The Dissertation Committee for Amanda Morgan Lanza Certifies that this is the approved version of the following dissertation:

Novel Tools for Engineering Eukaryotic Cells Using a Systems Level Approach

Committee:

Hal S. Alper, Supervisor

George Georgiou

Makkuni Jayaram

Jennifer Maynard

Daniel I.C. Wang

Novel Tools for Engineering Eukaryotic Cells Using a Systems Level Approach

by

Amanda Morgan Lanza, B.S.Ch.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 2013

Dedication

To my parents:

For all the sacrifices you made and love you’ve given.

Acknowledgements

The accomplishment of this work has been achieved in a large part through the support and guidance of my research advisor, Professor Hal Alper. I would like to thank him for many thoughtful scientific discussions and a healthy interplay of ideas that has resulted not only in me completing the projects described herein, but also in my development as an independent and driven researcher. Hal has always supported my participation in conferences and seminars. Because of his recommendations, I have been privileged to be considered for and awarded many fellowships and other honors during my years at the University of Texas.

I would also like to thank my thesis committee members, who have each been supportive and helpful to me. First, I would like to acknowledge Professor Jennifer Maynard for always making herself available to me. Jennifer has provided both friendship and advice on both personal and professional matters and helped me to extend my professional network. Second, I would like to thank Professor Makkuni Jayaram (Jay) for a warmth and kindness that is not always easy to find. He has graciously given me materials, resources, and time in his labs to help with my experiments. Furthermore, I have benefitted from our discussions and admire his scientific curiosity. Next, I would like to thank Professor George Georgiou who has on many occasions made himself available despite a busy schedule. His insightful questions have helped guide my research, and career oriented discussions have been especially useful in this last year. I have enjoyed many a conversation about skiing and appreciate him reminding me of the importance of a good vacation, even while in graduate school. Finally, I would like to thank Professor Daniel Wang, who graciously agreed to serve on my thesis committee despite many other obligations and a significant geographical distance between us.

Professor Wang has long served as a mentor for me and remains an important compass as I transition from student to professional. I appreciate the many, unique opportunities that have come from his generosity over the years, as well as the direct, blunt and always useful feedback he gives me. His approval is hard to earn, but well worth the effort.

I would like to acknowledge many others that have directly contributed to the success of my experiments: Richard Salinas for guidance regarding flow cytometry, Joseph Hanson for teaching me how to do southern blots, Dr. Supriya Pai for her thoughtful discussions on RT-PCR, Dr. Scott Hunicke-Smith for his input regarding microarray processing, Dr. Taejoon Kwon for help with microarray data analysis, Keng-Ming Chang for teaching me how to do immunofluorescence assays, and Joyce Ho and Yubin Park for help with matrix drift calculations.

I would also like to acknowledge financial support that I have received throughout the years. My work was been partially supported by Shire Human Genetic Therapies, a National Science Foundation Graduate Research Fellowship, a Harrington Dissertation Fellowship, a P.E.O. Graduate Scholarship and a Cockrell School of Engineering Thrust Fellowship.

I would like to thank fellow members of the Alper Lab: Joseph Abatemarco, John Blazeck, Joseph Cheng, Nathan Crook, Kathleen Curran, Andrew Hill, Rebecca Knight, John Leavitt, Sun Mi Lee, Leqian Liu, Heidi Redden, Dr. Jie Sun, and Eric Young. I have also had the pleasure of supervising three excellent undergraduate researchers: Timothy Dyess, Do Soon Kim and Lindsey Rey. Their influence in my work has been significant and is much appreciated.

I have also been helped by many others within the ChE Department. To Professor Roger Bonnecaze, thank you for supporting my endeavors to develop leadership within the department and for writing several references on my behalf. To Jim Smitherman,

thank you for promptly fixing many important pieces of equipment, even when you didn't have to, you saved me a lot of stress. To Randy Rife and Patrick Danielewski, thank you for helping me solve many IT problems over the years. Finally, Kevin Haynes, Eddie Ibarra, T Stockman, and Tammy McDade of the UT Chemical Engineering Department were always supportive in a variety of capacities and a pleasure to work with day in and day out.

Graduate research can sometimes be a lonely process, and I would like to thank my close friends outside of the Alper Lab who have made my time in Austin very enjoyable (in no particular order): Dr. Landry Khounlavong, Dr. Collin Smith, Douglas French, Peter Frailie, Zachary Smith, William Kelton, William, Emily and Jackson Liechty, Dr. Mary Caldorera-Moore, Dr. David Van Wagener, Dr. Ross Dugas, Paul Abel, Dr. Alexander Voice, Ramiro Palma, and Dylan Kipp. I want to thank the members of the CHEetahs flag football team that I have been privileged to play with (and captain) over the years: Dr. Jamie Sutherland, Misty Heitsch, Dr. Jennifer Pai, Dr. Stephanie Freeman, Jessica Goodfellow, Laurene Dobrowolski, Dr. Danielle Smith, Dr. Diana Van Blarcom, Dr. Andrea Miller, Dr. Alyson Sagle, Dr. Elizabeth Costner, Tyne Johns, Julie Cushen, Jennifer Knipe, Michelle Robinson, Karen Scida, Reika Katsumata, Stephanie Steichen, Amanda Paine, Ellen Wagner, Paola Gonzalez, Catherine Silvestri, Victoria Cotham, and Joyce Ho.

Most importantly, I would like to thank my loving and supportive family for always believing in me and being there when I needed them. To my parents, Susan and Robert, you have given me everything I could possibly need to become a successful and happy adult. I appreciate all the sacrifices you have made for me over the years and feel incredibly fortunate to have parents like you. I want to especially thank you for all the times you have helped me solve problems while in graduate school, it has been a long

road and I am so excited to celebrate it with you. To my incredible sister Jennifer, thank you for being the kind of sibling I aspire to be. You are loving, generous, and a great listener. Your tenacity and resilience is inspiring and I will always look up to you, big sister. I would like to thank all of my grandparents for their unconditional love, but especially Cyrus Schoen, for I do not believe there is any individual more proud of my accomplishments than him. Finally, I would like to thank my loving and supportive fiancé, Thomas Lewis. You have been my rock throughout graduate school, suffering through many research setbacks and frustrations, and always offering much needed comfort, support, and laughter. I am so grateful to have met you, my best friend, in graduate school.

Novel Tools for Engineering Eukaryotic Cells Using a Systems Level Approach

Amanda Morgan Lanza, PhD

The University of Texas at Austin, 2013

Supervisor: Hal S. Alper

Engineered cellular systems are a promising avenue for production of a wide range of useful products including renewable fuels, commodity and specialty chemicals, industrial enzymes, and pharmaceuticals. Achieving this breadth of biological products is facilitated by the diversity of organisms found in nature. Using biological and engineering principles, this diversity can be harnessed to make efficient and renewable bio-based products. Such advancements rely upon our ability to modify host genetics and metabolism. This work focuses on the development of new biotechnological tools which enable cellular engineering, and the implementation of these tools in eukaryotic systems.

Mammalian cell engineering has important implications in protein therapeutics and gene therapy. One major limitation, however, is the ability to predictably control gene expression. We address this challenge by examining critical aspects of gene expression in human cells. First, we evaluate the impact of selection markers, a common mammalian expression element, on cell line development. In doing so, we determine that Zeocin is the best selection agent for human cells. Next, we identify loci across the genome that support high level expression of recombinant DNA and demonstrate their advantage for stable integration. Finally, we optimize a Cre recombinase based

methodology that enables efficient retargeting of genomic loci. Collectively, this work augments the current genetic toolbox for human cell lines.

Beyond basic gene expression, there is interest in understanding global interactions within the cell and how they relate to phenomena including gene regulation, expression and disease states. Although our tools are not yet sufficient to study these phenomena in many hosts, methods can be developed in lower eukaryotes and then adapted for more complex hosts later. We demonstrated two methods in *S. cerevisiae* that utilize a systems-level approach to understand complex phenotypes. First, we developed condition-specific codon optimization that utilizes systems biology information to optimize gene sequence in a condition-specific manner. Additionally, we developed a Graded Dominant Mutant Approach which can be used to dissect multifunctional proteins, understand epigenetic factors, and quantitatively determine protein-DNA interactions. Both can be implemented in many cellular hosts and expand our ability to engineer complex phenotypes in eukaryotic cell systems.

Table of Contents

List of Tables xvi

List of Figures xvii

Chapter 1: Introduction and Background 1

1.1 Cell and Metabolic Engineering 1

1.2 Predicted Gene Expression in Mammalian Cellular Hosts 2

1.2.1 Traditional cell line development is inefficient 3

1.2.2 Selection markers as a component of mammalian expression constructs 4

1.2.3 The influence of integration locus on gene expression 6

1.2.4 Site-specific genetic editing techniques for mammalian hosts 8

1.3 Advancements in Cell Engineering Through Systems Biology 11

1.3.1 The interplay between recombinant DNA and systems biology 11

1.3.2 The role of codon optimization in host engineering 13

1.3.3 Methods for understanding multi-functional proteins 15

Chapter 2: Evaluating the Influence of Selection Markers on Obtaining Selected Pools and Stable Cell Lines using Human HT1080 Cells 18

2.1 Chapter Summary 18

2.2 Introduction 19

2.3 Results and Discussion 20

2.3.1 Zeocin selection results in pools with highest GFP fluorescence 20

2.3.2 Zeocin selection enables the generation of better stable populations 26

2.3.3 Zeocin selection aids in the identification of stable, enriched GFP expression at the clonal level 30

2.4 Concluding Remarks 34

Chapter 3: Identifying High Transcription Loci in the Human Genome 36

3.1 Chapter Summary 36

3.2 Introduction 36

3.3 Results and Discussion 38
3.3.1 Isolation of stable, high expression recombinant cell lines 38
3.3.2 Determination of high-expression integration loci in the human genome 44
3.3.3 Expression mapping reveals minimal disruption of endogenous gene expression 45
3.3.4 Site-specific targeting of Grik1 intron 5 demonstrates superior transgene expression 51
3.4 Concluding Remarks 55
Chapter 4: Exploring the Cre/*lox* System for Targeted Integration into the Human Genome 57
4.1 Chapter Summary 57
4.2 Introduction 58
4.3 Results and Discussion 60
4.3.1 Evaluation of mutant lox sites for improved swapping efficiency60
4.3.2 Sequential introduction of target DNA and Cre increases swapping efficiency 67
4.3.3 Modeling the delayed introduction of Cre DNA to improve swapping efficiency 71
4.3.4 Determining the optimal ratio of Cre and target DNA 72
4.4. Concluding Remarks 76
Chapter 5: A Method for Condition-Specific Codon Optimization for Improved Heterologous Gene Expression 77
5.1: Chapter Summary 77
5.2: Introduction 77
5.3: Results and Discussion 79
5.3.1: Developing a Condition-Specific Codon Usage Bias 79
5.3.2: Condition-specific optimization of eGFP for high expression outperforms wild-type and control variants 84
5.3.3: Stationary Phase Optimization of CatA 90
5.3.4: Organization of transcription factors suggests codon usage is linked to gene regulation 95

5.4: Concluding Remarks.....100

Chapter 6: Linking yeast Gcn5p catalytic function and gene regulation using a quantitative, graded dominant mutant approach.....102

6.1 Chapter Summary.....102

6.2 Introduction.....103

6.3 Results And Discussion.....104

6.3.1: Identifying Gcn5p dominant mutants.....104

6.3.2 *gcn5-F221A* competitively inhibits the catalytic function of Gcn5p in a dose-responsive manner.....107

6.3.3: Gcn5p graded dominant mutants competitively inhibit global histone acetylation at H3K18.....113

6.3.4: Combining expression profiling with a graded dominant mutant approach reveals novel Gcn5p targets and function.....115

6.3.5: Perturbation control experiments for *gcn5-F221A* mutant.....121

6.3.6: Chromatin DB analysis determines histone modifications using a graded dominant mutant approach.....124

6.3.7: Gene ontology analysis reveals new cellular processes impacted by Gcn5p acetylation.....125

6.3.8: Graded dominant mutant approach can interface with phenotypic and genetic assays.....125

6.4 Concluding Remarks.....129

Chapter 7: Conclusions and Major Findings.....131

Chapter 8: Proposals for Future Work.....135

Chapter 9: Materials and Methods.....140

9.1 Common Materials and Methods.....140

9.1.1 Conditions and media for human cell growth.....140

9.1.2 Conditions and media for microbial growth.....140

9.1.3 Transfection and selection of human cell populations.....141

9.1.4 Isolation of human cell clonal populations.....141

9.1.5 Flow cytometry for human cell populations.....141

9.1.6 Flow cytometry for yeast populations.....142

9.2 Materials and Methods for Chapter 2 143
9.2.1 Plasmid Construction 143
9.2.2 MIC_{75} Measurements 144
9.2.3 Cell Line Development 144
9.2.4 Real-Time PCR Measurements 145
9.3 Materials and Methods for Chapter 3 145
9.3.1 Plasmid Construction 145
9.3.2 Sterile FACS Sorting 146
9.3.3 Methods for identifying integration loci 146
9.3.4 Real-Time PCR 147
9.3.5 Site-Specific Retargeting 148
9.4 Materials and Methods for Chapter 4 149
9.4.1 Plasmid Construction 149
9.4.2 Cell Line Development 150
9.4.3 Copy Number Assay 151
9.4.4 Measuring Cre recombinase performance 152
9.4.5 Southern Blot 153
9.5 Materials and Methods for Chapter 5 155
9.5.1 Microarray Data Analysis 155
9.5.2: Plasmid construction 155
9.5.3: CatA Activity Assay 156
9.5.4: Muconic acid production 156
9.5.5: Matrix Drift Analysis 157
9.6 Materials and Methods for Chapter 6 158
9.6.1: Strain and Plasmid Construction 158
9.6.2: Growth Experiments 160
9.6.3: Yeast Immunofluorescence 161
9.6.4: Gene Expression Microarrays 162
9.6.5: Real Time PCR 163
9.6.6: Growth Analysis for Synthetic Lethal Genes 165

9.6.7: TEF Promoter Engineering.....165
Appendix A: Primers 168
Appendix B: Integration Loci Identification and Validation.....172
Appendix C: Codon Optimized Gene Variants 176
Appendix D: Python Scripts 182
CodonUsageBias.....182
GeneDesigner.....185
Map_Draw 189
Appendix E: Chromatin DB systems analysis of microarray data 192
Appendix F: Cytoscape BiNGO systems analysis of microarray data 194
References.....201

List of Tables

Table 2.1: MIC_{75} values for HT1080 and HEK293 21

Table 2.2: Zeocin clones result in no false positives 30

Table 3.1: High transcription loci are distributed throughout the genome 45

Table 5.1: Control codon usage table 86

Table 5.2: High expression codon usage table 88

Table 5.3: Drift calculations using the Frobenius Matrix Norm 98

Table 6.1: Impact of *gcn5-F221A* on the growth rate of *GCN5* synthetic lethal genes. 127

Table 6.2: Putative *GCN5*-Dependant growth inhibitors tested for an impact with mutant 129

Table 9.2: Reaction conditions for error-prone PCR 166

Table A.1: Primers from Chapter 2 168

Table A.2: Primers from Chapter 3 168

Table A.3: Primers from Chapter 4 170

Table A.4: Primers from Chapter 6 171

Table A.5: Microarray genes identified as over represented for cellular processes 197

List of Figures

Figure 2.1: Transgene constructs for comparison of four mammalian selection agents. 21

Figure 2.2: Zeocin selection enables a higher percentage of GFP expression and greater stability.....23

Figure 2.3: Zeocin enables higher GFP expression at the transcriptional level.....29

Figure 2.4: Zeocin selection identifies better candidate cell lines 32

Figure 2.5: Zeocin-resistant single cell clones exhibit high level, stable GFP expression 33

Figure 3.1: Dual-selection transgene constructs for high expression clones 39

Figure 3.2: Establishing recombinant HT1080 populations 41

Figure 3.3: Isolated single cell clones exhibit high protein and mRNA expression43

Figure 3.4: Expression maps for protein-coding sequences surrounding integration loci.....47

Figure 3.5: Targeted integration into the Grik1 loci results in elevated transgene expression 53

Figure 4.1: Mammalian expression vectors for dual fluorescent screen 61

Figure 4.2: A dual fluorescent screen to determine Cre-mediated swapping and excision 62

Figure 4.3: Without a swapping target, Cre recombinase results exclusively in excision. 65

Figure 4.4: Swapping and excision rates for three mutant *lox-loxP* pairings. 66

Figure 4.5: Delayed introduction of Cre DNA improves swapping efficiency. 69

Figure 4.6: Time-dependent swapping behavior can be mathematically modeled.72

Figure 4.7: Increased quantities of Cre DNA improves net recombination. 73
Figure 4.8: Increased transfection of Cre vector does not improve swapping compared to excision activity. 75
Figure 5.1: Condition-specific codon optimization utilizes systems level information and codon context 80
Figure 5.2: High expression codon optimization matrix 85
Figure 5.3: Optimization for a high expression condition results in eGFP expression exceeding the wild-type 90
Figure 5.4: Stationary phase codon optimization matrix 92
Figure 5.5: Optimization for stationary phase results in CatA variants that are improved at late growth 93
Figure 5.6: Drift of transcription-factor codon matrices reveals diverse codon usage relative to the control matrix.....96
Figure 6.1: *GCN5* complementation assay to determine potential dominant mutants. 105
Figure 6.2: Measuring native and mutant Gcn5p mRNA expression levels.....107
Figure 6.3: *gcn5-F221A* can impart a graded phenotype as detected by starvation assays 108
Figure 6.4: Evaluating competitive inhibition by *gcn5-F221A* using a synthetic construct.....111
Figure 6.5: Global acetylation at H3K18 is attenuated by expression of mutant *gcn5-F221A*.....114
Figure 6.6: Global H3K18 acetylation is attenuated by expression of mutant *gcn5-E173A*.....115

Figure 6.7: Expression analysis comparing a graded dominant mutant of *gcn5-F221A* to *gcn5Δ*. 117
Figure 6.8: Gene expression heat maps for select genes illustrating unique traits.119
Figure 6.9: RT-PCR of select graded genes confirms microarray findings.....122
Figure 6.10: Expression of *gcn5-F221A* does not influence the native *GCN5* promoter 123
Figure 6.11: Decreased growth rate caused by cycloheximide treatment of S288C is linked with Gcn5p acetylation activity. 128
Figure 9.1: Mathematical Definition of Frobenius Matrix Norm 157
Figure 9.2: Sequence of additional low strength TEF promoters 167

Chapter 1: Introduction and Background

1.1 Cell and Metabolic Engineering

Since the advent of recombinant DNA technology in the 1970s[1], cellular hosts have been used to produce a wide range of useful products including therapeutics, antibiotics, biofuels, specialty chemicals and other small molecules. Bio-based processes are often renewable, environmentally friendly and cost effective, making them an attractive alternative to chemical synthesis. As petroleum-based resources continue to dwindle and society's need for more complex and diverse products increases, the demand for biological processes to make these products will increase.

In order to meet the increasing demand for bio-based products, technology must be developed that enables us to manipulate and engineer a wide variety of cellular hosts. Model organisms such as *Escherichia coli* and *Saccharomyces cerevisiae* have been used to commercially produce many products because they are relatively easy to culture and molecular biology tools for these organisms are well established. Despite their success, these model organisms cannot be used for all bio-based processes. The characteristics of a product, such as toxicity and post-translational modifications, often dictate which cellular hosts can be used on an industrial scale. Although a vast diversity of organisms exist in nature, most of them are not culturable and no biotechnological tools have been developed for them. The development of new tools and approaches that enable controlled, robust manipulation of cellular hosts is a key challenge for cell and metabolic engineering.

1.2 Predicted Gene Expression in Mammalian Cellular Hosts

The advancement of biotechnological tools is particularly important in mammalian cells and other eukaryotic hosts because of their important roles in medicine, human health and the production of protein therapeutics. Mammalian cells remain the predominant host for producing antibodies and other protein therapeutics based on advantageous post-translational modifications[2-7], reduced immunogenicity, and the establishment of an infrastructure of mammalian cell cultivation and bioprocess engineering at pharmaceutical companies. Our capacity to culture and engineer mammalian cell systems for protein production has rapidly expanded in past decades and has raised the importance of mammalian bioprocess engineering efforts. This improvement is most apparent in the ever-increasing titers of monoclonal antibodies that have gone from 50 mg/L to upwards of 5 g/L in just over two decades[8]. Moreover, the production of complex proteins in mammalian hosts continues to be a successful approach and accounts for a good number of the over 9,700 clinical drug candidates in industry pipelines annually[9]. Collectively, these advancements have led to increases in the quality, quantity and complexity of recombinant products.

Despite these advantages and many improvements in cell engineering technology, mammalian cells lack many of the useful characteristics inherent in bacterial and simple eukaryotic hosts. Specifically, these cells are unable to autonomously replicate plasmid DNA. A long term, stable production cell therefore requires integration of heterologous DNA into the host cell genome and the subsequent isolation of a high expressing cell line. Furthermore, homologous recombination is very inefficient[8], making targeted integration challenging. This makes the cell line development process time consuming, labor intensive, unreliable and expensive and involves the screening of thousands of potential cell lines[8,10-12]. Advancements that eliminate this screening process would

provide significant cost and time savings, in addition to functioning as a useful genetic tool in human cell engineering[11].

Furthermore, issues of cell productivity, cell stability, cost of goods and services, and speed of development have put new demands on the field. In general, the cost of bringing a drug to the market is quite high[13] as a result of significant R&D, clinical testing, and failure rates. While improved cell engineering tools cannot solve clinical testing and failure rates, they can improve the speed of R&D as well as reduce cost of goods. To this end, there is a need to develop novel tools that facilitate predicted gene expression in recombinant mammalian cell lines.

1.2.1 Traditional cell line development is inefficient

The vast majority of protein therapeutics are produced in recombinant mammalian cell hosts including Chinese hamster ovary (CHO)[7,14-16], mouse myeloma (NS0)[15,16], human embryo kidney (HEK293) cells[14,17] and human sarcoma (HT1080) cells[17-19]. While advancements in mammalian cell technology have resulted in improved titers, quality and techniques, the process of developing a cell line with sufficient production capability remains both time and labor intensive[5,8,11,20]. In a typical cell line development program, a transgenic construct containing both the transgene of interest and a selection marker is introduced, usually by illegitimate (random) integration[21-23], into the host cell genome. While it is well-known that integration locus can strongly influence gene expression[3,8,10,24,25], most cell line development programs still rely on illegitimate integration followed by selection and amplification over site-specific recombination approaches. After integration, a selection agent is used to kill off a significant portion of the cell population with either no or low expression of the transgene. The success of this selection can be enhanced by using an IRES element[26,27] to transcriptionally link the

selection marker and the transgenic protein genes. In some instances, the selection pressure is continuously increased to enable cells to amplify transgene copy number[5] and likewise protein expression.

Regardless of the approach used, the pools of cells that survive selection are subsequently diluted to single cell lines and expanded to produce homogenous, clonal populations, as many as 10,000 at once[5,23,28,29], which are then subjected to a series of selection strategies. This typically involves multiple assessment stages where clones are evaluated at each stage for characteristics including growth, production, and stability characteristics[8] and low performers are removed from the data set. This methodology assumes the final, desirable cell lines will perform well for each assessment, despite variations in culture conditions and feed strategies.

This cell line development process is time consuming, with a fast-tracked scenario, under ideal conditions, estimating at least 70 days between transfection and identification of a candidate cell line[11]. Additionally, screening of thousands of clones for a variety of characteristics is laborious and costly. Furthermore, there are instances where selection strategy can falsely remove top performers and promote low performers[12]. There is clear motivation to streamline and improve the efficiency and success rate of the cell line development process[11,20,30]. The development of tools that enable controlled and predictable recombinant DNA expression in mammalian cell systems will do much to advance cell line development.

1.2.2 Selection markers as a component of mammalian expression constructs

Regulating gene expression in mammalian hosts involves not only the gene of interest, but many other genetic elements that collectively enable and enhance synthetic gene circuits. Some of these elements previously used in mammalian expression

constructs include enhancers, promoters, internal ribosome entry sites (IRES), ubiquitous chromatin opening elements (UCOEs), scaffold/matrix attachment regions (S/MARs), micro RNAs (miRNAs), and selection markers, which can be combined with both recombinant and native genes to enhance genetic engineering efforts in these cells[31]. The utility of such genetic elements has been clearly demonstrated in microbes[32-36]; however, adaptation and adoption across mammalian genomes is in its early stages[37].

One critical component of most mammalian expression constructs is a selection marker. This genetic element enables cells expressing the transgenic construct at sufficient levels to survive exposure to an otherwise toxic agent during a selection phase. The effectiveness of this selection phase directly influences the quality of the selected pool and the resulting single cell clones, making it a critical component of cell line development.

Although several selection markers are available for mammalian cells, the most commonly used is DHFR in which cell line production typically involves identification of increased copy number loci using methotrexate (MTX) selection[3,8,10,38]. In DHFR screening, cells are exposed to progressively higher concentrations of MTX, which inhibits folic acid metabolism. The DHFR gene and the gene of interest are co-transfected, thus the surviving population is able to over express DHFR, which usually indicates higher expression of the gene of interest. This system is commonly used to identify cell line candidates.

Despite its popularity, the DFR/MTX system, and other auxotrophic systems like glutamine synthetase (GS), has some drawbacks. MTX selection favors those cell lines with multiple copies, even hundreds of copies, of the DHFR gene[10,25]. Once selective pressure is removed, these cell lines are typically unstable and lose many or most copies of the heterologous DNA[8,10,39]. The cytogenetic effects of MTX, as well as additional

cost, prohibits continued use of the selective drug[8] in growth media. Additionally, the DHFR system does not work in all cell lines and can even result in decreased productivity of the desired gene[25]. This is likely a result of increased expression of DHFR while the desired gene is either lost or unexpressed. Finally, although deletion strains are available for CHO, these diploid deletion cell lines have not been established for most other cellular hosts, making it ineffective to use an auxotrophic selection system.

Fortunately, a variety of eukaryotic antibiotic selection markers and corresponding resistance genes are available and can be used in mammalian expression. Some common antibiotics include blasticidin, geneticin, hygromycin B, mycophenolic acid, neomycin, puromycin, and Zeocin. In many cases of cell line development, these agents and corresponding markers are used interchangeably despite the wide variations in stringency that are known[8,10,25,40]. Furthermore, these antibiotics act through several different modes of action resulting in both rapid and slow cell death. Because of these factors, and the near ubiquitous presence of selection markers in transgenic constructs, it is probable that selection marker choice plays an important role in mammalian cell line development. In other model eukaryotes including plants[41-43], yeasts[44,45], and insects[46], studies comparing selection marker performance have been conducted. Analogous studies of selection markers in mammalian cells would likely contribute directly to predicted gene expression efforts.

1.2.3 The influence of integration locus on gene expression

Stable, long term expression of heterologous DNA requires integration into the host cell genome because mammalian cells lack the ability to autonomously replicate plasmids. Like many other chromosome-related phenomena[39,47-54], recombinant gene expression has been shown to be heavily dependent on the site of genomic

integration[3,8,10,12,25]. The non-random distribution of such events across the genome indicates that a novel structural environment or specific genetic elements must be present[55]. The biases of retroviral integration sites in the genome further illustrate that not all loci are accessible or capable of sustained, stable transcription. Retroviral integration occurs at a non-random frequency and exhibits an integration bias with defined motifs and preference for CpG islands, regions of high gene density, and regions near transcription start sites and transcription factor binding sites[51,56-59]. As an example, HIV has been shown to preferentially integrate into actively transcribed genes in an attempt to identify transcriptional hot spots[52,60].

While non-viral integration appears to be a distributed, random process aided by native recombination mechanisms[61,62], not all genomic regions are conducive to expression. Integration into euchromatin, lightly packed gene rich regions, is most likely to favor expression[10]. Due to the proximity to essential genes, these regions are often actively recruiting transcription machinery, which the integrated transgene can take advantage of. Alternatively, integration into heterochromatin is unlikely to confer transcription capacity, as these regions are often silenced by histone deacetylation, histone methylation and promoter methylation[10].

Stable integration of recombinant DNA is commonly used for transgenic studies, stable cell line development, and gene therapy applications. Despite the widely accepted importance of integration locus on recombinant DNA expression[3,8,10,12,25], and its utility in advancing predicted gene expression efforts, limited information is available about desirable genomic integration sites. Attempts to determine the exact genetic location of transgene insertions have only been performed in isolated cases for a particular germline or cell line with interesting characteristics[63,64], while for most high expressing cell lines, no effort is made. Alternatively, pre-determined criteria, such as 'Good Safe Harbours'[65]

have been applied *a priori* to identify potentially useful integration sites before targeting them and measuring expression[66]. Although these guidelines have utility in gene therapy applications, such an approach is inherently biased and does not sample any significant portion of the genome. No global study has been conducted in mammalian cells to more comprehensively identify loci capable of supporting high expression of heterologous DNA. The closest related dataset that does exist is global expression analysis using microarray data. However, this dataset is limited in its utility because it exclusively encompasses protein coding sequences and lacks information about non-coding regions which can confer high expression.

Studies comparing genomic integration sites have been conducted in other model organisms including *E. coli*[67], *S. cerevisae*[68] and zebrafish[69]. A recent study in yeast examined 20 genomic integration sites and found more than an 8 fold difference in expression levels for these sites[68]. A similar or even greater expression range could exist in mammalian genomes and identifying loci suitable for integration is an important step in developing better tools for stable recombinant cell lines, especially if these sites can be retargeted[70]. By identifying and mapping transcriptionally active areas, existing and future technologies can be used to deliver a gene of interest to those loci.

