
XylAintron	pGALmTy1-Ty1-MCS-XylAintronnosite-TEF1	IntronSiteF	IntronSiteR
XylA3intron	pGALmTy1-Ty1-MCS-XylA3intronnosite-TEF1	IntronSiteF	IntronSiteR
Spt15intron	pGALmTy1-Ty1-MCS-Spt15ntronnosite-TEF1	IntronSiteF	IntronSiteR
Spt15-300intron	pGALmTy1-Ty1-MCS-Spt15-300-intronnosite-TEF1	IntronSiteF	IntronSiteR
XylAlcnoRT	pGALmTy1-Ty1-MCS-XylA-TEF1 (low copy)	His3AIgenomeflankF	ARTR
XylA3lcnoRT	pGALmTy1-Ty1-MCS-XylA3-TEF1 (low copy)	His3AIgenomeflankF	ARTR
SPT15lcnoRT	pGALmTy1-Ty1-Spt15-TEF1 (low copy)	His3AIgenomeflankF	ARTR
SPT15-300lcnoRT	pGALmTy1-Ty1-Spt15-300-TEF1 (low copy)	His3AIgenomeflankF	ARTR
XylAlcintnoRT	pGALmTy1-Ty1-MCS-XylAintron-TEF1 (low copy)	His3AIgenomeflankF	ARTR
XylA3lcintnoRT	pGALmTy1-Ty1-MCS-XylA3intron-TEF1 (low copy)	His3AIgenomeflankF	ARTR
SPT15lcintnoRT	pGALmTy1-Ty1-Spt15intron-TEF1	His3AIgenomefl	ARTR

	(low copy)	ankF	
SPT15-300lcintnoRT	pGALmTy1-Ty1-Spt15-300intron-TEF1 (low copy)	His3AIgenomeflankF	ARTR
HXT7	S. cerevisiae genome	LTRHXT7F	XRHXT7R
XR	p416-TEF-BM-XR-1-GPD-BM-XDH-BM-XK	XRF	TKC8HXT7R
GPD	S. cerevisiae genome	TKC8GPDF	LADGPDR
LAD	p423-GPD-BM-LAD	LADF	TKC1LADR
TEF	S. cerevisiae genome	TKC1TEFF	LXRTEFR
LXR	p424-GPD-BM-LXR	LXRF	TKC5LXRR
ENO2	S. cerevisiae genome	TKC5ENO2F	XDHENO2R
XDH	p416-TEF-BM-XR-1-GPD-BM-XDH-BM-XK	XDHF	TKC6XDHR
PGI	S. cerevisiae genome	TKC6PGIF	XKPGIR
XK1	p416-TEF-BM-XR-1-GPD-BM-XDH-BM-XK	XKF	TKC22XKR
XK2	XK1	XKF	RTTKC22R
XDH2	XDH	XDHF	MCSTKC6R
ENO2-2	S. cerevisiae genome	TKC1ENO2F	XDHENO2R
HXT7-XR	HXT7, XR	LTREndF	GPDTKC8R
GPD-LAD	GPD, LAD	TKC8GPDF	TEFTKC1R
TEF-LXR	TEF, LXR	TKC1TEFF	TKC5LXRR2
ENO2-XDH	ENO2, XDH	TKC5ENO2F	PGITKC6R
PGI-XK	PGI, XK2	TKC6PGIF	RTMCSR
ENO2-XDH2	ENO2, XDH2	TKC5ENO2F	RTMCSTCK6R
GPD-LAD2	GPD, LAD	TKC8GPDF	ENO2TKC1R
ENO2-2-XDH	ENO2-2, XDH	TKC1ENO2F	PGITKC6R
ENO2-2-XDH2	ENO2-2, XDH2	TKC1ENO2F	RTMCSTCK6R
Cargo-eGFP	p423-GPD-eGFP	RT-eGFPf	RT-eGFPr
Cargo-LacZ-BB	pGALmTy1-Ty1-Cargo3	Lacz-*URA3*-BBf	Lacz-*URA3*-BBr
Cargo-eGFP-LacZ-ins	p423-GPD-eGFP	LacZ-eGFPf	RT-eGFPr
mStraw-YFP	pGALmTy1-Ty1-mStrawberry-P2A-YFP	mStraw-YFPf	mStraw-YFPr
NoOri-mStraw-YFP	pGALmTy1-Ty1-mStrawberry-P2A-YFP (high-copy)	AfterOriF	BeforeOriR

Appendix Table A6-1: PCR fragments used to assemble the plasmids used in this study.

Plasmid Name	PCR Fragments Used For Assembly
pGALmTy1H-HIV	HIVPBS, HIVPPT, MidTy1-HIV, pGALmBackbone
pGALmTy1-HHT	RTmutBackbone, HH-, --T
pGALmTy1-HTH	RTmutBackbone, H--, -T-, --H
pGALmTy1-HTT	RTmutBackbone, H--, -TT
pGALmTy1H-HHT	RTmutBackboneH, HH-, --T
pGALmTy1H-HTH	RTmutBackboneH, H--, -T-, --H
pGALmTy1H-HTT	RTmutBackboneH, H--, -TT
pGALmTy1-Ty1-*CAN1*	pGALmTy1-Ty1 BB, *CAN1*
pGALmTy1-HIV-*CAN1*	pGALmTy1-HIV BB, *CAN1*
pGALmTy1-Ty1-CYC	pGALmTy1-Ty1, CYC
pGALmTy1-Ty1-TEF	pGALmTy1-Ty1, TEF
pGALmTy1-Ty1-GPD	pGALmTy1-Ty1, GPD
pGALmTy1-HIV-CYC	pGALmTy1-HIV, CYC
pGALmTy1-HIV-TEF	pGALmTy1-HIV, TEF
pGALmTy1-HIV-GPD	pGALmTy1-HIV, GPD
pGALmTy1-Ty1-XylA	pGALmTy1-Ty1, *XylA*
pGALmTy1-Ty1-XylA3	pGALmTy1-Ty1, *XylA*3
pGALmTy1-Ty1-Spt15	pGALmTy1-Ty1, *Spt15*
pGALmTy1-HIV-XylA	pGALmTy1-HIV, *XylA*
pGALmTy1-HIV-XylA3	pGALmTy1-HIV, *XylA*3
pGALmTy1-HIV-Spt15	pGALmTy1-HIV, *Spt15*
pGALmTy1-Ty1-XylA-TEF1	pGALmTy1-Ty1-XylA, *XylA*-TEF1 cassette
pGALmTy1-Ty1-XylA3-TEF1	pGALmTy1-Ty1-XylA3, *XylA*3-TEF1 cassette
pGALmTy1-Ty1-Spt15-TEF1	pGALmTy1-Ty1-Spt15, *Spt15*-TEF1 cassette
pGALmTy1-HIV-XylA-TEF1	pGALmTy1-HIV-XylA, *XylA*-TEF1 cassette
pGALmTy1-HIV-XylA3-TEF1	pGALmTy1-HIV-XylA3, *XylA*3-TEF1 cassette
pGALmTy1-HIV-Spt15-TEF1	pGALmTy1-HIV-Spt15, *Spt15*-TEF1 cassette
PGALMTY1-HIV-Intstop1	HIV-instop1
PGALMTY1-HIV-Intstop2	HIV-instop2
PGALMTY1-HIV-Intstop3	HIV-instop3
PGALMTY1-HIV-Intstop4	HIV-instop4
PGALMTY1-HIV-Intstop5	HIV-instop5
PGALMTY1-HIV-Intdel1	HIV-intdel1

PGALMTY1-HIV-Intdel2	HIV-intdel2
PGALMTY1-HIV-Intdel3	HIV-intdel3
PGALMTY1-HIV-Intdel4	HIV-intdel4
PGALMTY1-HIV-Intdel5	HIV-intdel5
PGALMTY1-HIV-Intdel6	HIV-intdel6
PGALMTY1-TY1-Intstop1	TY1-instop1
PGALMTY1-TY1-Intstop2	TY1-instop2
PGALMTY1-TY1-Intstop3	TY1-instop3
PGALMTY1-TY1-Intstop4	TY1-instop4
PGALMTY1-TY1-Intstop5	TY1-instop5
PGALMTY1-TY1-Intdel1	TY1-intdel1
PGALMTY1-TY1-Intdel2	TY1-intdel2
PGALMTY1-TY1-Intdel3	TY1-intdel3
PGALMTY1-TY1-Intdel4	TY1-intdel4
PGALMTY1-TY1-Intdel5	TY1-intdel5
PGALMTY1-TY1-Intdel6	TY1-intdel6
p415-TEF-XylA	p415-TEF, XmaIXylAXhoI
p415-TEF-XKS	p415-TEF, XbaI*XKS1*XhoI
p415-GPD-*XKS1*-Tkc6	p415-GPD, *XKS1*-Tkc6 2
pGALmTy1-Ty1-XylA3-*XKS1*	pGALmTy1-Ty1-MCS-XylA3-TEF1, Tkc6-*XKS1*-GPD-Tkc1 2
pGALmTy1-Ty1-XylA-*XKS1*	pGALmTy1-Ty1-MCS-XylA-TEF1, EcoRI-Tkc6-*XKS1*-GPD-Tkc1
pGALmTy1-Ty1-XylA-*XKS1*-HR	pGALmTy1-Ty1-MCS-XylA3-TEF1, Tkc6-*XKS1*-GPD-Tkc1 HR3
p423-TEF-Spt15	p423-TEF, SacI-Spt15-XhoI
pGALmTy1-Ty1-Spt15-300-TEF1	pGALmTy1-Ty1-TEF, *Spt15*-300
pGALmTy1-Ty1-Tkc1-XylA-TEF1	pGALmTy1-Ty1-MCS, TEF1, *XylA*, Tkc1
pGALmTy1-Ty1-Tkc1-XylA3-TEF1	pGALmTy1-Ty1-MCS, TEF1, *XylA*3, Tkc1
pGALmTy1-Ty1-Tkc6-*XKS1*-GPD	pGALmTy1-Ty1-MCS, GPD, *XKS1*, Tkc6
pGALmTy1-Ty1-Tkc6-*XKS1*-GPD-Tkc6-XylA-TEF1	pGALmTy1-Ty1-Tkc1-XylA-TEF1, GPD-*XKS1*-Tkc6 cassette
pGALmTy1-Ty1-Tkc6-*XKS1*-GPD-Tkc6-XylA3-TEF1	pGALmTy1-Ty1-Tkc1-XylA3-TEF1, GPD-*XKS1*-Tkc6 cassette
pGALmTy1-Ty1 (low-copy)	CEN6/ARSH, NoOri
pGALmTy1-Ty1-Cargo1	RT-*URA3*-BB, Cargo-eGFP
pGALmTy1-Ty1-Cargo2	RT-*URA3*-BB, Cargo-LacZ
pGALmTy1-Ty1-Cargo3	RT-*URA3*-BB, Cargo-LacZ
pGALmTy1-Ty1-XylA-TEF1 (low-	CEN6/ARSH, NoOri-XylA

copy)	
pGALmTy1-Ty1-Spt15-TEF1 (low-copy)	CEN6/ARSH, NoOri-Spt15
pGALmTy1-Ty1-XylA3-TEF1 (low-copy)	CEN6/ARSH, NoOri-XylA3
pGALmTy1-Ty1-Spt15-300-TEF1 (low-copy)	CEN6/ARSH, NoOri-Spt15-300
pGALmTy1-Ty1-YFP (no linker)	Enzyme-digested pGALmTy1-Ty1, YFP-Ty1
pGALmTy1-HIV-YFP (no linker)	Enzyme-digested pGALmTy1-HIV, YFP-HIV
pGALmTy1-Ty1-ara3gene	pGALmTy1-Ty1-MCS BB, HXT7-XR, GPD-LAD2, ENO2-2-XDH2
pGALmTy1-Ty1-aranoLXR	pGALmTy1-Ty1-MCS BB, HXT7-XR, GPD-LAD2, ENO2-2-XDH, PGI-XK
pGALmTy1-Ty1-aranoXKS	pGALmTy1-Ty1-MCS BB, HXT7-XR, GPD-LAD, TEF-LXR, ENO2-XDH2
pGALmTy1-Ty1-ara5gene	pGALmTy1-Ty1-MCS BB, HXT7-XR, GPD-LAD, TEF-LXR, ENO2-XDH, PGI-XK
pGALmTy1-Ty1-Cargo2	RT-*URA3*-BB, Cargo-*CAN1*
pGALmTy1-Ty1-Cargo4	Cargo-LacZ-BB, Cargo-eGFP
pGALmTy1-Ty1-TEF-mStrawberry-intron-P2A-YFP (low-copy)	CEN6/ARSH, NoOri-mStraw-YFP

Appendix Table A6-3: Plasmids generated through recombination cloning

Strain Name	Parent Strain	Fragment Used for Construction
BY4741 Δ*cac2*	BY4741	CAC2 cassette
BY4741 Δ*cac3*	BY4741	CAC3 cassette
BY4741 Δ*hir3* ΔAPL2	BY4741 Δ*hir3*	APL2 cassette
BY4741 Δ*hir3* Δ*mre11*	BY4741 Δ*hir3*	MRE11 cassette
BY4741 ΔAPL2 Δ*mre11*	BY4741 ΔAPL2	MRE11 cassette
BY4741 Δ*hir3* Δ*cac2*	BY4741 Δ*hir3*	CAC2 cassette
BY4741 Δ*hir3* Δ*cac3*	BY4741 Δ*hir3*	CAC3 cassette
CEN.PK2 Δ*CAN1*	CEN.PK2	CPKCAN cassette
CEN.PK2 Δ*cac2*	CEN.PK2	CAC2 cassette
CEN.PK2 Δ*cac3*	CEN.PK2	CAC3 cassette
CEN.PK2 ΔAPL2	CEN.PK2	APL2 cassette
CEN.PK2 Δ*hir3*	CEN.PK2	BYHIR cassette
CEN.PK2 Δ*mre11*	CEN.PK2	MRE11 cassette
CEN.PK2 ΔICE2	CEN.PK2	ICE2 cassette
CEN.PK2 Δ*rrm3*	CEN.PK2	RRM3 cassette
CEN.PK2 ΔMRC1	CEN.PK2	CPKMRC cassette
CEN.PK2 ΔCKB2	CEN.PK2	CPKCKB cassette

BY4741 Δ*rrm3*/Δ*gre3*/*XKS1*	BY4741 Δ*rrm3*	GRE3KO+XKS cassette
BY4741 Δ*rrm3*/Δ*gre3*	BY4741 Δ*rrm3*	GRE3KO cassette

Appendix Table A6-4: Strains generated in this study

Fragment Name	Template	Enzyme I	Enzyme II
pGALmTy1-Ty1-MCS BB	pGALmTy1-Ty1-MCS (low copy)	NotI	SbfI

Appendix Table A6-5: Restriction fragments used to assemble the plasmids used in this study.

Appendix A7

Plasmid Name	Source
p416-GPD-YFP	(37)
p416-CYC-YFP	(37)
pYES2.1-Dcr1	(141)
pYES2.1-Ago1	(141)
p413-GPD	(181)
p415-GPD	(181)
p424-GPD	(181)
p414-CYC	(181)
p414-TEF	(181)
p414-GPD	(181)
p423-GPD	(181)
p41K-GPD	(239)
p416-UASCLBUASCITUASTEFpGPD-CAD1	(251)

Appendix Table A7-1: Plasmids obtained for this study

Strain	Genotype	Source
***S. cerevisiae* BY4741**	*MATa; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0*	EUROSCARF
***S. cerevisiae* BY4741 ΔTRP1**	*MATa; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0; trp1::KanMX*	*Saccharomyces* Knockout Collection Database
***S. cerevisiae* CEN.PK2-1C**	*MATa; ura3-52; trp1-289; leu2-3,112; his3Δ1; MAL2-8^{C}; SUC2*	EUROSCARF
***S. cerevisiae* Sigma 10560-4A**	*MATa; ura 3-52; trp1::hisG; leu2::hisG; his3::hisG*	Gerald R. Fink Laboratory
***S. cerevisiae* BY4741 ΔADE3**	*MATa; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0; ade3::loxP*	*Saccharomyces* Knockout Collection Database

Appendix Table A7-2: Strains obtained for this study

Primer Name	Primer Sequence
ScDcr1F	GG-ACTAGT-ATGAATAGAGAAAAAAGCGCCG
ScDcr1R	CCCCG-CTCGAG-TCAATGGTGATGGTGATGATG
ScAgo1F	GG-ACTAGT-ATGTCATCCAATTCGGAGGA
ScAgo1R	CCCCG-CTCGAG-TCATATGTAGTACATGATGTCAGTG
YFPfrag1SpeIF	GG-ACTAGT-CATGGCCAACCTTAGTCAC
YFPfrag1EcoRIR	CG-GAATTC-CATGTTGTTTCATATGATCTGGGT
YFPfrag1XhoIF	CCCCG-CTCGAG-CATGGCCAACCTTAGTCAC
YFPfrag3SpeIF	GG-ACTAGT-CATGCCAGAAGGTTATGTTCAA
YFPfrag3EcoRIR	CG-GAATTC-CATGATGTAAACATTGTGAGAGTTATAG
YFPfrag3XhoIF	CCCCG-CTCGAG-CATGCCAGAAGGTTATGTTCAA
YFPfrag3SalIR	TAACGC-GTCGAC-CATGATGTAAACATTGTGAGAGTTATAG
SmallAIF	CCCCG-GAATTC-GTATGTTAATATGGACTAAAGGAGGCTTTTCCCGGGGAATTTTTACTAA
SmallAIR	TAACGC-GTCGAC-CTGTTATAAATAATACCATTTGTTAGTAAAAATTCCCCGGGAAAAGCC
YFP-6SpeF	GG-ACTAGT-AGGTGAAGGTGATGCTACTT
YFP-6EcoRIR	CG-GAATTC-TGACTTCAGCTCTGGTCTTGT
YFP-6SalIF	TAACGC-GTCGAC-TGACTTCAGCTCTGGTCTTGT
YFP-6XhoIR	CCCCG-CTCGAG-AGGTGAAGGTGATGCTACTT
ADE3-1SpeF	GG-ACTAGT-GATTGAACATTGACCCGGACA
ADE3-1EcoRIR	CG-GAATTC-ACTCTTCCAATACGTTCCTTCATG
ADE3-1SalIF	TAACGC-GTCGAC-ACTCTTCCAATACGTTCCTTCATG
ADE3-1XhoIR	CCCCG-CTCGAG-GATTGAACATTGACCCGGACA
TKC6-p4XXF	TTTTTTTTATATAACTGTCTAGAAATAAATTTTTTCAAA-CTCGAGCAATTCGCCCTATAG
GPD rev	ATCCGTCGAAACTAAGTTCTGG
GPDDicerF	TTTAAAACACCAGAACTTAGTTTCGACGGATACTAGT-ATGAATAGAGAAAAAAGCGCCGA
TKC1Dicer	TACTCTTTATTTCTAGACAGTTATATATATATACCCGG-