1.2.4 Site-specific genetic editing techniques for mammalian hosts

Genetic engineering is required to transform mammalian host cells into super-producers of proteins. Specifically, efficient mammalian cell engineering requires precise, site-specific genome editing techniques to enable the expression of heterologous genes and deletion of unwanted genes at known loci. In most microbial systems, this is efficiently achieved via homologous recombination. Unfortunately, homologous recombination is a rare and inefficient process in mammalian cells[8]. Alternative

approaches have been and continue to be developed which typically rely on targeted double strand breaks that trigger DNA repair mechanisms. As part of the repair process, non-homologous end joining can occur, which has a probability of resulting in both loss of nucleotides (deletion events) and incorporation of DNA constructs (integration events). There are a variety of enzymes naturally capable of performing these double strand breaks and many can be modified to increase or introduce site specificity.

Cre and FLP recombinase and ΦC31 integrase have long been utilized in this capacity[29,71,72] and continue to be a popular area of innovation[73-76]. Each of these enzymes targets genomic regions based on recognition of specific sequences, which reduces their flexibility. Cre recombinase was prolifically utilized in early mouse recombineering efforts[77], resulting in the development of hundreds of cell lines with the *loxP* targeting sequence integrated at specific sites. However, extending this effort to all mammalian hosts is impractical. There are several hundred sites that are naturally recognized by ΦC31 integrase and this can lead to undesired heterogeneous integrations[78-80]. While useful in some applications, this characteristic makes ΦC31 integrase unreliable for site-specific integration at a single, unique locus. Endonucleases are another class of enzymes that can be engineered to recognize specific genomic sequences and perform cleavages[81].

Zinc finger nucleases (ZFNs) do not require generic targeting sequences and are modular in assembly, allowing greater flexibility in their targeting[6]. ZFNs facilitate both genomic integrations and gene knockouts[6]. Custom ZFNs can be ordered through companies such as Sangamo BioSciences and have been demonstrated in a variety of cell types and applications[82], including the rapid and efficient deletion of genes[83,84]. Recently, zinc-finger recombinases (ZFR) were developed by fusing zinc finger domains and serine recombinases, and utilized in human cells to deliver reporter genes at specific

loci[22]. Although this method requires pre-insertion of ZFR recognition sites, DNA damage responses are circumvented and thus higher levels of specificity are achieved[22]. Transcription activator-like effector nucleases (TALENs) are also modular in nature and can be built to recognize any DNA sequence[85,86]. Efficient endogenous deletions[87] and gene insertions[88] were recently demonstrated in human cells using TALEN architecture. The type II bacterial CRISPR system has also been engineered to function in conjunction with custom RNA sequences to create targeted double strand breaks in human and mouse cells[89,90]. This approach is more easily adapted than ZFNs or TALENs and targeting rates have been demonstrated to be similar or even better[89].

Artificial chromosomes present an alternative technology that does not require integration into the host genome. This technology can support large quantities of recombinant DNA and has been demonstrated to generate monoclonal antibody expressing CHO cells exhibiting high productivity[28]. Human artificial chromosomes (HACs), which act as a small, 47th chromosome, can be used to introduce up to 10 megabases of foreign DNA into host cells[91,92]. HACs, however, are only mitotically stable for six months and could be subjected to regulation and silencing much sooner. The recent discovery of small, circular microDNAs in mammalian tissues represents another genetic avenue that could be engineered to complement and extend existing transgene expression technologies[93].

These tools collectively provide flexibility and precision in genome editing and represent significant improvements over standard practices such as homologous recombination or illegitimate integration. Further innovation and discovery in this area will prove to be valuable in the larger goal of predictable mammalian gene expression.

1.3 Advancements in Cell Engineering Through Systems Biology

The advent of high-throughput biology has led to a rapid acceleration in the ability to obtain systems-level information about living organisms including genomes, transcriptomes, proteomes, metabolomes, epigenetic states, and transcription factor binding profiles. This global information, integrated with computational approaches for analysis and model-based prediction, has led to an enormous and transformative understanding of biomolecular networks in a field termed 'systems biology'[94]. Combining these global measurements has led to robust, high resolution information about cellular responses and metabolism[95]. Recent advances allow dissection of apoptotic signals[96], elucidation of mechanisms of complex phenotypes[97], construction of predictive, genome-scale metabolic models[98,99], the modeling of complex microbiomes[100], the cataloguing of complexity through metabolomics[101], and the study of complex signal cascades[102,103]. Collectively, these advancements are driven by the need to understand biological systems at the quantitative level, and exemplify the breadth of knowledge enabled by systems biology. These examples serve to illustrate the power of a top-down approach to understand cellular function and underlying design principles. Furthermore, many of these techniques are generic tools and can easily be applied to other cellular hosts for which basic gene expression methodologies are available.

1.3.1 The interplay between recombinant DNA and systems biology

Successful implementation of systems biology across a variety of biological techniques and cellular hosts requires recombinant DNA technology, or the assembly and expression of well-characterized, heterologous genetic parts. Despite being relatively young disciplines, important and transformative contributions to biotechnology have come from both systems biology and recombinant DNA technology. However, a new age of understanding and advancement lies at the intersection of these approaches, as

demonstrated by the recent dramatic increase in the number of studies utilizing recombinant DNA techniques in conjunction with systems biology[104-111]. The precision resulting from this synergy eliminates much of the uncertainty and failure associated with biological design and allows for more meaningful conclusions to be drawn from experimental studies. There is much to be gained from the development of models which capture ever-increasing biological complexity. Moreover, the ability to precisely perturb these systems will enable further high-resolution insight into biological networks. Such advances hold great promise for deciphering the mechanisms underlying development and disease[112,113] and will enable further development of versatile, robust, and useful organisms for industrial biotechnology. While sufficient tools are not currently available to dissect many of these phenomena in higher eukaryotes like human cells, lower eukaryotes can be used to develop appropriate and adaptable methods.

The merger of recombinant DNA technology and systems biology will provide unique opportunities to both study and build cellular function. Specifically, synthetic perturbations can enable higher resolution systems analysis. In this regard, modifying genes via precise, recombinant DNA-enabled control may yield higher-resolution data compared with standard, traditional coarse genetic approaches such as gene deletions. Higher resolution datasets including causative linkages, cellular localization, and controlled regulation are uniquely enabled at this intersection. This information will certainly upgrade the quality of genome scale modeling efforts. Likewise, cataloging cellular interactions can help predict failures of designed synthetic circuits by predicting (and avoiding) component cross-talk and interference. Understanding the global interactions and overlap between cellular components can facilitate better design of synthetic circuits that can be isolated from endogenous cellular control. By borrowing advances in computational tools to model cellular systems, it will eventually be possible

to model and predict synthetic circuit design. Collectively, these prospective advances help mitigate the great complexity and uncertainty currently impeding the study and design of cellular systems[114]. Along this vein, there exist many recombinant DNA techniques that can be augmented and improved by utilizing widely available -omics data and a systems biology approach.

1.3.2 The role of codon optimization in host engineering

One common recombinant DNA technique which could be improved through a systems biology approach is codon optimization. Codon optimization refers to the rational redistribution of synonymous codon usage for improved expression of a protein. Proteins, which perform nearly all cellular functions, are built from just twenty amino acids, designated by a corresponding three base pair codon. Although there are only twenty unique amino acids, there are sixty-one codons, which results in synonymous codon usage for eighteen of the twenty amino acids (methionine and tryptophan being the exceptions). Interestingly, the distribution of synonymous codons for a given amino acid is not uniform, resulting in both rare and abundant codons. Furthermore, the preference for particular codons throughout the genome varies across all organisms[115-117]. This species-specific deviation is defined as codon usage bias (CUB)[118] and is known to influence translational efficiency[119]. CUB can be uniquely determined for a given organism's genome or subset of genes, and used for codon optimization of heterologous genes[116-118]. The traditional approach to codon optimization involves the removal of rare codons and replacement with more abundant codons, which often increases non-native gene expression[116]. To this end, codon optimization has emerged as a popular synthetic biology tool to improve heterologous gene expression across a variety of host organisms, and has applications in metabolic and cellular engineering[116,120,121].

Codon optimization of heterologous genes requires a methodology for identifying a CUB. For most organisms, a CUB is determined using the Codon Usage Tabulated from GenBank (CUTG)[122]. CUTG determines a CUB using all of the annotated protein coding genes of the host organism. This process is easily carried out for any fully sequenced organism and is the primary source of codon bias information being used to optimize heterologous genes. Currently, CUTG has CUB information for 35,799 organisms. This process can be done commercially by companies including Blue Heron, GeneArt and DNA 2.0. However, although codon optimization is generically applied to a variety of cellular hosts, it fails to take into account any systems level information previously collected for the host strain or growth conditions.

There are alternative approaches for determining CUB, including codon adaptation index (CAI)[123,124], codon bias index (CBI) and the effective number of codons (N_c)[125]. From these approaches, online optimization programs have been developed and are freely available[126-128]. Much of the research involving these alternative approaches is focused around endogenous gene expression, which greatly limits the utility of the work. Furthermore, these alternative approaches have predominantly been explored in prokaryotic hosts and there are very few examples of their use in optimizing heterologous gene expression.

While a CUTG approach to determining CUB often results in improved gene expression and translational efficiency, this is not always the case. There are many instances where traditional codon optimization does not lead to improved expression compared to a wild-type, unmodified sequence[129-132]. In fact, in a survey of 44 synthetic genes manufactured by Blue Heron, 32% of the "optimized" synthetic genes expressed at lower levels than the wild-type, indicating that their algorithms are not ideal for all conditions and hosts[130]. One possible explanation for this could be that traditional codon

optimization neglects to take into account variation in charged tRNA abundance known to result from changes in environmental factors including growth condition and cell-cycle[133-135]. Furthermore, despite the fact that much of an organism's protein coding genes are lowly expressed and minimal evolutionary pressure has been present to drive efficient natural evolution[118,136], the CUTG approach assumes that using all of a genome's protein coding information, as opposed to a subset, provides the best information for codon optimization. These factors suggest that applying a systems biology perspective to CUB identification, and subsequently codon optimization, could result in a better, more robust methodology for codon optimization.

1.3.3 Methods for understanding multi-functional proteins

Another area that would greatly benefit from an approach combining systems biology and recombinant DNA technology is the study of multi-functional proteins. These proteins often act globally within the cell, impacting hundreds of targets. Establishing high-resolution, causative mapping of all protein functions, interactions, and cell responses is a critical facet underlying success in genetics, drug discovery, and molecular biotechnology[137]. Some of this work has been done, utilizing methodologies including chromatin immunoprecipitation (ChIP)[138], ChIP sequencing[139] and yeast two-hybridization[140]. Although these approaches can provide useful information, they are *in vitro* techniques and therefore the conditions used are not very representative of a normal cell environment. Many of the proteins we wish to study are multifunctional and contain diverse functionalities including protein and DNA interactions and catalytic activity. These proteins are often non-essential and act globally in critical roles including epigenetic modification, signaling cascades, and transcriptional regulation.

The typical approach to studying these proteins *in vivo* is gene deletion followed by characterization of phenotypic changes. This approach has been demonstrated to cause pleiotropic effects that can often lead to misinformation. This is particularly problematic for multifunctional proteins, where it is difficult to directly link one particular function of these proteins (such as catalytic activity or a specific protein-protein interaction) to downstream gene regulation, as knockouts remove the entire protein and thus all of its functions, thereby creating an environment for non-natural associations or activity compensations that confound data analysis. In this regard, gene knockout studies probe cellular response and compensation, not necessarily precise protein function. Ultimately, this results in the collection of inaccurate information that is incorporated into metabolic and systems models.

Some alternative strategies to gene deletion are available[141-146]. The use of molecular analogues to competitively inhibit biological function is a commonplace technique that allows for controllable or graded activity. Epigenetic inhibitors include nucleotide analogues to inhibit methyltransferase activity[147,148] and other small molecules such as valproic acid[149], butyrates, and hydroxamic acids[150,151] to inhibit methylation and acetylation. However, these small molecules are difficult to design *de novo*, lack single target specificity, are limited in their concentration ranges, and often have a lower than anticipated response rate[151]. In some respect, dominant mutations (a mutant allele that acts competitively with the wild-type allele) act as specific inhibitors. Dominant mutations are used in classic genetics to assess gene function, improve tolerances and drug resistances[152-154] and characterize disease states[155-158].

The development of new synthetic approaches that enable the dissection of individual protein functionalities would be useful. It would enable the discovery of new gene targets for proteins of interest, as well as the determination of causative linkages

between protein-DNA interactions. Such a methodology would be implementable and useful in a variety of protein classes and eukaryotic hosts and would contribute directly to an expansion of systems biology information.

Novel tools are needed for a variety of cellular hosts. The development of these tools enables the production of useful compounds from bio-based processes. Furthermore, it facilitates our understanding of cellular interactions and how these interactions can be controlled to enable complex phenotypes and regulate disease states. With these goals in mind, I have developed novel tools for engineering eukaryotic cells using a systems level approach. Many of these tools are specific to mammalian cells and enable more precise, controlled gene expression in this host. Other tools were developed in *S. cerevisiae*, because we currently lack the technology necessary to robustly explore complex phenotypes in many higher eukaryotes. These tools focus on the development of complex phenotypes using systems level information. Collectively, the tools and approaches described herein utilize a systems level approach, improve our ability to flexibly engineer eukaryotic cellular hosts, and have many biotechnology and health applications.

Chapter 2: Evaluating the Influence of Selection Markers on Obtaining Selected Pools and Stable Cell Lines using Human HT1080 Cells

2.1 Chapter Summary

Stable and constitutive gene expression in mammalian cells can be controlled and influenced by a variety of genetic elements. However, the impact of these genetic elements has not been fully characterized. One such genetic element used to isolate recombinant populations is a selection marker. Selection markers are a nearly ubiquitous element in mammalian expression cassettes. While several selection systems exist for use in mammalian cell lines, no previous study has comprehensively evaluated their performance in the isolation of recombinant populations and cell lines. Here we examine four antibiotics, hygromycin, neomycin, puromycin, and Zeocin, and their corresponding selector genes, using a green fluorescent protein (GFP) as a reporter in two model cell lines, HT1080 and HEK293. We identify Zeocin as the best selection agent for cell line development in human cells. In comparison to the other selection systems, Zeocin is able to identify populations with higher fluorescence levels, which in turn leads to the isolation of better clonal populations and less false positives. Further, Zeocin-resistant populations exhibit better transgene stability in the absence of selection pressure compared to other selection agents. All isolated Zeocin-resistant clones, regardless of cell type, exhibited GFP expression. By comparison, only 79% of hygromycin-resistant, 47% of neomycin-resistant and 14% of puromycin-resistant clones expressed GFP. Based on these results, we would rank Zeocin > hygromycin ~ puromycin > neomycin for cell line development in human cells.

2.2 INTRODUCTION

Selection markers are known to be a critical part of expression vector design[159,160]. Several selection markers (and corresponding agents) are widely available for mammalian cells; however, their efficacy has not been compared in a single study. Previous studies, each acknowledging their variations in stringency and stability for different selection agents[8,10,25,40], support the need for a direct, unbiased evaluation of selection systems. We chose to examine the performance of four common antibiotics (Zeocin, hygromycin, neomycin and puromycin)[10,20] that can be utilized in nearly all recombinant mammalian cell lines, as resistance is only imparted by a single gene.

The modes of action of these commonly used antibiotics are quite different. Zeocin, a bleomycin analogue, is a small molecule that binds and cleaves DNA. The resistance gene encodes a protein that binds the antibiotic to prevent it from acting on DNA. False positives were shown to be rare but previous studies have suggested that Zeocin is not fully detoxified and chronic exposure during prolonged selection could cause mutagenesis and adaptive responses in clonal selection[161]. Hygromycin is an aminoglycoside that kills eukaryotic cells by binding ribosomal components and inhibiting translation[162]. The resistance gene, which encodes for a kinase that inactivates hygromycin via phosphorylation, was first demonstrated for cultured mammalian cells in 1985 with rare false positive rates[162]. Neomycin is also an aminoglycoside and the resistance gene has similar kinase activity[163]. Although widely used, neomycin has been shown to induce changes in global gene expression[163]. Puromycin is an aminonucleoside that disrupts translation. Part of the molecule resembles a charged tRNA and is able to bind a growing polypeptide chain and prematurely terminate it[164,165]. This mechanism is not specific to eukaryotes. Puromycin resistance is conferred by the puromycin N-acetyl

transferase (*pac*) gene natively found in *Streptomyces alboniger*. Resistance was first demonstrated in mammalian cells in 1986 using VERO cells[166]. Puromycin has also been linked to a breakdown of the polysomes[165].

No prior study has tested the efficacy of all of these selection agents in the same context and cell lines for a mammalian protein production host. Specifically, we were interested in determining the influence of the selection marker on the quality of selected populations, the stability of those populations, and the quality of clonal populations. We chose to focus on two human cell systems, HEK293 and HT1080, which have important industrial applications. Although this study was limited to human cell lines, such an evaluation could be extended to other cell lines including CHO.

2.3 Results and Discussion

In this study, we evaluate the efficacy of hygromycin, neomycin, puromycin and Zeocin in the human cell lines, HT1080 and HEK293 on the basis of selected pool quality, stability, and success-rate of isolating single cell lines. To establish the same genetic context for this comparison, we created four mammalian expression cassettes that were identical in sequence with the exception of the resistance gene (Figure 2.1). These constructs each contain the constitutive immediate-early enhancer CMV promoter, followed by the human optimized hrGFP fluorescent reporter gene, and the selection marker gene under study. The wild-type EMCV- IRES was placed between the two genes to link their transcription levels.

2.3.1 Zeocin selection results in pools with highest GFP fluorescence

These selection markers were first tested on the basis of selected pool quality. To assess this facet, healthy wild-type HT1080 and HEK293 cells were transfected in

batches with one of the four linearized plasmids shown in Figure 2.1. Three days after transfection, cells were treated with the respective selection agent at a concentration corresponding to the MIC_{75} level for non-transfected, parental cells as determined through experimental trials outlined in Materials and Methods. These values are shown in Table 2.1.

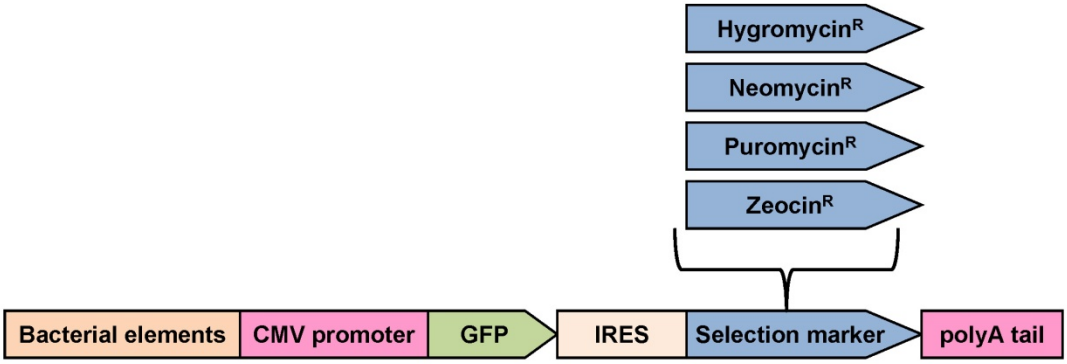
Table 2.1: MIC_{75} values for HT1080 and HEK293

Selection Marker	**HT1080 (µg/mL)**	**HEK293 (µg/mL)**
Hygromycin	50	85
Neomycin	50	170
Puromycin	0.7	1.0
Zeocin	70	60

MIC_{75} values were determined for two human cell lines, HT1080 and HEK293, and four antibiotics: hygromycin, neomycin, puromycin and Zeocin. These concentrations were used for stable library selection, with the exception of neomycin when applied to HEK293 cells. This library was unable to stabilize after more than 40 days at the MIC_{75} value and the concentration was eventually lowered to 42.75 µg/mL.

In order to establish a selected population, each pool was passaged routinely under selective pressure until cell viability rose above 90% (roughly lasting until 25 days post transfection). These selected populations, established in duplicate, were profiled for fluorescent hrGFP protein expression using flow cytometry. The percentage of the population expressing GFP was determined by comparing fluorescence profiles to that of a wild-type control population (auto fluorescence).

Figure 2.1: Transgene constructs for comparison of four mammalian selection agent

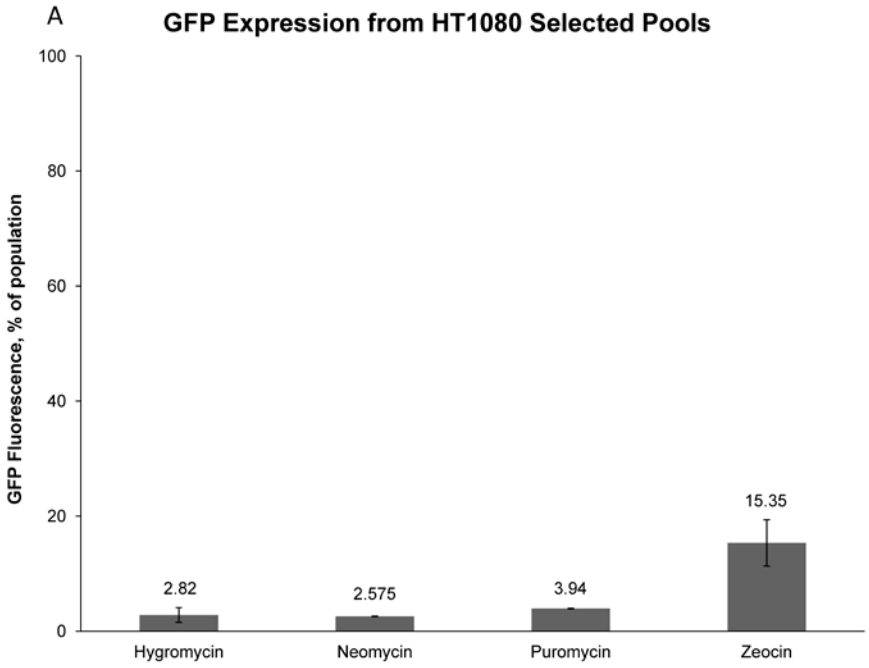


Four transgene DNA cassettes were designed for expression in the human cell lines, HT0180 and HEK293. Expression of the GFP and selector genes was driven by the CMV promoter and a single cistron was enabled by an IRES site. A separate cassette was made for each of the hygromycin, neomycin, puromycin and Zeocin resistance genes. Bacterial elements, including the F1 origin of replication and an ampicillin marker, allowed for plasmid maintenance in bacteria

At this early time point in the cell line development process, we find that hrGFP fluorescence was not detected in the vast majority of the resistant populations, except the HEK293 Zeocin selected population. It is important to note that these values are in stark contrast to the values seen shortly after the transfection where the percent of GFP positive cells can be upwards of 70-85% with our transfection conditions. Thus, despite the selection used, a significant fraction of the population had a silenced transgene when it became integrated into the genome. Nevertheless, the Zeocin selected libraries had the best fluorescence profiles, with 15.35% of the HT1080 population and 89.8% of the HEK293 population exhibiting GFP fluorescence above control values. This fraction was significantly higher than any of the other selection systems (p-value = 0.029 and 0.007 for HT1080 and HEK293 respectively). In HT1080, puromycin was the second best antibiotic followed by hygromycin with no statistical difference between the two (Figure 2.2a & c). In HEK293, hygromycin performed slightly better than puromycin but there was no real statistical difference between the two (Figure 2.2a & c). Regardless of cell

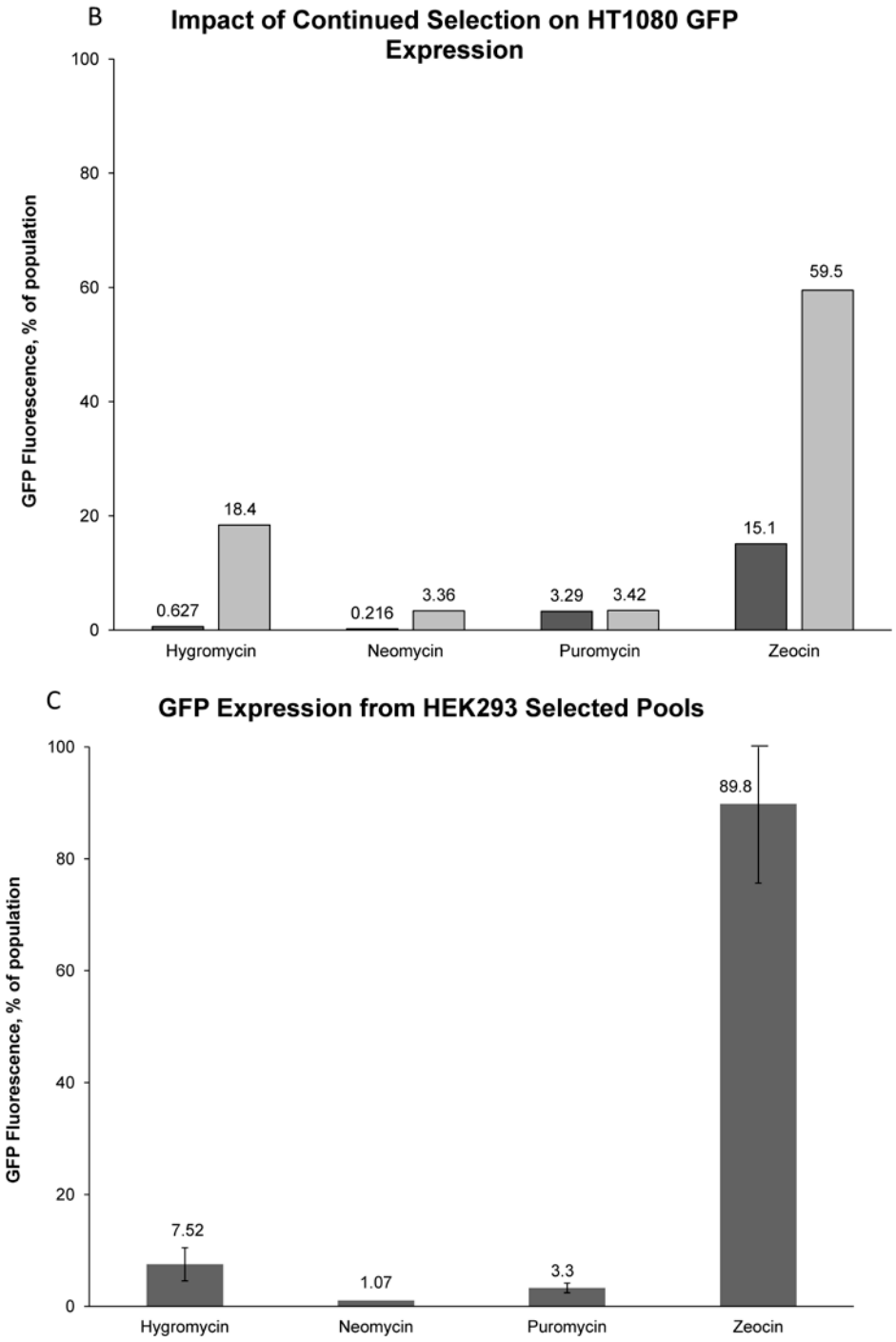
types, neomycin selection resulted in the lowest percentage of the population expressing GFP. In HEK293, neomycin significantly underperformed both hygromycin and puromycin (p-values = 0.046 and 0.034 respectively).

Figure 2.2: Zeocin selection enables a higher percentage of GFP expression and greater stability.



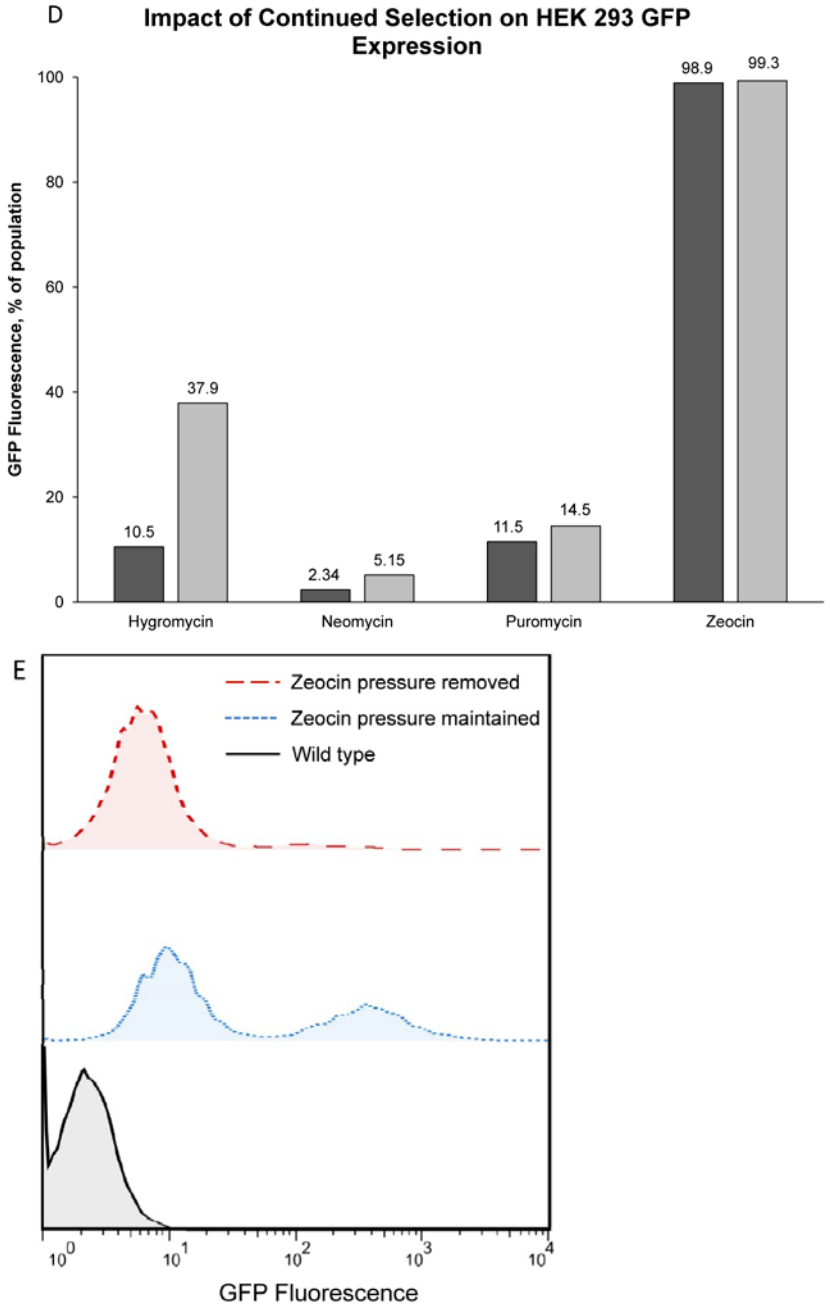
Four stable pools were established in duplicate in both cell lines after transfecting DNA constructs and treatment with the corresponding selection agent. **A.** In HT1080, the GFP expression levels, as a percentage of the total population, were measured using flow cytometry.

Figure 2.2 (continued)



B. GFP expression was measured for the HT1080 populations with prolonged selection pressure (white) and without prolonged selection pressure (grey) using flow cytometry. **C.** In HEK293, the GFP expression levels, as a percentage of the total population, were measured using flow cytometry.

Figure 2.2 (continued)



D. GFP expression was measured for the HEK293 populations with prolonged selection pressure (white) and without prolonged selection pressure (grey) using flow cytometry. **E.** GFP expression profiles of the Zeocin population without prolonged selection (red dashed) and with prolonged selection (blue dotted) were compared to a wild-type control population (black). The standard deviations (±) of duplicate trials are indicated by error bars in A and C. A single trial was conducted for the stability analysis (B and D), thus error bars are not presented.

Based on these results, it is evident that Zeocin selection identifies resistant cell pools with a greater percentage of the population expressing GFP, as compared to the three other selective markers. These results strongly support the use of Zeocin as the selective agent of choice, as it is likely these improved pools will result in better single cell clones (described in later sections).

2.3.2 Zeocin selection enables the generation of better stable populations

While the level of enrichment in selected population is an important first aspect in cell line development, it is perhaps superseded by the quality of stable, long-term expression of transgenes. Frequently, excision or silencing of the integrated transgene can occur over time[8], especially when selection pressure is relieved, which results in undesirable, decreased expression of a gene of interest. While previous reports suggest that some selection agents contribute to a wide variation in stable expression[5,8,10,14,25], no previous study has compared these four selection markers simultaneously in a single cell line. Therefore, we sought to evaluate the stability of GFP expression within the selected pools identified above. To do so, each of these pools were split such that one half of the pool was cultured under continued selection pressure while the other half was maintained without selection over the course of one month (roughly 60 generations). After the month of growth, fluorescent profiles (as determined relative to the auto fluorescence signal of a wild-type population) were once again analyzed and compared using flow cytometry.

By comparing the populations subcultured over the month with and without antibiotic pressure (either hygromycin, neomycin, puromycin or Zeocin), we observed that continued selection generally results in a higher percentage of the population expressing GFP, as shown in Figures 2.2b and d**,** compared with Figures 2.2a and c. This

result is expected since continued selection is known to delay or inhibit excision and silencing of transgenic DNA[10] and potentially results in gene amplification events. This trend holds for all of the HT1080 populations except puromycin-selected, as shown in Figures 2.2a and b. This effect is particularly prominent for the HT1080 Zeocin selected library. An additional one month of antibiotic pressure increased the percent of the population expressing GFP to 59.5% from an original value of 15.35% (Figure 2.2b). Hygromycin saw a similar upwards shift from 2.82 to 18.4% and the neomycin population shifted from 2.58 to 3.36% GFP expressing. A similar trend was observed in the case of the HEK293 selected populations, as shown in Figures 2.2c and d**.** The largest observed shift was in the hygromycin resistant population, which shifted from 7.52% to 37.9% GFP expressing after one month of prolonged selection. The puromycin resistant population shifted upwards from 7.9% to 14.5% GFP expressing, neomycin resistant population from 1.07 to 5.14% GFP expressing and Zeocin resistant population from 89.8% to 99.3% GFP expressing.

Furthermore, in the case of the HT1080 Zeocin population, prolonged selection resulted not only in an increase in the percent of positive cells with GFP, but also in an enhancement of absolute GFP expression level. The GFP expression profiles of the Zeocin-resistant populations (both with and without prolonged selection) are shown in Figure 2.2e. These histograms show a bimodal distribution of GFP expression for both conditions; a portion of the Zeocin-resistant populations have low GFP expression and a second portion of the population exhibits distinctly higher GFP expression. This finding suggests prolonged selection may result in a further enrichment of gene expression. While this can easily be monitored using a reporter protein like GFP, most proteins of interest cannot be easily measured. Therefore, prolonged selection could be advantageous prior to single cell cloning.

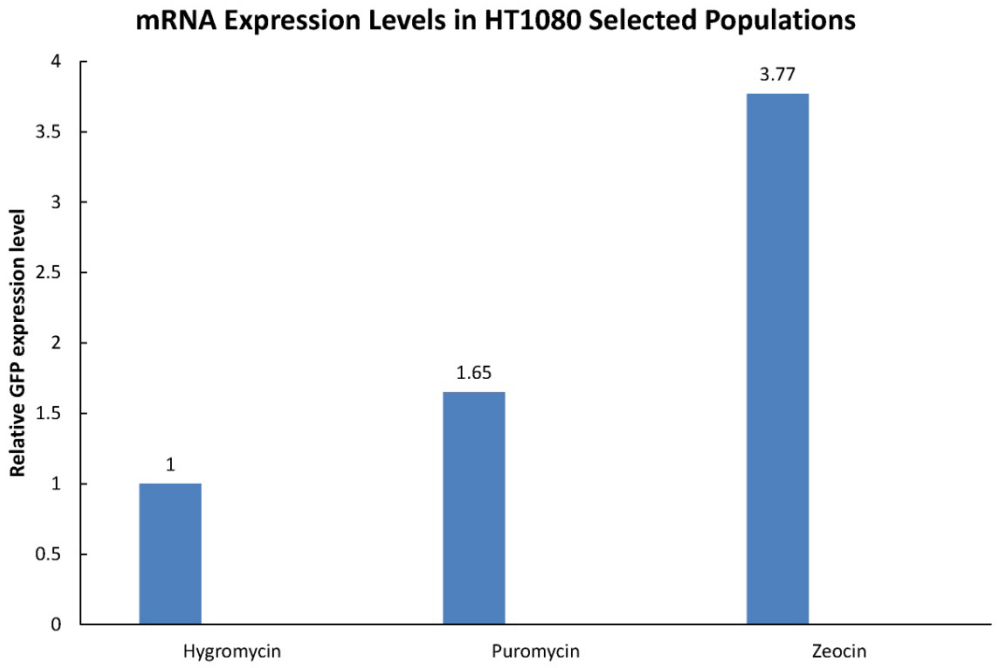
In HT1080 cells, for the majority of the selection systems, removal of the selection pressure over the course of the month resulted in a decrease in the percent of the population expressing GFP. In the case of the hygromycin pool, the percentage of the population expressing GFP decreased from 2.82 to 0.63%. Likewise, for neomycin the percentage of the population expressing GFP decreased from 2.58 to 0.22% (comparing Figure 2.2a to Figure 2.2b). In contrast, the Zeocin and puromycin selected population appeared to be extremely stable with very little loss of GFP expression or intensity occurring. The Zeocin selected pool had a negligible change in the percentage of population expressing GFP, going from 15.35% to 15.1%. Likewise, the mean fluorescence levels of these subpopulations within the Zeocin enriched pool remained unchanged even in the absence of selection pressure. Similarly, the puromycin selected pool only decreased from 3.94% to 3.29% of the population expressing GFP (comparing Figure 2.2a to Figure 2.2b). This observed stability in the absence of Zeocin selection is particularly important in cell line development applications.

In contrast, for HEK293 cells, the removal of selection pressure for one month actually resulted in an increased percentage of GFP expressing cells compared to the initial pool. Comparing Figures 2.2c and d, the hygromycin pool went from 7.52% to 10.5% GFP expressing, neomycin from 1.07% to 2.34%, puromycin from 3.30% to 11.5%, and for Zeocin from 89.8% to 98.9%. This result is unexpected but may be a result of increased HEK293 cell health associated with the removal of selection pressure. This may especially apply in the case of the neomycin selected population, in which it was not possible to establish a stable population at the pre-transfection determined MIC_{75}, as mentioned in the Materials and Methods. In the presence of even lowered concentrations of neomycin, HEK293 cells exhibited an unusual cell morphology and elevated levels of cell debris. Collectively, these results highlight the benefit of

prolonged selection as well as the differences of the selection markers with respect to their efficiency of selection and the stability of resulting pools.

As a secondary test, RNA was extracted from each of the HT1080 selected populations (Figure 2.2a, excluding neomycin) and GFP mRNA levels were measured using Real Time PCR. In the case of HT1080, the hygromycin population had the lowest GFP mRNA expression; puromycin had 1.65 times more GFP mRNA expression and Zeocin had 3.77 times more GFP mRNA expression. These results trend closely with flow cytometry measurements depicted in Figure 2.2a, which reflect protein expression levels. These results are summarized in Figure 2.3.

Figure 2.3: Zeocin enables higher GFP expression at the transcriptional level



Whole cell mRNA was extracted from the HT1080 hygromycin, puromycin and Zeocin selected populations. GFP expression levels were determined using RT-PCR and normalized using the comparative Ct method and the RPS11 housekeeping gene. Relative GFP expression levels were observed to be highest in the Zeocin selected population, followed by the puromycin and then hygromycin populations.

2.3.3 Zeocin selection aids in the identification of stable, enriched GFP expression at the clonal level

The final step in recombinant stable cell line development is the isolation and expansion of single cell clones with high level transgene expression. The impact of selection marker on single cell cloning has not been previously addressed, thus we were interested in isolating clones from our selected populations and evaluating resulting fluorescence levels. To do so, limited dilution cloning was utilized and cell cultures were prepared into 96-well culture plates for each of the four resistant HT1080 populations and the HEK293 Zeocin resistant population, as described in the Materials and Methods. In total, six to eight weeks of non-selective culturing was achieved for each clone. Eighty-two of the eighty-four single cell clones successfully grew and were expanded and evaluated (one hygromycin and one puromycin resistant clone did not survive). GFP fluorescent profiles and mean expression levels were determined using flow cytometry. Clones exhibiting fluorescence levels higher than the range exhibited by a wild-type population (auto-fluorescence) were determined to be expressing GFP. Mean fluorescence values (in relative fluorescence units) were measured for all clones.

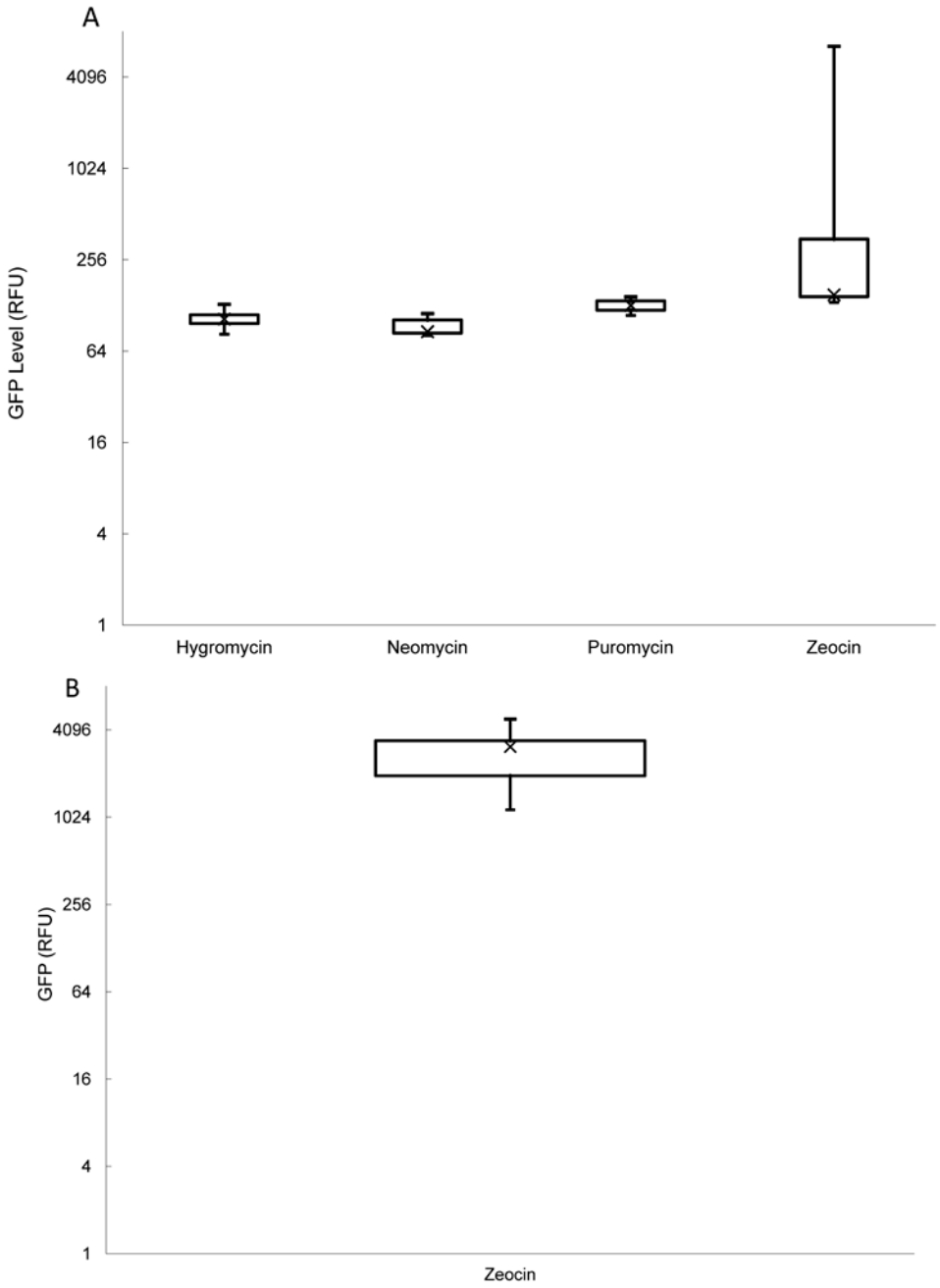
Table 2.2: Zeocin clones result in no false positives

Selection Agent	**GFP positive clones, count**	**Total clones, count**
HT1080 Clones		
Hygromycin	11	14
Neomycin	7	15
Puromycin	2	14
Zeocin	15	15
HEK293 Clones		
Zeocin	24	24

Between 14 and 24 clones were isolated from stably selected HT1080 populations and the Zeocin-selected HEK293 population. Each clonal population was expanded and GFP fluorescence profiles determined using flow cytometry. Positive GFP expression was compared to an untransfected, wild-type cell line.

The results for each set of clones are summarized in Table 2.2. Comparing the clonal populations isolated from each selected library, we see that all of the Zeocin selected clones (both HT1080 and HEK293) exhibit GFP expression and there are no false positives. Single cell cloning from each of the other HT1080 populations did result in false positives. Only eleven of the fourteen hygromycin clones (79%) and seven of the fifteen (47%) neomycin clones show some GFP expression. The puromyin selected clones performed the worst, with only two of the fourteen clones (14%) having any GFP expression. Additionally, the level of GFP expression is important in cloning, and these values for clones from each selected population are summarized in Figures 2.4a and b with box and whisker plots. The median fluorescence for each set of HT1080 clones was very similar, 111 RFU for hygromycin, 103 RFU for neomycin, 137 RFU for puromycin and 150 for Zeocin. However, the Zeocin selected clones have the largest range (6398 RFU) and correspondingly four clones with fluorescence above 500 RFU. None of the other HT1080 selected populations (hygromycin, neomycin or puromycin) produced clones with fluorescence levels above 150 RFU. Similarly, for the HEK293 Zeocin clones we see a high range (3710 RFU) and all of the clones have fluorescence levels exceeding 500 RFU. Again, these findings are consistent with our previous conclusions that Zeocin selection outperforms the three other agents evaluated. All of these clones were expanded over an extended period of time without selection pressure, which further illustrates that Zeocin selection can result in a stable, enriched, high-expressing populations.

Figure 2.4: Zeocin selection identifies better candidate cell lines



Single cell clones were isolated from the selected pools. After expansion, clonal populations were examined for GFP expression, with fluorescence data shown in box and whisker plots. **A**. In HT1080, clones were isolated from all four selected pools. While median values (x) were similar across the sets of clones, the Zeocin clones have a large range of expression, indicative of several highly expressing clones. **B**. In HEK293, clones were only isolated from the Zeocin population. The median GFP value was more than an order of magnitude higher than in HT1080.

Fluorescence of several of the best performing clones (isolated from the HT1080 and HEK293 Zeocin-resistant pools) are depicted in Figure 2.5. The best single cell clone isolated from the HT1080-zeocin selection population pool (specifically, clone 6 in Figure 2.5a) exhibits both high fluorescence (a 2.05 fold increase over transient expression) and a unimodal distribution. All five of the HEK293 clones depicted in Figure 2.5b exhibit a unimodal distribution and four have fluorescent levels exceeding that of transient expression. The best HEK293-zeocin clone (clone 5 in Figure 2.5b) has a 2.60 fold increase over transient expression. These results demonstrate that from a Zeocin-resistant population, we successfully isolated clones exhibiting high expression of GFP and low variance in expression, both characteristics that are desirable in candidate cell lines and important characteristics in successful cell line development.

Figure 2.5: Zeocin-resistant single cell clones exhibit high level, stable GFP expression

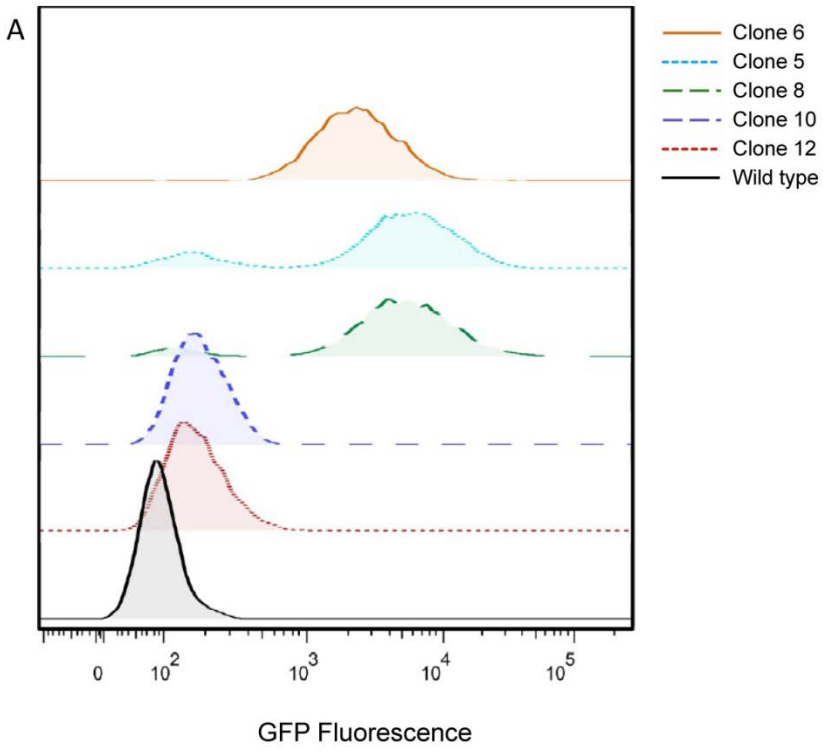
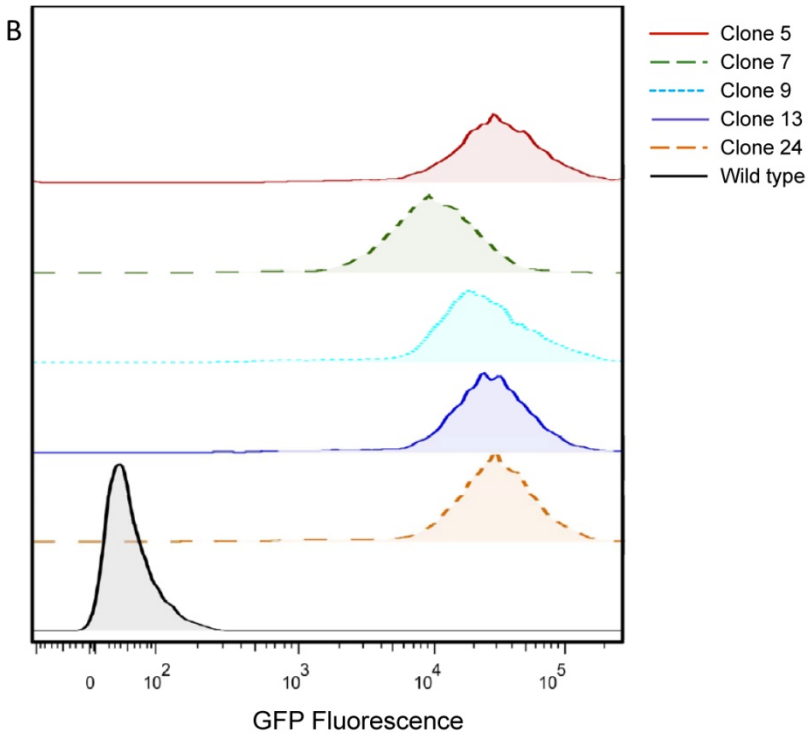


Figure 2.5 (continued)



The fluorescent profiles for select Zeocin-resistant clones are shown compared to control population (black solid). **A**. In HT1080, clone 6 exhibits both high GFP expression and a stable, unimodal distribution. **B**. In HEK293, all five depicted clones exhibit high GFP expression and a stable, unimodal distribution.

2.4 Concluding Remarks

For the first time in human cell lines, four common selection markers (and the corresponding resistance genes) were consistently and comprehensively evaluated on the basis of selected pool quality, stability, and resulting single-cell clones. This study clearly demonstrates that compared to the other three selection systems, Zeocin is the best selection agent for the establishment of recombinant cell populations in human cell lines. We evaluated selection marker performance in both HT1080 and HEK293 cells and saw

similar trends regardless of cell type, demonstrating these results are generic to human cell lines.

We have shown that Zeocin is able to identify pools with higher recombinant protein expression levels (through GFP fluorescence), which in turn leads to the isolation of better clonal populations and less false positives. Moreover, the Zeocin-resistant populations appear to be relatively stable when cultured in the absence of selective pressure. Based on these results, we believe Zeocin would be a good selection agent for the identification of high transcription loci throughout the genome. As a distant second, we identify hygromycin as the next best selection agent, followed closely by puromycin. Neomycin performed the worst and showed limited success in establishing a stable transgenic population in HEK293. These results strongly suggest that for genetic engineering in human cell lines, Zeocin is the best selection agent, and likely a good starting point for engineering other mammalian cell types. Furthermore, this finding suggests that antibiotics that work through mechanisms similar to that of Zeocin (the bleomycin family, a class of molecules that initiate DNA double strand breaks) could be of interest and a useful starting point for the development of better mammalian selection systems.

Finally, we demonstrated that a slightly prolonged selection after cell recovery can further increase production levels. While the establishment of candidate cell lines is a time-consuming process, we have demonstrated that the initial choice of a selection agent can strongly influence the quality of eventual clonal populations. This is an important finding for the areas of biotechnology and cell line development because selection conditions and selective genes are an important component in mammalian expression constructs and the establishment of transgenic populations.

Chapter 3: Identifying High Transcription Loci in the Human Genome

3.1 Chapter Summary

Mammalian gene expression and stability are strongly influenced by the genomic locus of integration. Here, we seek to identify productive loci within the human genome that will result in stable, high expression of heterologous DNA. Using an unbiased, random integration approach and a green fluorescent reporter construct, we identify ten single-integrant, recombinant human cell lines that are stable, high-expressors. From these cell lines, eight corresponding integration loci were identified. These loci are concentrated in both non-protein coding regions and intronic regions of protein coding genes. Expression mapping of the surrounding genes reveals minimal disruption of endogenous gene expression. Finally, we demonstrate that targeted integration at one of the identified loci, the 5th intron of the GRIK1 gene on chromosome 21, results in superior expression compared to the standard, illegitimate integration approach. The information identified here can be used in conjunction with site-specific genomic editing techniques, which are continually advancing, to retarget these advantageous integration loci. Such improvements in site-specific genomic editing techniques can result in flexible, predictable and robust cell line engineering which can reduce both the cost and time to identify candidate cell lines.

3.2 Introduction

One major limitation with mammalian cell hosts is their inability to autonomously replicate plasmid DNA. A long term, stable production cell therefore requires integration of heterologous DNA into the host cell genome and the subsequent isolation of a high expressing cell line. This process is time consuming, labor intensive, unreliable and

expensive and involves the screening of thousands of potential cell lines[8,10-12]. Advancements that eliminate this screening process would require techniques that enable heterologous DNA to be efficiently integrated in a site-specific manner, and the identification of genomic sites that support stable, high level expression of foreign DNA. These technologies would provide significant cost and time savings in BioPharma cell line development, facilitate gene therapy, and function as useful genetic tools in mammalian cell engineering[11].

In the past few years, many exciting technologies have been identified that facilitate site-specific integration in mammalian cell lines[167], often using enzymes to execute double-strand breaks at specific sites within the genome. These breaks then increase the likelihood that heterologous DNA constructs will be incorporated into the genome during non-homologous end joining (NHEJ) DNA repair[168]. Before these genome editing tools can be effectively employed however, it is necessary to know where to integrate a transgene to ensure high level, stable expression. Although it is well established that integration site is important to expression[3,8,10,12,25], limited information is available about desirable sites. Attempts to determine the exact genetic location of transgene insertions have been performed in isolated cases for cell lines with interesting characteristics[63,64,66]. In other cases, specific exonic sites have been targeted and evaluated, but this is a biased approach[66,169]. No unbiased, global study has been conducted in human cells to more comprehensively identify loci capable of supporting high expression of heterologous DNA.

Identifying loci suitable for integration is an important step in developing better tools for stable recombinant cell lines, especially if these sites can be retargeted[70]. This study seeks to identify transcriptionally active areas, demonstrate improved expression and stability compared to illegitimate integration, and map the surrounding expression

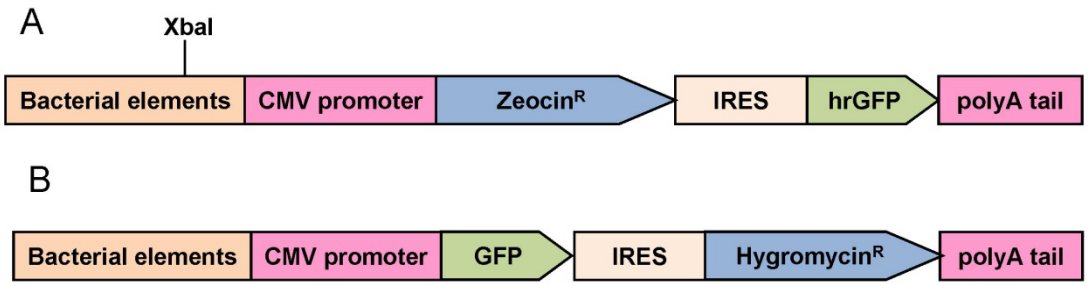
landscape. Studying productive integration loci in the human genome is an important advancement that can be coupled with existing technologies to advance our abilities to predictably engineer human cell lines. This work is novel and addresses an unmet need at the forefront of human genome and biologics research.

3.3 Results and Discussion

3.3.1 Isolation of stable, high expression recombinant cell lines

Despite advancements in our ability to target integration to specific genomic loci and our understanding that integration loci strongly impacts transgene expression levels, very little information is known about which sites are advantageous to target. Industrial cell line development is often still conducted through random integration of transgenic DNA, followed by a laborious screening process. Many commercial technologies continually exploit a small number of integration loci that have previously been demonstrated to be easily targeted, but very little thought is given to the productivity of these sites. Furthermore, only a small number exonic regions of protein coding sequences are considered[170,171], giving no consideration to the majority of protein coding and all non-coding regions of the genome for integration.

Figure 3.1: Dual-selection transgene constructs for high expression clones

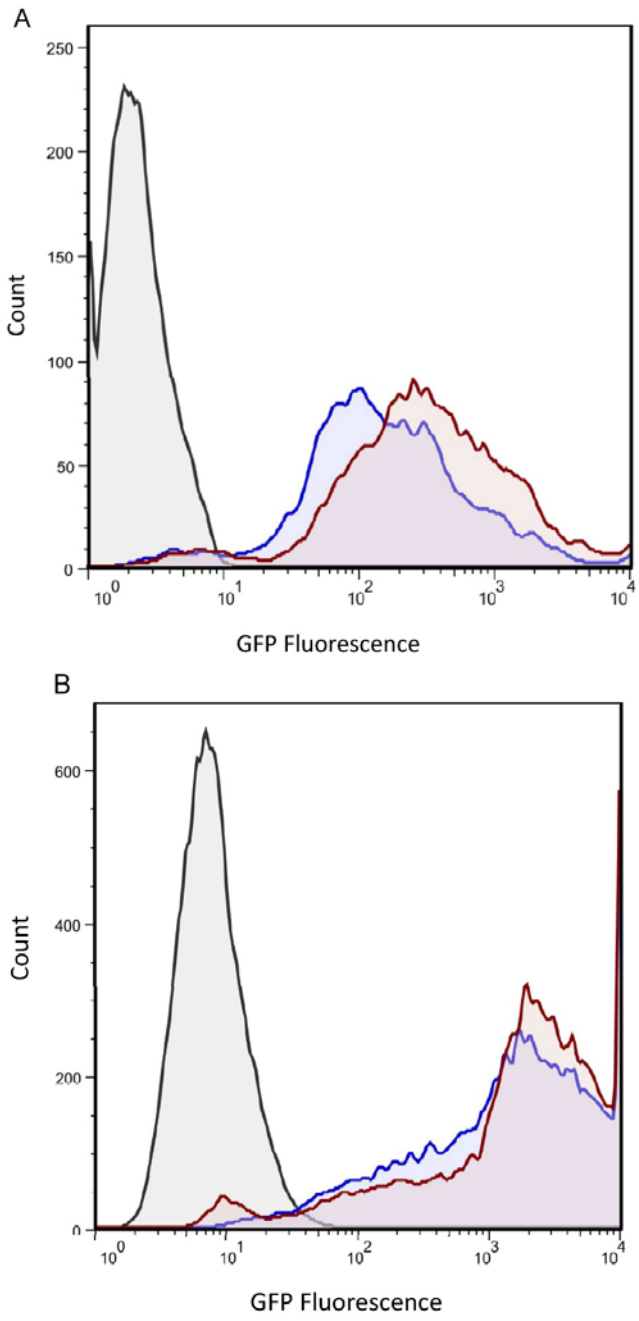


HT1080 cells were transfected with two heterologous constructs, each containing a single promoter and IRES to allow for simultaneous expression of two genes that enable dual expression. **A.** The pIRES-hrGFP construct contains the Zeocin resistance gene in the first cistron and a human optimized GFP gene in the second cistron. **B.** The pHL-GFP contruct contains a GFP gene in the first cistron and the hygromycin resistance gene in the second position.