R	GTCAATGGTGATGGTGATGATG
TKC1TEFF	CATTGACCCGGGTATATATATATAACTGTCTAGAAATAAAGAGTATC ATCTTTCAAAGCGGCCG-CATAGCTTCAAAATGTTTCTACTCCT
AgoTEFR	TTCTCCTCCGAATTGGATGACATGCATGC- AAACTTAGATTAGATTGCTATGCTTTCTTTC
TEFAgoF	TAGAAAGAAAGCATAGCAATCTAATCTAAGTTTGCATGC- ATGTCATCCAATTCGGAGGAG
TKC6AgoR	ATTTCTAGACAGTTATATAAAAAAAAAATCGAT- TCATATGTAGTACATGATGTCAGTGAC
BefPromF	GTGAAAGTTTGCGGCTTGCAGAGCACAGAGGCCGCAGAATGT- GAACAAAAGCTGGAGCTC
AftMarkerR	ACACCATTTGTCTCCACACCTCCGCTTACATCAACACCA- AAATAGGCGTATCACGAGGC
YFPRTPCRF	TTCTGTCTCCGGTGAAGGTGAA
YFPRTPCRR	TAAGGTTGGCCATGGAACTGGCAA
ALG9RTPCRF	ATCGTGAAATTGCAGGCAGCTTGG
ALG9RTPCRR	CATGGCAACGGCAGAAGGCAATAA
ADE3RTPCRF	CTAATGCTGTGGTCTTGGTTG
ADE3RTPCRR	AGTGTATGCGGAAGGTAAAGG
MCS-Fwd-SpeI	G-ACTAGT-ATGTCTAAAGGTGAAGAATTATTCACTGG
MCS-Rev-2	CCCCG-CTCGAG-TTATTTGTACAATTCATCCATACCATGGG

Appendix Table A7-3: Primers used in this study (IDT)

Fragment Name	Template	Forward Primer/Restriction Enzyme	Reverse Primer/Restriction Enzyme
Dicer	pYES2.1-Dcr1	ScDcr1F	ScDcr1R
Ago2	pYES2.1-Ago1	ScAgo1F	ScAgo1R
YFPfrag100-F	p416-GPD-YFP	YFPfrag1SpeIF	YFPfrag1EcoRIR
YFPfrag100-R	p416-GPD-YFP	YFPfrag1XhoIF	YFPfrag1EcoRIR
YFPhp100	p424-GPD-YFPhp100	SpeI	KpnI
YFPfrag200-F	p416-GPD-YFP	YFPfrag3SpeIF	YFPfrag3EcoRIR
Intron	None	SmallAIF	SmallAIR
YFPfrag200-R	p416-GPD-YFP	YFPfrag3XhoIF	YFPfrag3SalIR
YFPhp200	p424-GPD-YFPhp200	SpeI	KpnI

YFPfrag240-F	p416-GPD-YFP	YFP-6SpeF	YFP-6EcoRIR
YFPfrag240-R	p416-GPD-YFP	YFP-6SalIF	YFP-6XhoIR
ADE3frag-F	*S. cerevisiae* BY4741 genome	ADE3-1SpeF	ADE3-1EcoRIR
ADE3frag-R	*S. cerevisiae* BY4741 genome	ADE3-1SalIF	ADE3-1XhoIR
ADE3hp	p414-GPD-ADE3hp	SpeI	KpnI
TKC6-p415-GPD	p415-GPD	TKC6-p4XXF	GPD rev
GPD-Dicer-TKC1	pYES2.1-Dcr1	GPDDicerF	TKC1DicerR
TKC1-TEF-Ago2	p416-TEF	TKC1TEFF	AgoTEFR
TEF-Ago2-TKC6	pYES2.1-Ago1	TEFAgoF	TKC6AgoR
TRP-CYC-YFP-TRP	p416-CYC-YFP	BefPromF	AftMarkerR
TRP-GPD-YFP-TRP	p416-GPD-YFP	BefPromF	AftMarkerR
YFP	p416-GPD-YFP	MCS-Fwd-SpeI	MCS-Rev-2

Appendix Table A7-4: DNA fragments generated in this study

Plasmid Name	Backbone	Insert	Rest. Enz. 1	Rest. Enz. 2
p413-GPD-Dicer	p413-GPD	Dicer	SpeI	XhoI
p415-GPD-Ago2	p415-GPD	Ago2	SpeI	XhoI
p424-GPD-YFPfrag100	p424-GPD	YFPfrag100-F	SpeI	EcoRI
p424-GPD-YFPhp100	p424-GPD-YFPfrag100	YFPfrag100-R	EcoRI	XhoI
p424-CYC-YFPhp100	p424-CYC	YFPhp100	SpeI	KpnI
p424-GPD-YFPfrag200	p424-GPD	YFPfrag200-F	SpeI	EcoRI
p424-GPD-YFPfrag200intron	p424-GPD-YFPfrag200	Intron	EcoRI	SalI
p424-GPD-YFPhp200	p424-GPD-YFPfrag200intron	YFPfrag200-R	SalI	XhoI
p414-GPD-YFPhp200	p414-GPD	YFPhp200	SpeI	KpnI
p41K-GPD-YFPhp200	p41K-GPD	YFPhp200	SpeI	KpnI
p414-GPD-YFPfrag240	p414-GPD	YFPfrag240-F	SpeI	EcoRI
p414-GPD-YFPfrag240intron	p414-GPD-YFPfrag240	Intron	EcoRI	SalI

p414-GPD-YFPhp240	p414-GPD-YFPfrag240intron	YFPfrag240-R	SalI	XhoI
p414-GPD-ADE3frag	p414-GPD	ADE3frag-F	SpeI	EcoRI
p414-GPD-ADE3fragintron	p414-GPD-ADE3frag	Intron	EcoRI	SalI
p414-GPD-ADE3hp	p414-GPD-ADE3fragintron	ADE3frag-R	SalI	XhoI
p414-TEF-ADE3hp	p414-TEF	ADE3hp	SpeI	KpnI
p414-CYC-ADE3hp	p414-CYC	ADE3hp	SpeI	KpnI
p413-GPD-YFP	p413-GPD	YFP	SpeI	XhoI
p423-GPD-YFP	p413-GPD	YFP	SpeI	XhoI
p41K-GPD-YFP	p41K-GPD	YFP	Spe	XhoI

Appendix Table A7-5: Plasmids generated through restriction enzyme cloning

Plasmid Name	DNA Fragments
p415-GPD-Dicer-TEF-Ago2	TKC6-p415-GPD, GPD-Dicer-TKC1, TKC1-TEF-Ago2, TEF-Ago2-TKC6

Appendix Table A7-6: Plasmids generated through homologous recombination cloning

Strain Name	Parent Strain	PCR Fragment
BY4741 TRP1::CYC-YFP	*S. cerevisiae* BY4741	TRP-CYC-YFP-TRP
BY4741 TRP1::GPD-YFP	*S. cerevisiae* BY4741	TRP-GPD-YFP-TRP

Appendix Table A7-7: Strains generated through genome editing

Strain No.	Plasmid 1	Plasmid 2	Plasmid 3	Plasmid 4	Host	Growth Medium
1	p416-GPD-YFP	p415-GPD-Ago2	p413-GPD-Dicer	p424-CYC-YFPhp100	BY4741 delTRP	YSC -His-Ura-Leu-Trp
2	p416-GPD-YFP	p415-GPD-Ago2	p413-GPD-Dicer	p424-GPD-YFPhp100	BY4741 delTRP	YSC -His-Ura-Leu-Trp
3	p416-GPD-YFP	p415-GPD-Ago2	p413-GPD-Dicer	p424-GPD-YFPhp20	BY4741 delTRP	YSC -His-Ura-

				0		Leu-Trp
4	None	p415-GPD-Ago2	p413-GPD-Dicer	p424-GPD-YFPhp200	BY4741 TRP1::GPD-YFP	YSC - His-Ura-Leu-Trp
5	None	p415-GPD-Ago2	p413-GPD-Dicer	p414-GPD-YFPhp200	BY4741 TRP1::GPD-YFP	YSC - His-Ura-Leu-Trp
6	None	p415-GPD-Ago2	p413-GPD-Dicer	p41K-GPD-YFPhp200	BY4741 TRP1::GPD-YFP	YSC - His-Ura-Leu +G418 (2g/L)
7	p416-CYC-YFP	p415-GPD-Ago2	p413-GPD-Dicer	p424-CYC-YFPhp100	BY4741 delTRP	YSC - His-Ura-Leu-Trp
8	p416-CYC-YFP	p415-GPD-Ago2	p413-GPD-Dicer	p424-GPD-YFPhp100	BY4741 delTRP	YSC - His-Ura-Leu-Trp
9	p416-CYC-YFP	p415-GPD-Ago2	p413-GPD-Dicer	p424-GPD-YFPhp200	BY4741 delTRP	YSC - His-Ura-Leu-Trp
10	None	p415-GPD-Ago2	p413-GPD-Dicer	p424-GPD-YFPhp200	BY4741 TRP1::CYC-YFP	YSC - His-Ura-Leu-Trp
11	None	p415-GPD-Ago2	p413-GPD-Dicer	p414-GPD-YFPhp200	BY4741 TRP1::CYC-YFP	YSC - His-Ura-Leu-Trp
12	None	p415-GPD-Ago2	p413-GPD-Dicer	p41K-GPD-YFPhp20	BY4741 TRP1::CYC-YFP	YSC - His-Ura-

				0		Leu +G418 (2g/L)
13	p416-CYC-YFP	p415-GPD-Ago2	p413-GPD-Dicer	p424-GPD	BY4741 delTRP	YSC - His- Ura- Leu- Trp
14	p416-GPD-YFP	p415-GPD-Dicer-TEF-Ago2	None	p414-GPD-YFPhp240	BY4741 delTRP	YSC - Ura- Leu- Trp
15	p416-GPD-YFP	p415-GPD-Dicer-TEF-Ago2	None	p414-GPD-YFPhp240	CEN.PK2	YSC - Ura- Leu- Trp
16	p416-GPD-YFP	p415-GPD-Dicer-TEF-Ago2	None	p414-GPD-YFPhp240	Sigma 10560-4A	YSC - Ura- Leu- Trp
17	p416-GPD-YFP	p415-GPD-Dicer-TEF-Ago2	None	p414-GPD-ADE3hp	BY4741 delTRP	YSC - Ura- Leu- Trp
18	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	None	BY4741 delADE3	YSC - Ura- Leu- Trp
19	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-CYC	BY4741 delTRP	YSC - Ura- Leu- Trp
20	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-CYC-ADE3	BY4741 delTRP	YSC - Ura- Leu- Trp
21	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-	None	p414-TEF	BY4741 delTRP	YSC - Ura- Leu-

		TEF-Ago2				Trp
22	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-TEF-ADE3hp	BY4741 delTRP	YSC - Ura-Leu-Trp
23	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-GPD	BY4741 delTRP	YSC - Ura-Leu-Trp
24	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-GPD-ADE3hp	BY4741 delTRP	YSC - Ura-Leu-Trp
25	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-GPD-YFPhp200	BY4741 delTRP	YSC - Ura-Leu-Trp
26	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-CYC	CEN.PK2	YSC - Ura-Leu-Trp
27	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-CYC-ADE3	CEN.PK2	YSC - Ura-Leu-Trp
28	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-TEF	CEN.PK2	YSC - Ura-Leu-Trp
29	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-TEF-ADE3hp	CEN.PK2	YSC - Ura-Leu-Trp
30	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-	None	p414-GPD	CEN.PK2	YSC - Ura-Leu-Trp

		Ago2				
31	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-GPD-ADE3hp	CEN.PK2	YSC -Ura-Leu-Trp
32	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-GPD-YFPhp200	CEN.PK2	YSC -Ura-Leu-Trp
33	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-CYC	Sigma 10560-4A	YSC -Ura-Leu-Trp
34	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-CYC-ADE3	Sigma 10560-4A	YSC -Ura-Leu-Trp
35	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-TEF	Sigma 10560-4A	YSC -Ura-Leu-Trp
36	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-TEF-50ADE3hp	Sigma 10560-4A	YSC -Ura-Leu-Trp
37	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-GPD	Sigma 10560-4A	YSC -Ura-Leu-Trp
38	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-GPD-ADE3hp	Sigma 10560-4A	YSC -Ura-Leu-Trp
39	p416-UASCLBUASCITUASTEFpGPD-CAD1	p415-GPD-Dicer-TEF-Ago2	None	p414-GPD-YFPhp200	Sigma 10560-4A	YSC -Ura-Leu-Trp

40	p413-GPD-YFP	None	None	None	BY4741	YSC -His
41	p423-GPD-YFP	None	None	None	BY4741	YSC -His
42	None	P415-GPD	P413-GPD	P414-GPD	BY4741 TRP1::GPD-YFP	YSC -His-Ura-Leu-Trp
43	None	p415-GPD-Ago2	p413-GPD-Dicer	P414-GPD	BY4741 TRP1::GPD-YFP	YSC -His-Ura-Leu-Trp
44	None	p415-GPD-Ago2	p413-GPD-Dicer	p414-GPD-YFPhp200	BY4741 TRP1::GPD-YFP	YSC -His-Ura-Leu-Trp
45	None	p415-GPD-Dicer-TEF-Ago2	P413-GPD	P414-GPD	BY4741 TRP1::GPD-YFP	YSC -His-Ura-Leu-Trp
46	None	p415-GPD-Dicer-TEF-Ago2	P413-GPD	p414-GPD-YFPhp200	BY4741 TRP1::GPD-YFP	YSC -His-Ura-Leu-Trp
47	p41K-GPD-YFP	None	None	None	BY4741 TRP1::GPD-YFP	YSC +G418 (2g/L)

Appendix Table A7-8: Strains generated through plasmid transformation

Figure	Category	Series	Strain No.
7-2: Gene knockdowns attained by each design cycle	Strong YFP Expression	Design Cycle 0	1
		Design Cycle 1	2
		Design Cycle 2	3
		Design Cycle 3	4
		Design Cycle	5

		4	
		Design Cycle 5	6
	Weak YFP Expression	Design Cycle 0	7
		Design Cycle 1	8
		Design Cycle 2	9
		Design Cycle 3	10
		Design Cycle 4	11
		Design Cycle 5	12
		No Hairpin	13
7-3: Growth Rate of Yeast Expressing the RNAi System	No RNAi		42
	Dual RNAi Plasmid		43
	Dual RNAi Plasmid + YFPhp		44
	Single RNAi Plasmid		45
	Single RNAi Plasmid + YFPhp		46
7-4A: Distribution of Knockdown Level in Strains Expressing Hairpins from High and Low Copy Plasmids		High Copy	3
		Low Copy (Auxotrophic)	5
		Low Copy (Antibiotic)	6
7-4B: Variance in the Copy Number of High and Low Copy Plasmids	High Copy Auxotrophic		40
	Low Copy Auxotrophic		41
	Low Copy Antibiotic		47
7-5: Gene knockdown in alternate strains of yeast	BY4741		14
	CEN.PK2		15
	Sigma 10560-4A		16
	Sham Hairpin		17
7-6: Rapid Prototyping of gene knockdowns conferring increased itaconic acid (IA) production in multiple yeast strains	BY4741 ADE3 KO	No Hairpin	18
	BY4741 Weak Hairpin	No Hairpin	19
		With Hairpin	20
	BY4741 Medium Hairpin	No Hairpin	21
		With Hairpin	22
	BY4741 Strong Hairpin	No Hairpin	23
		With Hairpin	24
	BY4741 Sham Hairpin	No Hairpin	23
		With Hairpin	25

	CEN.PK Weak Hairpin	No Hairpin	26
		With Hairpin	27
	CEN.PK Medium Hairpin	No Hairpin	28
		With Hairpin	29
	CEN.PK Strong Hairpin	No Hairpin	30
		With Hairpin	31
	CEN.PK Sham Hairpin	No Hairpin	30
		With Hairpin	32
	Sigma Weak Hairpin	No Hairpin	33
		With Hairpin	34
	Sigma Medium Hairpin	No Hairpin	35
		With Hairpin	36
	Sigma Strong Hairpin	No Hairpin	37
		With Hairpin	38
	Sigma Sham Hairpin	No Hairpin	37
		With Hairpin	39
7-7: Downregulation of ADE3 mRNA	BY4741	No Hairpin	23
		Sham Hairpin	25
		Strong Hairpin	24
		Gene Knockout	18
	CEN.PK	No Hairpin	30
		Sham Hairpin	32
		Weak Hairpin	27
		Medium Hairpin	29
		Strong Hairpin	31
	Sigma	No Hairpin	37
		Sham Hairpin	39
		Strong Hairpin	38

Appendix Table A7-9: Strains used in experiments described in this study

Name	
1B-5-1	TGCGGGGTACTCTTGCTATCGAATTCTTTTTTTTTTTTTTTTTTTTTTTT GAAAATGAATGAATTGATGCGCTTACTACTTACTTACATACGGTTTTTA TTCAAGTATATTATCATTAACATTAGTTGGTTAGACCAATGACACCACA GGCTGGTCTTGGACCGGCATTACCAGTCTTCAAAGATTCTTCAGTGTCA CCCTTACCTAAGTCATCTTGGCCGGCGTGGATAACGACGCTTCTGCCTA CAACGGAGGTAGGACCGATAAGCTTGATCAAAGAGTCCTTGAAGGAGC CCTTGGCCACACCATTTTCGTCCGTCTTTACGTTACCCATGTCACCGACA

	TGTCTGACTTCNTCAGTTGGAGCACCATGTGTCTTTTTGAAAGGATTGA AGTGAGGACCA
1B-6-1	TAGATTCGTCTCCAAGTTGGCTGAAGAAAAAATCAGAGCTGCTGGTGG TGTTGTTGAATTGATCGCTTAAGCGCATCAACAAAAACTCTATGTATTT TCCAATAAATTATATATCTTCAGTTTAATCTAATTCAACATCTACTTCTG TATTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGAA TTCNATAGCAAGAGTACCCCGCA
1B-8-1	CCTGACTTGCACGCTTGGCACGGTGACTGGCCATTGCCAGTTAAGCTAC CATTAGTCGGTGGTCACGAAGGTGCCGGTGTCGTTGTCGGCATGGGTGA AAACGTTAAGGGCTGGAAGATCGGTGACTACGCCGGTATCAAATGGTT GAACGGTTCTTGTATGGCCTGTGAATACTGTGAATTGGGTAACGAATCC AACTGTCCTCACGCTGACTTGTC
1B-12-1	CTTTTGAACAGATTCGTCGTCGAATCTTCTACCAATCAAACGCTTAGCA TCGAAGACAGTGTTTTCTG
1B-13-1	TTGGACTTCTTCGCCAGAGGTTTGGTCAAGTCTCCAATCAAGGTTGTCG GCTTGTCTACCTTGCCAGAAATTTACGAAAAGATGGAAAAGGGTCAAA TCGTTGGTAGATACGTTGTTGACACTTCTAAATAAGCGAATTTCTTATG ATTTATGATTTTTATTATTAAATAAGTTATAAAAAAAATAAGTGTATAC AAAAAAAAAAAAAAAAAAAAAAAAAGAATTCGATAGCAAGAGTACCC CGCA
1B-14-1	GTGCCGCTAACTTGAGAGTCTTGAACTCCTACTGGGTTAACCAAGATTC TACTTACAAGTACTTCGAAGTTATCTTGGTCGACCCTCAACACAAGGCT ATCAGAAGAGATGCTCGTTACAACTGGATCTGTGACCCAGTTCACAAG CACCGTGAAGCTAGAGGTTTGACTGCCACTGGTAAGAAATCCAGAGGT A
1B-17-1	TGCGGGGTACTCTTGCTATCGAATTCTTTTTTTTTTTTTTTTTTTTTAA AATGATAT
1B-19-1	CATCCTGGTCTATATAAAGACCTGGTTGACCTGGAGCGCTAAGTTTTCT TCCTTTTGGATTCAAAGGTACGTTTAATTTGGGTATTTTCTCAGTTAATG AGTTTTCCATACCTAATTTCATGTATTTACCTCTTTGATATTTGATTTCTA AGCCAAGGG
1B-21-1	CCGAATTCTGCTTCGGTATGATAGGAAGAGCCGACATCGAAGAATCAA AAAGCAATGTCGCTATGAG
1B-22-1	CATTAAAGAACTTAGAATAAGAAAGCGACACCGGCAGCGACGGCAGC ACCGAAGACACCAGCACCCAAAGCGTTGGAAGCAGCAGCACCAGTGG AGATCTTGGTGCTGGTGGTGTTGGAACCGTTGGTACGGTTGGAACCATT GGTAGAAACGTTAGCA
1B-23-1	TTTATGATTTTTATTATTAAATAAGTTATAAAAAAAATAAGTGTATACA AATTTTAAAGTGACTCTTAGGTTTTAAAAAAAAAAAAAAAA
1B-24-1	TTTACCAAAAACGTCTTAGAACGGTCTGACGGCACTGTACTTGCACCTA TCGTAAAATATGGAGACTTTTACTGGGTATCTAAAAAGTACTTGCTTCC ATCAAATATCTCCGTACCCACCATCAATAATGTCCATACAAGTGAAAGT ACACGCAAATATCCTTATCCTTTCATTCATCGAATGCTTGCGCATGCCA ATGCACAGACAAATTCGATACTCACTTAAAAATAA
1B-28-1	TGACAATCAAAACCATACCAACACCCATGTTGAAGGTTCTTAAAATGTC ATCGTGTGGAACATTACCTGCTTGACCAAACCATTTGAAGACACGGGGT