We sought to conduct an unbiased survey of the human genome to identify genomic loci that afforded stable, high-level heterologous gene expression. As genomic integration of recombinant DNA in human cell lines has been shown to be a random process aided by native recombination mechanisms[61,62], we used a random integration strategy with a transgenic reporter construct to remove biases and explore the entire genome. We first established reporter constructs, shown in Figure 3.1, which contain both an antibiotic selection marker and fluorescent reporter gene (GFP) expressed with the CMV promoter in a single cistron. The human sarcoma cell line, HT1080, was transfected with the pIRES-hrGFP reporter construct (Figure 3.1a), as described in the Materials and Methods, and recombinant populations expressing the transgene were first identified through Zeocin antibiotic selection. We used Zeocin as our selection agent because previous work had demonstrated its superiority in establishing recombinant human cell populations[172]. These recombinant populations were initially identified using two different Zeocin concentrations: 100 and 250 μg/mL. At a later point, a Zeocin-resistant HT1080 population was established with 50 μg/mL selection. Hygromycin-resistant single cell clones were similarly established by Shire Human Genetic Therapies

using the pHL-GFP reporter construct (Figure 3.1b). Expression of the GFP reporter gene was measured using flow cytometry. Each population shows a broad range in expression, as designated by high coefficient of variance. Additionally, the more stringent selection at 250 μg/mL identified a population with enriched GFP expression. The GFP expression profiles of the recombinant populations established with 100 and 250 ug/mL are shown in Figure 3.2a.

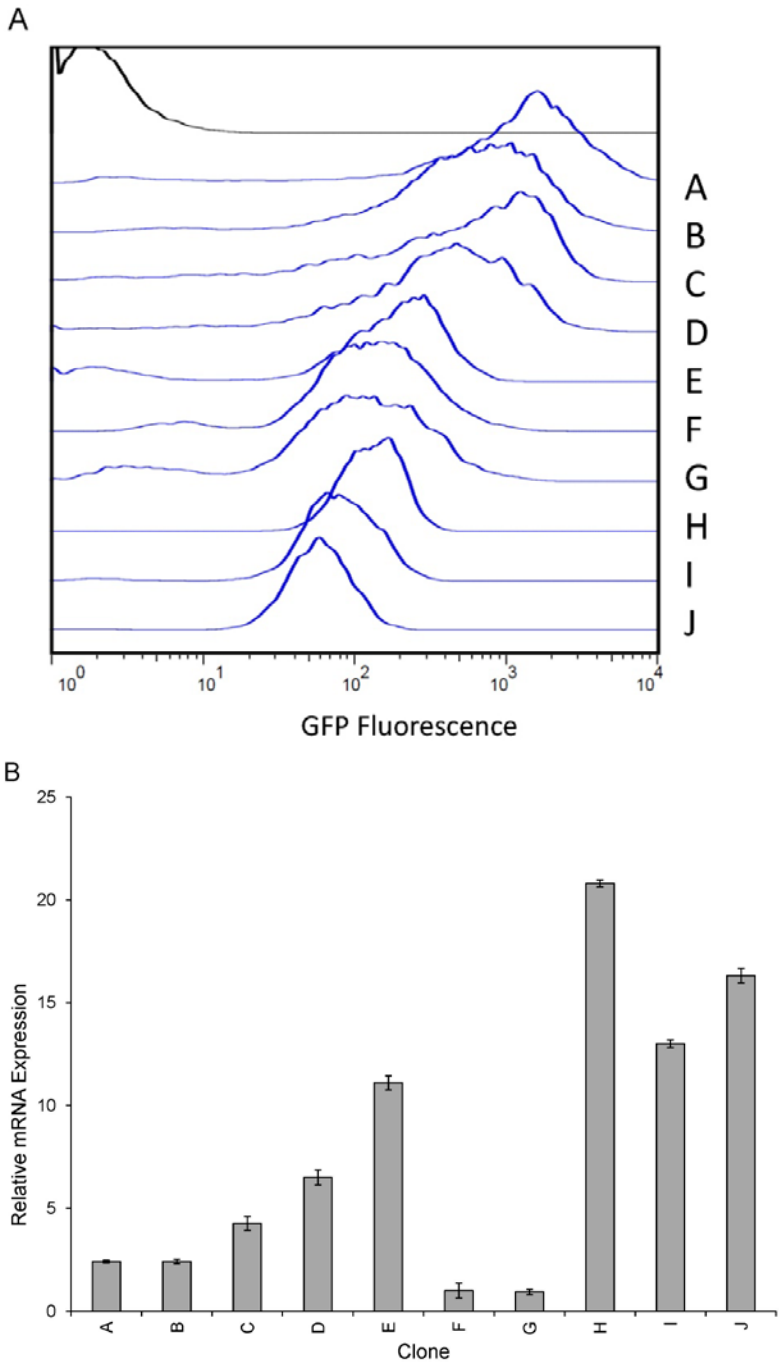
Figure 3.2: Establishing recombinant HT1080 populations



HT1080 recombinant populations were selected using both antibiotic resistance (**A**) and high GFP expression (**B**) as criteria. **A.** Antibiotic selection was first applied, using Zeocin concentrations of 100 (blue) or 250 (red) µg/mL. The GFP expression of the resultant populations are shown compared to untransfected HT1080 (black). **B.** FACS Aria sorting was next applied to isolate the top 10-15% of the resistant populations. The GFP expression of the populations following two rounds of FACS Aria sorting are shown for the 100 (blue) or 250 (red) µg/mL resistant populations and untransfected HT1080 (black).

These recombinant populations established with 100 and 250 ug/mL were further enriched using FACS Aria technology to isolate a sub-population expressing GFP at high levels. As discussed in the Materials and Methods, we selected the top 10-15% of these Zeocin-resistant populations. These profiles are shown in Figure 3.2b. This enrichment was performed prior to dilution cloning, which was used to isolate single cells and establish homogenous populations. In order to identify stable clone lines, the expansion of single cell clones, which extended over a period of 6-8 weeks, was performed without any antibiotics in the media. After expansion, the transgene copy number was determined for each of the clones using a previously established protocol[173]. We continued to work with ten clone lines in which only a single integration event had occurred. GFP expression of these clone lines was evaluated using both flow cytometry to measure protein expression (Figure 3.3a) and RT-PCR to measure mRNA expression (Figure 3.3b). Each of these clones has a stable expression profile and mRNA levels are very high relative to the endogenous 40S ribosomal protein, encoded by the RPS11 gene. These results indicate that these clone lines represent integration loci that are supporting stable, high-level transgene expression.

Figure 3.3: Isolated single cell clones exhibit high protein and mRNA expression



Ten single cell clonal populations were isolated from the recombinant populations and protein (**A**) and mRNA expression (**B**) were measured. **A.** GFP expression profiles for clonal populations (A-J) were measured using flow cytometry and are shown compared to untransfected HT1080 (top). **B.** Relative mRNA expression of the clonal populations (A-J) was measured by RT-PCR for the first cistron gene. mRNA expression levels are normalized to clone F.

3.3.2 Determination of high-expression integration loci in the human genome

After isolating these ten clonal populations, we next sought to identify where in the genome the GFP construct had integrated. Genomic DNA was extracted for each clone and a variety of PCR techniques were used to isolate and amplify genomic DNA adjacent to the GFP construct. These fragments were then identified using standard sequencing techniques and a BLAST search of the publicly available genome sequence. Each integration site was re-confirmed through a positive PCR reaction in which a primer specific to the GFP construct and a primer matching the amplified, genomic sequence were used together. The details for these confirmations are included in Appendix B for each clone.

Identification of these integration events would be greatly aided by high-throughput sequencing techniques. However, given the size of the human genome, this is an expensive approach. Therefore, we employed a variety of low throughput methodologies, including TAIL PCR, inverse PCR and plasmid recovery. Each is explained in the Materials and Methods. Using these approaches, we were able to identify the integration loci of our ten GFP-expressing clone lines. These integration sites are summarized in Table 3.1. We see that these integration loci are distributed throughout the human chromosomes. Interestingly, none of the integration sites are in exonic regions of protein coding genes. We do see three integration events into intronic regions. Although not previously explored, it makes sense that intronic regions afford high-level expression because of their proximity to promoter and transcription factor binding sites. In many cases, however, we see that integration events have occurred very far from the nearest protein-coding regions. This is a surprising result and clearly demonstrates that regions outside of protein-coding sequence are hospitable towards heterologous gene expression.

Additionally, we identified two integration sites through duplicate, independent integration events. Clones I and J both arose from integration into the 5th intron of the GRIK1 gene on chromosome 21. Clones A and B both arose from integration into an unplaced genomic contig. Unfortunately, very little information is available about this genomic contig, including its chromosome, because this is a region of high redundancy that has not yet been resolve. Of these ten clones, the highest expresser (from mRNA measurements, Figure 3.3b), was integrated on chromosome 14 in the IGHG2 gene. This region of the genome is rich in immunoglobulin proteins, which are closely spaced.

Table 3.1: High transcription loci are distributed throughout the genome

Clone	Chromosome	Intron	Nearest gene	Nearest gene function
A	Unknown		Unplaced genomic contig (3980 bp)	
B	Unknown		Unplaced genomic contig (3980 bp)	
C	18	26	DCC	Netrin 1 receptor
D	5		SEMA6A, 31kb downstream	Transmembrane domain
E	4		SPINK2, 9kb upstream	Serine peptidase inhibitor
F	15	1	SV2B	Synaptic vesicle glycoprotein
G	7		SEMA3A, 78 kb upstream	Secreted neuronal protein
H	14		IGHG2	Immunoglobulin heavy constant gamma 2
I	21	5	GRIK1	Glutamate receptor, neuronal
J	21	5	GRIK1	Glutamate receptor, neuronal

From ten stable, high expressing clones we identified eight integration loci using PCR-based low throughput methodologies. Each site was confirmed using primers matching the transgene and genomic locus, which produced a positive band but lack of band with wild-type gDNA. Each locus is discussed in detail in Appendix B.

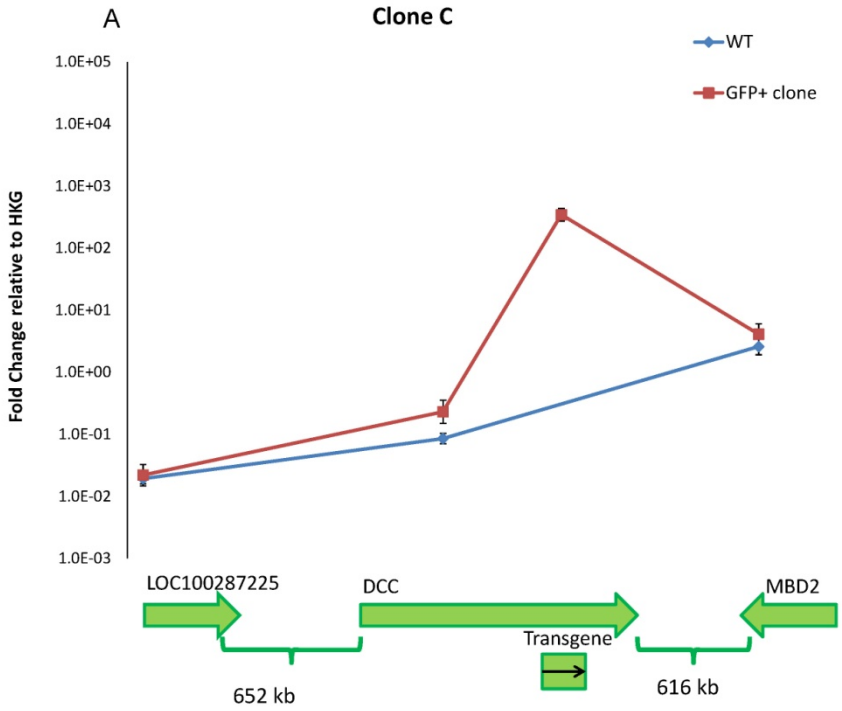
3.3.3 Expression mapping reveals minimal disruption of endogenous gene expression

Next, we sought to determine what we could learn from these integration sites that may guide rational identification of future integration loci. In particular, we examined

the expression profiles of surrounding protein coding genes. We were interested in benefits to transgene expression provided by the surrounding genomic DNA, as well as perturbations that may be caused by integration. Perturbations are specifically important in gene therapy applications, where 'harmless' integration loci must be chosen such that surrounding cancer related genes are not impacted[65]. Previous studies have indicated cases where integration events caused no impact to surrounding genes, as well as cases where it resulted in an undesirable phenotype[66,174].

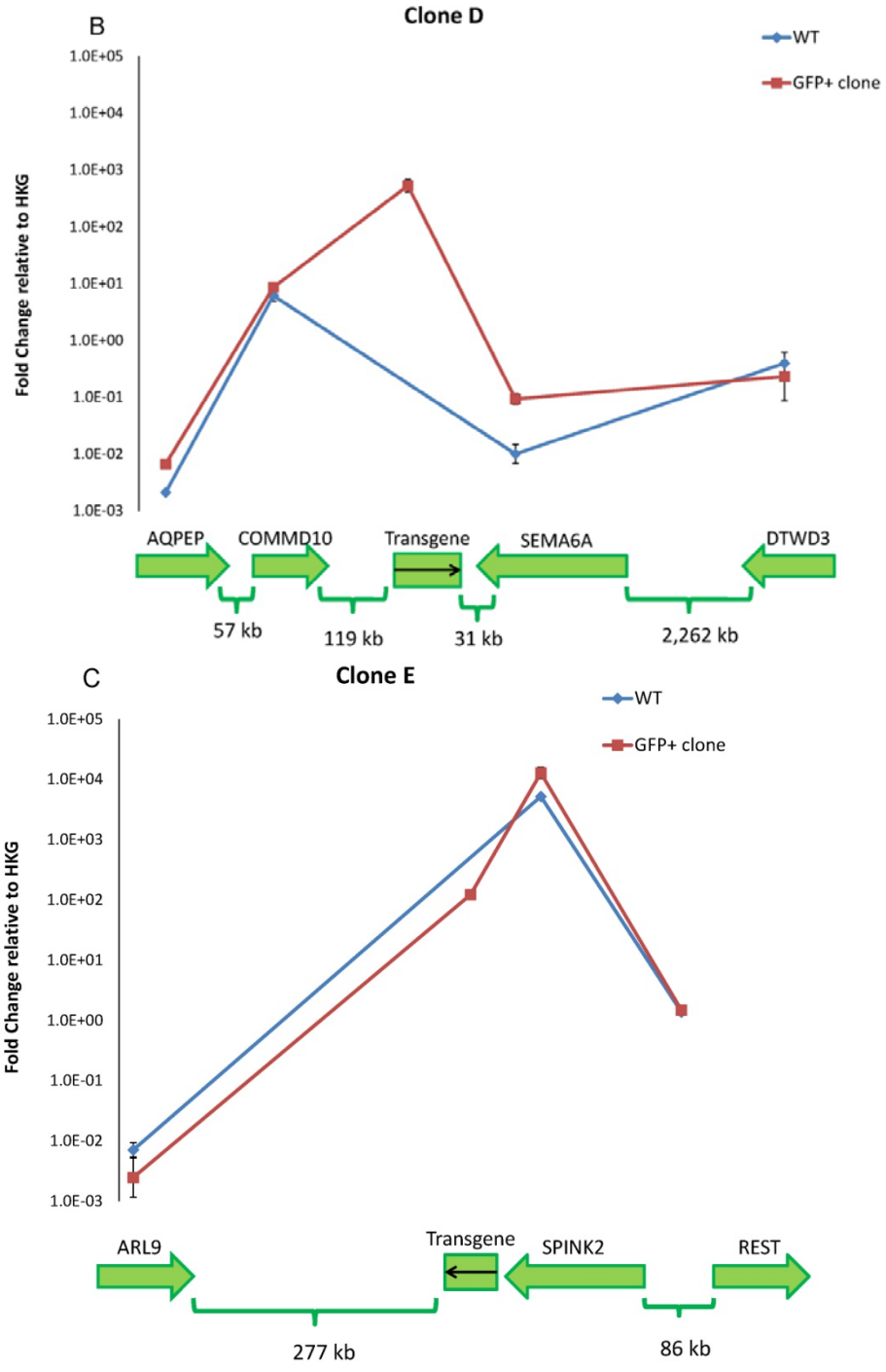
Expression levels of protein coding genes were determined using RT-PCR conducted with whole cell RNA, as described in the Materials and Methods. Expression of each gene was compared to RPS11 for both the GFP positive clone and wild-type HT1080. The resulting expression maps for all clones (excluding those integrated in the unplaced human genomic contig for which no information is available) are shown in Figure 3.4. Universally, we see that for protein coding sequences distantly located from the site of integration, there is no expression difference between the wild-type and GFP positive clone. This indicates minimal expression perturbation caused by transgene integration and agrees with other findings that transgene integration does not impact neighboring gene expression[66]. Negligible invasion of endogenous expression can be observed for clones C, E, F, and I in Figures 3.4a, c, d and g. In a few cases, we see changes in expression of endogenous genes closest to the integrated transgene. For clone D (Figure 3.4b), expression in the GFP positive clone of SEMA6A is elevated compared to a wild-type gene. This could be caused by the presence of the strong CMV promoter in the integration cassette, which would recruit transcription factors to the region.

Figure 3.4: Expression maps for protein-coding sequences surrounding integration loci



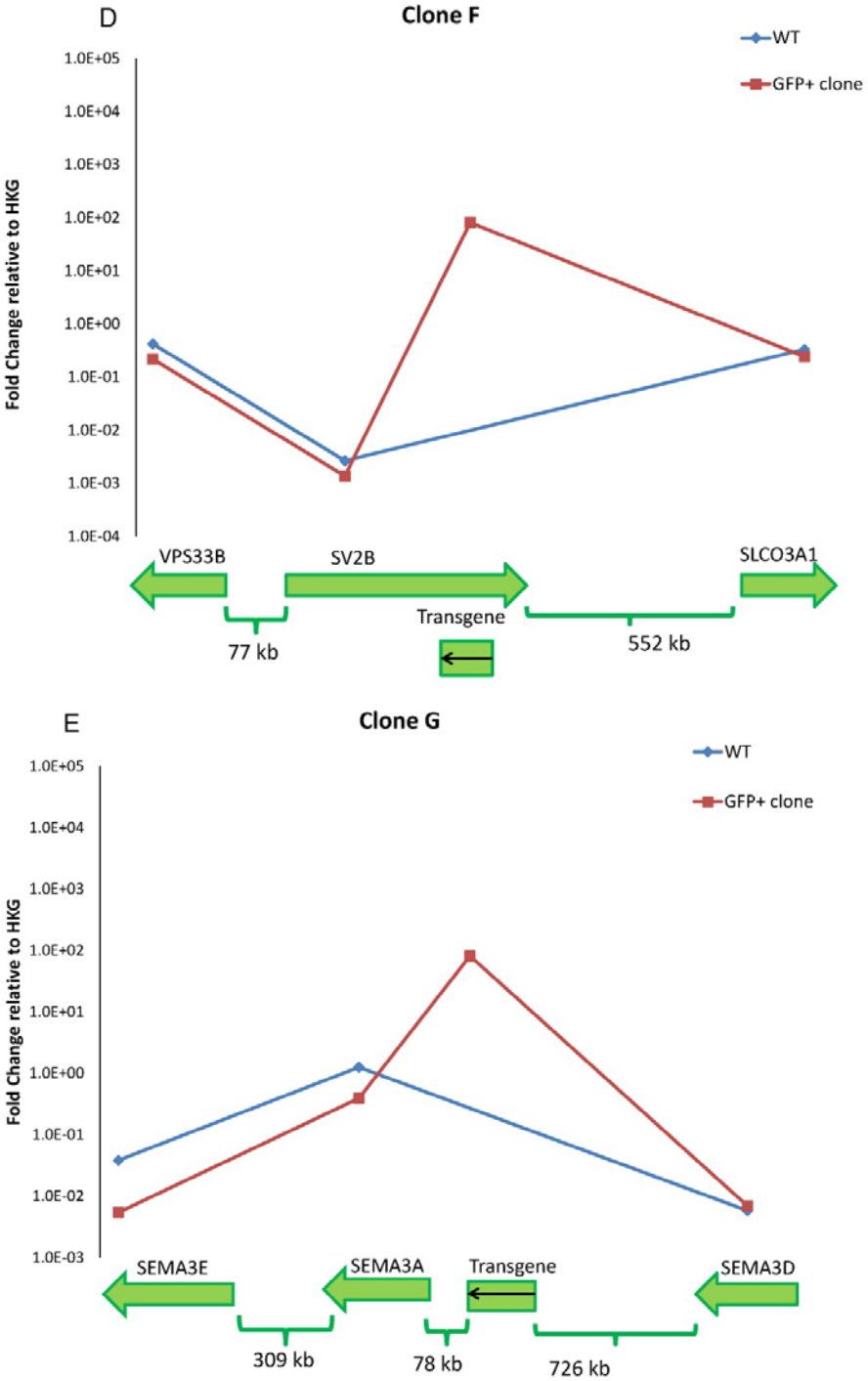
Using RT-PCR and whole-cell RNA, mRNA expression was determined for the integrated transgene and surrounding protein-coding genes. Fold change in mRNA expression was measured relative to RPS11, an endogenous gene. The black arrow indicates the promoter direction for the transgene. Error bars indicate standard deviation from RT-PCR triplicates. **A.** mRNA expression profile for clone C on chromosome 18, including the transgene and endogenous genes DCC, MBD2 and uncharacterized locus 100287225.

Figure 3.4 (continued)



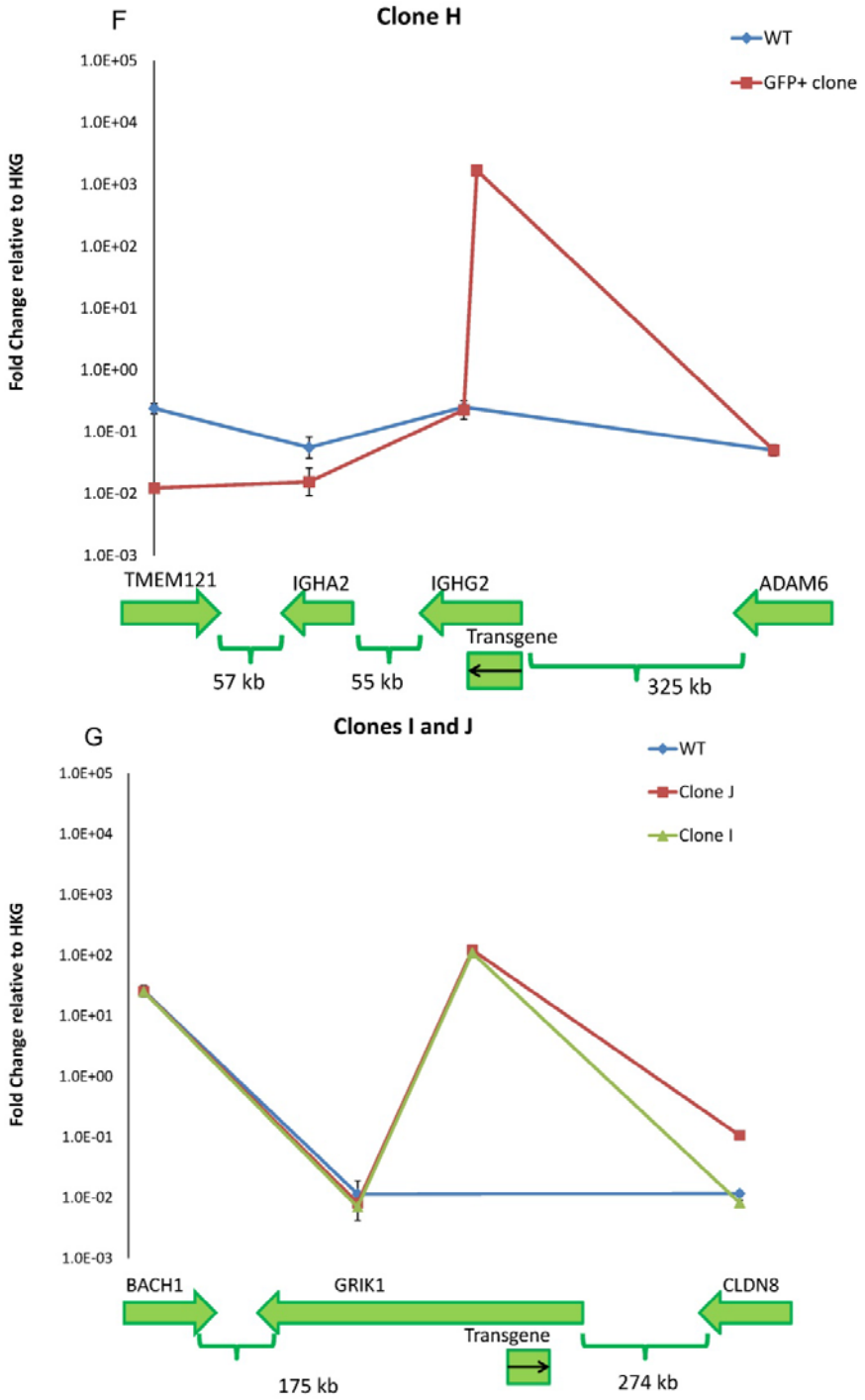
B. mRNA expression profile for clone D, integrated on chromosome 5, including the transgene, AQPEP, COMMD10, SEMA6A and DTWD3. **C.** mRNA expression profiles for clone E, integrated on chromosome 4, including the transgene, ARL9, SPINK2 and REST.

Figure 3.4 (continued)



D. mRNA expression profile for clone F, integrated on chromosome 15, including the transgene and endogenous genes VPS33B, SV2B and SLCO3A1. **E.** mRNA expression profile for clone G, integrated on chromosome 7, including the transgene, SEMA3E, SEMA3A, and SEMA3D.

Figure 3.4 (continued)



F. mRNA expression profile for clone H, integrated on chromosome 14, including the transgene, TMEM121, IGHA2, IGHG2 and ADAM6. **G.** mRNA expression profiles for clones I and J integrated on chromosome 21, including the transgene and endogenous genes BACH1, GRIK1 and CLDN8.

With the exception of the integration site for clone E, we see that expression of the transgene is significantly elevated relative to the surrounding genes, which in most cases are lowly expressed. However, in the case of the clone E integration event (Figure 3.4c), which is 9kb downstream of the SPINK2 gene, we observe that SPINK2 expression exceeds that of the transgene expression. In this case, we hypothesize that the transgene is piggy-backing off of endogenous gene expression, which is creating a hospitable expression environment. The orientation of the transgene, which matches that of SPINK2, further supports this. Again, clone D (Figure 3.4b) may be benefitting from the highly expressed, adjacent COMMD10 gene. Although it is 119kb from the transgene, the promoter of the COMMD10 gene and GFP construct are similarly oriented.

3.3.4 Site-specific targeting of Grik1 intron 5 demonstrates superior transgene expression

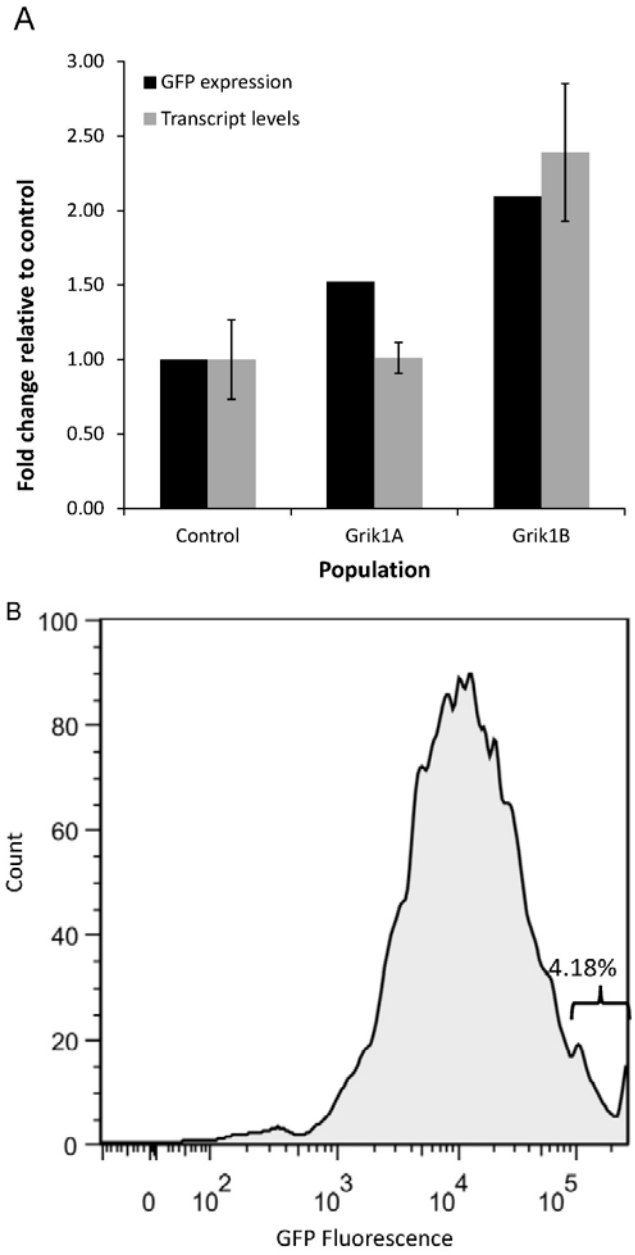
Finally, we sought to demonstrate the impact of combining the high-transcription loci identified here with site-specific genomic targeting techniques. As an alternative to random integration followed by tedious and resource-intensive clonal screening, we can use pre-existing technologies to initiate double-strand breaks (DSBs) at these advantageous loci and efficiently deliver a transgenic construct to that site. Although many technologies exist to perform these DSBs, we utilized the CRISPR system, which was recently demonstrated to be a flexible, highly efficient method for mammalian genome editing[89]. This method combines the Cas9 protein with an editable crRNA-tracrRNA fusion transcript to site-specifically cleave DNA. We selected the 5th intron of the Grik1 gene on chromosome 21 as the target of choice. We were interested in replicating this integration site because of the high mRNA expression we observed in

clones I and J, as well as the independent occurrence of two clones with the same integration locus.

Tests were conducted in both HT1080 and HEK293 cells. Control cells were transfected with a mammalian expression cassette containing both hrGFP and Zeocin driven by the CMV promoter (Figure 2.1) and the hCas9 plasmid, which encodes for the cas9 protein. Because these reactions lacked targeting sequence, they function as illegitimate integrations.

Additionally, two targeted integrations were performed in which a guide RNA (gRNA) construct was transfected with the hCas9 and hrGFP cassettes. The gRNA encodes a crRNA-tracrRNA fusion transcript driven by the U6 polymerase III promoter and can be modified to include a specific 23 nucleotide region of homology. We designed two distinct gRNA constructs specific to our locus of interest, named Grik1A and Grik1B. Both were selected to minimize off-targeting effects using the criteria outlined by previous researchers[89]. The Grik1A construct contains the targeting sequence GCTATTTTAGATATATAGCAAGG and was designed to cut within the previously identified integration locus. The Grik1B construct contains the targeting sequence GTGGGGGTTATACCACTCGTAGG and was designed to cut 65 base pairs away from the previous site. Seventy-two hours after the transfection event, cells were subjected to selection at MIC_{75} levels until viability recovered to greater than 90%. The heterogenous populations were then evaluated for both mRNA and protein expression. Flow cytometry was used to evaluate GFP protein expression and RT-PCR was used to determine mRNA expression levels.

Figure 3.5: Targeted integration into the Grik1 loci results in elevated transgene expression



A mammalian expression cassette expressing GFP and Zeocin was transfected into HT1080 cells in a random (control) and targeted fashion (Grik1A/B) using the CRISPR system. Following antibiotic selection, heterogeneous populations were evaluated. **A.** Flow cytometry was used to measure mean GFP expression (black) and RT-PCR was used to determine transcript levels (gray). Error bars reflect the standard deviation of RT-PCR technical triplicates. **B.** A sub-population expressing GFP at high levels was gated for and used to estimate Grik1 targeting efficiency. This sub-population is 4.18% of the total Grik1B population, shown here.

Compared to the control, the targeted transfections in HT1080 resulted in increased GFP expression levels. A FACS Fortessa was used to evaluate 10,000 cells from each population. The mean fluorescence of the Grik1A targeted population was 1.52 fold higher than the control population and the Grik1B targeted population was 2.09 fold higher. We also evaluated mRNA expression levels of the integrated construct using RNA extracted from the control populations as well as the Grik1A and B targeted populations. Although we observed no difference in transgene transcripts levels between the control and Grik1A targeted populations, we did see 2.39 fold higher transcript levels in the Grik1B targeted population, which aligns closely with our GFP expression data. These results are summarized in Figure 3.5a.

Using flow cytometry, the expression profiles of the selected populations were examined. Both the Grik1A and Grik1B targeted populations contained a clear, sub-population with very high GFP expression in the upper region of the histogram. We believe this sub-population represents successful Grik1 integration events. This sub-population is shown for Grik1B in Figure 3.5b. Gating for these events, we calculated that 2.29 and 4.18% of the total population for the Grik1A and B cases respectively fall in this region. In the control case, less than 1% of the population is in this region. Based on these results, we estimate that our targeted integration efficiency is 2-4% depending on the gRNA construct used and the Grik1B construct is more efficient than the Grik1A construct. These efficiencies, as well as variation between constructs, are similar to previously reported values for other human cell types[89].

No enrichment of the HEK293 targeted populations was observed using either flow cytometry or RT-PCR. This is likely a result of low transfection efficiencies for this cell line (less than 30%). We are currently optimizing a protocol to alleviate this bottleneck before proceeding with additional experiments in HEK293.

Collectively, the results in HT1080 cells confirm that the Grik1 integration locus supports high level expression of heterologous DNA. Furthermore, the targeted populations we evaluated are heterogeneous and we expect a much larger expression difference on the clonal level. We are currently isolating single cell clones to carry out this comparison. In measuring both transcript and GFP levels of the mixed populations, we observed more than a two-fold increase in expression in targeted populations compared to random integration. However, in examining the sub-population which we believe includes Grik1 integration, we see more than an order of magnitude increase in expression levels. These results clearly demonstrate the advantage of targeted integration in cell line development and can be combined with other approaches, such as gene duplication, to achieve further increases in protein expression.

3.4 CONCLUDING REMARKS

Here, using an unbiased, random integration approach, we identify ten recombinant human cell lines with stable, high-level heterologous gene expression. Each was confirmed to have a single copy of the transgenic cassette. Using low-throughput methodologies, we have identified the corresponding integration loci, which occur in intronic regions of protein coding sequences and non-protein coding regions. These results indicate the importance of non-protein coding regions for heterologous gene expression, despite the fact that previous studies have focused exclusively on exonic regions. Expression maps for each integration loci demonstrated that in most cases, negligible perturbation has been caused to surrounding genes. This is an important observation that suggests that many of these identified loci good have gene therapy applications[65]. Finally, we demonstrate that targeted integration at one of the identified

loci, the 5th intron of the GRIK1 gene on chromosome 21, results in superior expression compared to the standard, illegitimate integration approach. Although locus of integration is well known to influence gene expression, little effort has been made to identify desirable loci. This work demonstrates that desirable genomic integration sites can be identified for human cell lines. Furthermore, coupling this information with site-specific genomic editing techniques, which are continually advancing, is advantageous to cell line development. Retargeting these advantageous integration loci can significantly reduce the time, labor and materials associated with cell line development. Additionally, this approach can be extended to other mammalian cell lines used for industrial protein production, including CHO and NS0. This is an important finding for the areas of biotechnology and cell line.

Chapter 4: Exploring the Cre/*lox* System for Targeted Integration into the Human Genome

4.1 Chapter Summary

Once productive integration loci have been identified for mammalian cell hosts, it is important to be able to efficiently target recombinant DNA to those sites. A variety of site-specific genomic editing techniques are being developed that enable these precise manipulations. Cre recombinase is a genome editing tool suitable for site-specific integrations in mammalian genomes. This enzyme is commonly used for deletions and insertions in the mouse genome and has been extended to other model organisms including CHO and human cells. Despite its utility, the efficiency of transgenic swapping events compared to excision remains limited. Here we sought to identify important parameters and limiting factors that influence swapping propensity in this system, especially when using one wild-type *loxP* site. To modulate and increase the occurrence of swapping events, we identify two novel parameters. First, we identify the *lox*FAS-*loxP* pairing, a sequence never before used in mammalian systems, as the best choice for increasing swapping events in human cell lines. Second, for the first time we implicate the importance of delayed introduction of Cre DNA for optimal swapping efficiency. This same modification could potentially be of use to other systems catalyzing trimolecular reactions such as ΦC31 integrase and FLP recombinase where we hypothesize that transport of the exchange cassette is likewise initially rate limiting. The total number of recombination events, but not the ratio of swapping to excision, was found to be influenced by the quantity of Cre DNA transfected. Through this study, we obtain Cre-mediated swapping frequencies of 8 to 12% without antibiotic enrichment, which represents nearly an order of magnitude increase over prior reports in literature.

4.2 INTRODUCTION

Integrases and recombinases are widely-used DNA-modifying enzymes that enable site-specific genome modifications in many hosts including mammalian cells. These enzymes, including Cre recombinase[175], facilitate genome editing functions such as deletions, insertions, inversions and exchanges based on recognition of short, palindromic sequences that are typically integrated into the host's genome[176]. These enzymes can be used to specifically retarget productive genomic loci in contrast to current practices that rely on randomly sampling the genome each time a stable cell line is desired. While many such enzymes have been utilized in mammalian cells, Cre recombinase is the most popular in the literature and thus the focus of this study. Cre recombinase has several distinct advantages: it does not require protein factors[177], has high expression in mammalian systems[80], is more efficient than both FLP recombinase[178] and ΦC31 integrase[179], and pseudo Cre recognition sites (*loxP*) do not interfere with recombinase activity in the mammalian genome[180]. Despite the fact that Cre recombinase is one of the most widely-used tools for precise, site-specific editing of mammalian genomes, its efficiency for transgene swapping and integration still remains limited[77,181].

Cre recombinase can mediate both intramolecular (excision or inversion of DNA fragment) and intermolecular (exchange of two DNA fragments) recombination[77]. Cre recombinase natively recognizes the *loxP* site, a thirty-four base pair palindromic sequence with an 8-base pair asymmetric spacer region[77,80,181-185], and acts upon the neighboring DNA sequences. Intermolecular recombination, often the desired outcome and hereafter referred to as 'swapping,' is accomplished by translocation between two DNA fragments with corresponding *lox* sites[80,186]. In contrast, intramolecular recombination, hereafter referred to as 'excision,' is the preferred function of Cre

recombinase[181] and involves removal of genetic material between two *lox* sites. Of these two functions, Cre-mediated swapping is desirable for site-specific integration and is the focus of this study.

Cre recombinase was first recognized as a mammalian genome editing tool in 1988[175]. Since then, adaptations have expanded the utility of the Cre/*lox* system in mammalian cells beyond mouse cell lines[181,183-185,187,188] to include Chinese Hamster Ovary (CHO)[189], flies[190], pig[191] and human cells[192]. Prior work has developed engineered variants with optimized performance in both CHO[193] and mouse cells[188]. Through protein engineering, variants able to recognize new target sequences have been developed[194,195]. Other studies have evaluated the impact of transfection methods and absolute amount of transgene DNA on recombination efficiency[192]. Many efforts to improve the preference of the Cre/*lox* system for swapping have focused on altering the *lox* site sequence. The use of mutated *lox* sites can significantly reduce excision and increase swapping efficiency[182,183,185,192], and have been tested for gene integration *in vitro*[196,197]. However, many applications still utilize at least one copy of the wild-type *loxP* site, often as an artifact of pre-established cell lines[187].

Despite improvements to the Cre/*lox* system, the frequency of targeted swapping still remains lower than desired and varies significantly with genomic site of interest[184,187]. Prior to antibiotic selection, Cre recombination typically results in site-specific targeting at frequencies of less than 1% of cells surviving transfection[77,181]. Furthermore, while many studies examine individual parameters influencing Cre-mediated intermolecular recombination, there is a surprising lack of research evaluating the parameters simultaneously for improving swapping efficiency. In this study, we sought to better understand Cre/*lox* swapping by examining multiple mutant *lox* pairings, relative and absolute amounts of transgene DNA, and timing of transfections.

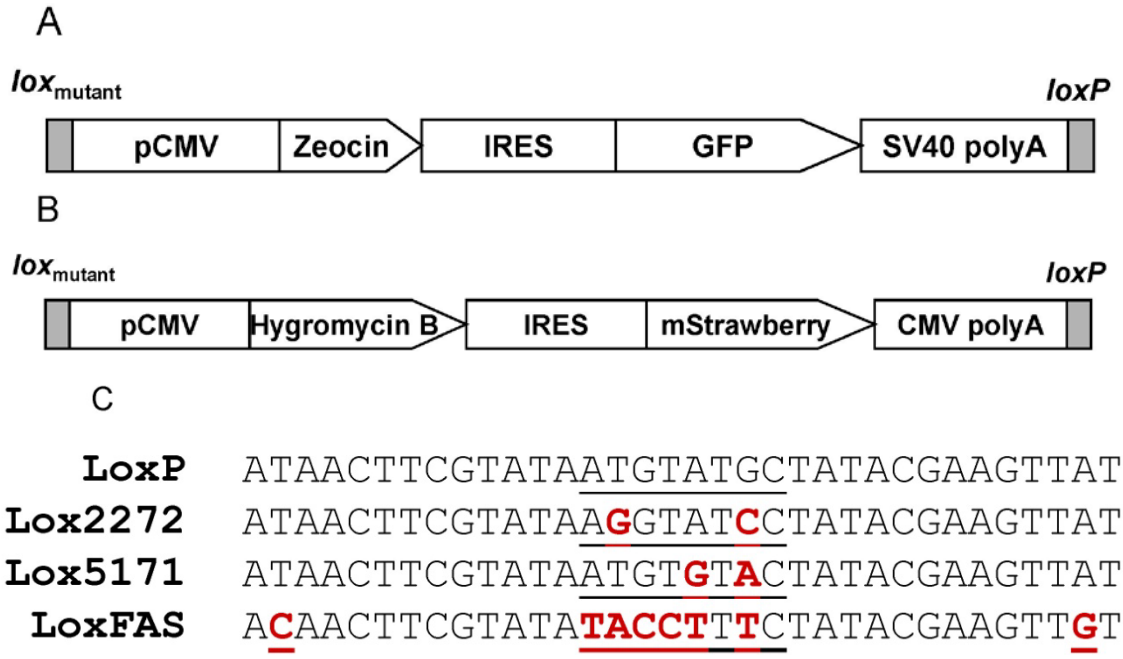
4.3 RESULTS AND DISCUSSION

4.3.1 Evaluation of mutant lox sites for improved swapping efficiency

Heterologous *lox* site pairings have previously been shown to improve swapping efficiency[77,80,182,184,185] and reduce excision[198], whereas the *loxP-loxP* pairing results almost exclusively in excision events. While several combinations of mutant sites have been tested *in vivo* using bacterial[182] or mouse cells[185], we were interested in determining an optimal pairing for use in human cell lines. Three mutant *lox* sites were paired with wild-type *loxP* for this experiment: *lox*2272, *lox*5171 and *lox*FAS (Figure 4.1). The *lox*2272 and *lox*5171 are frequently used variants each containing two mutations in the spacer region and have previously been shown to have a higher efficiency for swapping compared to a *lox* site with a single mutation[183]. The *lox*FAS sequence occurs natively in *S. cerevisiae* and contains 8 mutations compared to wild-type *loxP*. To date, *lox*FAS has rarely been utilized in Cre recombination despite one *in vivo* phagemid study suggesting that 99.8% of excision was lost when paired with *loxP*[182]. This work represents the first functional test of this mutant site in a mammalian cell system.

To evaluate the impact of these pairings on Cre-mediated swapping rates in human cells, we utilized a dual fluorescent screening system. Specifically, this assay facilitates detection of excision that cannot be assessed using standard antibiotic selection techniques. This screen was validated through the use of a Southern Blot for one of the clones and resembles a similar approach previously used to identify successful RMCE events[199].

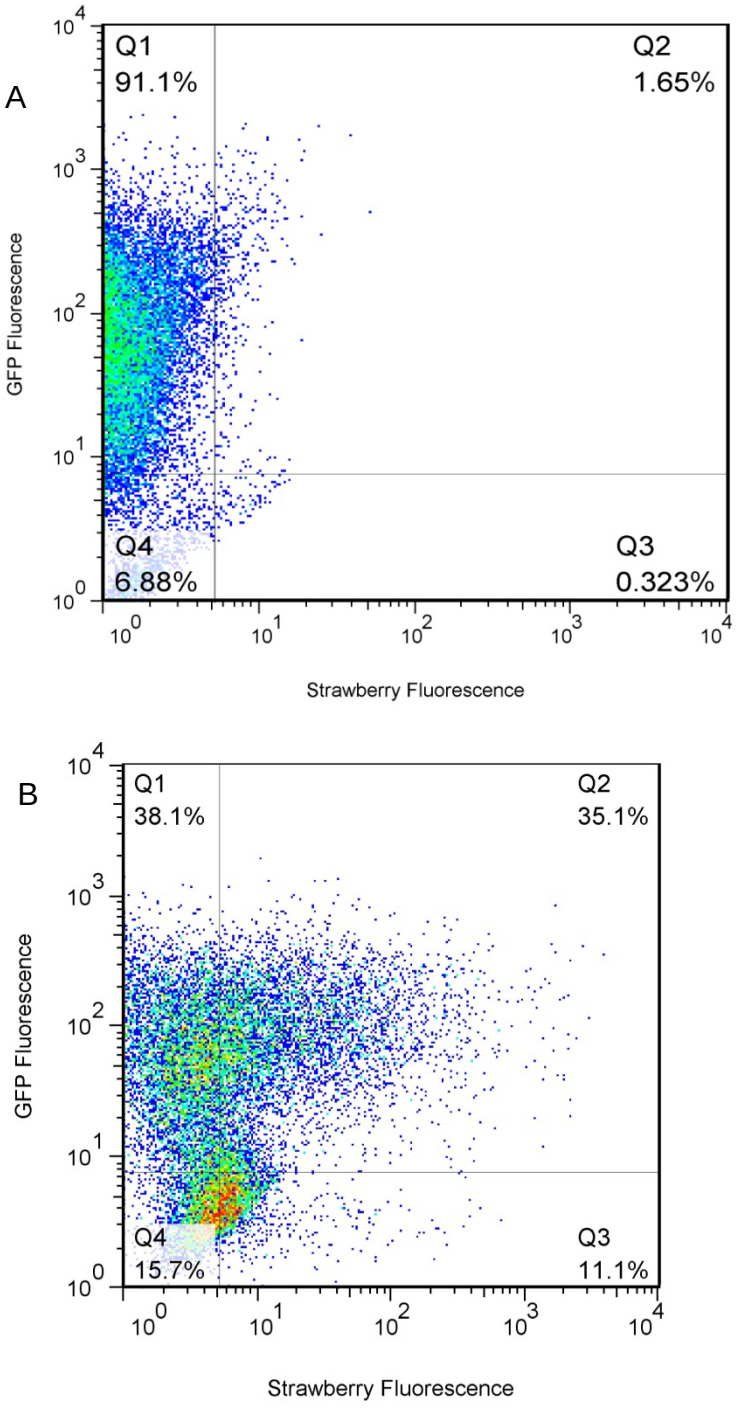
Figure 4.1: Mammalian expression vectors for dual fluorescent screen



Two mammalian expression vectors were constructed: **A.** The pIRES-hrGFP vector contains a Zeocin resistance gene and human optimized GFP gene, both under the control of the CMV promoter and flanked by a mutant *lox* site preceding the promoter and *loxP* site following the GFP gene. **B.** A Strawberry construct is similarly flanked by a mutant and wild-type *lox* site. **C.** The *lox*2272 and *lox*5171 sites both contain two mutations localized to the asymmetric, 8 base pair spacer region (underlined) of the lox sequence. The *lox*FAS site contains 8 mutations, with the majority of the spacer region being altered, as well as two mutations to the palindromic region of the sequence.

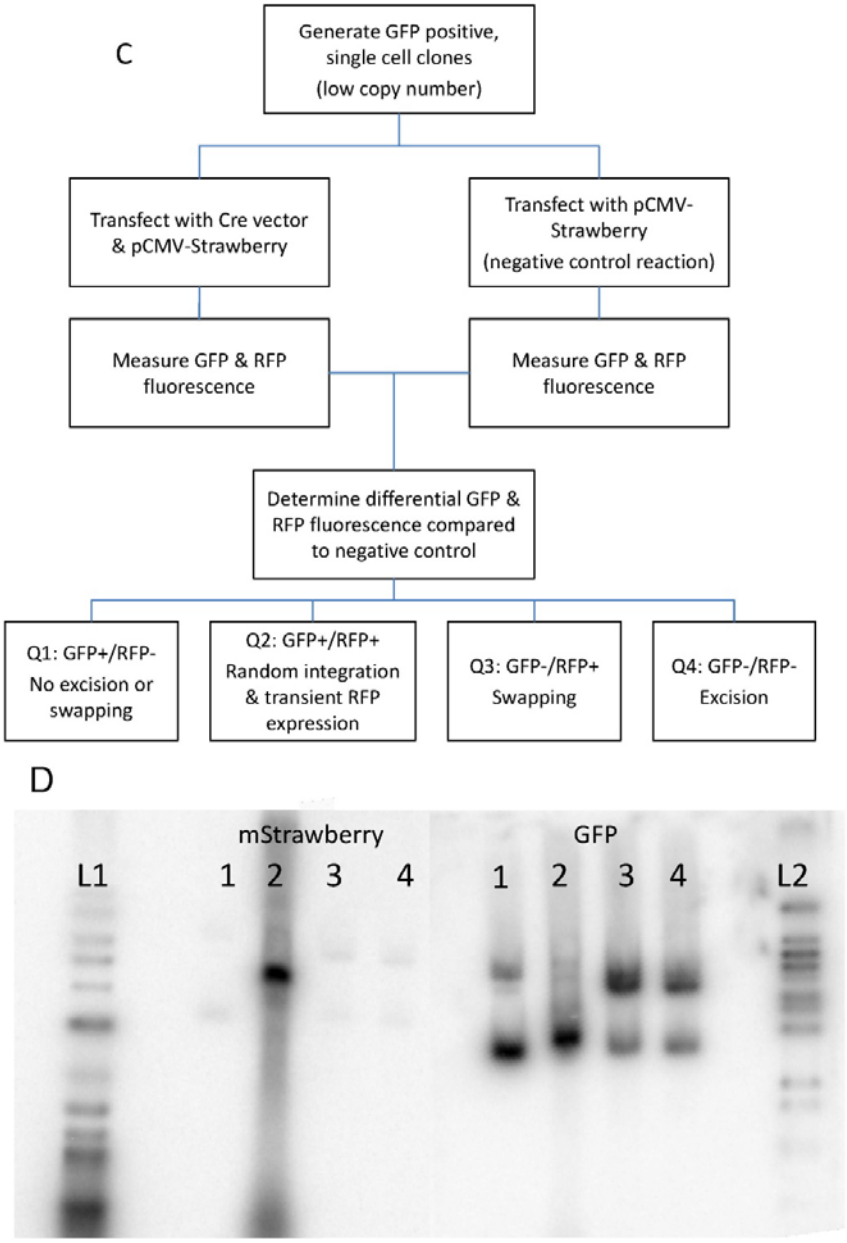
First, we constructed the pIRES-hrGFP vector (Figure 4.1). The operon is flanked on both sides by a mutant *lox* site and a *loxP* site, and was randomly integrated into the HT1080 genome. Based on its superior performance in HT1080 populations, the Zeocin selection marker was used. Cells were treated with Zeocin and subjected to single cell cloning. In total, clones from 3 distinct libraries were selected to cover the 3 mutant *lox* sites used in this study (FAS, 2272 and 5171). We surveyed multiple clones for each *lox* site, allowing us to deconvolute the importance of integration site and *lox* sequence on recombination efficiency. Collectively, 5 clones from the *lox*FAS pool, 5 with *lox*5171, and 8 with *lox*2272 sequence were selected to test swapping and excision efficiency.

Figure 4.2: A dual fluorescent screen to determine Cre-mediated swapping and excision



An example of flow cytometry patterns for the dual fluorescent screen are shown here. **A.** Prior to transfection with the Strawberry construct, high levels of GFP expression are seen (most cells in Q1) and the non-fluorescent population is small (Q4). **B.** After transfection with the Strawberry construct and Cre vector, a portion of the population expresses the Strawberry protein (Q2 and Q3).

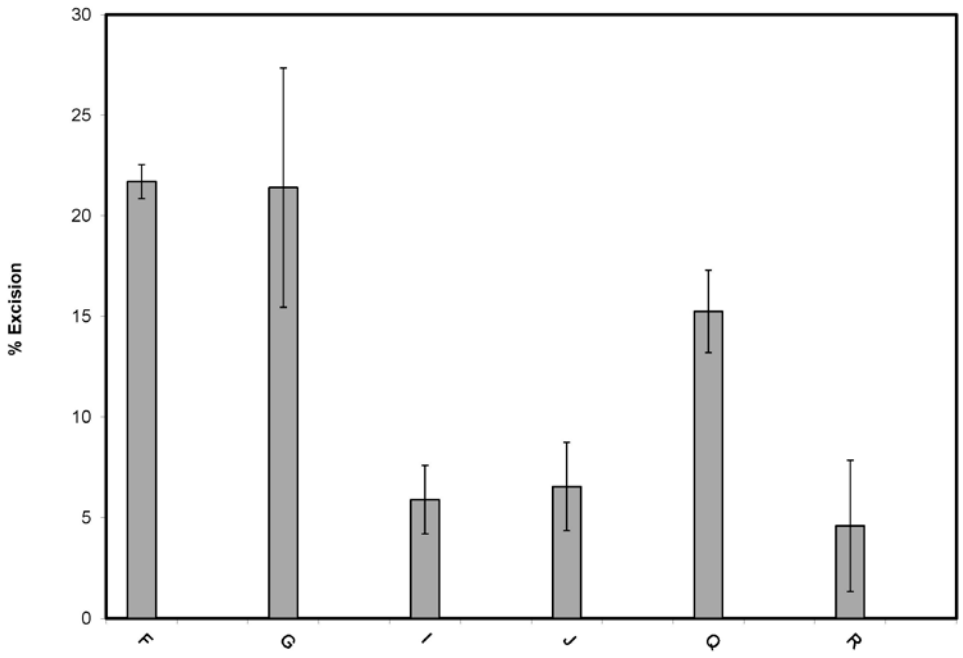
Figure 4.2 (continued)



C. A flow chart illustrates how swapping and excision frequencies were measured. **D.** A southern blot was conducted on clone R (described in Materials and Methods) to validate the dual fluorescent assay and demonstrate swapping. Digested DNA was loaded in the following order: GFP positive (1), Strawberry-expressing sorted population (2), Strawberry and GFP-expressing, sorted population (3), and GFP positive clone R transfected with the Strawberry construct but no Cre recombinase (4). The left half of the membrane was exposed to the mStrawberry probe, and the right half to the GFP probe. The two bands present for samples 1, 3 and 4 on the right indicate two integrations of the GFP construct. The prominent band for sample 2 on the left (and lack of the equivalent band on the right) indicates site-specific swapping of the fluorescent constructs.

A second transgene construct containing the Strawberry gene (Figure 4.1) was used to visualize recombination events in 18 GFP-expressing cell lines. Specifically, we attempted to replace the GFP construct with the Strawberry construct bearing the same *lox* site combinations for each clone. Clones underwent a co-transfection using 30μg of Strawberry construct and 5μg of Cre construct, and swapping and excision rates were simultaneously measured using a dual fluorescent screen (Figure 4.2). The reported values correct for false positives using a control reaction, conducted for each cell line and condition (Figure 4.2). This dual fluorescent assay was validated for site-specific swapping activity in one clone using a Southern Blot and P^{32}-labeled probes specific to the GFP and Strawberry genes (Figure 4.2). In particular, these results suggest that, even for a cell line with multiple integrations, the cell population in quadrant 3 represents site-specific swapping. Moreover, under the conditions of this study, random integration is negligible and the dual fluorescence region (quadrant 2) was a result of transient expression of Strawberry. We conjecture that the presence of a single Strawberry band indicates that clone R contains only one full (or functional) copy of the GFP construct, and the second integration could contain mutations or be lacking one of the *lox* sites necessary for targeting. In the absence of a swapping target, transfection of Cre vector alone results exclusively in excision (Figure 3).

Figure 4.3: Without a swapping target, Cre recombinase results exclusively in excision.

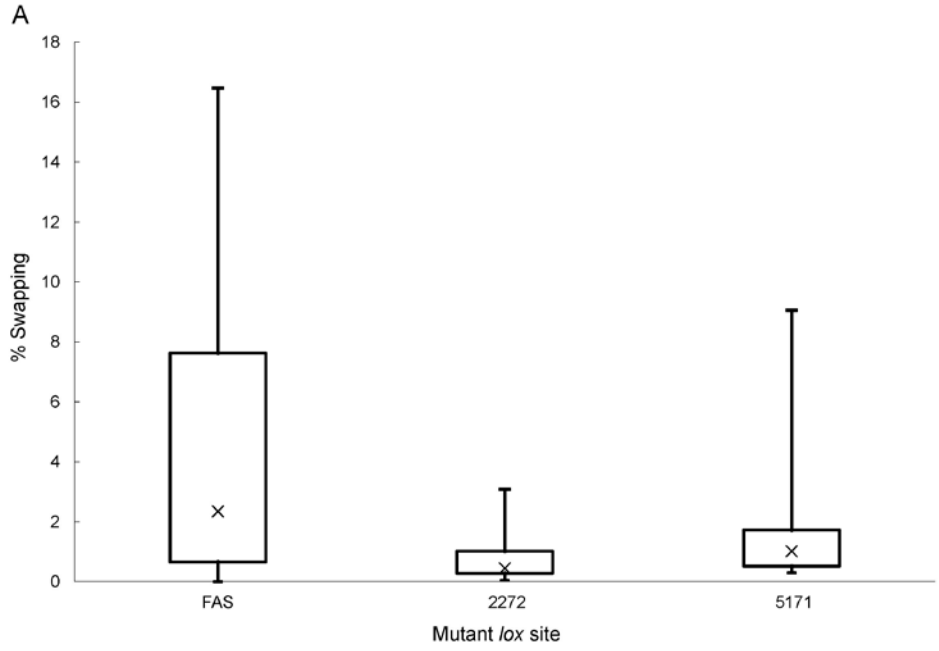


Excision was measured for 6 different cell lines representing three lox pairings in duplicate. GFP expressing cell lines were transfected with 10 µg of a Cre recombinase vector and compared to untransfected cell lines 72 hours later. In the absence of a swapping target, Cre recombinase results exclusively in excision, which varied from 5-20%.