	ACTTCCCAGGTTGACATATCAACGCGGGCCTGTAGGTGGTCTGGAATAG CTCTTGGGATATTCTCTACTAAACCACCACCTGTTATATGAGCTAAACC TAGTAGTAGTCTTTGTCTAATTGATGGCAATAATTGCTTGACGTAAATT TTTGTTGGTTCAAGAATACCTTCACCTAACGTCTTAGATTCATCCCATGG ACATGGAGCGTCCCATGGTAATGCTACATGTTGAATAATTTTTCTAACC AAAGAGAAACCATTAGAATGAACACCGCTAGAGGCGAGACCCAGAAG AACATCTCCTGC
1B-29-1	TGCGGGGTACTCTTGCTATCGAATTCTTTTTTTTTTTTTTTTTTTTTTTG GATATTGAGATAAATTTTCCTTCAATTAANATAATAAAACATGTTATAT AAAATCANACAAAATAATATGTAAATTTTTAACGTATTATA
1B-30-1	GGTGGTGNTTCTTTGAAGCCAGAATTTGTTGATATCATCAACTCTAGAA ACTAAGATTAATATAATTATATAAAAATATTATCTTCTTTTCTTTATATC TAGTGTTATGTAAAATAAATTGATGACTACGGAAAAAAAAAAAAAAAA AAAAAAGAATTCGATAGCAAGAGTACCCCGCAG
1B-32-1	CCCGAAATTTACGAAAAGATGGAAAAGGGTCAAATCGTTGGTAGATAC GTTGTTGACACTTNTAAATAAGCGAATTTCTTATGATTTATGATTTTTAT TATTAAATAAGTTATAAAAAAAATAAGTGTATACAAATTTTAAAGNGA CTNTTNGGTTTTAAAAAAAAAAAAAAAAAAAAAAAAAAAGAATTCG ATAGCAAGAGTACCCCGCA
1B-33-1	TTTTTATACTTTTTCCTTTTAGATATATGTACTTTTGGCTTAATTTAATAT AATTAACTATAAAAAAAAAAAAAAAAAAAAAAAAAAGAATTCGATAGC AAGAGTACCCCGCA
IB-1-1	TAGATTCGTCTCCAAGTTGGCTGAAGAAAAAATCAGAGCTGCTGGTGGT GTTGTTGAATTGATCGCTTAAGCGCATCAACAAAAACTCTATGTATTTTC CAATAAATTATATATCTTCAGTTTAATCTAATTCAACATCTACTTCTGTAT TATTTCTAAAAAAAAAAAAAAAAAAAAAAAAAAGAATTCGATAGCAAGA GTACCCCGCA
IB-6-1	TAACGCCCCGGGCGAATATCGTAGGTTGGATGCGATGCGATTGCAAAAA AAAAAAAAAAAAAAAAAAAAGAATTCGATAGCAAGAGTACCCCGCAG
IB-7-1	CCCAACTAATGGTATGCAAGCATTTTTATATGTGTGAAAAAAAAAAAAA AAAAAAAAAAAAAAGAATTCAATAGCAANANTACCCCGCA
IB-9-1	CTGCGGGGTACTCTTGCTATCGAATTCTTTTTTTTTTTTTTTTTTTTTT TTTTTTTTAAGTGTACACGTTGAGTTTATTGTTTTATTTCCCCTACATATA TATACATATATATGAAATTACTTTACGTACGTATAAGCTTTGTTCAGTCA TCATGAACCANTGTCTTTTCGTACTGTTCTAAGGACATTAAACCCTCNAC CTGTTCCACATTAACGCCCTCACCAAGCTTCATTTTGACTAGCCAGCCGT CACCCATAGGATCTTCGTTCACCACACCTGGATTTTCCTCAANATTANTG TTAATTTCCTCTACGGTACCATCGGC
IB-12-1	CTTGGATTTGATAATCAAGATAGGGTCCACTTTATCTCAGTTGTCCTTTC AAGAGAAGGTTTATCCAGTTTTATTGAGTGATGCGAACTTTCCAGTATTA TTAAAAAAAGCCACCATCTGCCTGATCGATAATTTAGACACATTGAAGC AAAAAGTTAAACGCTCTGATTTCTTGGAAAATATACTCAAGCCGCTATTT AATTATGTTTTACATGACTCCGAAAGTCACAT
IB-16-1	TTCTGGACAGTAAGTAGTATATTCAGTGACTACAGTGTAGTCAGTGGTGT AGTCAGTAGATGGGTTGTGTTTAGTTGGTGGGTCTATTGCCGATATGTTC CCAAGTGAAACAAGAGGTAAGGCTATTGCTTTCTTCGCTTTTGCTCCTTA

	CGTTGGTCCCGTTGTTGGTCCACTAGTTAACGGTTTTATTTCCGTTTCTAC CGGACGTATGGG
IB-19-1	ATTTGTTACTTATTTTTTATTAACTAGCTTTGGGGGAGAGCCATGGAAAA TAGCACTCGGTCTTGTGGCGG
IB-20-1	TAACGCCCCGGGCGAATATCGTAGGTTGGATGCGATGCGATTGCAAAAA AAAAAAAAAAAAAAAAAGAATTCGATAGCAAGAGTACCCCGCAG
IB-21-1	CTTCTTTGGGGCTTCAGTGGTCAATGGGCACCAGGTGGTGTATTGAGTGA TAACGTCATCGACGGTGACGGT
IB-24-1	TGTAAATCAACATCAGCCTTCTTACAGACCAAGTTGGAGTAACGACGAC CAACACCCTTGATAGTGGTCAAAGCGTAAACGATCTTAATGTTACCGTCA ACGTTAGTGTTCAACAAACGTAAAATGTGTTGGAAGGAACCTTGTTCTTG GACAACTAAAGACATCTTTTATCCGCTCTTGTGTATACGTTCCGCAATCG CATCGCATCCAACCTACGATATG
IB-25-1	CATTCCATCTAAAGCTAAATATTGGCGAGAGACCGATAGCGAACAAGTA CAGTGATGGAAAAAAAAAAAAAAAAAAAAAAAGAACAACGGCACGC AATGTTGCTGTGACAACCGTAGGCAGATAACTTGGCTTTTTTTAC
IB-26-1	TTAACATTAGTTGGTTAGACCAATGACACCACAGGCTGGTCTTGGACCG GCATTACCAGTCTTCAAAGATTCTTCAGTGTCACCCTTACCTAAGTCATC TTGGCCGGCGTGGATAACG
IB-28-1	TCTGTCCNTGTCCAAGTGTAACTTAGTGATAACAAGCTTGGATGGGTGCA AGTTAATTGGAACGGAAGCACCGTTGACCTTTTCCTTGGTGACCTTGTCA ACTTGAACAGCAAACTTCAATCTGTAAACAGATGAAATCTTACCTTCTTG ACCCTTCTTGGAACCACGAACAACCAAGACTTCATCGTCTCTTCTGATTG GCAAAGCCTTGATACCATATTGAGCTG
IB-32-1	CNNNNCNTACTTTAACGACGCTCAAAGACAAGCTACCAAGGATGCCGGT GCCATTTCTGGTTTGAACGTTTTGCGTATCATCAACGAACCTACTGCCGC TGCTG
IB-33-1	CTCTTAGCTCTTTCAGCAGCAGTTCTCAATCTTCTCAAAGCTCTGGCATC GTCGGAGATGTCCAAACCAGTCTTCTTCTTGAATTCAGCCTTGAAGTGTT CCAACAAGTTGGTGTCGAAATCTTGACCACCCAAGTGAGTGTTACCGGA AGTAGATTTAACAGTGTAAACACCACCAGCAATGTGCAACAAGGAAACA TCGAAAGTACCACCACCCATGCGGGGTACTCTTGCTATCGAATTCTTTTT TTTTTTTT
IB-35-1	ACTTTATATTTAATATCTAGATATTACATAATTTCCTCTCTAATAAAATAT CATTAATAAAATAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAATTTC

Appendix Table A7-10: Knockdown cassettes confirmed to improved the growth rate of BY4741 in 1-butanol an isobutanol.

APPENDIX B: SOFTWARE WRITTEN IN THIS WORK

Appendix B1: Software Written for Chapter 2

Readme for MATLAB scripts

Table of Contents:

1. The Purpose of the Scripts

Computational redesign of native or synthetic promoters for altered nucleosome affinity.

2. Contact Information

Hal Alper
The University of Texas at Austin
200 E. Dean Keeton Street, Stop C0400
Austin, TX 78712-1589
CPE 5.408 / phone: (512) 471-4417 / fax: (512) 471-7060
E-mail: halper@che.utexas.edu

3. List and Description of Scripts

affinity.m
Takes a DNA sequence and computes nucleosome affinity values for each nucleotide.

containsforbidden.m
This script looks for instances of user-defined DNA motifs in a DNA sequence. Motifs can include degenerate bases.

gccontent.m
Computes the GC content of a sequence

gcprofile.m
Calculates the GC contents of each 100bp sliding window of input DNA sequence.

maxprom.m
This program will take a promoter and iteratively decrease predicted nucleosome occupancy in user-defined basepair increments until the occupancy can no longer be decreased.

nucleomin.m
Nucleomin takes a sequence as input and searches all n-nucleotide variants of the starting sequence to find the one with the minimum predicted nucleosome affinity, with the requirement that the sequence is also synthesizable and contains no additional or fewer transcription factor binding sites. n is user-defined.

problemrank.m
Notes the positions of the input DNA sequence which contain particular DNA motifs and ranks them from lowest nucleotide to highest nucleotide.

randprom.m
Initializes a random DNA sequence for a synthetic promoter and generates a list of sequences within the promoter which must be conserved during the design process based on user specifications.

randseq.m
Makes a random DNA sequence of the length and GC content specified

remforbidden.m
Tries to remove as many matches to a set of DNA motifs as possible from an input sequence. Users may also specify the locations of bases which may not be changed during this process.

seqarea.m
Computes the cumulative affinity score for a DNA sequence.

seqcheck.m
The sole purpose of this program is to make sure that a sequence can be synthesized by IDT's gblocks. It was sufficient at the time of writing but some features of it may no longer be necessary as synthesis technology improves.

synthprom.m
This function takes a general outline for a promoter and makes a synthetic nucleosome optimized promoter.

4. Installation and Setup

Setup instructions provided for Windows systems.

1) Obtain a copy of MATLAB (tested on r2011b) with the bioinformatics toolbox (tested on r2013b) installed

2) Copy the scripts listed above into the MATLAB working directory

3) Download a copy of the FORTRAN code for NuPoP (as of 6/28/2013 it was located at http://nucleosome.stats.northwestern.edu/ as "NuPoP_F")

4) Edit the FORTRAN code for NuPoP_F as follows:

Replace the following in npred.f90:

REPLACE:

```
implicit none
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

integer i,lfn,mlL,rep,species,order; character*80 fileName; character*3 tpc
real*8  freqL1(4),tranL1(4,4),tranL2(16,4),tranL3(64,4),tranL4(256,4),Pd(500,11)
real*8  freqN4(64,4),tranN4((147-4)*256,4),freqN1(147,4),tranN1(584,4)
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

open(1,file='yourpath/NuPoP_F/profile/freqL.txt')
read(1,*) freqL1; close(1)
open(1,file='yourpath/NuPoP_F/profile/tranL.txt')
do i=1,4; read(1,*) tranL1(i,:); end do; close(1)
open(1,file='yourpath/NuPoP_F/profile/tranL2.txt')
do i=1,16; read(1,*) tranL2(i,:); end do; close(1)
```

```
                open(1,file='yourpath/NuPoP_F/profile/tranL3.txt')
                do i=1,64; read(1,*) tranL3(i,:); end do; close(1)
                open(1,file='yourpath/NuPoP_F/profile/tranL4.txt')
                do i=1,256; read(1,*) tranL4(i,:); end do; close(1)

                open(1,file='yourpath/NuPoP_F/profile/147freqN.txt')
                do i=1,147; read(1,*) freqN1(i,:); end do
                close(1)
                open(1,file='yourpath/NuPoP_F/profile/147tranN.txt')
                do i=1,584; read(1,*) tranN1(i,:); end do
                close(1)

                open(1,file='yourpath/NuPoP_F/profile/146-149freqN4.txt')
                do i=1,64; read(1,*) freqN4(i,:); end do; close(1)
                open(1,file='yourpath/NuPoP_F/profile/146-149tranN4.txt')
                do i=1,(147-4)*256; read(1,*) tranN4(i,:); end do; close(1)

                open(1,file='yourpath/NuPoP_F/profile/Pd.txt')
                do i=1,500; read(1,*) Pd(i,1:11); end do; close(1)

                write(*,'(a)') 'Please input'
                write(*,'(a)',advance='no') '  File name of DNA sequence (FASTA)                    :  ';
read*,fileName
                mlL=500
                write(*,'(a)',advance='no') '  Order of Markov model (1 or 4)                       :  '; read*,order
                if(order/=1.and.order/=4) then; print*,'1 or 4 should be inputed! stop.'; stop; end if
                rep=1
                print*,' '
                write(*,'(a)') 'Select the species from the following list:'
                print*,'1=Human          2=Mouse          3=Rat'
                print*,'4=Zebrafish      5=D. melanogaster  6=C. elegans'
                print*,'7=S. cerevisiae  8=C. albicans      9=S. pombe'
                print*,'10=A. thaliana   11=Maize          0=Other'
                print*,' '
                write(*,'(a)',advance='no') 'Input the lable of selected species                  :  ';
read*,species
                print*,' '
                write(*,'(a)') 'Predicting.....'

                lfn=len_trim(fileName)
                if(order==1) then
                        call vtbfb(lfn,trim(fileName),freqL1,tranL1,freqN1,tranN1,mlL,rep,species,Pd)
                else if(order==4) then
                        call
vtbfbNL4(lfn,trim(fileName),freqL1,tranL1,tranL2,tranL3,tranL4,freqN4,tranN4,mlL,rep,species,Pd)
                end if

                write(*,'(a)') '              Done.'
              end

              WITH:
```

```
implicit none
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

integer  i,lfn,mlL,rep,species,order;  character*80  fileName,stringorder,stringspecies;
character*3 tpc
real*8  freqL1(4),tranL1(4,4),tranL2(16,4),tranL3(64,4),tranL4(256,4),Pd(500,11)
real*8  freqN4(64,4),tranN4((147-4)*256,4),freqN1(147,4),tranN1(584,4)
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

open(1,file='yourpath/NuPoP_F/profile/freqL.txt')
read(1,*) freqL1; close(1)
open(1,file='yourpath/NuPoP_F/profile/tranL.txt')
do i=1,4; read(1,*) tranL1(i,:); end do; close(1)
open(1,file='yourpath/NuPoP_F/profile/tranL2.txt')
do i=1,16; read(1,*) tranL2(i,:); end do; close(1)
open(1,file='yourpath/NuPoP_F/profile/tranL3.txt')
do i=1,64; read(1,*) tranL3(i,:); end do; close(1)
open(1,file='yourpath/NuPoP_F/profile/tranL4.txt')
do i=1,256; read(1,*) tranL4(i,:); end do; close(1)

open(1,file='yourpath/NuPoP_F/profile/147freqN.txt')
do i=1,147; read(1,*) freqN1(i,:); end do
close(1)
open(1,file='yourpath/NuPoP_F/profile/147tranN.txt')
do i=1,584; read(1,*) tranN1(i,:); end do
close(1)

open(1,file='yourpath/NuPoP_F/profile/146-149freqN4.txt')
do i=1,64; read(1,*) freqN4(i,:); end do; close(1)
open(1,file='yourpath/NuPoP_F/profile/146-149tranN4.txt')
do i=1,(147-4)*256; read(1,*) tranN4(i,:); end do; close(1)

open(1,file='yourpath/NuPoP_F/profile/Pd.txt')
do i=1,500; read(1,*) Pd(i,1:11); end do; close(1)

CALL GETARG(1,fileName)
CALL GETARG(2,stringorder)
CALL GETARG(3,stringspecies)

read(stringorder,*) order
read(stringspecies,*) species

mlL=500

if(order/=1.and.order/=4) then; print*,'1 or 4 should be inputed! stop.'; stop; end if
rep=1


lfn=len_trim(fileName)
if(order==1) then
        call vtbfb(lfn,trim(fileName),freqL1,tranL1,freqN1,tranN1,mlL,rep,species,Pd)
else if(order==4) then
```

```
                call
vtbfbNL4(lfn,trim(fileName),freqL1,tranL1,tranL2,tranL3,tranL4,freqN4,tranN4,mlL,rep,species,Pd)
              end if

            end
```

5) Replace the string "yourpath" with the directory in which NuPoP_F is located

6) Rename the file to "Npred2.f90" and compile Npred2.f90 as Npred2.exe using the instructions provided in the manual included with NuPoP_F. See the NuPoP_F manual for more detailed information as to the installation of NuPoP.

7) Add the directory containing Npred2.exe to your system's path.

You're now ready to begin designing promoters!

5. Usage

1) Pick a promoter. Promoters must be designed including 200bp upstream and 100bp downstream of its genomic or plasmid context. This will ensure that the nucleosome affinity values calculated for promoter variants will be comparable to one another. Note the nucleotide positions of the start and the end of the promoter.

2) Annotate the transcription factor binding sites and note the nucleotides covered by the binding sites. A particularly user-friendly repository is the Yeast Promoter Atlas http://ypa.ee.ncku.edu.tw/

3) Annotate any sequences you would not want introduced into the designed promoter. These sequences, if present in the wild-type promoter, will not be altered.