Significant variation was seen in clone-to-clone swapping and excision efficiency (Figure 4.4) indicating that integration locus impacts the accessibility and effective activity of Cre recombinase. This result echoes prior work demonstrating that chromatin state can influence recombination efficiency[175,184]. Despite this variation, the *lox*FAS-*loxP* pairing exhibited statistically significantly higher swapping efficiency compared to either the *lox*2272-*loxP* ($P<0.05$) and *lox*5171-*loxP* ($P=0.05$) pairings. Under these initial test conditions, the *lox*FAS-*loxP* pairing had a median swapping efficiency of 2.35%, whereas the *lox*2272-*loxP* and *lox*5171-*loxP* pairings have median efficiencies of 0.44 and 1.01% respectively (Figure 4.4). Excision efficiencies had a similar trend with median values of 9.21, 1.43 and 3.25% for the *lox*FAS, *lox*2272 and *lox*5171-*loxP* pairings respectively (Figure 4.4). These results implicate the *lox*FAS site as a novel

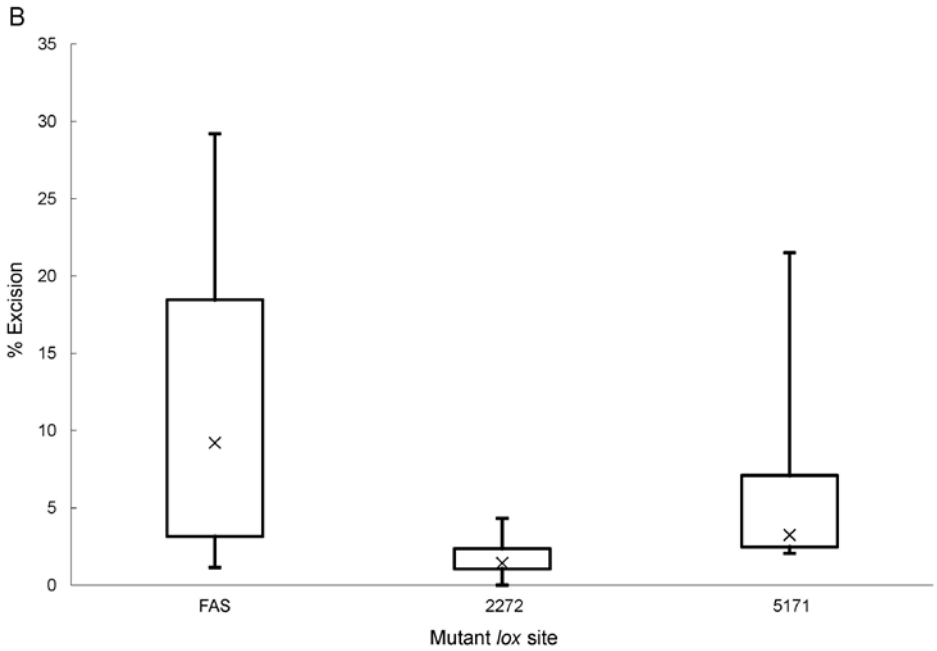
mutant site for mammalian cell recombination. Moreover, it demonstrates that heavily mutated *lox* sites can potentially be more efficient than sites closer to wild-type sequence. On the basis of these statistically higher recombination frequencies, we selected the *lox*FAS-*loxP* pairing for further study.

Figure 4.4: Swapping and excision rates for three mutant *lox*-*loxP* pairings.



Box and whisker plots are used to depict the distribution of swapping and excision efficiencies measured using 18 different cell lines, 8 containing a *lox*2272 site, 5 containing a *lox*5171 site and 5 containing a *lox*FAS site. **A.** Swapping efficiencies varied significantly regardless of mutant lox site, indicating the importance of integration locus on site accessibility. Despite variability, *lox*FAS exhibited statistically significant ($P<0.05$) higher swapping than *lox*2272.

Figure 4.4 (continued)



B. A similar trend was observed regarding excision efficiencies, with *lox*FAS statistically outperforming both *lox*2272 and *lox*5171 ($P<0.05$).

4.3.2 Sequential introduction of target DNA and Cre increases swapping efficiency

In nearly all prior reports, the Cre vector and swapping cassette are co-transfected, resulting in lower than desired swapping and high levels of excision. Mechanistically, a swapping event requires both Cre recombinase and a target cassette (here, Strawberry construct) to be present at the chromosomal integration locus of the first cassette (here, GFP construct). From a probabilistic standpoint, this three-body event is less likely to occur than excision, which only requires Cre recombinase to interact with the integrated cassette. Cre recombinase is known to be actively transported to the nucleus[80,200] and is relatively small in size, thus, we hypothesized that transport of the Strawberry construct was limiting swapping efficiency. This tri-molecular mechanistic problem is not limited to Cre recombinase, and applies generically to all integrases, recombinases and transposons performing transgene swapping at a specific locus. For these reasons, we

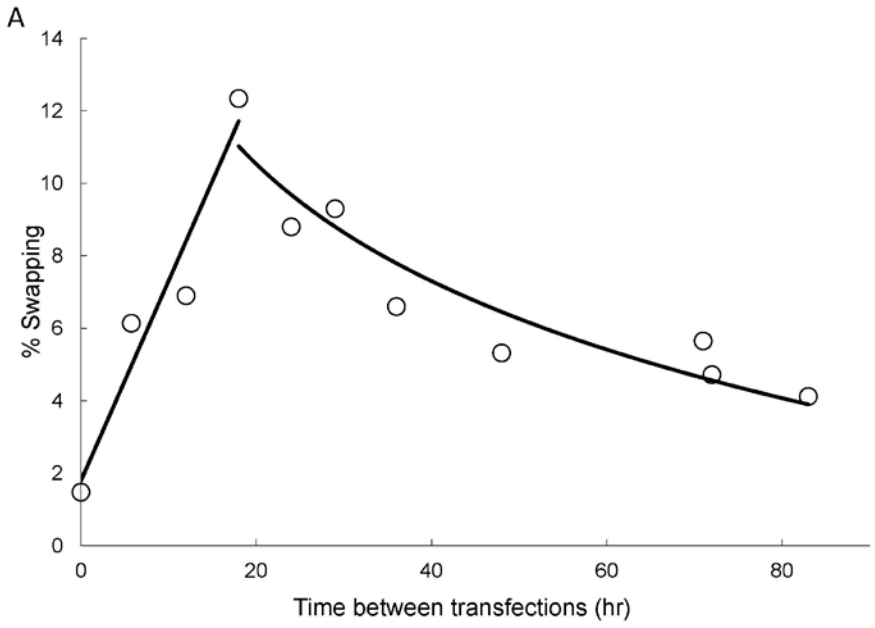
examined the impact of a sequential transfection scheme in which the target DNA was transfected first to allow more time for nuclear transport prior to introduction of Cre recombinase. Twenty-five μg of linear, Strawberry construct (bearing the loxFAS-loxP pairing) was transfected at time zero, followed by 5μg of the Cre vector at a later time ranging from 0 to 83 hours. Swapping and excision efficiency were measured as a function of transfection gap time and compared to a control (Figure 4.2c).

We observed that delaying transfection of the Cre vector, and therefore Cre recombinase expression, increases swapping efficiency (Figure 4.5). While we believe this phenomena applies generically to any *lox* pairing, we chose to study the impact of delayed introduction of Cre with two *lox*FAS clones in an effort to achieve the highest possible swapping efficiencies. The experiments indicate an optimal time of 18 to 24 hours between transfections, where swapping efficiencies could be increased from a median of 2.35% to values between 8.2 and 12.3%, all without antibiotic selection. We also observed the lowest efficiencies when the Cre vector and Strawberry construct were co-transfected, strongly indicating that a time delay of Cre addition (even as great as 83 hours) improves swapping. This is an important observation, and suggests that for optimal results, a swapping cassette and Cre vector should not be simultaneously introduced.

We observed that swapping is controlled by two regimes: an early, linear phase in which we hypothesize transport of the exchange cassette is limiting, and a later, logarithmic decay phase in which DNA degradation and dilution due to cell divisions dominate. We tested two separate cell lines, and following maximal swapping, observed the same decay constant (-4.76 ± 0.14 1/hr). This indicates a conserved rate process within the cell (namely, growth dilution and DNA/protein decay). The rate of swapping during early times, however, varied between the two samples (0.19 and 0.55%

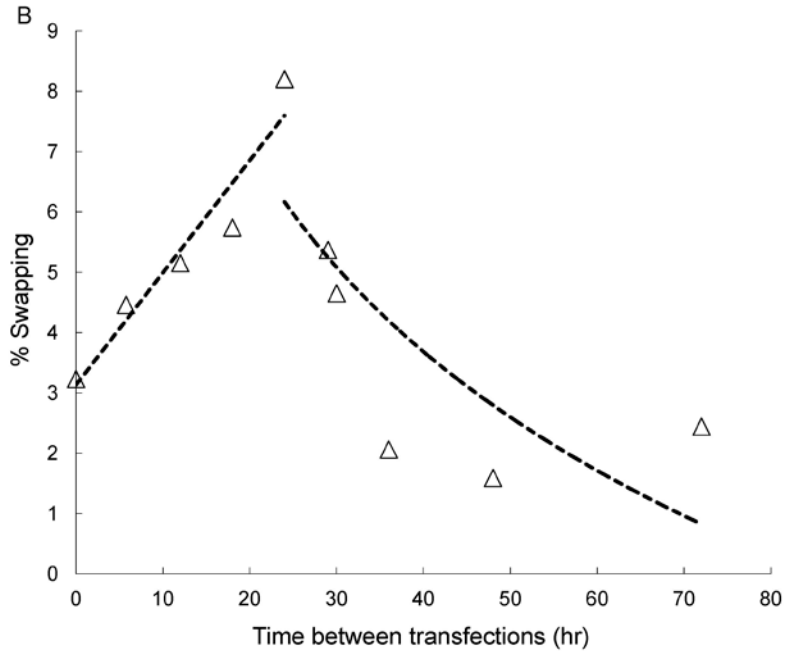
swapping/hr). This variation in slope reflects the variation in swapping efficiency expected at different chromosomal loci. A higher initial swapping rate in this linear phase likely indicates a more accessible chromosomal location. These observations are shown in Figure 4.5 for two cell lines.

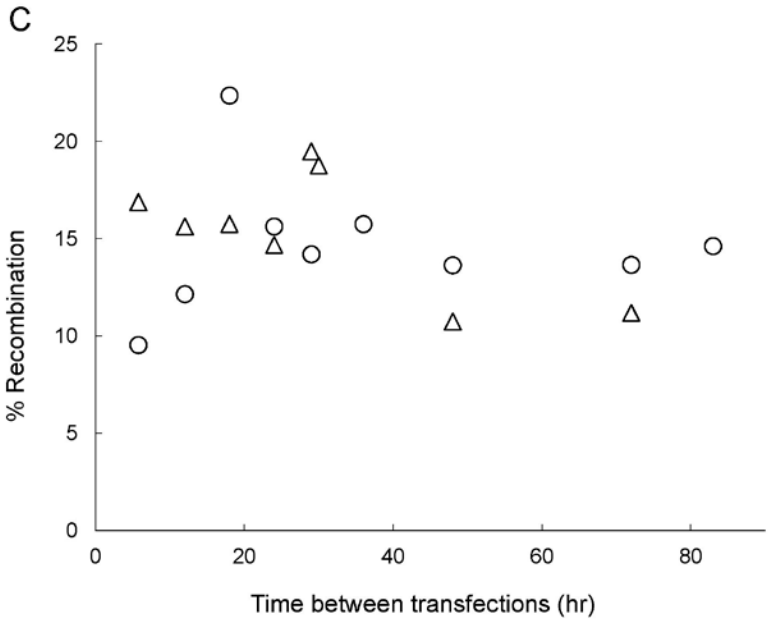
Figure 4.5: Delayed introduction of Cre DNA improves swapping efficiency.



The impact on swapping of delayed transfection of Cre DNA was measured for two loxFAS clones (**A, B**). In both cases (○ = R, Δ = Q), maximal swapping efficiency occurs between 18 and 24 hours.

Figure 4.5 (continued)





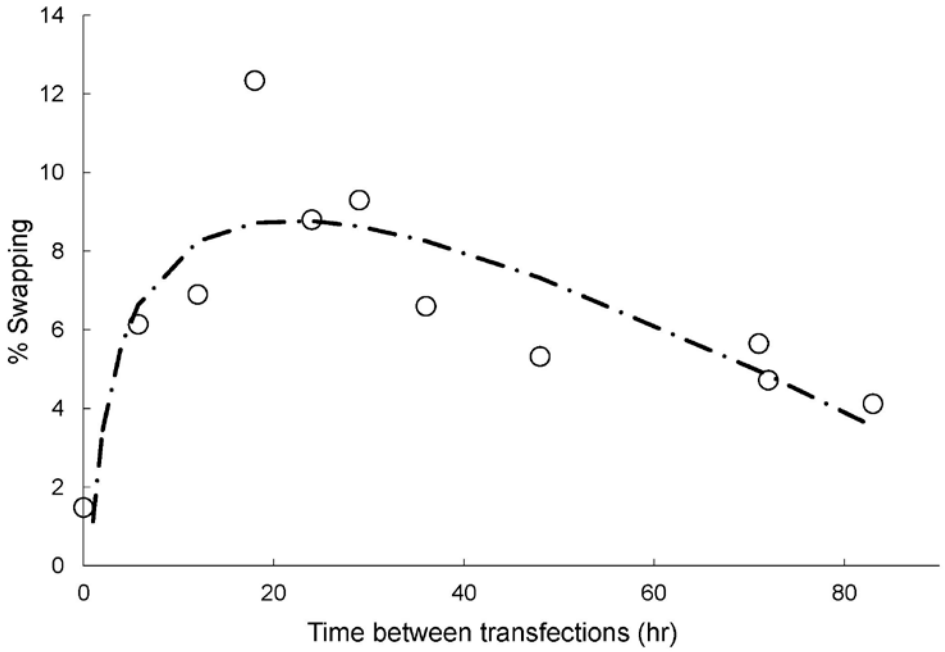
C. Total recombination (excision and swapping) was measured as a function of delayed transfection of Cre DNA. While total recombination is relatively constant, the proportion of swapping to excision decreases with time.

We also examined total Cre recombination events (excision and swapping together), compared to time between addition of the Strawberry construct and Cre vector (Figure 4.5). Total recombination efficiency was relatively constant for each cell line with a median value of 14.4%. This indicates that while the propensity for swapping versus excision activity can be influenced by sequential transfection, the total Cre activity is fixed for a specific amount of Cre vector transfected. As time between transfections increases, the ratio of swapping to excision activity decreases, likely due to degradation of the Strawberry construct and dilution by cell divisions.

4.3.3 Modeling the delayed introduction of Cre DNA to improve swapping efficiency

Based on these heuristic experimental observations, we developed a mathematical model that captures both the initial, linear time dependence in which swapping efficiency increases, followed by the later, logarithmic decay of swapping efficiency. This behavior can be mathematically described: $\mathbf{S = At + Bln(t) + C}$ where S is swapping efficiency, t is time between the two transfection events (in hours), and A, B and C are constants. These parameters are cell line specific and using best fit analysis, we determined these values for one cell line with a *lox*FAS-*loxP* pairing, for which A is -0.164, B is 3.593 and C is 1.284. This model of time-dependent swapping is shown in Figure 4.6 compared to corresponding experimental data.

Figure 4.6: Time-dependent swapping behavior can be mathematically modeled.



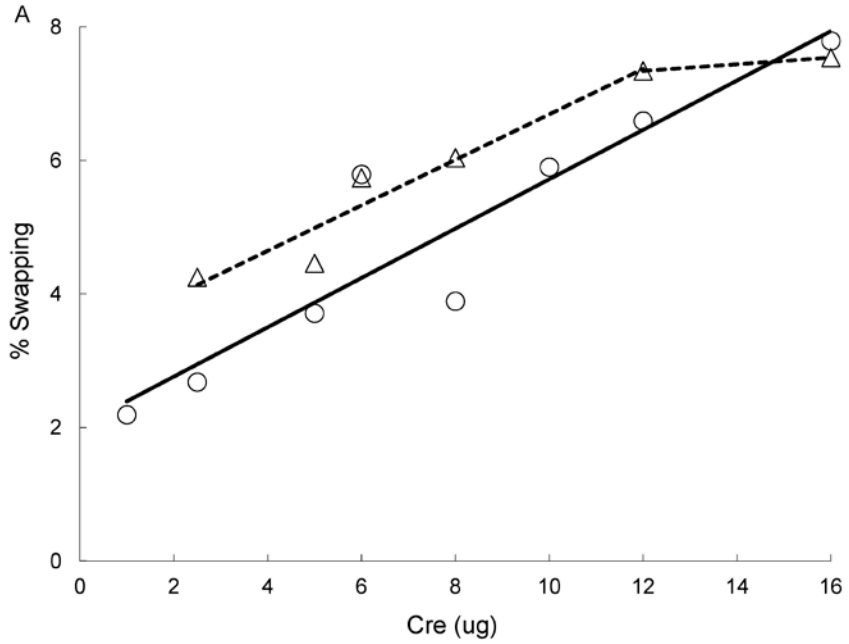
This behavior can be mathematically described, where swapping, S = At + Bln(t) + C. For one cell line (○), these parameters were fit as S = -0.164 t + 3.593 ln(t) + 1.284 and are shown compared to corresponding experimental data.

4.3.4 Determining the optimal ratio of Cre and target DNA

There is evidence that in the absence of a *lox* target site, expression of Cre recombinase can cause chromosomal rearrangements and DNA damage[180,182], thus using minimal amounts of Cre is desirable. This motivated us to examine the impact of the ratio of swapping cassette to Cre expression vector. To evaluate this phenomenon in the sequential transfection setting, 25μg of the Strawberry construct was transfected at time zero, and 23 hours later, the Cre vector was transfected in amounts ranging from 1 to 16 μg. Swapping and excision efficiency were measured and are shown in Figure 4.7. We observed that both increased linearly with the amount of Cre transfected. This phenomenon is likely due to increased Cre recombinase transfection and expression efficiency. Interestingly, the rate of increase in swapping efficiency does not exhibit site

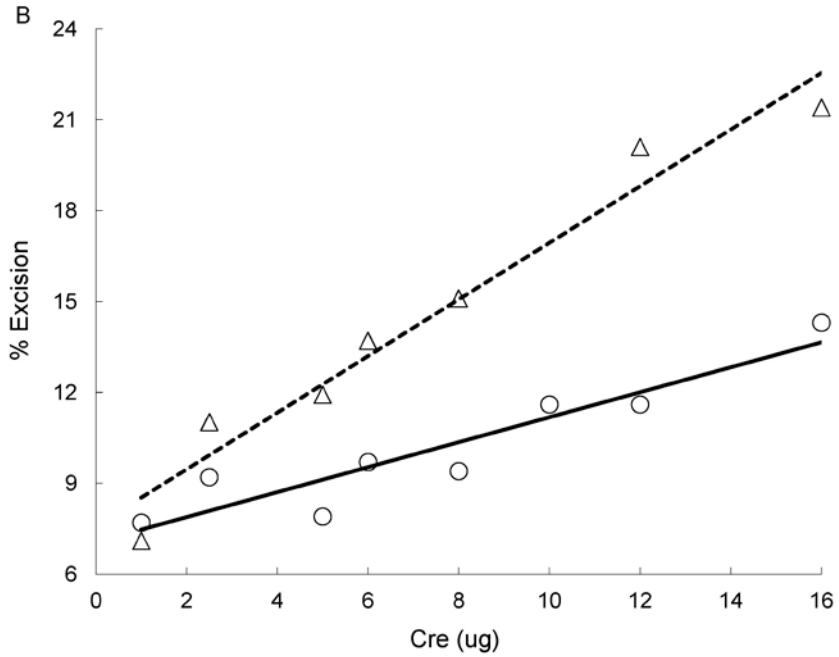
dependence, as both cell lines have a rate of approximately 0.35% recombination per μg Cre transfected. Excision efficiency does exhibit site dependence, with the rate of one clone being higher than the other (0.93 and 0.41% excision per μg Cre respectively). This difference illustrates the variability in recombination accessibility of cell lines, but that with swapping (a three-body event), many factors contribute to rate dependence. However, in the case of excision, this process is mainly guided by the accessibility of chromosomal sites and Cre recombinase levels.

Figure 4.7: Increased quantities of Cre DNA improves net recombination.



The impact of increasing Cre DNA levels on swapping (**A**) and excision (**B**) was measured for two clones (○ = R, Δ = Q). **A**. Swapping increased linearly with Cre DNA, with a similar slope for both clones tested. One of the clones (Δ) exhibited a maximum increase in swapping efficiency at 12μg of Cre vector.

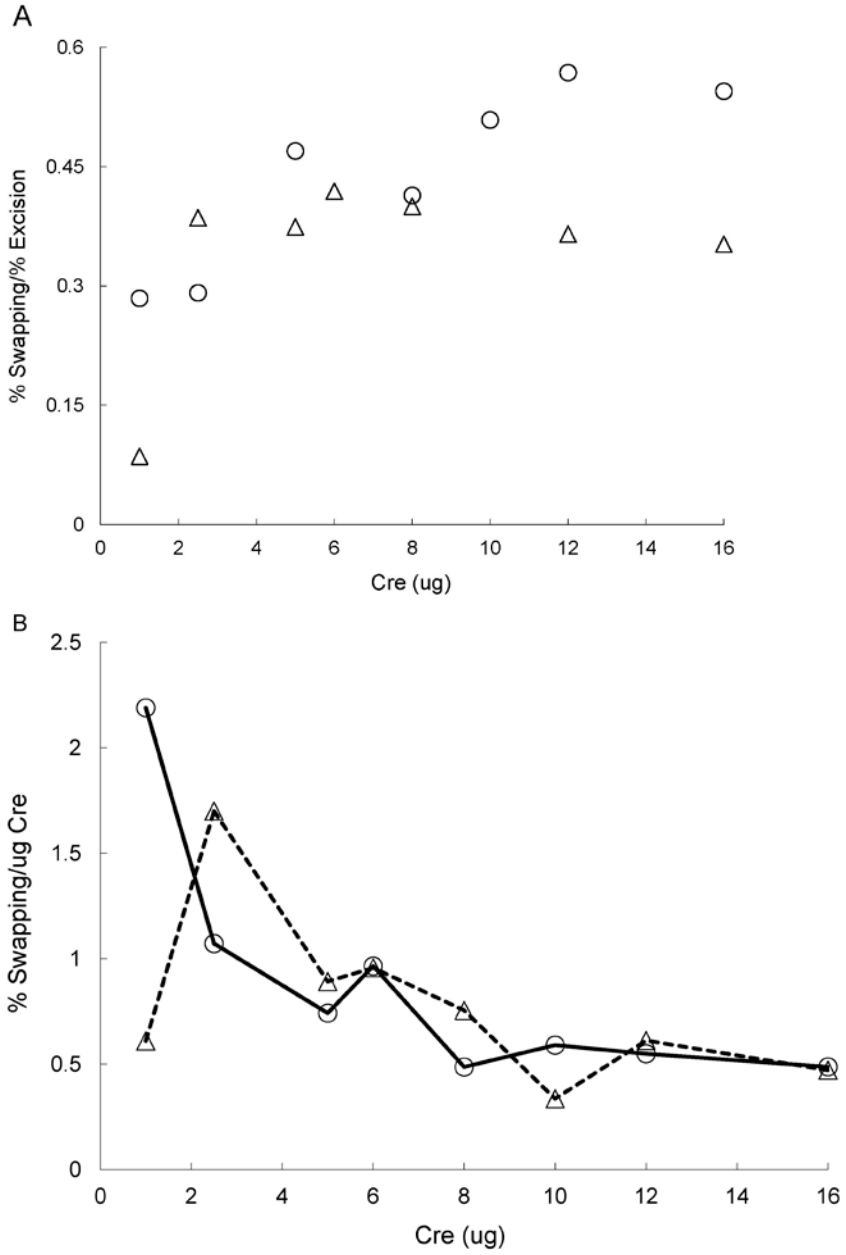
Figure 4.7 (continued)



B. Excision also increased linearly with Cre DNA, and slope varied for the two clones (○ = 0.41, Δ = 0.93).

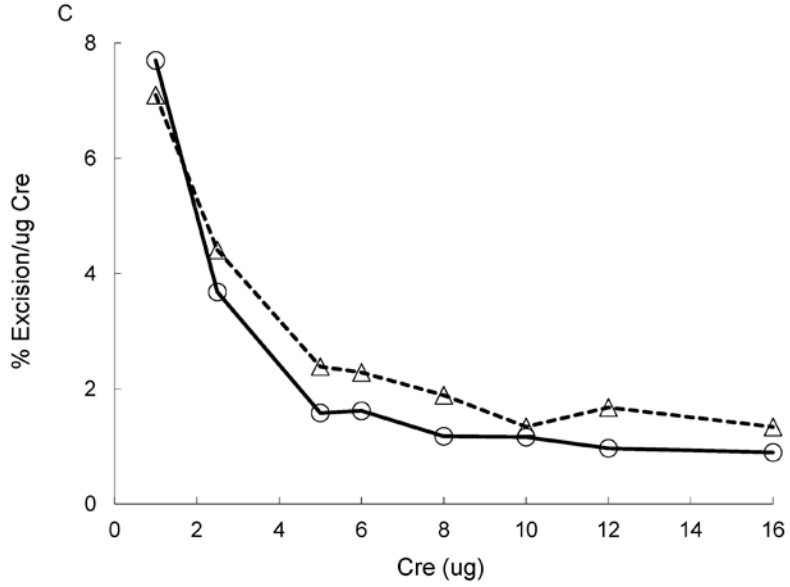
While adding more Cre vector increases net recombination, it does not result in improved or biased swapping efficiency. In fact, as more Cre vector is transfected, the ratio of swapping to excision remains fairly constant (Figure 4.8). This indicates that increasing the amount of transfected Cre vector does not favor either excision or swapping activity, only the total amount of recombination events that occur within the cell. We normalized recombination efficiency with respect to the amount of Cre vector transfected (Figure 4.8), and observed that increasing the amount of Cre transfected has a diminishing impact on both swapping and excision, and per microgram of DNA, recombination efficiency actually decreases. These results were observed in both cell lines tested and thus are independent of integration locus.

Figure 4.8: Increased transfection of Cre vector does not improve swapping compared to excision activity.



The impact of transfected Cre DNA on excision and swapping was measured for two clones (○, Δ). Twelve million cells were transfected with 25µg linear Strawberry DNA. Twenty-three hours later, Cre DNA was transfected in amounts ranging from 1 to 16µg. **A**. While transfecting more Cre vector increased net recombination, it does not preferentially increase either swapping or excision efficiency. The ratio of swapping to excision activity is relatively constant for Cre vector levels above 2.5µg. **B, C**. Swapping and excision were normalized per µg of Cre vector, illustrating that the rate of both events significantly decreases as more Cre is added to the system.

Figure 4.8 (continued)



4.4. CONCLUDING REMARKS

The Cre/*lox* system is an important site-specific genome editing tool with significant applications in human cell lines. Despite wide adoption of Cre recombinase, there has previously been a lack of research simultaneously examining many of the variables shown to influence Cre activity. As a result, swapping efficiencies have typically been reported as being extremely low (<1%). This work identifies a set of optimal parameters that resulted in the highest swapping efficiencies ever reported (upwards of 12%). We identify the *lox*FAS-*loxP* pairing as a better choice, compared to either the *lox*5171 or *lox*2272-*loxP* pairings, and we implicate the importance of delayed introduction of Cre DNA for optimal swapping efficiency. As many other genome editing enzymes require trimolecular interactions, these findings can be extended to other recombinases and integrases. The swapping frequencies identified here can greatly improve the prospects of using this genome editing tool in mammalian cell systems.

Chapter 5: A Method for Condition-Specific Codon Optimization for Improved Heterologous Gene Expression

5.1: CHAPTER SUMMARY

Heterologous gene expression is an important biotechnology tool that enables metabolic engineering and the production of non-natural biologics in a variety of host organisms. The translational efficiency of heterologous genes can often be improved by optimizing synonymous codon usage for the host organism. Traditional approaches for optimization neglect to take into account many factors known to influence synonymous codon distributions. Here we define an alternative approach for codon optimization that utilizes systems level information and codon context for the condition under which heterologous genes are being expressed. Furthermore, we utilize a stochastic algorithm to generate multiple variants of a given gene. We demonstrate improved translational efficiency using this condition-specific codon optimization approach with two heterologous genes, eGFP and CatA, expressed in *S. cerevisiae*. Gene variants were optimized for conditions of high expression and stationary phase expression. Compared to wild-type genes and traditional approaches, we observe improved protein expression, up to three times higher, for both genes.

5.2: INTRODUCTION

Codon optimization, which is a rational redistribution of synonymous codons in a gene sequence, can result in improved translational efficiency and gene expression[116,120,121]. Because different organisms exhibit diverse codon usage, codon optimization has emerged as a powerful tool to improve heterologous gene expression. In doing so, it can often relieve pathway bottlenecks and improve overall flux, making this approach a critical part of both metabolic and cellular engineering.

Typical approaches for codon optimization utilize whole genome information to determine which synonymous codons are rare and should be replaced or abundant and should be used frequently. While this methodology has resulted in some significant successes regarding gene expression, it can often result in gene variants with lower expression levels than wild-type sequences[129-132]. These failures are problematic and likely result from an oversimplified approach to codon optimization. In particular, traditional approaches fail to take into account changes in tRNA abundance known to result from changes in environmental factors including growth condition and cell-cycle[133-135,201]. Furthermore, despite the fact that much of an organism's protein coding genes are lowly expressed and minimal evolutionary pressure has been present to drive efficient natural evolution[118,202], the traditional optimization approach assumes that using all of a genome's protein coding information, as opposed to a subset, provides the best information for codon optimization. Finally, traditional approaches examine each codon individually, as opposed to adjacent codon pairs (known as codon context), the importance of which was recently demonstrated[203-205].