4) Build input files. For the TEF promoter, we enter the DNA sequence of the promoter itself plus 200bp upstream and 100bp downstream as follows:

```
TEF='GGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTACCTCACTCATTAG
GCACCCCAGGCTTTACACTTTATGCTTCCGGCTCCTATGTTGTGTGGAATTGTGAGCGGATAAC
AATTTCACACAGGAAACAGCTATGACCATGATTACGCCAAGCGCGCAATTAACCCTCACTAAA
GGGAACAAAAGCTGGAGCTCATAGCTTCAAAATGTTTCTACTCCTTTTTTACTCTTCCAGATTT
TCTCGGACTCCGCGCATCGCCGTACCACTTCAAAACACCCAAGCACAGCATACTAAATTTCCC
CTCTTTCTTCCTCTAGGGTGTCGTTAATTACCCGTACTAAAGGTTTGGAAAAGAAAAAAGAGA
CCGCCTCGTTTCTTTTTCTTCGTCGAAAAAGGCAATAAAAATTTTTATCACGTTTCTTTTTCTTG
AAAATTTTTTTTTTGATTTTTTTCTCTTTCGATGACCTCCCATTGATATTTAAGTTAATAAACGG
TCTTCAATTTCTCAAGTTTCAGTTTCATTTTTCTTGTTCTATTACAACTTTTTTTACTTCTTGCTC
ATTAGAAAGAAAGCATAGCAATCTAATCTAAGTTTTCTAGAACTAGTATGTCTAAAGGTGAAG
AATTATTCACTGGTGTTGTCCCAATTTTGGTTGAATTAGATGGTGATGTTAATGGTCACAAATT
TTCTGTCT';
```

For its nucleotides covered by transcription factor binding sites, we enter:

```
TEFforbidden=[281:291 334:343 377:383 443:484];
```

For the sequences which will not be introduced or removed from the designed promoter, make a cell array containing the relevant motifs. We used the TF consensus list found at yeastract.com, in addition to the start codon and TATA box for our studies.

These input files are included in Sample Data.mat

Example MATLAB Commands:

Optimize TEF in 1bp steps:

```
[TEFproms,TEFareas,TEFcurves]=maxprom(TEF,TEFstart,TEFend,1,TEFforbidden,forbiddenseq
s);
```

Design Psynth1 in 1bp steps:
[psynth1proms,psynth1areas,psynth1curves]=synthprom(psynth1params,psynth1start,psynth1end,1,forbiddenseqs);

For each command, the first output is a list of nucleosome-optimized promoters, starting from the wild-type (or seed) sequence, and proceeding in 1bp steps toward a variant with reduced predicted nucleosome affinity. The second output is the corresponding cumulative affinity score for each promoter, and the third output is the nucleosome affinity curves used to compute the cumulative affinity score for each promoter.

As the programs are running, they will periodically display a progress indicator which describes how far along the program is in computing the current mutation.

6. Changelog
No changes yet!

nucleomin.m (MATLAB)

```
function
[maxsequence,maxarea]=nucleomin(sequence,prombeg,promend,numchanges,forbiddensites,forbiddenseqs
)
%nucleomin takes a sequence as input and computes the n-nucleotide variant
%with the minimum nucleosome affinity that is also synthesizable and also
%does not have any additional or fewer transcription factor binding sites. n is user-defined.

%input sequence must be uppercase strings

%forbiddenseqs must be a cell array with each motif in the first row.
%motifs specified in forbiddenseqs will neither be created or destroyed.
%This is for things like TATA boxes or other general purpose transcription
%factors which may be present.  Also ATGs if you like.

%forbiddensites is a row vector of positions that you don't want the
%program to mutate.  For example, things like transcription factor binding
%sites.

%numchanges tells the program how far to search from the parent sequence to
%find an improved promoter.  numchanges=1 searches all single mutants,
%numchanges=2 searches all double mutants, etc...

%Prombeg and promend specify the positions of the beginning and end of the
%promoter in "sequence"  We recommend prombeg be at least 200.

forbiddenruns={'AAAAAAAAA','CCCCCC','TTTTTTTTT','GGGGGG'};
runsstart=containsforbidden(sequence(prombeg:promend),forbiddenruns);
%IDT doesn't like these sequences, so we're making note of where they are.
%For determining if a sequence is synthesizable

maxaffinities=affinity(sequence);
learningcurve=maxaffinities(1:25);
%This is the affinity of the sequence we're starting from.

refforbidden=containsforbidden(sequence,forbiddenseqs);
```

```
%We also save the locations of anything in forbiddenseqs.  For determining
%if a sequence contains any extra transcription factor binding sites.

tomutate=pick(prombeg:promend,numchanges,'r');
basechanges=str2digit(dec2base(0:4^numchanges-1,4,numchanges));
%generates a worklist for all the bases to mutate during the search
%for an improved promoter.  For each entry in tomutate, basechanges is a
%worklist for what to mutate those bases to in its search.

n=1;
testarea=[];
tic
%n and tic are just there if you are impatient and want to see the progress
%of nucleomin in real time. Also initializing testarea.

for i=1:size(tomutate,1)
   %cycles through all the positions needing to be randomized

   for j=1:size(basechanges,1)
      %cycles through all possible bases at the randomized positions

      badseq=0;
      %badseq is 1 if sequence has an issue and should be thrown out,
      %badseq is 0 otherwise.

      testseq=sequence;
      %this is the sequence we're going to be mutating

      unicom=[tomutate(i,:)' basechanges(j,:)'];

      if length(unique(tomutate(i,:)'))==size(unique(unicom,'rows'),1)
         % the previous two lines are for making sure that for more than
         % nbp mutations at a time (n>1), that the (<n)bp mutants are
         % also computed and without unnecessary repetitions

         for k=1:numchanges
            %making the specified mutations to testseq


            if sum(forbiddensites==tomutate(i,k))==0
               %makes sure we're not going to mutate anything in
               %forbiddensites

               if basechanges(j,k)==0
                  if sequence(tomutate(i,k))=='A'
                     badseq=1;
                     %prevents us from mutating to the same base, which
                     %would eat up time.
                  else
                     testseq(tomutate(i,k))='A';
                     %make the mutation
                  end
```

```
            elseif basechanges(j,k)==1
               if sequence(tomutate(i,k))=='C'
                  badseq=1;
               else
                  testseq(tomutate(i,k))='C';
               end
            elseif basechanges(j,k)==2
               if sequence(tomutate(i,k))=='T'
                  badseq=1;
               else
                  testseq(tomutate(i,k))='T';
               end
            elseif basechanges(j,k)==3
               if sequence(tomutate(i,k))=='G'
                  badseq=1;
               else
                  testseq(tomutate(i,k))='G';
               end
            end
         else
            badseq=1;
         end
      end
   else
      badseq=1;
   end
   if badseq==0
      testforbidden=containsforbidden(testseq,forbiddenseqs);
      %looks for forbidden motifs in the mutated sequence

      isok=seqcheck(testseq(prombeg:promend),sequence(prombeg:promend),runsstart);
      %makes sure IDT can synthesize the mutated sequence.

      if isequal(refforbidden,testforbidden)==0||isok==0
         badseq=1;
         %A sequence is bad if it contains a different number of
         %forbidden motifs than the starting sequence or if it
         %cannot be synthsized by IDT.
      end
   end
   if badseq==0;
      testaffinity=affinity(testseq);
      testarea(n)=seqarea(learningcurve,testaffinity,prombeg-73,promend-73);
      testseqs{n}=testseq;
      %if the sequence is ok, this will compute the nucleosome
      %affinity under the promoter and add this area to the list of
      %mutants
      n=n+1;
     percentdone=((i-1)*size(basechanges,1)+j)/(size(tomutate,1)*size(basechanges,1))*100
     timeleft=toc/(percentdone/100)-toc
%  you can enable the previous two lines if you are impatient and want to
%  see progress of nucleomin.
```

```
      end
   end
end
[maxarea,i]=min(testarea);
maxsequence=testseqs{i};
%finds the promoter with the minimum nucleosome affinity and returns it.
Toc
```

maxprom.m (MATLAB)

```
function
[proms,areas,curves]=maxprom(sequence,prombeg,promend,numchanges,forbiddensites,forbiddenseqs)
%This program will take a promoter and iteratively make improvements to it
%in 'numchanges' increments until it can no longer be improved.  See the
%functions this program calls for what each of the input arguments are.

%Outputs the promoter, summed affinity area, and the affinity values along
%the promoter for each iteration in the optimization

proms{1}=sequence;
refaffinity=affinity(sequence);
curves(1,:)=refaffinity;
learningcurve=refaffinity(1:25);
areas(1)=seqarea(learningcurve,refaffinity,prombeg-73,promend-73);
%initializes things

[proms{2},~]=nucleomin(sequence,prombeg,promend,numchanges,forbiddensites,forbiddenseqs);
testaffinity=affinity(proms{2});
[areas(2),curves(2,:)]=seqarea(learningcurve,testaffinity,prombeg-73,promend-73);
%does the first iteration of promoter improvement

[proms{3},~]=nucleomin(proms{2},prombeg,promend,numchanges,forbiddensites,forbiddenseqs);
testaffinity=affinity(proms{3});
[areas(3),curves(3,:)]=seqarea(learningcurve,testaffinity,prombeg-73,promend-73);
%and the second

[proms{4},~]=nucleomin(proms{3},prombeg,promend,numchanges,forbiddensites,forbiddenseqs);
testaffinity=affinity(proms{4});
[areas(4),curves(4,:)]=seqarea(learningcurve,testaffinity,prombeg-73,promend-73);
%and the third

n=4;
while isequal(proms{n},proms{n-2})==0
   %continues to make improvements while we aren't stuck in a local
   %minimum.
   n
   n=n+1;
   [proms{n},~]=nucleomin(proms{n-1},prombeg,promend,numchanges,forbiddensites,forbiddenseqs);
   testaffinity=affinity(proms{n});
   [areas(n),curves(n,:)]=seqarea(learningcurve,testaffinity,prombeg-73,promend-73);
   %makes improvements
```

```
    areas
    %things without semicolons are just so you can see how far along it is.
    % Feel free to take these out if you're hardcore.

    save('maxpromdata.mat')
    % just in case your computer crashes you can start from where you left
    % off.
End
```

randprom.m (MATLAB)

```
function [sequence,forbiddensites]=randprom(params,gc)
%makes an intial random sequence for the synthetic promoter and generates
%forbiddensites from the nonrandom (user-specified) regions of params.

forbiddensites=[];
sequence=[];
for n=1:length(params)
    if ischar(params{n})==1
        a=length(sequence)+1;
        sequence=strcat(sequence,params{n});
        b=length(sequence);
        forbiddensites=[forbiddensites a:b];
    else
        sequence=strcat(sequence,randseq(params{n},gc));
    end
end
```

problemrank.m (MATLAB)

```
function problemsites=problemrank(sequence,forbiddenseqs)
%Notes the positions of sequence containing a motif found in forbiddenseqs
%and ranks them from lowest nucleotide to highest nucleotide.
badsites=containsforbidden(sequence,forbiddenseqs);
problemsites=[];
for n=1:length(badsites)
    if isempty(badsites{n})==0
        problemsites=[problemsites badsites{n}];
    end
end
problemsites=sort(problemsites);
```

gcprofile.m (MATLAB)

```
function GC=gcprofile(seq)
%calculates GC contents in each 100bp sliding window of seq.
len=length(seq);
if len<100
    GC=gccontent(seq);
else
    for n=1:len-99
        GC(n)=gccontent(seq(n:n+99));
    end
```

```
end
```

containsforbidden.m (MATLAB)

```
function forbiddensites=containsforbidden(sequence,forbiddenseqs)
%This script looks for instances of motifs in forbiddenseqs in seq.  motifs
%can include degenerate bases.

for n=1:size(forbiddenseqs,2)
   %cycles through everything in forbiddenseqs

   forbiddenregexp1=seq2regexp(forbiddenseqs{1,n}(1),'Ambiguous',false);
   forbiddenregexp2=seq2regexp(forbiddenseqs{1,n}(2:length(forbiddenseqs{1,n})),'Ambiguous',false);
   forbiddensites{n}=regexp(sequence,strcat(forbiddenregexp1,'(?=',forbiddenregexp2,')'));
   %regexp "consumes" a sequence as it checks for matches so I'm just
   %taking the first nucleotide of the motif and looking ahead to see if
   %it finds the rest after this basepair.  Then I save the positions or matches in
   %forbiddensites
End
```

affinity.m (MATLAB)

```
function affinities=affinity(sequence)
%takes 'sequence' and computes nucleosome affinity values for each
%nucleotide.

%'sequence' is simply a string of nucleotides in uppercase.

fastawrite('sequence.txt','sequence',sequence);
%makes a FASTA-formatted text file containing the DNA sequence.  This
%function is part of the bioinformatics toolbox

system('Npred2 sequence.txt 4 7');
% system('Npred2 sequence.txt 4 1');
%Uses Npred2 to calculate the nucleosome affinities at each base.  This
%MATLAB script forces a 4th-order Markov model and uses data for S. cerevisiae.
%In this command, '4' designates the order of the Markov model and '7' designates the
%organism (in this case, yeast).  Other organisms are found in the manual for NuPoP_F.

%Note that Npred2 must be added to the system path for this script to run.
%If it is not, simply edit the above system command to point to the right
%directory.

%Npred2 is nearly equivalent to the script 'Npred' published in the below reference,
%the only difference being the ability to accept variables from the command
%line.  See the attached Readme

%Xi, L., Fondufe-Mittendor, Y., Xia, L., Flatow, J., Widom, J. and Wang, J.-P.,
%Predicting nucleosome positioning using a duration Hidden Markov Model,
%BMC Bioinformatics, 2010, doi:10.1186/1471-2105-11-346.

system('del "sequence.txt"');
```

```
%deletes temp input file

fid=fopen('sequence.txt_Prediction4.txt');
data=textscan(fid,'%s %s %s %s %s');
fclose(fid);
system('del "sequence.txt_Prediction4.txt"');
%gets data from temp output file and deletes the file

affin=str2double(data{1,5}(2:length(data{1,5}),1));
%makes a vector of affinities from extracted data

affinities=affin(~isnan(affin));
%removes NaNs from affinity
```

gccontent.m (MATLAB)

```
function GC=gccontent(seq)
%computes the gc content of seq
numSeq = double(nt2int(seq));
baseNum = [sum(numSeq == 1) sum(numSeq == 2) sum(numSeq == 3) sum(numSeq == 4)];
GC = 100 * ((baseNum(2) + baseNum(3)) / length(numSeq));
```

randseq.m (MATLAB)

```
function sequence=randseq(n,gc)
%makes a random sequence of the length specified by n and with an average
%gc content of gc
randnum=randi(100,[1,n]);
for i=1:n
   if randnum(i)>=1&&randnum(i)<gc/2
      sequence(i)='C';
   elseif randnum(i)>=gc/2&&randnum(i)<gc
      sequence(i)='G';
   elseif randnum(i)>=gc&&randnum(i)<(gc+(100-gc)/2)
      sequence(i)='T';
   else
      sequence(i)='A';
   end
end
```

synthprom.m (MATLAB)

```
function [proms,areas,curves]=synthprom(params,prombeg,promend,numchanges,forbiddenseqs)
%this function takes a general outline for a promoter (specified in params)
%and makes a synthetic nucleosome optimized promoter.  All variables are
%the same for nucleomin except for params.  Params is a cell array whose
%contents are either numbers or DNA sequences.  numbers represent length of
%random (nucleosome optimizable) unspecified DNA sequences, and DNA
%sequences are TFBSs or anything else you want to keep constant during the
%optimization.  Put each segement in the order you want it to appear.

%Example: {{3},{'AGTAGCA'},{7}} is NNNAGTAGCANNNNNNN
```

```
gc=35;
%GC content of randomly generated portions of the promoter.  Yeast is
%around 35% but if you are in a different organism you can change that
%here.

[sequence,forbiddensites]=randprom(params,gc);
%makes an intial random sequence for the synthetic promoter and generates
%forbiddensites

sequencefix=remforbidden(sequence,forbiddensites,forbiddenseqs);
%removes anything in forbiddenseqs (like TFBSs) randomly generated in sequence, unless
%they're contained in forbiddensites.

[proms,areas,curves]=maxprom(sequencefix,prombeg,promend,numchanges,forbiddensites,forbiddenseqs);
%Takes the initialized sequence and performs a nucleosome optimiztion.
```

remforbidden.m (MATLAB)

```
function sequencefix=remforbidden(sequence,forbiddensites,forbiddenseqs)
%Tries to remove as many motifs found in forbiddenseqs as possible from
%sequence given that no bases in forbiddensites can be changed

isbetter=1;
tic
while isbetter==1;
   problemsites=problemrank(sequence,forbiddenseqs);
   %notes which bases contain motifs in forbiddenseqs
   isbetter=0;
   for i=problemsites
      %iterates through the problem bases
      for j=0:3
         %iterates through all basepair changes
         if isbetter==0;
            % if we haven't removed a motif yet
            badseq=0;
            testseq=sequence;
            if sum(forbiddensites==i)==0
               %makes sure we aren't mutating a forbidden site
               if j==0
                  if testseq(i)=='A'
                     badseq=1;
                  else
                     testseq(i)='A';
                  end
               elseif j==1
                  if testseq(i)=='C'
                     badseq=1;
                  else
                     testseq(i)='C';
                  end
               elseif j==2
                  if testseq(i)=='T'
```

```
                badseq=1;
            else
                testseq(i)='T';
            end
        elseif j==3
            if testseq(i)=='G'
                badseq=1;
            else
                testseq(i)='G';
            end
        end
    else
        badseq=1;
    end
    if badseq==0
        testsites=problemrank(testseq,forbiddenseqs);
        %counts the forbidden motifs in the new sequence
        if length(testsites)<length(problemsites)
            isbetter=1;
            sequence=testseq;
            %if the new sequence contains less motifs than the
            %original, discard the parent and save the good one
            %for the next round
        end
    end
  end
 end
end
sequencefix=sequence;
%this sequence should contain the minimum number of forbidden motifs given
%that we can't change anything in forbiddensites.
Toc
```

seqcheck.m (MATLAB)

```
function isok=seqcheck(seq,parent,runsstart)
%The sole purpose of this program is to make sure that a sequence can be
%synthesized by IDT's gblocks.  It was sufficient at the time of writing but some
%features of it may no longer be necessary as synthesis technology
%improves.

%initiates things
isok=1;
complement=seqcomplement(seq);
len=length(seq);
GCstart=gcprofile(parent);
forbiddenruns={'AAAAAAAAA','CCCCCC','TTTTTTTTT','GGGGGG'};

%check total GC content
GC=gccontent(seq);
if GC>75||GC<=25
```

```
   isok=0;
end

%check GC content every 100bp
if isok==1
   GC=gcprofile(seq);
   if max(GC)>80||min(GC)<24 %checks if GC content is not within acceptable range
      if min(GC)<min(GCstart) %is minimum GC content of new sequence lower than parent?
         isok=0;
      elseif min(GC)==min(GCstart) %is minimum GC content of new sequence equal to parent?
         if sum(GC==min(GC))>=sum(GCstart==min(GCstart)) %is there not less of the minimum GC
value than for the parent?
            isok=0;
         end
      end
   end
end

%Check for homopolymers which are too long
if isok==1
   seqforbidden=containsforbidden(seq,forbiddenruns);
   if
isempty(seqforbidden{1})==0||isempty(seqforbidden{2})==0||isempty(seqforbidden{3})==0||isempty(seqf
orbidden{4})==0 %Are there runs?
      for n=1:4
         if isempty(seqforbidden{n})==1
            moreruns(n)=0;
         else
            moreruns(n)=(sum(seqforbidden{n})<sum(runsstart{n}));
         end
      end
      if sum(moreruns)==0 % Are there equal or more runs in the new sequence than the parent sequence?
         isok=0;
      end
   end
end


% check for hairpins
if isok==1
   %create the dot matrix and rotate it
   for n=1:len
      hpdot(n,:)=complement==seq(n);
   end
   hpdot=rot90(hpdot);
   %search through all the diagonals for runs of "dots" of a certain
   %length.  higher than 8bp hairpin is bad if GC content is greater than
   %80%, higher than 11bp hairpin is always bad
   for n=-length(seq)+1:length(seq)-1
      if isok==1;
         a=diag(hpdot,n);
         dia=a(1:ceil(length(a)/2));
```

```
        hpsmall=strfind(dia',ones(1,9));
        hpbig=strfind(dia',ones(1,12));
        if size(hpbig)~=0
           isok=0;
        elseif size(hpsmall)~=0
           for m=1:size(hpsmall)
              if n<=0
                 GC=gccontent(seq(hpsmall(m):hpsmall(m)+6));
              else
                 GC=gccontent(seq(len-hpsmall(m)-5:len-hpsmall(m)+1));
              end
              if GC>80
                 isok=0;
              end
           end
        end
     end
  end
end

%check for repeats or inverted repeats in each 100bp subsequence
done=0;
if isok==1
  %create the dot matrices
  for m=1:len
     repdot(m,:)=seq==seq(m);
  end
  %straighten out the diagonals for only those nucleotides within a 100bp
  %window of one another
  for m=1:95
     repdiags(:,m)=padarray(diag(repdot,m),length(diag(repdot,1))-length(diag(repdot,m)),'post');
  end
  %iterate through all relevant repeat lengths
  for l=4:50
     reps=zeros(100-l,len);
     if isok==1&&done==0
        %look for the relevant repeat length within a 100bp window and
        %save all instances of the repeat to reps as its location (if
        %the repeat would extend beyond a 100bp window then it isn't
        %counted, hence the 100-l)
        for m=1:100-2*l
           occur=strfind(repdiags(:,m+l-1)',ones(1,l));
           if size(occur,1)>0
              reps(m,:)=padarray(occur,[0 len-length(occur)],'post');
           end
        end
        %and filter out any repeats which occur within l bp of one
        %another
        if sum(sum(reps))>0
           filteredreps=reps;
           for n=1:100-2*l+1
              a=~ismember(filteredreps(n+1:n+l-1,:),filteredreps(n,:));
```

```
            filteredreps(n+1:n+l-1,:)=a.*filteredreps(n+1:n+l-1,:);
          end
          res = [(1:len)' histc(filteredreps(:), 1:len)];
          sortedres = sortrows(res, -2);
          %check if the repeats cover more than 30bp (ie 30% of the 100bp
          %window)
          if (sortedres(1,2)+1)*l>=30&&sortedres(1,2)>0
            isok=0;
          end
        else
          done=1;
          %stops calculation early if it can be.
        end
      end
    end
end
```

seqarea.m (MATLAB)

```
function [area,affinitiesscaled]=seqarea(learningcurve,affinities,prombeg,promend)
%scales an input affinity curve so that it is comparable to another
%affinity curve, specified by 'learningcurve'

%Npred does this weird thing where it scales its affinity output depending
%on the max and min affinities.  This becomes a problem when promoter
%modifications change the extrema of the affinity curve.  This program
%effectively "unscales" this affinity curve and "rescales" it to the same
%scale as a reference curve.  To do so, it does a simple linear regression
%between the first 25 affinity values of each curve.  In order for this to
%be valid, the first 25bp of each generating seqence must be the same.  We
%recommend putting 200bp of context before and 100bp of context after each
%promoter, as detailed in "Notes for MATLAB".  prombeg and promend specify
%the locations of the first and last nucleotide of the promoter of interest
%in 'affinities', for calculation of areas.


[~,m,b]=regression(affinities(1:25)',learningcurve');
%Performs regression between learningcurve and affinities

affinitiesscaled=affinities.*m+b;
%Uses regression to scale affinities

area=sum(affinitiesscaled(prombeg:promend));
%computes nucleosome area under promoter region.
```

Appendix B2: Software Written for Chapter 7

transmutratefit.m (MATLAB script)

```
function [ft,rm]=transmutratefit(f1m,f2m,f3m,L1,L2,fi1,fi2)
  function a=fun(x)
```

```
      a=(f2m-(x(1)*(1+(1-fi1)*x(2)-x(2))^L1*(1+(1-fi2)*x(2)-x(2))^L2))^2+(f1m-(1-x(1)+x(1)*(1-(1+(1-
fi1)*x(2)-x(2))^L1)*(1-(1+(1-fi2)*x(2)-x(2))^L2)))^2;
   end
[x,~,~,~]=fsolve(@fun,[(1-f1m),(f3m/((1-f1m)*L1*fi1))]);
ft=x(1);
rm=x(2);
end
```

calceverything.sh (shell script)

```
#!/bin/bash

# first argument is forward read, second is reverse, third is a space-delimited list of barcodes and output
files
# fourth is a space-delimited list of output files and their corresponding templates, fifth is output file. see
examples.

START=$(date +%s)

echo "Unzipping first file..."
gunzip -c $1 > fwdread.fastq
echo "Unzipping second file..."
gunzip -c $2 > revread.fastq

pandaseq -f fwdread.fastq -r revread.fastq -d bfsmrk -L 256 -l 244 -t 0.9 -N > align.fasta

rm -f fwdread.fastq
rm -f revread.fastq

echo "demultiplexing reads..."

perl Fakefastq.pl align.fasta align.fastq

rm -f align.fasta

sabre se -m 0 -f align.fastq -b $3 -u failed.fastq

rm -f failed.fastq
rm -f align.fastq

fsuffix="F.fastq"
rsuffix="R.fastq"
F="F"

while read line
do
        filename=$(echo "$line" | awk '{print $2}')
        noext="${filename%.*}"
        last=${noext: -1:1}
        base="${noext%?}"
        mkdir tempalign
        mv $filename tempalign/input.fastq
```

```
		rm -f $filename
		cp $3 tempalign/barcodeinfo.txt
		cd tempalign
		fastx_reverse_complement -i input.fastq -o inputrc.fastq
		rm -f input.fastq
		sabre se -m 0 -f inputrc.fastq -b barcodeinfo.txt -u turbofailed.fastq
		if [ "$last" == "$F" ]
		then
			cp $base$rsuffix ../$base$fsuffix
		else
			cp $base$fsuffix ../$base$rsuffix
		fi
		cd ..
		rm -r tempalign
done < $3

echo "Preparing Templates..."

ls *.fasta > templateout

while read line
do
		filename=$(basename "$line")
		extension="${filename##*.}"
		filename="${filename%.*}"
		ssaha2Build -solexa -save "$filename" "$line"
done < templateout

mkdir qiime_inputs
mkdir $5
align="align"
qiimeext=".fna"

while read line
do
		filename=$(echo "$line" | awk '{print $1}')
		templatestring=$(echo "$line" | grep -o "\S\+.fasta")
		templatename=$(basename "$templatestring")
		extension="${templatename##*.}"
		templatename="${templatename%.*}"
		filenamef=$filename$fsuffix
		filenamer=$filename$rsuffix
		#cat $filenamef$trim > $filename
		cat $filenamef $filenamer > $filename
		rm -f $filenamef
		rm -f $filenamer
		echo "preparing template for qiime"
		bash fastqtofna.sh "$filename" "qiime_inputs/$filename$qiimeext"
		echo "aligning a read to its template"
		ssaha2 -solexa -disk 1 -align 1 -save "$templatename" "$filename" > $filename$align
		rm -f "$filename"
		echo "counting mutations"
```

```
        bash spectrumcalc.sh $filename$align $5
        rm -f $filename$align
done < $4

echo "Analyzing Phylogenetic Groups"
cat qiime_inputs/* >> $5/qiimeseqs.fna
pick_de_novo_otus.py -i $5/qiimeseqs.fna -p parameters.txt -o $5/taxonomies099

echo "cleaning stuff up..."

rm -r qiime_inputs
rm -f otu_table.biom

base=".base"
body=".body"
head=".head"
name=".name"
size=".size"

while read line
do
        filename=$(basename "$line")
        extension="${filename##*.}"
        filename="${filename%.*}"
        rm -f $filename$base
        rm -f $filename$body
        rm -f $filename$head
        rm -f $filename$name
        rm -f $filename$size
done < templateout

rm -f templateout

echo "Making Phylogenetic Groups look Pretty"
cd $5/taxonomies099
plot_rank_abundance_graph.py -i otu_table.biom -s '*' -o otu_graph.pdf
convert_biom.py -i otu_table.biom -o otu_table_with_taxonomy.txt -b --header_key taxonomy --
process_obs_metadata taxonomy

END=$(date +%s)
ELAPSED=$((END - START))
echo "Elapsed Time is"
echo "$ELAPSED"
echo "Seconds"
echo "Job Done!  Hooray!"
```

trimquals.sh (shell script)

```
#!/bin/bash

#first argument is input file, second argument is output file
```

```
numlines=`wc -l $1 | cut -f1 -d ' '`
carat=">"
n=1

while read line
do
	m=$(( ($n-2) % 4 ))
	if [ $m = 0 ];
	then
		seqnum=$(( (n-2)/4 ))
		echo $prevline >> $2
		echo $line >> $2
	fi
	n=$(($n + 1))
	prevline=$line
done < $1
```

fastqtofna.sh (shell script)

```
#!/bin/bash

#first argument is input file, second argument is output file

numlines=`wc -l $1 | cut -f1 -d ' '`
carat=">"
underscore="_"
newline="\n"
space=" "
other="orig_bc=ATACGA new_bc=ATACGA bc_diffs=0"
n=1

while read line
do
	m=$(( ($n-2) % 4 ))
	if [ $m = 0 ];
	then
		seqnum=$(( (n-2)/4 ))
		echo $carat$1$underscore$seqnum$space$prevline$space$other >> $2
		echo $line >> $2
	fi
	n=$(($n + 1))
	prevline=$line
done < $1
```

spectrumalc.sh (shell script)

```
#!/bin/bash
#first argument is input, second is output folder
grep -o "Sbjct[[:space:]]\{2,\}[[:digit:]]\+[[:space:]][GCAT-]\+" $1 > "$1sbjct"
grep -o "Query[[:space:]]\{2,\}[[:digit:]]\+[[:space:]][GCAT-]\+" $1 > "$1query"

python seqcat.py "$1sbjct" "$1sbjctcat"
```

```
python seqcat.py "$1query" "$1querycat"
python mutspectrum.py "$1sbjctcat" "$1querycat" $2/$1spectrum
rm -f "$1sbjct"
rm -f "$1query"
rm -f "$1sbjctcat"
rm -f "$1querycat"
```

seqcat.py (Python)

```
# 1st arg is input, 2nd is output file

import sys

file = open(sys.argv[1],"rb")

import re
n=1
seq=[]
f=open(sys.argv[2],'w')
for line in file:
        seqline=re.search('[GATC\-]+',line)
        seq.append(seqline.group(0))
seqstr=''.join(seq)
f.write(seqstr + '\n')
f.close()
file.close()
```

mutspectrum.py

```
#!/bin/python

import sys
subjectfile = open(sys.argv[1],"rb")
queryfile = open(sys.argv[2],"rb")

from collections import defaultdict

spectrum={'match':0, 'cta':0, 'tta':0, 'gta':0, 'atc':0, 'ttc':0, 'gtc':0, 'att':0, 'ctt':0, 'gtt':0, 'atg':0, 'ctg':0, 'ttg':0,
'ains':0, 'cins':0, 'tins':0, 'gins':0, 'adel':0, 'cdel':0, 'tdel':0, 'gdel':0}

for qline,sline in zip(queryfile, subjectfile):
        for qnuc,snuc in zip(qline,sline):
                if qnuc==snuc:
                        spectrum['match']=spectrum['match']+1
                elif qnuc=='A' and snuc=='C':
                        spectrum['cta']=spectrum['cta']+1
                elif qnuc=='A' and snuc=='T':
                        spectrum['tta']=spectrum['tta']+1
                elif qnuc=='A' and snuc=='G':
                        spectrum['gta']=spectrum['gta']+1
                elif qnuc=='A' and snuc=='-':
                        spectrum['ains']=spectrum['ains']+1
```

```
                elif qnuc=='C' and snuc=='A':
                        spectrum['atc']=spectrum['atc']+1
                elif qnuc=='C' and snuc=='T':
                        spectrum['ttc']=spectrum['ttc']+1
                elif qnuc=='C' and snuc=='G':
                        spectrum['gtc']=spectrum['gtc']+1
                elif qnuc=='C' and snuc=='-':
                        spectrum['cins']=spectrum['cins']+1
                elif qnuc=='T' and snuc=='A':
                        spectrum['att']=spectrum['att']+1
                elif qnuc=='T' and snuc=='C':
                        spectrum['ctt']=spectrum['ctt']+1
                elif qnuc=='T' and snuc=='G':
                        spectrum['gtt']=spectrum['gtt']+1
                elif qnuc=='T' and snuc=='-':
                        spectrum['tins']=spectrum['tins']+1
                elif qnuc=='G' and snuc=='A':
                        spectrum['atg']=spectrum['atg']+1
                elif qnuc=='G' and snuc=='C':
                        spectrum['ctg']=spectrum['ctg']+1
                elif qnuc=='G' and snuc=='T':
                        spectrum['ttg']=spectrum['ttg']+1
                elif qnuc=='G' and snuc=='-':
                        spectrum['tins']=spectrum['tins']+1
                elif qnuc=='-' and snuc=='A':
                        spectrum['adel']=spectrum['adel']+1
                elif qnuc=='-' and snuc=='C':
                        spectrum['cdel']=spectrum['cdel']+1
                elif qnuc=='-' and snuc=='T':
                        spectrum['tdel']=spectrum['tdel']+1
                elif qnuc=='-' and snuc=='G':
                        spectrum['gdel']=spectrum['gdel']+1

queryfile.close()
subjectfile.close()

outfile=open(sys.argv[3],"w")

for key, value in spectrum.iteritems():
        outfile.write('%s\t%s\n' % (key,str(value)))
```

Nt_Count.py (python)

```
# 1st arg is input, 2nd is output file
from collections import defaultdict
from collections import Counter
import sys

d = defaultdict()

file = open(sys.argv[1],"rb")
```

```
def initial():
	for line in file: #addressing each line
		n = line.rstrip()
		i = 0
		for x in n: # for each letter in the line
			i += 1
			if not d.has_key(i):
				d[i] = {}
			if not d[i].has_key(x):
				d[i][x] = 1
			elif d[i].has_key(x):
				d[i][x] += 1

initial()

file.close()

f=open(sys.argv[2],"w")

for key, value in d.iteritems():
	for subkey, subvalue in value.iteritems():
		f.write('%s\t%s\t%s\n' % (key,subkey,subvalue))

f.close()
```

templateinfo.txt (example)

```
Amp1    Amp.fasta
Amp2    Amp.fasta
intron1   URA.fasta
intron2   URA.fasta
intron3   URA.fasta
intron4   URA.fasta
intron5   URA.fasta
intron6   URA.fasta
intron7   URA.fasta
intron8   URA.fasta
intron9   URA.fasta
intron10 URA.fasta
intron11 URA.fasta
intron12 URA.fasta
intron13 URA.fasta
intron14 URA.fasta
intron15 URA.fasta
intron16 URA.fasta
intron17 URA.fasta
intron18 URA.fasta
intron19 URA.fasta
intron20 URA.fasta
```

barcodeinfo.txt (example)

ACTGATGCCAACTTACTTCTGACAACG Amp1F.fastq
GCTACCCCGCCTCCATCCAGTC Amp1R.fastq
CGTACGGCCAACTTACTTCTGACAACG Amp2F.fastq
TGACATCCGCCTCCATCCAGTC Amp2R.fastq
GATACAATCGCGGCCGCCC intron1F.fastq
GGAACTAGAATGGGCAGACATTACGAATG intron1R.fastq
AGTCAAATCGCGGCCGCCC intron2F.fastq
TAACCGAGAATGGGCAGACATTACGAATG intron2R.fastq
AGCTTTATCGCGGCCGCCC intron3F.fastq
TACAAGAGAATGGGCAGACATTACGAATG intron3R.fastq
GGCTACATCGCGGCCGCCC intron4F.fastq
AAGCTAAGAATGGGCAGACATTACGAATG intron4R.fastq
ATACGAATCGCGGCCGCCC intron5F.fastq
CTGATCAGAATGGGCAGACATTACGAATG intron5R.fastq
TTACTGATCGCGGCCGCCC intron6F.fastq
AGTTCCAGAATGGGCAGACATTACGAATG intron6R.fastq
ACTTGAATCGCGGCCGCCC intron7F.fastq
GATCTGAGAATGGGCAGACATTACGAATG intron7R.fastq
ACATCTATCGCGGCCGCCC intron8F.fastq
AATCGTAGAATGGGCAGACATTACGAATG intron8R.fastq
GCCAATATCGCGGCCGCCC intron9F.fastq
CACTGTAGAATGGGCAGACATTACGAATG intron9R.fastq
AGAATCATCGCGGCCGCCC intron10F.fastq
GCCTAAAGAATGGGCAGACATTACGAATG intron10R.fastq
CTGCAGATCGCGGCCGCCC intron11F.fastq
ACATCGAGAATGGGCAGACATTACGAATG intron11R.fastq
ATCACGATCGCGGCCGCCC intron12F.fastq
CACGTAAGAATGGGCAGACATTACGAATG intron12R.fastq
TCACATATCGCGGCCGCCC intron13F.fastq
TATAGAAGAATGGGCAGACATTACGAATG intron13R.fastq
TGCAAAATCGCGGCCGCCC intron14F.fastq
GTGCCAAGAATGGGCAGACATTACGAATG intron14R.fastq
TGTTAGATCGCGGCCGCCC intron15F.fastq
ATAGAAAGAATGGGCAGACATTACGAATG intron15R.fastq
TCGAAGATCGCGGCCGCCC intron16F.fastq
GAATGAAGAATGGGCAGACATTACGAATG intron16R.fastq
TACAGCATCGCGGCCGCCC intron17F.fastq
TCTGAGAGAATGGGCAGACATTACGAATG intron17R.fastq
CTATACATCGCGGCCGCCC intron18F.fastq
AGCTAGAGAATGGGCAGACATTACGAATG intron18R.fastq
CGGAATATCGCGGCCGCCC intron19F.fastq
GCCATGAGAATGGGCAGACATTACGAATG intron19R.fastq
CACGATATCGCGGCCGCCC intron20F.fastq
GCTCATAGAATGGGCAGACATTACGAATG intron20R.fastq

URA.fasta (example)

>URA
GATACAATCGCGGCCGCCCATGTCTCTTTGAGCAATAAAGCCGATAACAAAATCTTTGTCGCT
CTTCGCAATGTCAACAGTACCCTTAGTATATTCTCCAGTAGATAGGGAGCCCTTGCATGACAA

TTCTGCTAACATCAAAAGGCCTCTAGGTTCCTTTGTTACTTCTTCTGCCGCCTGCTTCAAACCG
CTAACAATACCTGGGCCCACCACACCGTGTGCATTCGTAATGTCTGCCCATTCTAGTTCC