We hypothesize that codon optimization would be improved by evaluating synonymous codons using a subset of the total protein coding genes in a stochastic design that takes codon context into account. In particular, codon usage would be determined using only those genes upregulated under a specific environmental condition. We refer to this alternative approach as 'condition-specific codon optimization' and propose that heterologous genes are optimized using codon usage information corresponding to the environmental conditions under which the host organism will be grown and the gene will be expressed. Furthermore, we utilize a stochastic approach that incorporates codon context into optimized gene design. The method for this condition-specific codon

optimization, as well as several applications to heterologous gene expression, are described herein.

5.3: Results and Discussion

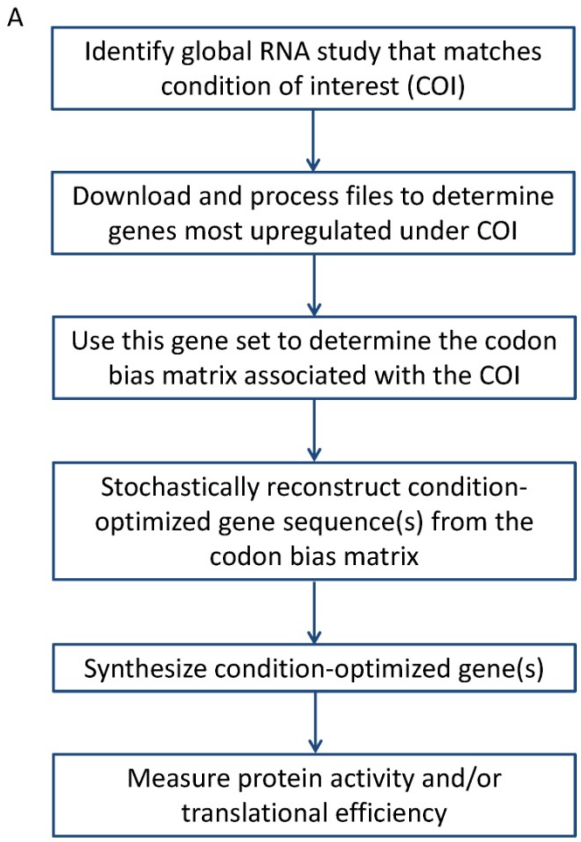
5.3.1: Developing a Condition-Specific Codon Usage Bias

The distribution and frequency of synonymous codons is often referred to as the codon usage bias (CUB). Traditional methods for determining CUB and codon optimizing a heterologous gene rely on information obtained by either all protein-coding genes or a subset of protein-coding genes. Alternatively, we propose that a CUB take into account the specific conditions under which the gene will be expressed. This approach is termed condition-specific codon optimization and utilizes those genes most upregulated under the growth condition of interest for the CUB. By doing so, the charged tRNA levels which fluctuate in response to environmental conditions are indirectly taken into account, resulting in improved gene expression over current methods. The steps for generating a condition-specific CUB are outlined in Figure 5.1a.

Before the CUB is generated, the condition(s) under which the heterologous gene of interest will be expressed must be identified and global expression data for the host should be obtained under the condition. Studies including but not limited to RNA microarray, RNA-seq and proteomics can be used. Thousands of such studies have previously been conducted and results can be freely accessed using databases such as Gene Expression Omnibus (GEO), the Center for Information Biology Gene Expression database (CIBEX) and Array Express. However, if a study has not previously been conducted for conditions of interest, a global RNA-seq experiment can be conducted and used as the starting point for condition-specific codon optimization. This has an

advantage over CUTG-based codon optimization, which requires a fully sequenced genome, therefore limiting cellular host options.

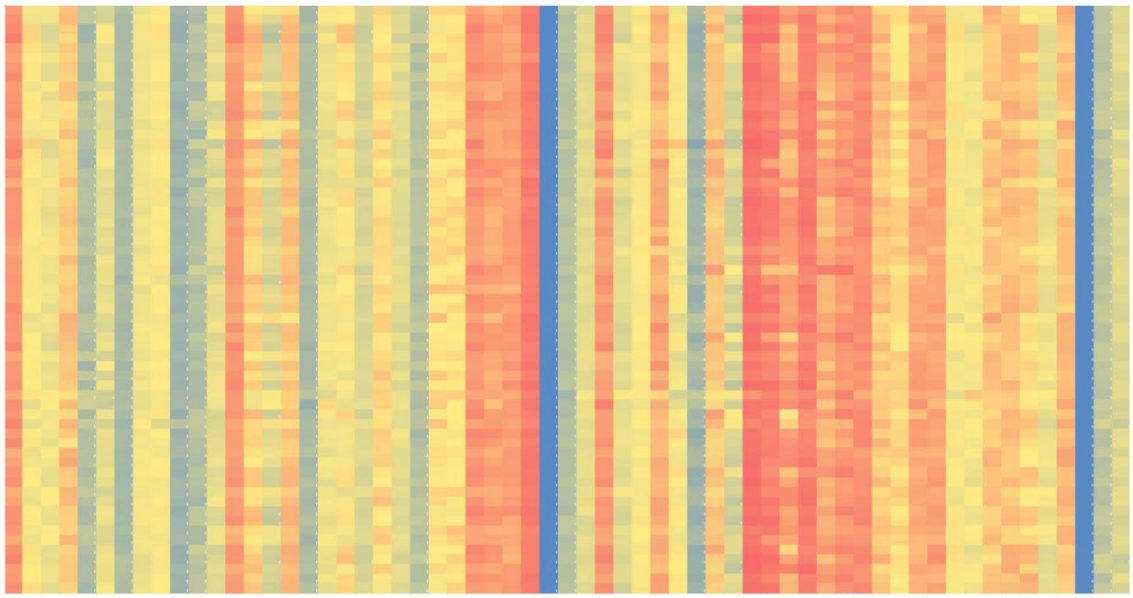
Figure 5.1: Condition-specific codon optimization utilizes systems level information and codon context



We utilize condition-specific gene expression information and codon context to generate optimized gene sequences in a stochastic manner. **A.** The steps taken to apply this approach for a given condition are outlined in this flow chart.

Figure 5.1 (continued)

B



B. The control codon matrix is compiled from 6,666 protein-coding genes in *S. cerevisiae* and serves as a point of comparison for condition-specific matrices. The first amino acid is indicated by the first column, and the second amino acid by the first row. The color indicates probability between 0 (red) and 1 (blue).

The global expression data set can then be analyzed to determine genes that are differentially upregulated under the condition of interest, as compared to a control condition. For the conditions examined in this study, we selected the top fifty to one hundred most differentially upregulated genes. Using the corresponding DNA sequences, codon frequency and probability can be determined for both individual codons and codon pair usage, or codon context. Previous studies suggest that codon context may be more important to gene optimization than individual codons[203-205] and that codon context directly correlates with translation elongation rate[206]. In particular, steric hindrance of charged tRNAs for adjacent codons can be avoided by taking adjacent codon pairing into account[204]. We generate a 'condition-specific codon usage table' with individual codon information, and keeping the importance of codon context in mind, a 'condition-specific

codon usage matrix' with codon context information. The python script used to generate the condition-specific tables and matrices is entitled 'CodonUsageBias' (Appendix D).

For comparison, we generated a control table and codon context matrix (hereafter referred to as the control matrix), which were assembled using the protein coding sequences of 6,666 *S. cerevisiae* genes. The control table is identical to the CUTG from GenBank for *S. cerevisiae,* which is used commercially. The control matrix is shown in Figure 5.1b with the y-axis representing the first codon and the x-axis representing the second codon in a pair. Each square represents the probability of a codon pair occurring given that the first codon is specified. The matrix has been gradient-colored such that blue represents a probability of one and red represents a probability of zero. The two solid blue columns correspond to a second codon of ATG (methionine) and TGG (tryptophan). Because there are no synonymous codons for these amino acids, the ATG and TGG codons will be used 100% of the time these amino acids are incorporated. We see that amongst the synonymous codons, this is not the case and some are used more frequently than others. Interestingly, we see very little diversity amongst each second codon regardless of the preceding codon, which can be determined by examining the columns of the matrix. Although some columns are shades of red (low probability) or yellow (medium probability) or blue (high probability), the color tends to be very consistent. This can be contrasted with the matrices made for specific conditions (Figures 5.2, and 5.4) where significant variation can be seen in nearly every column. Because this control matrix incorporates codon context for nearly all protein coding genes in *S. cerevisiae*, the probability values are an average of codon usage across the entire genome. As a result, the columns become indicative of the frequency of each codon, with the rarest indicated in red.

The condition-specific table and matrix can then be used as a starting point for optimization of heterologous genes that will be expressed under the condition of interest. In the simplest case, the condition-specific table can directly replace the standard table currently used for codon optimization and the most frequently occurring codon should be incorporated for each amino acid. Alternatively, DNA sequence can be optimized by considering each adjacent codon pairing in the chain. In this case, we elect to stochastically reconstruct the DNA sequence from the protein sequence utilizing the codon context probabilities stored in the condition-specific matrix and an algorithm developed in house.

This stochastic methodology is achieved using the python script entitled 'GeneDesigner' (Appendix D). Specifically, if the first two amino acids are methionine followed by cysteine, there are two possible corresponding DNA sequences: ATGTGT or ATGTGC. The condition-specific matrix stores the probability that each DNA sequence occurred in the genes upregulated for that condition. GeneDesigner selects the DNA sequence based on the corresponding probability. For example, if 60% of Met-Cys pairs are ATGTGT and only 40% are ATGTGC, GeneDesigner will stochastically select ATGTGT 60% of the time and ATGTGC 40% of the time that a Met-Cys pair is present in the peptide sequence. This approach allows us to easily generate many versions of a single gene, all with the same protein sequence. Furthermore, this approach biases the DNA sequences for codon pairs with a higher frequency, takes into account codon context, and minimizes the presence of rare codons while maintaining codon diversity. This is an important advantage because exclusive use of specific codons can result in bottlenecks with the formation of charged tRNA-amino acid complexes, thereby reducing translational efficiency[201].

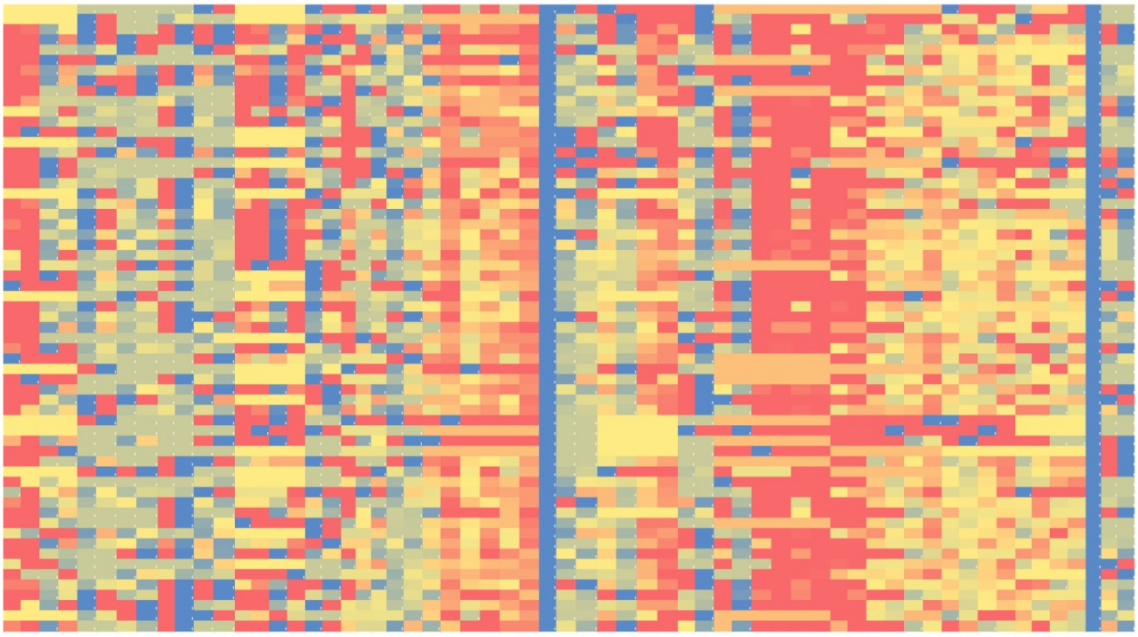
After GeneDesigner has been used to generate one or more condition-specific codon optimized sequences, the corresponding DNA can be synthesized and introduced into an expression system for the cellular host of interest. In the case where several sequences are used, a metric can then be applied to experimentally determine which variant performs the best. This can be done by directly measuring translational efficiency, measuring protein production or measuring crude protein activity. As a point of comparison, activity can be compared to gene variants optimized using the control matrix.

The steps for this condition-specific codon optimization are outlined in Figure 5.1a. In order to validate our hypothesis and this approach, we selected three heterologous genes of interest, applied the approach as outlined above, expressed the resulting optimized genes in *S. cerevisiae*, and measured expression.

5.3.2: Condition-specific optimization of eGFP for high expression outperforms wild-type and control variants

The first condition we sought to codon optimize for was constitutive high expression in *S. cerevisiae*. Nearly all codon optimization efforts to date in yeast have sought to optimize codon usage using rules derived from all of the protein coding genes in the genome. This is problematic because a large majority of the *S. cerevisiae* genome, as with other eukaryotic genomes, is lowly expressed[207]. Considering the scientific community is often interested in expressing heterologous genes constitutively and at the highest possible expression level, we sought to determine an alternative CUB that could be used for codon optimization.

Figure 5.2: High expression codon optimization matrix



The high expression codon optimization matrix is compiled from the 100 most highly expressed protein-coding genes in *S. cerevisiae*[207]. The first amino acid is indicated by the first column, and the second amino acid by the first row. The color indicates probability between 0 (red) and 1 (blue). The Frobenius norm of the difference between the control matrix and this high-expression matrix is 14.48.

The condition-specific table and matrix were assembled as described above and in the Materials and Methods using the 100 most highly expressed yeast genes when grown on YPD media[207]. The resulting table and matrix are shown in Figure 5.2. An *E. coli* optimized green fluorescent protein (eGFP), which is poorly expressed in *S. cerevisiae*, was selected as a reporter protein. Eight sequence variants of eGFP were generated and compared to the wild-type sequence. One variant was optimized using the control table (Table 5.1), one variant using the high expression table (Table 5.2), three variants using the control matrix (Figure 5.1b) and three variants using the high expression matrix (Figure 5.2). The sequences can be found in Appendix C. Each variant, including wild-type eGFP, was inserted into the p41K-GPD yeast expression vector and transformed into BY4741[44]. Three replicates of each variant were grown on YPD media and fluorescence

was screened in mid-log phase. The results of this test are shown in Figure 5.3 in grey. We observed that the eGFP variants generated using the high expression matrix are statistically better expressed than those variants generated using the control matrix (pvalue = 6.08e-6) and wild-type eGFP (pvalue = 1.36e-6). This demonstrates that optimizing codon usage specifically for high expression was effective. While all three of the high expression matrix-generated variants outperform the wild-type eGFP, only two of the three control matrix-generated variants outperform wild-type eGFP.

Table 5.1: Control codon usage table

Amino Acid	Codon	Count	Frequency per 1000	Fraction
Ala	GCG	18708	6.22	0.11
	GCA	48873	16.25	0.30
	GCT	60549	20.13	0.37
	GCC	36389	12.1	0.22
Cys	TGT	24477	8.14	0.62
	TGC	15123	5.03	0.38
Asp	GAT	112614	37.44	0.65
	GAC	60473	20.1	0.35
Glu	GAG	58013	19.29	0.30
	GAA	135489	45.04	0.70
Phe	TTT	80562	26.78	0.59
	TTC	54877	18.24	0.41
Gly	GGG	18301	6.08	0.12
	GGA	33737	11.22	0.23
	GGT	67010	22.28	0.45
	GGC	29538	9.82	0.20
His	CAT	41902	13.93	0.64
	CAC	23507	7.81	0.36
Ile	ATA	55646	18.5	0.28
	ATT	90225	29.99	0.46
	ATC	51005	16.96	0.26
Lys	AAG	90474	30.08	0.41
	AAA	127843	42.5	0.59
Leu	TTG	79567	26.45	0.28
	TTA	79072	26.29	0.28

Table 5.1 (continued)

	CTG	32360	10.76	0.11
	CTA	40619	13.5	0.14
	CTT	38311	12.74	0.13
	CTC	17438	5.8	0.06
Met	ATG	62753	20.86	1
Asn	AAT	109078	36.26	0.60
	AAC	73921	24.57	0.40
Pro	CCG	16471	5.48	0.12
	CCA	53446	17.77	0.41
	CCT	40872	13.59	0.31
	CCC	20970	6.97	0.16
Gln	CAG	37106	12.34	0.32
	CAA	80576	26.79	0.68
Arg	AGG	28378	9.43	0.21
	AGA	62777	20.87	0.47
	CGG	5765	1.92	0.04
	CGA	9673	3.22	0.07
	CGT	19062	6.34	0.14
	CGC	8225	2.73	0.06
Ser	AGT	43985	14.62	0.16
	AGC	30320	10.08	0.11
	TCG	26613	8.85	0.10
	TCA	57738	19.19	0.21
	TCT	70764	23.53	0.26
	TCC	42720	14.2	0.16
Thr	ACG	24689	8.21	0.14
	ACA	54561	18.14	0.31
	ACT	60493	20.11	0.34
	ACC	37621	12.51	0.21
Val	GTG	32469	10.79	0.19
	GTA	36737	12.21	0.22
	GTT	64463	21.43	0.38
	GTC	33842	11.25	0.20
Trp	TGG	31313	10.41	1
Tyr	TAT	57528	19.12	0.57
	TAC	43710	14.53	0.43

A control codon usage table was generated from the sequences of 6,666 protein-coding genes in *S. cerevisiae* using the CodonUsageBias python script. From these sequences, we are able to determine for all 64 codons the total count of each codon, frequency of occurrence per 1000 codons, and probability amongst synonymous codons (fraction).

Table 5.2: High expression codon usage table

Amino Acid	Codon	Count	Frequency per 1000	Fraction
Ala	GCG	19	1.02	0.04
	GCA	22	1.18	0.05
	GCT	299	16.01	0.65
	GCC	118	6.32	0.26
Cys	TGT	256	13.71	0.68
	TGC	119	6.37	0.32
Asp	GAT	258	13.81	0.53
	GAC	227	12.15	0.47
Glu	GAG	215	11.51	0.28
	GAA	553	29.61	0.72
Phe	TTT	212	11.35	0.46
	TTC	245	13.12	0.54
Gly	GGG	58	3.11	0.12
	GGA	47	2.52	0.10
	GGT	345	18.47	0.73
	GGC	25	1.34	0.05
His	CAT	140	7.5	0.55
	CAC	113	6.05	0.45
Ile	ATA	62	3.32	0.14
	ATT	195	10.44	0.44
	ATC	188	10.07	0.42
Lys	AAG	944	50.54	0.61
	AAA	592	31.7	0.39
Leu	TTG	788	42.19	0.26
	TTA	308	16.49	0.10
	CTG	773	41.39	0.25
	CTA	442	23.67	0.14
	CTT	523	28	0.17
	CTC	234	12.53	0.08
Met	ATG	332	17.78	1
Asn	AAT	341	18.26	0.50
	AAC	335	17.94	0.50
Pro	CCG	296	15.85	0.26
	CCA	610	32.66	0.53
	CCT	168	9	0.14
	CCC	86	4.6	0.07
Gln	CAG	198	10.6	0.36

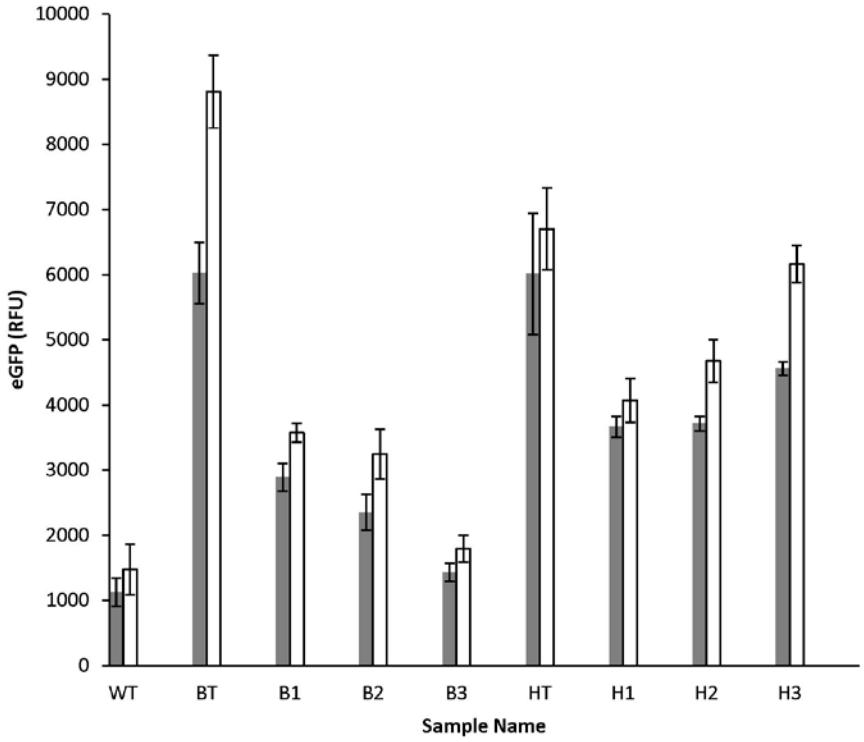
Table 5.2 (continued)

	CAA	348	18.63	0.64
Arg	AGG	427	22.86	0.37
	AGA	663	35.5	0.58
	CGG	10	0.54	0.01
	CGA	6	0.32	0.01
	CGT	34	1.82	0.03
	CGC	3	0.16	<0.01
Ser	AGT	208	11.14	0.11
	AGC	119	6.37	0.06
	TCG	360	19.28	0.19
	TCA	490	26.24	0.26
	TCT	472	25.27	0.25
	TCC	259	13.87	0.14
Thr	ACG	449	24.04	0.27
	ACA	504	26.99	0.30
	ACT	421	22.54	0.25
	ACC	300	16.06	0.18
Val	GTG	510	27.31	0.30
	GTA	390	20.88	0.23
	GTT	467	25	0.27
	GTC	340	18.2	0.20
Trp	TGG	433	23.18	1
Tyr	TAT	80	4.28	0.32
	TAC	169	9.05	0.68

A high expression codon usage table was generated from the sequences 100 most highly expressed protein-coding genes in *S. cerevisiae* using the CodonUsageBias python script. From these sequences, we are able to determine for all 64 codons the total count of each codon, frequency of occurrence per 1000 codons, and probability amongst synonymous codons (fraction).

We also tested the eGFP variants generated using the control and high expression table. While there was no statistical difference between the two conditions, expression of these variants were higher than any of the matrix-generated conditions. This result is not surprising, as the table-optimized variants lack codon diversity, but we recognize that there are many metabolic engineering applications that require high protein production, in which case we anticipate low codon diversity will become a bottleneck and decrease enzyme fitness[132,201].

Figure 5.3: Optimization for a high expression condition results in eGFP expression exceeding the wild-type



In addition to wild-type eGFP, eight variants were generated. The high expression variants were made from a codon usage table and matrix constructed using the 100 most highly expressed genes in yeast grown in rich media. Control variants were constructed from the standard usage table and control matrix (Figure 5.1b). eGFP expression was measured using flow cytometry for yeast grown in both YPD (grey) and minimal media (white). Biological triplicates were used to calculate standard deviations, indicated by error bars.

Finally, we measured GFP expression for all nine variants grown in minimal media. These results are shown in Figure 5.3 in white. We observed changes in the relative expression of these variants as we changed environmental context. This further demonstrates the importance of condition-specific codon optimization, and how small changes in growth conditions can influence protein expression.

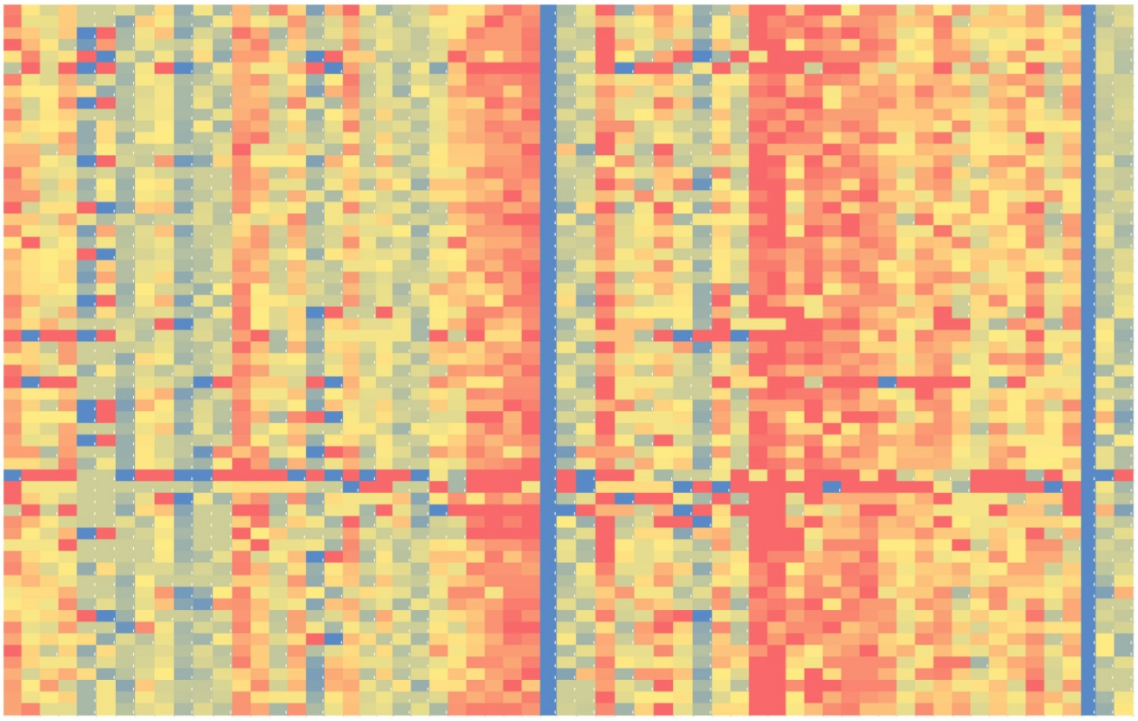
5.3.3: Stationary Phase Optimization of CatA

We were interested in extending the condition-specific codon optimization approach to a gene involved in a metabolic pathway. We selected the heterologous

expression of catechol 1,2-dioxygenase (CatA) from *Acinetobacter baylyi* in *S. cerevisiae*. CatA converts catechol to muconic acid, a useful polymer precursor[9] that is not natively produced by *S. cerevisiae*. In this case, we selected stationary phase growth for our condition of interest. In some cases, including the formation of toxic products, it is preferable to delay gene expression until stationary phase to increase product output. Therefore, this was a useful condition for us to explore. We developed our codon usage matrix as previously described using the 50 genes most differentially upregulated in yeast grown for three days as compared to exponential growth yeast[208], shown in Figure 5.4. This data is publicly available under GEO ID E-TABM-496. We designed three CatA variants using the stationary phase matrix and three using the control matrix (Figure 5.1b). Additionally, we included wild-type *A. baylyi* CatA and a variant that was codon-optimized for expression in *S. cerevisiae* by Blue Heron. These sequences are included in Appendix C. Expression of these 8 variants was determined using a protein activity assay as previously described[131] during various stages of growth (6, 18 and 24 hours). We calculated a V_{max} (mM/min*μg protein) for each variant, where a high V_{max} is indicative of higher CatA concentration. These results are summarized in Figure 5.5a.

In exponential phase (6 hours of growth), there is no statistical difference between the stationary derived variants, the control variants and wild-type CatA. However, the V_{max} for wild-type CatA is significantly higher than the Blue Heron optimized variant (p-value = 0.002). This illustrates again that traditional codon optimization approaches can often result in poor performance. Although the most highly expressed variant after 6 hours of growth is stationary #3, the lowest expressed variant is stationary #1.

Figure 5.4: Stationary phase codon optimization matrix

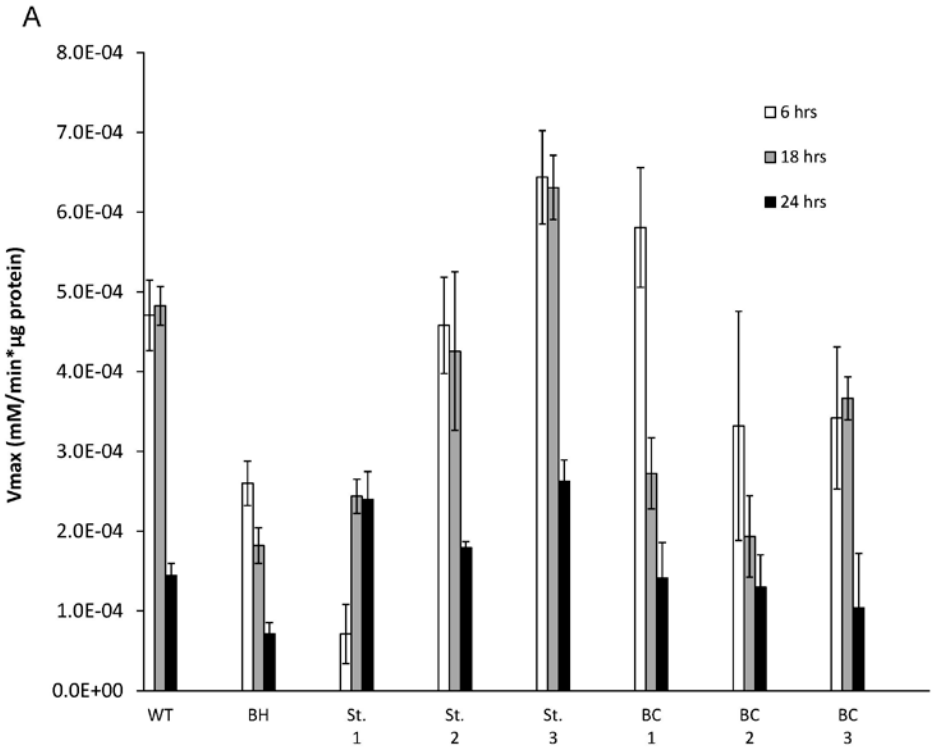


The stationary codon optimization matrix is compiled from the 50 most highly expressed protein-coding genes in *S. cerevisiae* grown for 3 days, compared to an exponential population[208]. The first amino acid is indicated by the first column, and the second amino acid by the first row. The color indicates probability between 0 (red) and 1 (blue). The Frobenius norm of the difference between the control matrix and this stationary phase matrix is 8.80.