Amp.fasta (example)

>Amp
GCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATG
GGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGA
CGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCG
AACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGG

References

1. Curran, K.A. and Alper, H.S. (2012) Expanding the chemical palate of cells by combining systems biology and metabolic engineering. *Metabolic Engineering*, **14**, 289-297.
2. Nevoigt, E. (2008) Progress in metabolic engineering of *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.*, **72**, 379-412.
3. Purnick, P.E.M. and Weiss, R. (2009) The second wave of synthetic biology: from modules to systems. *Nature Reviews Molecular Cell Biology*, **10**, 410-422.
4. Ro, D.K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A., Ham, T.S., Kirby, J. *et al.* (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, **440**, 940-943.
5. Verwaal, R., Wang, J., Meijnen, J.P., Visser, H., Sandmann, G., van den Berg, J.A. and van Ooyen, A.J. (2007) High-level production of beta-carotene in Saccharomyces cerevisiae by successive transformation with carotenogenic genes from Xanthophyllomyces dendrorhous. *Applied and environmental microbiology*, **73**, 4342-4350.
6. Curran, K.A., Leavitt, J.M., Karim, A.S. and Alper, H.S. (2013) Metabolic engineering of muconic acid production in Saccharomyces cerevisiae. *Metabolic Engineering*, **15**, 55-66.
7. Lee, S.M., Jellison, T. and Alper, H.S. (2012) Directed evolution of xylose isomerase for improved xylose catabolism and fermentation in the yeast Saccharomyces cerevisiae. *Applied and environmental microbiology*, **78**, 5708-5716.
8. Westfall, P.J., Pitera, D.J., Lenihan, J.R., Eng, D., Woolard, F.X., Regentin, R., Horning, T., Tsuruta, H., Melis, D.J., Owens, A. *et al.* (2012) Production of amorphadiene in yeast, and its conversion to dihydroartemisinic acid, precursor to the antimalarial agent artemisinin. *Proceedings of the National Academy of Sciences*.
9. Elowitz, M.B. and Leibler, S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**, 335-338.
10. Gardner, T.S., Cantor, C.R. and Collins, J.J. (2000) Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, **403**, 339-342.
11. Friedland, A.E., Lu, T.K., Wang, X., Shi, D., Church, G. and Collins, J.J. (2009) Synthetic Gene Networks That Count. *Science*, **324**, 1199-1202.
12. Tan, C., Marguet, P. and You, L. (2009) Emergent bistability by a growth-modulating positive feedback circuit. *Nature Chemical Biology*, **5**, 842-848.
13. Blazeck, J. and Alper, H.S. (2012) Promoter engineering: Recent advances in controlling transcription at the most fundamental level. *Biotechnology Journal*, Accepted/In Press.
14. Gama-Castro, S., Jimenez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M.I., Contreras-Moreira, B., Segura-Salazar, J., Muniz-

Rascado, L., Martinez-Flores, I., Salgado, H. *et al.* (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Research*, **36**, D120-124.

15. Gupta, R., Bhattacharyya, A., Agosto-Perez, F.J., Wickramasinghe, P. and Davuluri, R.V. (2010) MPromDb update 2010: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-seq experimental data. *Nucleic Acids Research*.
16. Yamamoto, Y.Y. and Obokata, J. (2008) ppdb: a plant promoter database. *Nucleic Acids Research*, **36**, D977-981.
17. Zhu, J. and Zhang, M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607-611.
18. Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W. and Sandelin, A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research*, **38**, D105-110.
19. Blazeck, J., Liu, L., Redden, H. and Alper, H. (2011) Tuning Gene Expression in *Yarrowia lipolytica* by a Hybrid Promoter Approach. *Applied and Environmental Microbiology*, **77**, 7905-7914.
20. Blazeck, J., Garg, R., Reed, B. and Alper, H.S. (2012) Controlling promoter strength and regulation in *Saccharomyces cerevisiae* using synthetic hybrid promoters. *Biotechnology and Bioengineering*.
21. Blazeck, J., Reed, B., Garg, R., Gerstner, R., Pan, A., Agarwala, V. and Alper, H. (2012) Generalizing a hybrid synthetic promoter approach in *Yarrowia lipolytica*. *Applied Microbiology and Biotechnology*, 1-16.
22. Amit, R., Garcia, Hernan G., Phillips, R. and Fraser, Scott E. (2011) Building Enhancers from the Ground Up: A Synthetic Biology Approach. *Cell*, **146**, 105-118.
23. Lam, F.H., Steger, D.J. and O'Shea, E.K. (2008) Chromatin decouples promoter threshold from dynamic range. *Nature*, **453**, 246-250.
24. Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A. and Segal, E. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, **30**, 521-530.
25. Sun, J., Shao, Z., Zhao, H., Nair, N., Wen, F., Xu, J.-H. and Zhao, H. (2012) Cloning and characterization of a panel of constitutive promoters for applications in pathway engineering in Saccharomyces cerevisiae. *Biotechnology and Bioengineering*, **109**, 2082-2092.
26. Alper, H., Fischer, C., Nevoigt, E. and Stephanopoulos, G. (2005) Tuning genetic control through promoter engineering. *Proc Natl Acad Sci U S A*, **102**, 12678-12683.
27. Nevoigt, E., Kohnke, J., Fischer, C.R., Alper, H., Stahl, U. and Stephanopoulos, G. (2006) Engineering of promoter replacement cassettes for fine-tuning of gene

expression in Saccharomyces cerevisiae. *Applied and environmental microbiology*, **72**, 5266-5273.
28. Peccoud, J., Blauvelt, M.F., Cai, Y., Cooper, K.L., Crasta, O., DeLalla, E.C., Evans, C., Folkerts, O., Lyons, B.M., Mane, S.P. *et al.* (2008) Targeted Development of Registries of Biological Parts. *PLoS ONE*, **3**, e2671.
29. Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, **288**, 911-940.
30. Gruber, A.R., Lorenz, R., Bernhart, S.H., Neubock, R. and Hofacker, I.L. (2008) The Vienna RNA websuite. *Nucleic Acids Research*, **36**, W70-74.
31. Zadeh, J.N., Steenberg, C.D., Bois, J.S., Wolfe, B.R., Pierce, M.B., Khan, A.R., Dirks, R.M. and Pierce, N.A. (2011) NUPACK: Analysis and design of nucleic acid systems. *Journal of Computational Chemistry*, **32**, 170-173.
32. Win, M.N., Liang, J.C. and Smolke, C.D. (2009) Frameworks for Programming Biological Function through RNA Parts and Devices. *Chemistry & Biology*, **16**, 298-310.
33. Isaacs, F.J., Dwyer, D.J., Ding, C., Pervouchine, D.D., Cantor, C.R. and Collins, J.J. (2004) Engineered riboregulators enable post-transcriptional control of gene expression. *Nature Biotechnology*, **22**, 841-847.
34. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406-3415.
35. Salis, H.M., Mirsky, E.A. and Voigt, C.A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology*, **27**, 946-950.
36. Na, D. and Lee, D. (2010) RBSDesigner: software for designing synthetic ribosome binding sites that yields a desired level of protein expression. *Bioinformatics*, **26**, 2633-2634.
37. Crook, N.C., Freeman, E.S. and Alper, H.S. (2011) Re-engineering multicloning sites for function and convenience. *Nucleic Acids Research*, **39**, e92.
38. Goldfless, S.J., Belmont, B.J., de Paz, A.M., Liu, J.F. and Niles, J.C. (2012) Direct and specific chemical control of eukaryotic translation with a synthetic RNA–protein interaction. *Nucleic Acids Research*.
39. Win, M.N. and Smolke, C.D. (2007) A modular and extensible RNA-based gene-regulatory platform for engineering cellular function. *Proceedings of the National Academy of Sciences*, **104**, 14283-14288.
40. Win, M.N. and Smolke, C.D. (2008) Higher-Order Cellular Information Processing with Synthetic RNA Devices. *Science*, **322**, 456-460.
41. Carothers, J.M., Goler, J.A., Juminaga, D. and Keasling, J.D. (2011) Model-Driven Engineering of RNA Devices to Quantitatively Program Gene Expression. *Science*, **334**, 1716-1719.
42. Kim, J.H., Lee, S.R., Li, L.H., Park, H.J., Park, J.H., Lee, K.Y., Kim, M.K., Shin, B.A. and Choi, S.Y. (2011) High cleavage efficiency of a 2A peptide derived

from porcine teschovirus-1 in human cell lines, zebrafish and mice. *PLoS One*, **6**, e18556.
43. Donnelly, M.L., Luke, G., Mehrotra, A., Li, X., Hughes, L.E., Gani, D. and Ryan, M.D. (2001) Analysis of the aphthovirus 2A/2B polyprotein 'cleavage' mechanism indicates not a proteolytic reaction, but a novel translational effect: a putative ribosomal 'skip'. *The Journal of general virology*, **82**, 1013-1025.
44. de Felipe, P., Hughes, L.E., Ryan, M.D. and Brown, J.D. (2003) Co-translational, intraribosomal cleavage of polypeptides by the foot-and-mouth disease virus 2A peptide. *J Biol Chem*, **278**, 11441-11448.
45. Niepmann, M. (2009) Internal translation initiation of picornaviruses and hepatitis C virus. *Biochim Biophys Acta*, **1789**, 529-541.
46. Paz, I., Abramovitz, L. and Choder, M. (1999) Starved Saccharomyces cerevisiae cells have the capacity to support internal initiation of translation. *J Biol Chem*, **274**, 21741-21745.
47. Witherell, G.W., Schultz-Witherell, C.S. and Wimmer, E. (1995) Cis-acting elements of the encephalomyocarditis virus internal ribosomal entry site. *Virology*, **214**, 660-663.
48. Hoffman, M.A. and Palmenberg, A.C. (1995) Mutational analysis of the J-K stem-loop region of the encephalomyocarditis virus IRES. *Journal of virology*, **69**, 4399-4406.
49. Van Der Velden, A., Kaminski, A., Jackson, R.J. and Belsham, G.J. (1995) Defective point mutants of the encephalomyocarditis virus internal ribosome entry site can be complemented in trans. *Virology*, **214**, 82-90.
50. Nakashima, N. and Uchiumi, T. (2009) Functional analysis of structural motifs in dicistroviruses. *Virus research*, **139**, 137-147.
51. Benton, B.M., Eng, W.K., Dunn, J.J., Studier, F.W., Sternglanz, R. and Fisher, P.A. (1990) Signal-mediated import of bacteriophage T7 RNA polymerase into the Saccharomyces cerevisiae nucleus and specific transcription of target genes. *Molecular and cellular biology*, **10**, 353-360.
52. Das, S., Ott, M., Yamane, A., Tsai, W., Gromeier, M., Lahser, F., Gupta, S. and Dasgupta, A. (1998) A Small Yeast RNA Blocks Hepatitis C Virus Internal Ribosome Entry Site (HCV IRES)-Mediated Translation and Inhibits Replication of a Chimeric Poliovirus under Translational Control of the HCV IRES Element. *J. Virol.*, **72**, 5638-5647.
53. Hertz, M.I. and Thompson, S.R. (2011) In vivo functional analysis of the Dicistroviridae intergenic region internal ribosome entry sites. *Nucleic Acids Research*.
54. Lippow, S.M. and Tidor, B. (2007) Progress in computational protein design. *Current Opinion in Biotechnology*, **18**, 305-311.
55. Kortemme, T. and Baker, D. (2004) Computational design of protein-protein interactions. *Current Opinion in Chemical Biology*, **8**, 91-97.

56. Samish, I., MacDermaid, C.M., Perez-Aguilar, J.M. and Saven, J.G. (2011) Theoretical and Computational Protein Design. *Annual Review of Physical Chemistry*, **62**, 129-149.
57. Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W. *et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, **487**, 545-574.
58. Siegel, J.B., Zanghellini, A., Lovick, H.M., Kiss, G., Lambert, A.R., St Clair, J.L., Gallaher, J.L., Hilvert, D., Gelb, M.H., Stoddard, B.L. *et al.* (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science*, **329**, 309-313.
59. Nannemann, D.P., Kaufmann, K.W., Meiler, J. and Bachmann, B.O. (2010) Design and directed evolution of a dideoxy purine nucleoside phosphorylase. *Protein Eng Des Sel*, **23**, 607-616.
60. Fleishman, S.J., Whitehead, T.A., Ekiert, D.C., Dreyfus, C., Corn, J.E., Strauch, E.M., Wilson, I.A. and Baker, D. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, **332**, 816-821.
61. Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popovic, Z. and players, F. (2010) Predicting protein structures with a multiplayer online game. *Nature*, **466**, 756-760.
62. Eiben, C.B., Siegel, J.B., Bale, J.B., Cooper, S., Khatib, F., Shen, B.W., Players, F., Stoddard, B.L., Popovic, Z. and Baker, D. (2012) Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nat Biotech*, **30**, 190-192.
63. Looger, L.L., Dwyer, M.A., Smith, J.J. and Hellinga, H.W. (2003) Computational design of receptor and sensor proteins with novel functions. *Nature*, **423**, 185-190.
64. Kaplan, J. and DeGrado, W.F. (2004) De novo design of catalytic proteins. *Proceedings of the National Academy of Sciences*, **101**, 11566-11570.
65. Jiang, L., Althoff, E.A., Clemente, F.R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J.L., Betker, J.L., Tanaka, F., Barbas, C.F. *et al.* (2008) De Novo Computational Design of Retro-Aldol Enzymes. *Science*, **319**, 1387-1391.
66. Rothlisberger, D., Khersonsky, O., Wollacott, A.M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J.L., Althoff, E.A., Zanghellini, A., Dym, O. *et al.* (2008) Kemp elimination catalysts by computational enzyme design. *Nature*, **453**, 190-195.
67. Zastrow, M.L., PeacockAnna, F.A., Stuckey, J.A. and Pecoraro, V.L. (2012) Hydrolytic catalysis and structural stabilization in a designed metalloprotein. *Nature Chemistry*, **4**, 118-123.
68. Rohl, C.A., Strauss, C.E., Misura, K.M. and Baker, D. (2004) Protein structure prediction using Rosetta. *Methods Enzymol*, **383**, 66-93.

69. Song, L.S. and Poulter, C.D. (1994) YEAST FARNESYL-DIPHOSPHATE SYNTHASE - SITE-DIRECTED MUTAGENESIS OF RESIDUES IN HIGHLY CONSERVED PRENYLTRANSFERASE DOMAIN-I AND DOMAIN-II. *Proc. Natl. Acad. Sci. U. S. A.*, **91**, 3044-3048.
70. Fasan, R., Chen, M.M., Crook, N.C. and Arnold, F.H. (2007) Engineered alkane-hydroxylating cytochrome P450(BM3) exhibiting nativelike catalytic properties. *Angew. Chem.-Int. Edit.*, **46**, 8414-8418.
71. Liang, J.C., Bloom, R.J. and Smolke, C.D. (2011) Engineering Biological Systems with Synthetic RNA Molecules. *Mol. Cell*, **43**, 915-926.
72. Ellington, A.D. and Szostak, J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818-822.
73. Crameri, A., Raillard, S.A., Bermudez, E. and Stemmer, W.P.C. (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature*, **391**, 288-291.
74. Crameri, A., Whitehorn, E.A., Tate, E. and Stemmer, W.P.C. (1996) Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nat. Biotechnol.*, **14**, 315-319.
75. Stemmer, W.P.C. (1994) RAPID EVOLUTION OF A PROTEIN IN-VITRO BY DNA SHUFFLING. *Nature*, **370**, 389-391.
76. Arnold, F.H. and Volkov, A.A. (1999) Directed evolution of biocatalysts. *Current Opinion in Chemical Biology*, **3**, 54-59.
77. Fasan, R., Meharenna, Y.T., Snow, C.D., Poulos, T.L. and Arnold, F.H. (2008) Evolutionary History of a Specialized P450 Propane Monooxygenase. *J. Mol. Biol.*, **383**, 1069-1080.
78. Alper, H., Moxley, J., Nevoigt, E., Fink, G.R. and Stephanopoulos, G. (2006) Engineering yeast transcription machinery for improved ethanol tolerance and production. *Science*, **314**, 1565-1568.
79. Alper, H., Fischer, C., Nevoigt, E. and Stephanopoulos, G. (2005) Tuning genetic control through promoter engineering. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 12678-12683.
80. Esvelt, K.M., Carlson, J.C. and Liu, D.R. (2011) A system for the continuous directed evolution of biomolecules. *Nature*, **472**, 499-U550.
81. Savinell, J.M. and Palsson, B.O. (1992) Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *Journal of Theoretical Biology*, **154**, 421-454.
82. Segrè, D., Vitkup, D. and Church, G.M. (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, **99**, 15112-15117.
83. Curran, K.A., Crook, N.C. and Alper, H.S. (2012) Using flux balance analysis to guide microbial metabolic engineering. *Methods in Molecular Biology*, **834**, 197-216.
84. Henry, C.S., Broadbelt, L.J. and Hatzimanikatis, V. (2007) Thermodynamics-based metabolic flux analysis. *Biophys J*, **92**, 1792-1805.

85. Beard, D.A., Babson, E., Curtis, E. and Qian, H. (2004) Thermodynamic constraints for biochemical networks. *Journal of Theoretical Biology*, **228**, 327-333.
86. Mavrovouniotis, M.L. (1990) Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnology and Bioengineering*, **36**, 1070-1082.
87. Fischer, E., Zamboni, N. and Sauer, U. (2004) High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived C-13 constraints. *Anal. Biochem.*, **325**, 308-316.
88. Antoniewicz, M.R., Kelleher, J.K. and Stephanopoulos, G. (2007) Elementary metabolite units (EMU): A novel framework for modeling isotopic distributions. *Metabolic Engineering*, **9**, 68-86.
89. Antoniewicz, M.R., Kelleher, J.K. and Stephanopoulos, G. (2006) Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. *Metabolic Engineering*, **8**, 324-337.
90. Jamshidi, N. and Palsson, B.Ø. (2010) Mass Action Stoichiometric Simulation Models: Incorporating Kinetics and Regulation into Stoichiometric Models. *Biophysical journal*, **98**, 175-185.
91. Canelas, A.B., Ras, C., ten Pierick, A., van Gulik, W.M. and Heijnen, J.J. (2011) An *in vivo* data-driven framework for classification and quantification of enzyme kinetics and determination of apparent thermodynamic data. *Metabolic Engineering*, **13**, 294-306.
92. Nolan, R.P. and Lee, K. (2011) Dynamic model of CHO cell metabolism. *Metabolic Engineering*, **13**, 108-124.
93. Smallbone, K., Simeonidis, E., Swainston, N. and Mendes, P. (2010) Towards a genome-scale kinetic model of cellular metabolism. *BMC Systems Biology*, **4**, 6.
94. Dugar, D. and Stephanopoulos, G. (2011) Relative potential of biosynthetic pathways for biofuels and bio-based products. *Nature Biotechnology*, **29**, 1074-1078.
95. Karr, Jonathan R., Sanghvi, Jayodita C., Macklin, Derek N., Gutschow, Miriam V., Jacobs, Jared M., Bolival, B., Assad-Garcia, N., Glass, John I. and Covert, Markus W. (2012) A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell*, **150**, 389-401.
96. Pramanik, J. and Keasling, J.D. (1997) Stoichiometric model of Escherichia coli metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnology and Bioengineering*, **56**, 398-421.
97. Förster, J., Famili, I., Fu, P., Palsson, B.Ø. and Nielsen, J. (2003) Genome-Scale Reconstruction of the *Saccharomyces cerevisiae* Metabolic Network. *Genome Research*, **13**, 244-253.
98. Schilling, C.H., Covert, M.W., Famili, I., Church, G.M., Edwards, J.S. and Palsson, B.O. (2002) Genome-Scale Metabolic Model of *Helicobacter pylori* 26695. *Journal of Bacteriology*, **184**, 4582-4593.

99. Edwards, J.S. and Palsson, B.O. (1999) Systems Properties of the *Haemophilus influenzae* Rd Metabolic Genotype. *Journal of Biological Chemistry*, **274**, 17410-17416.
100. Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V. and Palsson, B.O. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, **3**.
101. Dobson, P.D., Smallbone, K., Jameson, D., Simeonidis, E., Lanthaler, K., Pir, P., Lu, C., Swainston, N., Dunn, W.B., Fisher, P. *et al.* (2010) Further developments towards a genome-scale metabolic model of yeast. *BMC Systems Biology*, **4**, 145.
102. Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B. and Stevens, R.L. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, **28**, 977-U922.
103. Karp, P.D., Paley, S. and Romero, P. (2002) The Pathway Tools software. *Bioinformatics*, **18 Suppl 1**, S225-232.
104. Kumar, V.S. and Maranas, C.D. (2009) GrowMatch: An Automated Method for Reconciling *in Silico*/*in Vivo* Growth Predictions. *PLoS Computational Biology*, **5**, e1000308.
105. Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Forum:, a.t.r.o.t.S., Arkin, A.P., Bornstein, B.J., Bray, D. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524-531.
106. Alper, H., Jin, Y.-S., Moxley, J.F. and Stephanopoulos, G. (2005) Identifying gene targets for the metabolic engineering of lycopene biosynthesis in *Escherichia coli*. *Metabolic Engineering*, **7**, 155-164.
107. Kennedy, C.J., Boyle, P.M., Waks, Z. and Silver, P.A. (2009) Systems-level Engineering of Non-fermentative Metabolism in Yeast. *Genetics*.
108. Burgard, A.P., Pharkya, P. and Maranas, C.D. (2003) Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, **84**, 647-657.
109. Fong, S.S., Burgard, A.P., Herring, C.D., Knight, E.M., Blattner, F.R., Maranas, C.D. and Palsson, B.O. (2005) *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnology and Bioengineering*, **91**, 643-648.
110. Ng, C.Y., Jung, M.Y., Lee, J. and Oh, M.K. (2012) Production of 2,3-butanediol in *Saccharomyces cerevisiae* by *in silico* aided metabolic engineering. *Microbial Cell Factories*, **11**, 68.
111. Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S.Y. *et al.* MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research*, **34**, D511-D516.
112. Caspi, R., Altman, T., Dale, J.M., Dreher, K., Fulcher, C.A., Gilham, F., Kaipa, P., Karthikeyan, A.S., Kothari, A., Krummenacker, M. *et al.* (2010) The MetaCyc

database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, **38**, D473-479.
113. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, **40**, D109-114.
114. Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Sohngen, C., Stelzer, M., Thiele, J. and Schomburg, D. (2011) BRENDA, the enzyme information system in 2011. *Nucleic Acids Research*, **39**, D670-676.
115. Schellenberger, J., Park, J.O., Conrad, T.M. and Palsson, B.O. (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, **11**, 213.
116. Henry, C.S., Broadbelt, L.J. and Hatzimanikatis, V. (2010) Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. *Biotechnology and Bioengineering*, **106**, 462-473.
117. Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J.D., Osterhout, R.E., Stephen, R. *et al.* (2011) Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nature Chemical Biology*, **7**, 445-452.
118. Cho, A., Yun, H., Park, J.H., Lee, S.Y. and Park, S. (2010) Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Systems Biology*, **4**, 35.
119. Yousofshahi, M., Lee, K. and Hassoun, S. (2011) Probabilistic pathway construction. *Metabolic Engineering*, **13**, 435-444.
120. Carbonell, P., Planson, A.G., Fichera, D. and Faulon, J.L. (2011) A retrosynthetic biology approach to metabolic pathway design for therapeutic production. *BMC Systems Biology*, **5**, 122.
121. Schellenberger, J., Que, R., Fleming, R.M.T., Thiele, I., Orth, J.D., Feist, A.M., Zielinski, D.C., Bordbar, A., Lewis, N.E., Rahmanian, S. *et al.* (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature Protocols*, **6**, 1290-1307.
122. Lewis, N.E., Nagarajan, H. and Palsson, B.O. (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. *Nature Reviews Microbiology*, **10**, 291-305.
123. Patil, K.R., Rocha, I., Forster, J. and Nielsen, J. (2005) Evolutionary programming as a platform for *in silico* metabolic engineering. *BMC Bioinformatics*, **6**, 308.
124. Brochado, A.R., Matos, C., Moller, B.L., Hansen, J., Mortensen, U.H. and Patil, K.R. (2010) Improved vanillin production in baker's yeast through *in silico* design. *Microbial Cell Factories*, **9**, 84.
125. Asadollahi, M.A., Maury, J., Patil, K.R., Schalk, M., Clark, A. and Nielsen, J. (2009) Enhancing sesquiterpene production in *Saccharomyces cerevisiae* through *in silico* driven metabolic engineering. *Metabolic Engineering*, **11**, 328-334.

126. Cvijovic, M., Olivares-Hernandez, R., Agren, R., Dahr, N., Vongsangnak, W., Nookaew, I., Patil, K.R. and Nielsen, J. (2010) BioMet Toolbox: genome-wide analysis of metabolism. *Nucleic Acids Research*, **38**, W144-149.
127. Le Fèvre, F., Smidtas, S., Combe, C., Durot, M., d'Alché-Buc, F. and Schachter, V. (2009) CycSim - an online tool for exploring and experimenting with genome-scale metabolic models. *Bioinformatics*.
128. Yamada, T., Letunic, I., Okuda, S., Kanehisa, M. and Bork, P. (2011) iPath2.0: interactive pathway explorer. *Nucleic Acids Research*, **39**, W412-415.
129. Bates, J.T., Chivian, D. and Arkin, A.P. (2011) GLAMM: Genome-Linked Application for Metabolic Maps. *Nucleic Acids Research*, **39**, W400-405.
130. Alper, H., Miyaoku, K. and Stephanopoulos, G. (2005) Construction of lycopene-overproducing E. coli strains by combining systematic and combinatorial gene knockout targets. *Nature biotechnology*, **23**, 612-616.
131. Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J.D., Osterhout, R.E., Stephen, R. *et al.* (2011) Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nat Chem Biol*, **7**, 445-452.
132. Atsumi, S., Cann, A.F., Connor, M.R., Shen, C.R., Smith, K.M., Brynildsen, M.P., Chou, K.J., Hanai, T. and Liao, J.C. (2008) Metabolic engineering of Escherichia coli for 1-butanol production. *Metab Eng*, **10**, 305-311.
133. Liu, L., Redden, H. and Alper, H.S. Frontiers of yeast metabolic engineering: diversifying beyond ethanol and Saccharomyces. *Current Opinion in Biotechnology*.
134. Na, D., Yoo, S.M., Chung, H., Park, H., Park, J.H. and Lee, S.Y. (2013) Metabolic engineering of Escherichia coli using synthetic small regulatory RNAs. *Nature biotechnology*, **31**, 170-174.
135. Tomari, Y. and Zamore, P.D. (2005) Perspective: machines for RNAi. *Genes & Development*, **19**, 517-529.
136. Soutschek, J., Akinc, A., Bramlage, B., Charisse, K., Constien, R., Donoghue, M., Elbashir, S., Geick, A., Hadwiger, P., Harborth, J. *et al.* (2004) Therapeutic silencing of an endogenous gene by systemic administration of modified siRNAs. *Nature*, **432**, 173-178.
137. Dietzl, G., Chen, D., Schnorrer, F., Su, K.C., Barinova, Y., Fellner, M., Gasser, B., Kinsey, K., Oppel, S., Scheiblauer, S. *et al.* (2007) A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. *Nature*, **448**, 151-U151.
138. Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M. *et al.* (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature*, **421**, 231-237.
139. Chuang, C.F. and Meyerowitz, E.M. (2000) Specific and heritable genetic interference by double-stranded RNA in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.*, **97**, 4985-4990.

140. Mansoor, S., Amin, I., Hussain, M., Zafar, Y. and Briddon, R.W. (2006) Engineering novel traits in plants through RNA interference. *Trends in plant science*, **11**, 559-565.
141. Drinnenberg, I.A., Weinberg, D.E., Xie, K.T., Mower, J.P., Wolfe, K.H., Fink, G.R. and Bartel, D.P. (2009) RNAi in Budding Yeast. *Science*, **326**, 544-550.
142. Crook, N. and Alper, H.S. (2013) Model-based design of synthetic, biological systems. *Chem. Eng. Sci.*, **103**, 2-11.
143. Khalil, Ahmad S., Lu, Timothy K., Bashor, Caleb J., Ramirez, Cherie L., Pyenson, Nora C., Joung, J.K. and Collins, James J. (2012) A Synthetic Biology Framework for Programming Eukaryotic Transcription Functions. *Cell*, **150**, 647-658.
144. Zaslaver, A., Bren, A., Ronen, M., Itzkovitz, S., Kikoin, I., Shavit, S., Liebermeister, W., Surette, M.G. and Alon, U. (2006) A comprehensive library of fluorescent transcriptional reporters for Escherichia coli. *Nature methods*, **3**, 623-628.
145. Du, J., Yuan, Y., Si, T., Lian, J. and Zhao, H. (2012) Customized optimization of metabolic pathways by combinatorial transcriptional engineering. *Nucleic acids research*, **40**, e142.
146. Xi, L., Fondufe-Mittendorf, Y., Xia, L., Flatow, J., Widom, J. and Wang, J.P. (2010) Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics*, **11**, 346.
147. Chang, D.T., Huang, C.Y., Wu, C.Y. and Wu, W.S. (2011) YPA: an integrated repository of promoter features in Saccharomyces cerevisiae. *Nucleic acids research*, **39**, D647-652.
148. Abdulrehman, D., Monteiro, P.T., Teixeira, M.C., Mira, N.P., Lourenco, A.B., dos Santos, S.C., Cabrito, T.R., Francisco, A.P., Madeira, S.C., Aires, R.S. *et al.* (2011) YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic acids research*, **39**, D136-140.
149. Flagfeldt, D.B., Siewers, V., Huang, L. and Nielsen, J. (2009) Characterization of chromosomal integration sites for heterologous gene expression in Saccharomyces cerevisiae. *Yeast (Chichester, England)*, **26**, 545-551.
150. McIsaac, R.S., Oakes, B.L., Wang, X., Dummit, K.A., Botstein, D. and Noyes, M.B. (2013) Synthetic gene expression perturbation systems with rapid, tunable, single-gene specificity in yeast. *Nucleic acids research*, **41**, e57.
151. Blount, B.A., Weenink, T., Vasylechko, S. and Ellis, T. (2012) Rational Diversification of a Promoter Providing Fine-Tuned Expression and Orthogonal Regulation for Synthetic Biology. *PLoS ONE*, **7**, e33279.
152. Jeppsson, M., Johansson, B., Jensen, P.R., Hahn-Hagerdal, B. and Gorwa-Grauslund, M.F. (2003) The level of glucose-6-phosphate dehydrogenase activity strongly influences xylose fermentation and inhibitor sensitivity in recombinant Saccharomyces cerevisiae strains. *Yeast (Chichester, England)*, **20**, 1263-1272.

153. Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S. and Young, R.A. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717-728.
154. Swamy, K.B., Chu, W.Y., Wang, C.Y., Tsai, H.K. and Wang, D. (2011) Evidence of association between nucleosome occupancy and the evolution of transcription factor binding sites in yeast. *BMC evolutionary biology*, **11**, 150.
155. Lanza, A.M., Blazeck, J.J., Crook, N.C. and Alper, H.S. (2012) Linking Yeast Gcn5p Catalytic Function and Gene Regulation Using a Quantitative, Graded Dominant Mutant Approach. *PLoS ONE*, **7**, e36193.
156. Keasling, J.D. (1999) Gene-expression tools for the metabolic engineering of bacteria. *Trends Biotechnol.*, **17**, 452-460.
157. Christianson, T.W., Sikorski, R.S., Dante, M., Shero, J.H. and Hieter, P. (1992) Multifunctional Yeast High-Copy-Number Shuttle Vectors. *Gene*, **110**, 119-122.
158. Gustafsson, C., Govindarajan, S. and Minshull, J. (2004) Codon bias and heterologous protein expression. *Trends Biotechnol.*, **22**, 346-353.
159. Blake, W.J., Kaern, M., Cantor, C.R. and Collins, J.J. (2003) Noise in eukaryotic gene expression. *Nature*, **422**, 633-637.
160. Salis, H.M., Mirsky, E.A. and Voigt, C.A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.*, **27**, 946-U112.
161. Kudla, G., Murray, A.W., Tollervey, D. and Plotkin, J.B. (2009) Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science*, **324**, 255-258.
162. Paik, S.Y., Ra, K.S., Cho, H.S., Koo, K.B., Baik, H.S., Lee, M.C., Yun, J.W. and Choi, J.W. (2006) The influence of the nucleotide sequences of random Shine-Dalgarno and spacer region on bovine growth hormone gene expression. *J. Microbiol.*, **44**, 64-71.
163. Pickering, B.M. and Willis, A.E. (2005) The implications of structured 5 ' untranslated regions on translation and disease. *Semin. Cell Dev. Biol.*, **16**, 39-47.
164. McCarthy, J.E.G. (1998) Posttranscriptional control of gene expression in yeast. *Microbiol. Mol. Biol. Rev.*, **62**, 1492-1553.
165. Kozak, M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, **361**, 13-37.
166. Baim, S.B. and Sherman, F. (1988) Messenger-RNA Structures Influencing Translation in the Yeast *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **8**, 1591-1601.
167. Ringner, M. and Krogh, M. (2005) Folding free energies of 5'-UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLoS Comput. Biol.*, **1**, 585-592.
168. Kapp, L.D. and Lorsch, J.R. (2004) The molecular mechanics of eukaryotic translation. *Annu. Rev. Biochem.*, **73**, 657-704.

169. Partow, S., Siewers, V., Bjorn, S., Nielsen, J. and Maury, J. (2010) Characterization of different promoters for designing a new expression vector in Saccharomyces cerevisiae. *Yeast*, **27**, 955-964.
170. Muller, S., Sandal, T., Kamp-Hansen, P. and Dalboge, H. (1998) Comparison of expression systems in the yeasts Saccharomyces cerevisiae, Hansenula polymorpha, Klyveromyces lactis, Schizosaccharomyces pombe and Yarrowia lipolytica. Cloning of two novel promoters from Yarrowia lipolytica. *Yeast*, **14**, 1267-1283.
171. Kozak, M. (1989) Circumstances and mechanisms of inhibition of translation by secondary structure in eukaryotic messenger-RNAs. *Mol. Cell. Biol.*, **9**, 5134-5142.
172. Wang, L.J. and Wessler, S.R. (2001) Role of mRNA secondary structure in translational repression of the maize transcriptional activates L-C. *Plant Physiol.*, **125**, 1380-1387.
173. Short, J.D. and Pfarr, C.M. (2002) Translational regulation of the JunD messenger RNA. *J. Biol. Chem.*, **277**, 32697-32705.
174. van der Velden, A.W., van Nierop, K., Voorma, H.O. and Thomas, A.A.M. (2002) Ribosomal scanning on the highly structured insulin-like growth factor II-leader 1. *Int. J. Biochem. Cell Biol.*, **34**, 286-297.
175. Hoover, D.S., Wingett, D.G., Zhang, J., Reeves, R. and Magnuson, N.S. (1997) Pim-1 protein expression is regulated by its 5 '-untranslated region and translation initiation factor eIF-4E. *Cell Growth Differ.*, **8**, 1371-1380.
176. Dirks, R.M., Bois, J.S., Schaeffer, J.M., Winfree, E. and Pierce, N.A. (2007) Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.*, **49**, 65-88.
177. Hofacker, I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429-3431.
178. Serra, M.J. and Turner, D.H. (1995), *Energetics of Biological Macromolecules*. Academic Press Inc, San Diego, Vol. 259, pp. 242-261.
179. Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M. and Turner, D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 7287-7292.
180. Sikorski, R.S. and Hieter, P. (1989) A System of Shuttle Vectors and Yeast Host Strains Designed for Efficient Manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics*, **122**, 19-27.
181. Mumberg, D., Muller, R. and Funk, M. (1995) Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene*, **156**, 119-122.
182. Sheff, M.A. and Thorn, K.S. (2004) Optimized cassettes for fluorescent protein tagging in Saccharomyces cerevisiae. *Yeast (Chichester, England)*, **21**, 661-670.

183. Teste, M.A., Duquenne, M., Francois, J.M. and Parrou, J.L. (2009) Validation of reference genes for quantitative expression analysis by real-time RT-PCR in Saccharomyces cerevisiae. *BMC Mol. Biol.*, **10**.
184. Nevoigt, E., Kohnke, J., Fischer, C.R., Alper, H., Stahl, U. and Stephanopoulos, G. (2006) Engineering of promoter replacement cassettes for fine-tuning of gene expression in Saccharomyces cerevisiae. *Appl. Environ. Microbiol.*, **72**, 5266-5273.
185. Andersen, J.B., Sternberg, C., Poulsen, L.K., Bjorn, S.P., Givskov, M. and Molin, S. (1998) New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. *Appl. Environ. Microbiol.*, **64**, 2240-2246.
186. Sharp, P.M. and Li, W.H. (1987) The codon adaptation index - a measure of directinal synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281-1295.
187. Grote, A., Hiller, K., Scheer, M., Munch, R., Nortemann, B., Hempel, D.C. and Jahn, D. (2005) JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res.*, **33**, W526-W531.
188. Ellis, T., Wang, X. and Collins, J.J. (2009) Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat. Biotechnol.*, **27**, 465-471.
189. Dueber, J.E., Wu, G.C., Malmirchegini, G.R., Moon, T.S., Petzold, C.J., Ullal, A.V., Prather, K.L.J. and Keasling, J.D. (2009) Synthetic protein scaffolds provide modular control over metabolic flux. *Nat Biotech*, **27**, 753-759.
190. Beekwilder, J., van Rossum, H.M., Koopman, F., Sonntag, F., Buchhaupt, M., Schrader, J., Hall, R.D., Bosch, D., Pronk, J.T., van Maris, A.J. *et al.* (2014) Polycistronic expression of a beta-carotene biosynthetic pathway in Saccharomyces cerevisiae coupled to beta-ionone production. *J Biotechnol*.
191. Lanza, A., Curran, K., Rey, L. and Alper, H. (2014) A Condition-Specific Codon Optimization Approach for Improved Heterologous Gene Expression in Saccharomyces cerevisiae. *BMC Systems Biology*.
192. Donnelly, M.L., Hughes, L.E., Luke, G., Mendoza, H., ten Dam, E., Gani, D. and Ryan, M.D. (2001) The 'cleavage' activities of foot-and-mouth disease virus 2A site-directed mutants and naturally occurring '2A-like' sequences. *The Journal of general virology*, **82**, 1027-1041.
193. Zhu, J., Musco, M.L. and Grace, M.J. (1999) Three-color flow cytometry analysis of tricistronic expression of eBFP, eGFP, and eYFP using EMCV-IRES linkages. *Cytometry*, **37**, 51-59.
194. Zhou, W., Edelman, G.M. and Mauro, V.P. (2003) Isolation and identification of short nucleotide sequences that affect translation initiation in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, **100**, 4457-4462.
195. Zhou, W., Edelman, G.M. and Mauro, V.P. (2001) Transcript leader regions of two Saccharomyces cerevisiae mRNAs contain internal ribosome entry sites that function in living cells. *Proc Natl Acad Sci U S A*, **98**, 1531-1536.

196. Han, F. and Zhang, X. (2006) Internal initiation of mRNA translation in insect cell mediated by an internal ribosome entry site (IRES) from shrimp white spot syndrome virus (WSSV). *Biochem Biophys Res Commun*, **344**, 893-899.
197. Marom, L., Hen-Avivi, S., Pinchasi, D., Chekanova, J.A., Belostotsky, D.A. and Elroy-Stein, O. (2009) Diverse poly(A) binding proteins mediate internal translational initiation by a plant viral IRES. *RNA biology*, **6**, 446-454.
198. Ronfort, C., De Breyne, S., Sandrin, V., Darlix, J.L. and Ohlmann, T. (2004) Characterization of two distinct RNA domains that regulate translation of the Drosophila gypsy retroelement. *RNA (New York, N.Y.)*, **10**, 504-515.
199. Komar, A.A., Lesnik, T., Cullin, C., Merrick, W.C., Trachsel, H. and Altmann, M. (2003) Internal initiation drives the synthesis of Ure2 protein lacking the prion domain and affects [URE3] propagation in yeast cells. *EMBO J*, **22**, 1199-1209.
200. Reineke, L.C., Cao, Y., Baus, D., Hossain, N.M. and Merrick, W.C. (2011) Insights into the role of yeast eIF2A in IRES-mediated translation. *PLoS One*, **6**, e24492.
201. Komar, A.A., Gross, S.R., Barth-Baus, D., Strachan, R., Hensold, J.O., Goss Kinzy, T. and Merrick, W.C. (2005) Novel characteristics of the biological properties of the yeast Saccharomyces cerevisiae eukaryotic initiation factor 2A. *J Biol Chem*, **280**, 15601-15611.
202. Davies, M.V. and Kaufman, R.J. (1992) The sequence context of the initiation codon in the encephalomyocarditis virus leader modulates efficiency of internal translation initiation. *Journal of virology*, **66**, 1924-1932.
203. Hennecke, M., Kwissa, M., Metzger, K., Oumard, A., Kroger, A., Schirmbeck, R., Reimann, J. and Hauser, H. (2001) Composition and arrangement of genes define the strength of IRES-driven translation in bicistronic mRNAs. *Nucleic acids research*, **29**, 3327-3334.
204. Wilhelm, F.X., Wilhelm, M. and Gabriel, A. (2005) Reverse transcriptase and integrase of the Saccharomyces cerevisiae Ty1 element. *Cytogenet. Genome Res.*, **110**, 269-287.
205. Wilhelm, F.X., Wilhelm, M. and Gabriel, A. (2005) Reverse transcriptase and integrase of the Saccharomyces cerevisiae Ty1 element. *Cytogenet Genome Res*, **110**, 269-287.
206. Boeke, J.D., Garfinkel, D.J., Styles, C.A. and Fink, G.R. (1985) Ty elements transpose through an RNA intermediate. *Cell*, **40**, 491-500.
207. Bolton, E.C., Coombes, C., Eby, Y., Cardell, M. and Boeke, J.D. (2005) Identification and characterization of critical cis-acting sequences within the yeast Ty1 retrotransposon. *RNA-Publ. RNA Soc.*, **11**, 308-322.
208. Boutabout, M., Wilhelm, M. and Wilhelm, F.-X. (2001) DNA synthesis fidelity by the reverse transcriptase of the yeast retrotransposon Ty1. *Nucleic Acids Research*, **29**, 2217-2222.
209. Sharon, G., Burkett, T.J. and Garfinkel, D.J. (1994) EFFICIENT HOMOLOGOUS RECOMBINATION OF TY1 ELEMENT CDNA WHEN INTEGRATION IS BLOCKED. *Molecular and Cellular Biology*, **14**, 6540-6551.

210. Nissley, D.V., Garfinkel, D.J. and Strathern, J.N. (1996) HIV reverse transcription in yeast. *Nature*, **380**, 30-30.
211. Curcio, M.J. and Garfinkel, D.J. (1991) SINGLE-STEP SELECTION FOR TY1 ELEMENT RETROTRANSPOSITION. *Proc. Natl. Acad. Sci. U. S. A.*, **88**, 936-940.
212. Boutabout, M., Wilhelm, M. and Wilhelm, F.X. (2001) DNA synthesis fidelity by the reverse transcriptase of the yeast retrotransposon Ty1. *Nucleic Acids Res.*, **29**, 2217-2222.
213. Bryk, M., Briggs, S.D., Strahl, B.D., Curcio, M.J., Allis, C.D. and Winston, F. (2002) Evidence that SET1, a factor required for methylation of histone H3, regulates rDNA silencing in S-cerevisiae by a sir2-independent mechanism. *Curr. Biol.*, **12**, 165-170.
214. Radford, S.J., Boyle, M.L., Sheely, C.J., Graham, J., Haeusser, D.P., Zimmerman, L. and Keeney, J.B. (2004) Increase in Tyl cDNA recombination in yeast sir4 mutant strains at high temperature. *Genetics*, **168**, 89-101.
215. Lee, B.S., Lichtenstein, C.P., Faiola, B., Rinckel, L.A., Wysock, W., Curcio, M.J. and Garfinkel, D.J. (1998) Posttranslational inhibition of Ty1 retrotransposition by nucleotide excision repair transcription factor TFIIH subunits Ss12p and Rad3p. *Genetics*, **148**, 1743-1761.
216. Conte, D. and Curcio, M.J. (2000) Fus3 controls Ty1 transpositional dormancy through the invasive growth MAPK pathway. *Molecular microbiology*, **35**, 415-427.
217. Scholes, D.T., Banerjee, M., Bowen, B. and Curcio, M.J. (2001) Multiple regulators of Tyl transposition in Saccharomyces cerevisiae have conserved roles in genome maintenance. *Genetics*, **159**, 1449-1465.
218. Chan, J.E. and Kolodner, R.D. (2011) A Genetic and Structural Study of Genome Rearrangements Mediated by High Copy Repeat Ty1 Elements. *PLoS genetics*, **7**.
219. Kohlstaedt, L.A., Wang, J., Friedman, J.M., Rice, P.A. and Steitz, T.A. (1992) CRYSTAL-STRUCTURE AT 3.5 ANGSTROM RESOLUTION OF HIV-1 REVERSE-TRANSCRIPTASE COMPLEXED WITH AN INHIBITOR. *Science*, **256**, 1783-1790.
220. de Oliveira, T., Engelbrecht, S., van Rensburg, E.J., Gordon, M., Bishop, K., zur Megede, J., Barnett, S.W. and Cassol, S. (2003) Variability at human immunodeficiency virus type 1 subtype C protease cleavage sites: an indication of viral fitness? *J. Virol.*, **77**, 9422-9430.
221. Merkulov, G.V., Lawler, J.F., Eby, Y. and Boeke, J.D. (2001) Ty1 proteolytic cleavage sites are required for transposition: All sites are not created equal. *J. Virol.*, **75**, 638-644.
222. Wilhelm, M., Boutabout, M. and Wilhelm, F.X. (2000) Expression of an active form of recombinant Ty1 reverse transcriptase in Escherichia coli: a fusion protein containing the C-terminal region of the Ty1 integrase linked to the reverse transcriptase-RNase H domain exhibits polymerase and RNase H activities. *Biochem. J.*, **348**, 337-342.

223. Servant, G., Pinson, B., Tchalikian-Cosson, A., Coulpier, F., Lemoine, S., Pennetier, C., Bridier-Nahmias, A., Todeschini, A.L., Fayol, H., Daignan-Fornier, B. *et al.* (2012) Tye7 regulates yeast Ty1 retrotransposon sense and antisense transcription in response to adenylic nucleotides stress. *Nucleic Acids Research*, **40**, 5271-5282.
224. Kawakami, K., Pande, S., Faiola, B., Moore, D.P., Boeke, J.D., Farabaugh, P.J., Strathern, J.N., Nakamura, Y. and Garfinkel, D.J. (1993) A RARE TRANSFER RNA-ARG(CCU) THAT REGULATES TY1 ELEMENT RIBOSOMAL FRAMESHIFTING IS ESSENTIAL FOR TY1 RETROTRANSPOSITION IN SACCHAROMYCES-CEREVISIAE. *Genetics*, **135**, 309-320.
225. Curcio, M.J., Kenny, A.E., Moore, S., Garfinkel, D.J., Weintraub, M., Gamache, E.R. and Scholes, D.T. (2007) S-phase checkpoint pathways stimulate the mobility of the retrovirus-like transposon Ty1. *Molecular and Cellular Biology*, **27**, 8874-8885.
226. Belcourt, M.F. and Farabaugh, P.J. (1990) Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a 7 nucleotide minimal site. *Cell*, **62**, 339-352.
227. Patrick, W.M., Firth, A.E. and Blackburn, J.M. (2003) User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. *Protein Eng.*, **16**, 451-457.
228. Stumpp, S.N., Heyn, B. and Brakmann, S. (2010) Activity-based selection of HIV-1 reverse transcriptase variants with decreased polymerization fidelity. *Biological chemistry*, **391**, 665-674.
229. Kaushik, N., Chowdhury, K., Pandey, V.N. and Modak, M.J. (2000) Valine of the YVDD motif of moloney murine leukemia virus reverse transcriptase: role in the fidelity of DNA synthesis. *Biochemistry*, **39**, 5155-5165.
230. Kaushik, N., Singh, K., Alluru, I. and Modak, M.J. (1999) Tyrosine 222, a member of the YXDD motif of MuLV RT, is catalytically essential and is a major component of the fidelity center. *Biochemistry*, **38**, 2617-2627.
231. Halvas, E.K., Svarovskaia, E.S. and Pathak, V.K. (2000) Development of an in vivo assay to identify structural determinants in murine leukemia virus reverse transcriptase important for fidelity. *Journal of virology*, **74**, 312-319.
232. Halvas, E.K., Svarovskaia, E.S. and Pathak, V.K. (2000) Role of murine leukemia virus reverse transcriptase deoxyribonucleoside triphosphate-binding site in retroviral replication and in vivo fidelity. *Journal of virology*, **74**, 10349-10358.
233. Dapp, M.J., Heineman, R.H. and Mansky, L.M. (2013) Interrelationship between HIV-1 fitness and mutation rate. *Journal of molecular biology*, **425**, 41-53.
234. Shah, F.S., Curr, K.A., Hamburgh, M.E., Parniak, M., Mitsuya, H., Arnez, J.G. and Prasad, V.R. (2000) Differential influence of nucleoside analog-resistance mutations K65R and L74V on the overall mutation rate and error specificity of human immunodeficiency virus type 1 reverse transcriptase. *The Journal of biological chemistry*, **275**, 27037-27044.

235. Lwatula, C., Garforth, S.J. and Prasad, V.R. (2012) Lys66 residue as a determinant of high mismatch extension and misinsertion rates of HIV-1 reverse transcriptase. *The FEBS journal*, **279**, 4010-4024.
236. Youngren, S.D., Boeke, J.D., Sanders, N.J. and Garfinkel, D.J. (1988) Functional organization of the retrotransposon Ty from Saccharomyces cerevisiae: Ty protease is required for transposition. *Molecular and cellular biology*, **8**, 1421-1431.
237. Wilhelm, M. and Wilhelm, F.X. (2005) Role of integrase in reverse transcription of the Saccharomyces cerevisiae retrotransposon Ty1. *Eukaryotic cell*, **4**, 1057-1065.
238. Lee, S.-M., Jellison, T. and Alper, H.S. (2012) Directed Evolution of Xylose Isomerase for Improved Xylose Catabolism and Fermentation in the Yeast Saccharomyces cerevisiae. *Appl. Environ. Microbiol.*, **78**, 5708-5716.
239. Karim, A.S., Curran, K.A. and Alper, H.S. (2013) Characterization of plasmid burden and copy number in Saccharomyces cerevisiae for optimization of metabolic engineering applications. *FEMS Yeast Res.*, **13**, 107-116.
240. Garfinkel, D.J., Boeke, J.D. and Fink, G.R. (1985) Ty element transposition: reverse transcriptase and virus-like particles. *Cell*, **42**, 507-517.
241. T. Werpy, G.P. (2004) Top Value Added Chemicals from Biomass.
242. Hegemann, J.H. and Heick, S.B. (2011) Delete and repeat: a comprehensive toolkit for sequential gene knockout in the budding yeast Saccharomyces cerevisiae. *Methods in molecular biology (Clifton, N.J.)*, **765**, 189-206.
243. Hooshangi, S., Thiberge, S. and Weiss, R. (2005) Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 3581-3586.
244. Friedland, A.E., Lu, T.K., Wang, X., Shi, D., Church, G. and Collins, J.J. (2009) Synthetic gene networks that count. *Science*, **324**, 1199-1202.
245. Shirane, D., Sugao, K., Namiki, S., Tanabe, M., Iino, M. and Hirose, K. (2004) Enzymatic production of RNAi libraries from cDNAs. *Nat Genet*, **36**, 190-196.
246. Johnson, E. and Srivastava, R. (2013) Volatility in mRNA secondary structure as a design principle for antisense. *Nucleic acids research*, **41**, e43.
247. Sledz, C.A., Holko, M., de Veer, M.J., Silverman, R.H. and Williams, B.R.G. (2003) Activation of the interferon system by short-interfering RNAs. *Nature Cell Biology*, **5**, 834-839.
248. Huang, L., Jin, J., Deighan, P., Kiner, E., McReynolds, L. and Lieberman, J. (2013) Efficient and specific gene knockdown by small interfering RNAs produced in bacteria. *Nat Biotech*, **31**, 350-356.
249. Voineagu, I., Narayanan, V., Lobachev, K.S. and Mirkin, S.M. (2008) Replication stalling at unstable inverted repeats: Interplay between DNA hairpins and fork stabilizing proteins. *Proceedings of the National Academy of Sciences*, **105**, 9936-9941.
250. Yoshimatsu, T. and Nagawa, F. (1989) Control of gene expression by artificial introns in Saccharomyces cerevisiae. *Science*, **244**, 1346-1348.

251. Blazeck, J., Miller, J., Pan, A., Gengler, J., Holden, C., Jamoussi, M. and Alper, H.S. (2013) Metabolic Engineering of Saccharomyces cerevisiae for itaconic acid production. *Submitted/Under Review*.
252. Troutt, A.B., McHeyzer-Williams, M.G., Pulendran, B. and Nossal, G.J. (1992) Ligation-anchored PCR: a simple amplification technique with single-sided specificity. *Proceedings of the National Academy of Sciences*, **89**, 9823-9825.
253. Si, T., Luo, Y., Xiao, H. and Zhao, H. (2014) Utilizing an endogenous pathway for 1-butanol production in Saccharomyces cerevisiae. *Metabolic Engineering*, **22**, 60-68.
254. Bukau, B. and Horwich, A.L. (1998) The Hsp70 and Hsp60 chaperone machines. *Cell*, **92**, 351-366.
255. Planta, R.J. and Mager, W.H. (1998) The list of cytoplasmic ribosomal proteins of Saccharomyces cerevisiae. *Yeast (Chichester, England)*, **14**, 471-477.
256. Bermingham-McDonogh, O., Gralla, E.B. and Valentine, J.S. (1988) The copper, zinc-superoxide dismutase gene of Saccharomyces cerevisiae: cloning, sequencing, and biological activity. *Proc Natl Acad Sci U S A*, **85**, 4789-4793.
257. Elowitz, M.B. and Leibler, S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**, 335-338.
258. Kozak, M. (2005) A second look at cellular mRNA sequences said to function as internal ribosome entry sites. *Nucleic Acids Research*, **33**, 6593-6602.
259. Sambrook, J. and Russell, D.W. (2001) *Molecular cloning: a laboratory manual*. 3 ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
260. Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M. and Iyer, V.R. (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol*, **6**, e65.
261. Li, W.-Z. and Sherman, F. (1991) Two Types of TATA Elements for the CYC1 Gene of the Yeast *Saccharomyces Cerevisiae*. *Mol. Cell. Biol.*, **11**, 666-676.
262. Hahn, S., Hoar, E.T. and Guarente, L. (1985) Each of 3 TATA Elements Specifies a Subset of the Transcription Initiation Sites at the CYC-1 Promoter of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.*, **82**, 8562-8566.
263. Nagashima, K., Kasai, M., Nagata, S. and Kaziro, Y. (1986) Structure of the 2 genes coding for polypeptide chain elongation factor 1-alpha (EF-1-alpha) from *Saccharomyces cerevisiae*. *Gene*, **45**, 265-273.
264. Zhang, Z.H. and Dietrich, F.S. (2005) Mapping of transcription start sites in Saccharomyces cerevisiae using 5 ' SAGE. *Nucleic Acids Res.*, **33**, 2838-2851.
265. Lamprecht, M.R., Sabatini, D.M. and Carpenter, A.E. (2007) CellProfiler(TM): free, versatile software for automated biological image analysis. *BioTechniques*, **42**, 71-75.
266. Hall, B.M., Ma, C.X., Liang, P. and Singh, K.K. (2009) Fluctuation AnaLysis CalculatOR: a web tool for the determination of mutation rate using Luria-Delbruck fluctuation analysis. *Bioinformatics*, **25**, 1564-1565.

267. Ma, W.T., Sandri, G.V. and Sarkar, S. (1992) Analysis of the Luria-Delbrück Distribution Using Discrete Convolution Powers. *Journal of Applied Probability*, **29**, 255-267.
268. Lamprecht, M.R., Sabatini, D.M. and Carpenter, A.E. (2007) CellProfiler: free, versatile software for automated biological image analysis. *BioTechniques*, **42**, 71-75.
269. Masella, A.P., Bartram, A.K., Truszkowski, J.M., Brown, D.G. and Neufeld, J.D. (2012) PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics*, **13**, 31.
270. Patel, R.K. and Jain, M. (2012) NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS ONE*, **7**, e30619.
271. Ning, Z., Cox, A.J. and Mullikin, J.C. (2001) SSAHA: a fast search method for large DNA databases. *Genome research*, **11**, 1725-1729.
272. CLOPPER, C.J. and PEARSON, E.S. (1934) THE USE OF CONFIDENCE OR FIDUCIAL LIMITS ILLUSTRATED IN THE CASE OF THE BINOMIAL. *Biometrika*, **26**, 404-413.
273. Guldener, U., Heck, S., Fielder, T., Beinhauer, J. and Hegemann, J.H. (1996) A new efficient gene disruption cassette for repeated use in budding yeast. *Nucleic acids research*, **24**, 2519-2524.
274. Shao, Z., Zhao, H. and Zhao, H. (2009) DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Research*, **37**, e16.
275. Lanza, A.M., Kim, D.S. and Alper, H.S. (2013) Evaluating the influence of selection markers on obtaining selected pools and stable cell lines in human cells. *Biotechnology Journal*, **8**, 811-821.