As growth continues to the beginning of stationary phase, 18 hours, the CatA expression shifts. V_{max} for the majority of the control variants (#1 and #2), as well as the Blue Heron optimized variant, decreases between 6 and 18 hours of growth. By comparison, V_{max} for stationary #2 and #3 is unchanged between 6 and 18 hours, and for stationary #1 V_{max} actually increases. At this growth phase, V_{max} for the stationary variants is significantly higher than the control variants (p-value = 0.013) and the Blue Heron variant (p-value = 0.026). While the stationary phase variants either maintain or increase their V_{max}, the activity of the control variants decrease significantly across the board. At 18 hours, the control variants perform worse than the *A. baylyi* wild-type

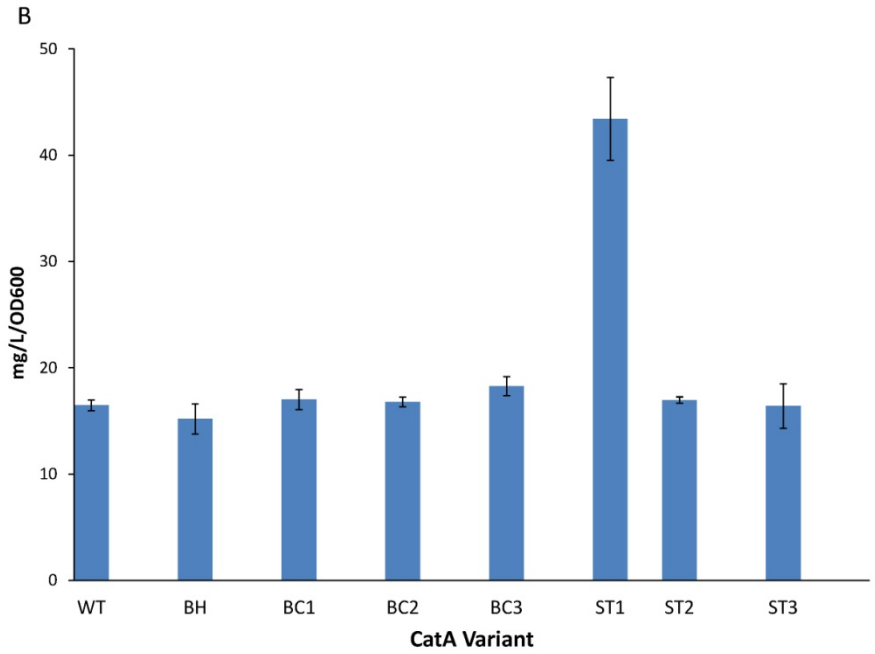
variant. These results clearly illustrate that condition-specific optimization for growth phase results in improved protein characteristics.

Figure 5.5: Optimization for stationary phase results in CatA variants that are improved at late growth



Eight CatA variants were generated, including wild-type and a version optimized by Blue Heron. The three stationary phase variants were made from a codon usage matrix (Figure 5.4) constructed using the 50 most highly expressed genes after three days of growth. The three control variants were constructed from the control matrix (Figure 5.1b). **A.** Cells expressing the CatA variants were grown for 6, 18 or 24 hours prior to bulk protein extraction. The V_{max} for conversion of catechol to muconic acid was determined for the bulk protein (mM/min*µg protein) using Lineweaver-Burke plots. A higher V_{max} is indicative of higher concentrations of CatA. Biological triplicates and technical triplicates were measured to determine standard deviations.

Figure 5.5 (continued)



B. Cells expressing the CatA variants were grown for 18 hours in 30mL before spiking the media with 1g/L of catechol. After 24 additional hours of growth, 1mL of supernatant was extracted and analyzed using HPLC, as previously described[131], to determine total muconic acid production. Normalized muconic acid levels (mg/L/OD600) are reported and standard deviation was determined using biological triplicates.

The contrast between the stationary and control variants is even more significant (p-value = 0.0005) at 24 hours. Interestingly, at 24 hours of growth, the variant with the highest V_{max} is stationary #1, which performed the worst in exponential phase. This is a further demonstration that codon optimization can result in genes whose expression is strongly influenced by the conditions under which they are expressed. Here, we were able to design three CatA variants which performed better than both a commercially-optimized sequence (Blue Heron variant) and three control CatA variants when grown in stationary phase.

Finally, we sought to measure product output using these eight CatA variants in an *in vivo* setting. Specifically, we grew yeast constitutively expressing each CatA variant for 18 hours and spiked the culture media with 1mg/mL catechol. Twenty-four

hours after the addition of catechol, supernatant was collected and analyzed using HPLC as previously described[131]. Total muconic acid was determined as the sum of *cis-cis* and *cis-trans* isomers. Results were normalized based on OD600 readings when supernatant was collected, as shown in Figure 5.5b. The stationary 1 variant outperforms all other conditions, resulting in 2.4 to 2.9 fold higher muconic acid levels. This mirrors the V_{max} measurements we made *in vitro* and is not unexpected. Further, it demonstrates the utility of condition-specific codon optimization in the context of product formation. Interestingly, we observed no difference between the other seven variants.

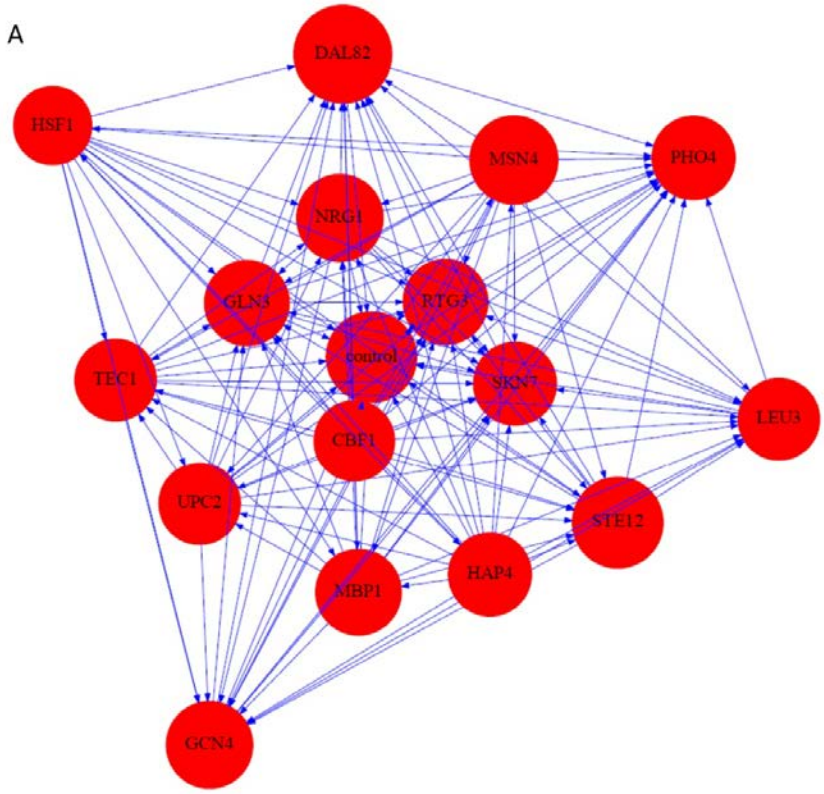
5.3.4: Organization of transcription factors suggests codon usage is linked to gene regulation

Our examination of two heterologous genes in *S. cerevisiae* (eGFP and CatA) under two different conditions demonstrates that the condition-specific codon optimization approach can result in improved protein expression. This likely reflects changes in tRNA expression levels that are demonstrated to arise under different growth conditions. We were interested in exploring additional underlying aspects of cell physiology that might explain this improvement. Utilizing a systems biology perspective, we concentrated on global transcription factors.

We examined sixteen global transcription factors for *S. cerevisiae*: CBF1, DAL82, GCN4, GLN3, HAP4, HSF1, LEU3, MBP1, MSN4, NRG1, PHO4, RTG3, SKN7, STE12, TEC1, and UPC2 and identified their genomic targets using data tabulated at yeastgenome.org. We were interested in defining a codon profile for those gene targets interacting with each global regulator and comparing it to our previously established control matrix that uses 6,666 protein coding genes. A codon usage matrix was established for the genetic targets of each transcription factor. In order to compare these matrices to our control condition, we sought to quantify drift or distance using the

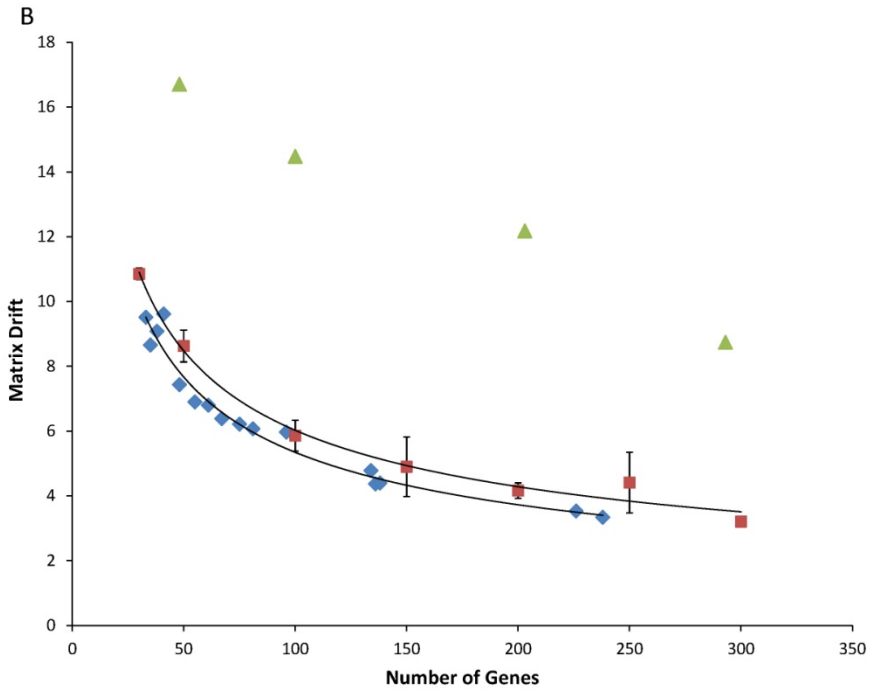
Frobenius matrix norm of the difference between the matrices. This metric is well established as a quantitative tool to determine drift between matrices of the same size[209]. A smaller Frobenius matrix norm indicates higher similarity between matrices, such that identical matrices have a norm of zero. The Frobenius matrix norm was calculated for all pairs of the 17 matrices (control and 16 transcription factors), the results of which are shown in Table 5.3. This data set was then used to generate a node-edge map representing drift differences between conditions (Figure 5.6a).

Figure 5.6: Drift of transcription-factor codon matrices reveals diverse codon usage relative to the control matrix



The genetic interaction targets for sixteen *S. cerevisiae* transcription factors were identified using yeastgenome.org. Using those corresponding gene target sequences, codon usage matrices were constructed for each transcription factor. **A.** Frobenius matrix norms were calculated for all matrix pairs, including the control matrix (Figure 5.1b) using MATLAB. The Frobenius norms represent drift between matrices and create the edges in this map between the nodes (transcription factors). The map was constructed using the Map_Draw python script (Appendix D).

Figure 5.6 (continued)



B. The matrix drift (as indicated by Frobenius norm) versus number of genes used to generate the codon usage matrices was plotted for each transcription factor (blue). For comparison, codon usage matrices were generated from a random sampling of genes (red). Both were fit with a power regression model. Standard deviation from five independent samples were used to generate error bars. Codon usage matrices from subsets of the most highly expressed genes are shown in green.

The two matrices that are the most similar are the control and genetic targets for RTG3. The control matrix and the genetic targets for CBF1 are also very similar. The most disparate matrices are the genetic targets for GCN4 and HSF1. This map clearly shows that overall, the genetic targets of transcription factors have very disparate codon usage. Furthermore, each transcription factor matrix is more similar to the control condition that to another transcription factor matrix. This further supports the averaging effect on codon usage caused by using all protein coding genes, as opposed to a subset. Additionally, these results suggest genes regulated by the same global factors may have similar codon usage patterns.

Table 5.3: Drift calculations using the Frobenius Matrix Norm

	CBF1	DAL82	GCN4	GLN3	HAP4	HSF1	LEU3	MBP1	MSN4	NRG1	PHO4	RTG3	SKN7	STE12	TEC1	UPC2	Control
CBF1	0	7.83	10.30	5.09	6.92	10.16	9.58	6.65	6.78	5.35	9.04	4.26	4.94	7.68	7.21	6.48	3.53
DAL82		0	11.94	8.23	9.48	12.10	11.71	9.08	9.32	8.03	11.41	7.79	7.91	10.01	9.69	9.59	7.43
GCN4			0	10.45	11.28	12.88	12.78	11.29	10.79	10.47	12.46	10.06	10.44	11.96	11.81	11.60	9.62
GLN3				0	7.39	10.54	10.30	7.47	7.41	6.32	9.66	4.99	5.75	8.20	8.23	7.51	4.78
HAP4					0	10.81	11.08	8.53	8.75	7.39	10.15	6.54	7.33	9.04	8.70	8.54	5.97
HSF1						0	12.66	11.37	10.86	10.74	12.77	10.08	10.23	11.37	11.54	11.53	9.51
LEU3							0	10.81	10.54	10.10	12.50	9.46	9.84	11.60	11.05	10.72	9.08
MBP1								0	8.12	7.26	10.55	6.74	6.97	9.05	8.58	8.21	6.21
MSN4									0	7.12	10.31	6.59	7.16	9.44	8.93	8.47	6.39
NRG1										0	9.19	5.13	5.80	8.10	7.98	7.16	4.37
PHO4											0	9.14	9.59	10.94	10.72	10.21	8.66
RTG3												0	4.98	7.58	7.22	6.40	3.34
SKN7													0	8.14	7.71	7.32	4.40
STE12														0	8.64	9.04	6.90
TEC1															0	8.57	6.80
UPC2																0	6.07
Control																	0

Codon usage matrices were generated for a control condition (Figure 5.1b) as well as for the gene targets of 16 transcription factors. Drift was determined by taking the Frobenius norm of the difference between two matrices. These values are reported above.

Next we looked at the relationship between the Frobenius matrix norm for each transcription factor relative to the control matrix and the number of genes that transcription factor is interacting with. We see that for the most part, as the number of genetic interactions increases, the Frobenius matrix norm or drift relative to the control condition decreases. This behavior is expected, as the inclusion of more genes in a codon context matrix will result in an averaging effect that begins to resemble the control matrix composed of all gene sequences. This data is shown in Figure 5.6b and closely fits a power regression model of the form $y = Ax^b$. As a comparison, we generated a corresponding control curve for which genes were selected at random and the corresponding Frobenius matrix norm was calculated. Control conditions of 30, 50, 100, 150, 200, 250 and 300 genes were selected in five independent events with the average matrix norm shown in Figure 5.6b. Again we see the power regression model fits well, however the two curves are distinct, indicating a difference between the genes selected at random and the genetic targets of transcription factors. An F test was used to determine that the two curves are statistically different with a p-value of 0.003.

A high drift relative to the control matrix may be indicative of evolutionary pressure on codon usage. This is especially evident when we consider the most highly transcribed genes in *S. cerevisiae*, as shown in green. This set of genes, which is arguably evolved for high expression, has very different codon usage than the control matrix. Interestingly, there are only 12 common genes between the top 100 most highly expressed genes and the approximately 1000 gene targets of the sixteen transcription factors. This may indicate that genes regulated by transcription factors are not highly expressed. We also see that the transcription factor curve sits below the curve of genes selected at random, indicating that genetic targets of these transcription factors have codon usage more closely resembling the control matrix than genes selected at random.

This is a surprising result and may indicate that the codon usage of genes regulated by transcription factors have evolved less, on average, than other parts of the genome.

5.4: CONCLUDING REMARKS

Here, we propose an alternative approach to traditional codon optimization methods that utilizes both systems-level information and codon context to generate condition specific variants stochastically. This approach, termed condition-specific codon optimization, takes into account environmental growth conditions that are known to influence tRNA abundance and therefore translational efficiency to determine codon frequency. In contrast, traditional codon optimization neglects codon context entirely and determines codon frequency using all genomic protein coding sequences. While traditional codon optimization can result in improved protein expression, this approach is not robust and often results in decreased translational efficiency. Alternatively, we validate the robustness of our approach through experiments under three disparate conditions.

We expressed two heterologous genes in S. cerevisiae, eGFP and CatA, under different growth conditions. In each case, compared to the wild-type and variants generated using traditional approaches, we observe improved protein expression in the condition-optimized cases. eGFP expression was successfully optimized for high expression in rich media and the CatA enzyme was optimized for high expression in stationary phase growth.

Codon optimization is an important biotechnology tool that enables recombinant DNA expression. As our ability and desire to produce chemicals in a renewable and environmentally-friendly capacity increases, more biological processes will be

developed. In doing so, the robust and high-level expression of foreign genes will be a crucial component of the process. Furthermore, many of these processes will be carried out under diverse, non-standard environmental conditions, including changes in temperature, pH, mineral concentration, carbon source, and oxygen concentration. The condition-specific approach to codon optimization can be used to identify gene variants with the ideal codon usage for these conditions.

This approach is simple, robust and generic. As a starting point, it utilizes global systems level expression data for the host strain and condition of interest. Much of this data is publicly available for common host organisms. In the case where it is not available *a priori*, a standard microarray or RNA-seq experiment can be conducted and utilized. From this data set, an algorithm can be used to extract codon usage, while preserving codon context, and to generate a codon usage matrix. From this matrix, a second algorithm can stochastically determine gene variants whose codon usage closely aligns with the codon usage matrix. This approach can be generically applied for any environmental condition in any sequenced host organism.

Chapter 6: Linking yeast Gcn5p catalytic function and gene regulation using a quantitative, graded dominant mutant approach

6.1 Chapter Summary

Establishing causative links between protein functional domains and global gene regulation is critical for advancements in genetics, biotechnology, disease treatment, and systems biology. This task is challenging for multifunctional proteins when relying on traditional approaches such as gene deletions since they remove all domains simultaneously. Here, we describe a novel approach to extract quantitative, causative links by modulating the expression of a dominant mutant allele to create a function-specific competitive inhibition. Using the yeast histone acetyltransferase Gcn5p as a case study, we demonstrate the utility of this approach and (1) find evidence that Gcn5p is more involved in cell-wide gene repression, instead of the accepted gene activation associated with HATs, (2) identify previously unknown gene targets and interactions for Gcn5p-based acetylation, (3) quantify the strength of some Gcn5p-DNA associations, (4) demonstrate that this approach can be used to correctly identify canonical chromatin modifications, (5) establish the role of acetyltransferase activity on synthetic lethal interactions, and (6) identify new functional classes of genes regulated by Gcn5p acetyltransferase activity—all six of these major conclusions were unattainable by using standard gene knockout studies alone. We recommend that a graded dominant mutant approach be utilized in conjunction with a traditional knockout to study multifunctional proteins and generate higher-resolution data that more accurately probes protein domain function and influence.

6.2 INTRODUCTION

The development of novel approaches for dissecting multifunctional proteins and their individual domain-target interactions *in vivo* would be a valuable expansion of the biotechnology toolbox. Such an approach would greatly increase our ability to collect systems level information about protein-DNA interactions and complement gene knockout techniques. Here, we demonstrate the capacity of a unique, graded dominant mutant approach to enable such goals.

This graded dominant mutant approach is illustrated with a systems biology study of the yeast histone acetyltransferase (HAT), Gcn5p. HAT proteins are important targets of genetic studies since they are critical for establishing acetylation of histones, which have long been recognized as a mark of euchromatin and an important activating genomic modification[210,211]. Gcn5p is a multifunctional protein with catalytic and binding domains (including Ada2 interaction and a bromodomain). A causative study of acetyltransferase activity thus requires a removal or reduction of catalytic function while maintaining native protein interactions. The yeast gene, *GCN5*, serves as a well-studied prototype[212-214] for transcription-associated HAT activity. Gcn5p has a known crystal structure[215] and direct homologues in higher eukaryotic systems. Only a small number of Gcn5p putative gene targets have been identified even though it is presumed that this HAT globally controls gene expression[216]. Moreover, as this HAT is nonessential like many epigenetic factors, inherent protein redundancy implies that other HAT proteins may compensate for Gcn5p in its absence and thus confound data relying on knockout studies alone.

Classically, dominant mutations have been widely used to probe gene function,[217] improve tolerances and drug resistances,[152,154,218] characterize disease states,[155-158] and map protein functional domains[219]. In this regard, small point mutations can abolish a

particular function in isolation without disrupting other protein activities. Here, we exploit the inhibitory nature of dominant mutations and demonstrate that varying the expression level of a non-catalytic dominant mutant in the presence of the native, wild-type allele can specifically isolate and titer the catalytic activity of the wild type protein. This can be achieved by identifying a mutant protein that lacks activity but still folds properly, thus functioning as a dominant mutant and causing competitive inhibition of a single, wild-type protein domain.

Despite the common use of dominant mutants, no prior study has paired these alleles with a promoter library to specifically and quantitatively grade an isolated protein function and collect systems level information. Here we demonstrate the power of a graded dominant mutant approach to isolate and causatively study the histone acetyltransferase catalytic activity of Gcn5p protein in the yeast *Saccharomyces cerevisiae* and in doing so, uncover previously unknown gene targets and functions of Gcn5p.

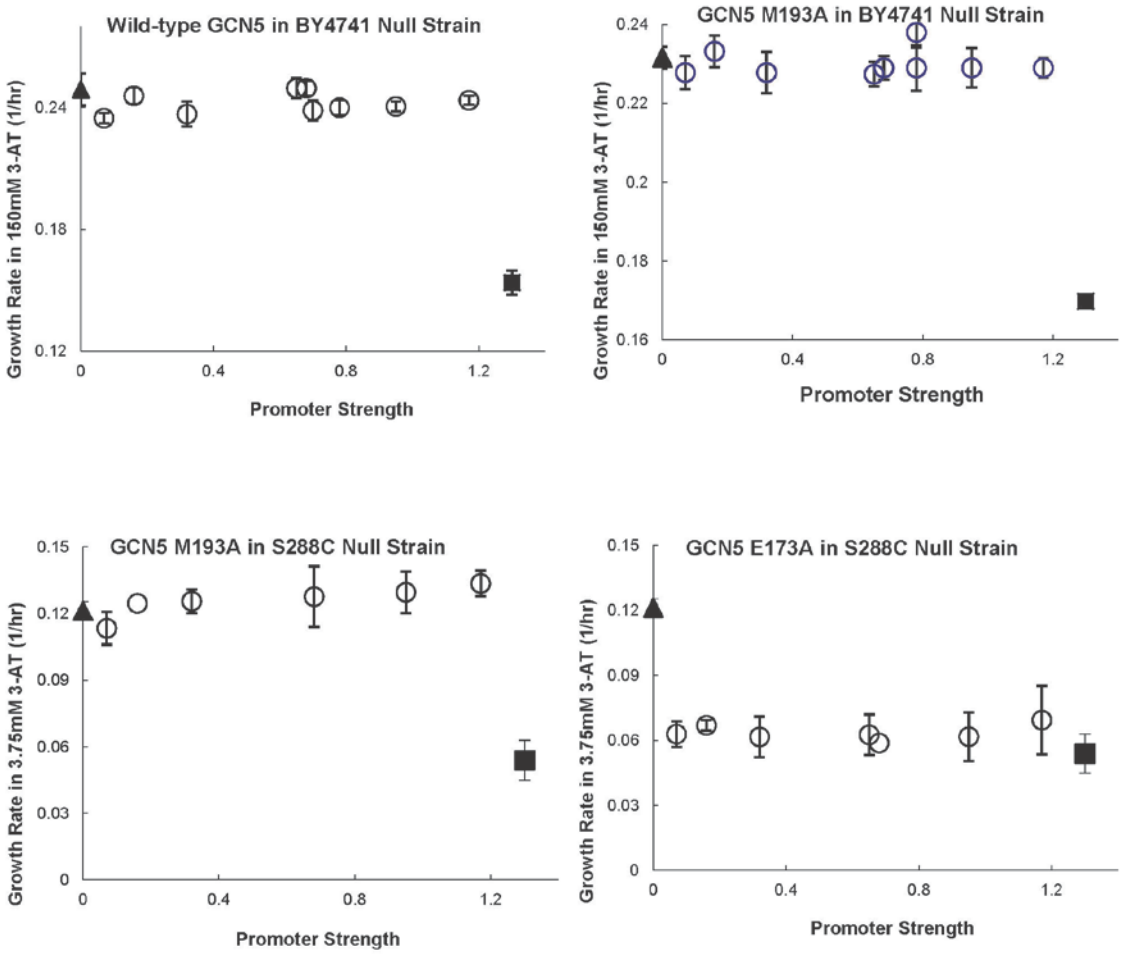
6.3 Results And Discussion

6.3.1: Identifying Gcn5p dominant mutants

We sought to identify a Gcn5p dominant mutant that lacked catalytic activity but was still able to fold like the wild-type allele. Previous studies conducted an alanine scan across the catalytic domain and measured acetylation activity *in vitro*[212,220]. Based on an absence of activity, we initially selected three mutations (E173A, M193A and F221A). We utilized a complementation assay in a gcn5 null strain transformed with mutant gcn5 on plasmids to test the *in vivo* activity of all three mutations. Activity was correlated

with growth in the presence of 5-aminotriazole, which is a known inhibitor of histidine synthesis that is naturally regulated by Gcn5p.

Figure 6.1: *GCN5* complementation assay to determine potential dominant mutants.



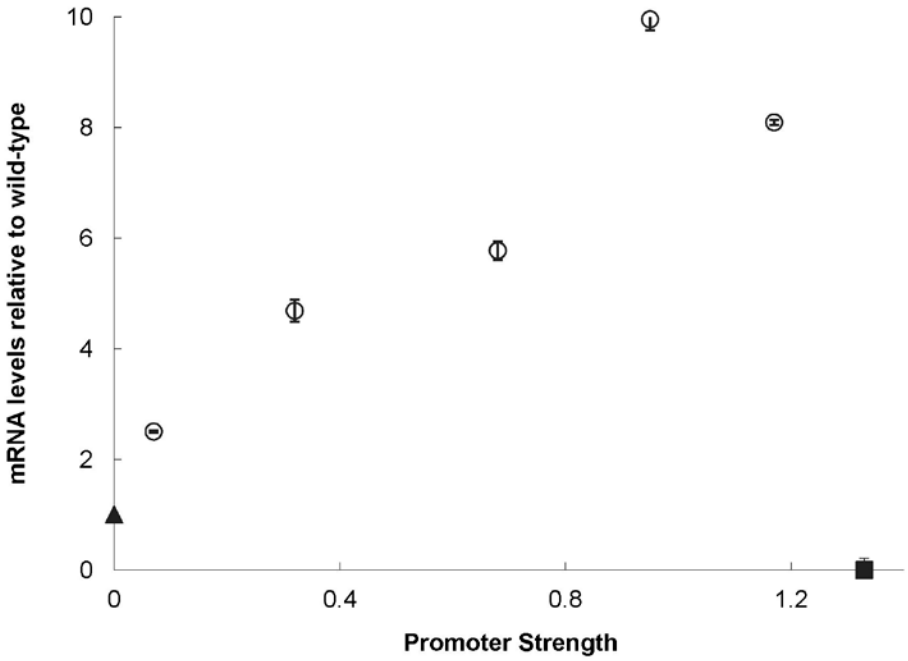
Using a BY4741 *gcn5Δ* strain, we expressed two Gcn5p mutants (*gcn5-M193A* and *F221A*) and wild-type *GCN5* with varying promoter strengths. Strains were grown in minimal media and growth rate was measured using a Bioscreen C. We compared the mutant growth rates to that of the native yeast. This process was repeated with an S288C *gcn5Δ* strain, in which we expressed two Gcn5p mutants (*gcn5-M193A* and *E173A*) with varying promoter strengths.

The *M193A* mutant exhibited the same growth rate as the control strain, indicating that catalytic activity was still intact. However, we observed that the *gcn5-E173A* and

mutants were unable to complement *gcn5Δ* and behaved similar to the knockout strain, indicating a complete loss of catalytic activity. Next, we determined the impact of the E173A and F221A mutations to Gcn5p's Gibbs Free energy using the Protein Interfaces, Surfaces and Assemblies[221] database. Both mutations resulted in no change to Gibbs Free energy, indicating a conservation of protein structure. These analyses established both *gcn5-F221A* and *gcn5-E173A* as dominant mutant candidates for Gcn5p-acetylation activity.

In order to create quantitative, graded expression of these dominant alleles, expression (and thus level of competitive inhibition) was modulated through the use of a promoter library. Expression of the mutant alleles was established by cloning *gcn5-F221A* and *gcn5-E173A* into centromeric yeast expression vectors under the control of a collection of mutant TEF-based promoters with previously established expression capacities[222,223]. This library resulted in a ratio of mutant to wild-type expression ranging from 2.5 fold with the weakest promoter to 8-10 fold with the strongest promoter (Figure 6.2**)**. While the majority of experiments were conducted with the *gcn5-F221A* mutant, several were done with both *gcn5-F221A* and *gcn5-E173A* to demonstrate robustness of the approach.

Figure 6.2: Measuring native and mutant Gcn5p mRNA expression levels



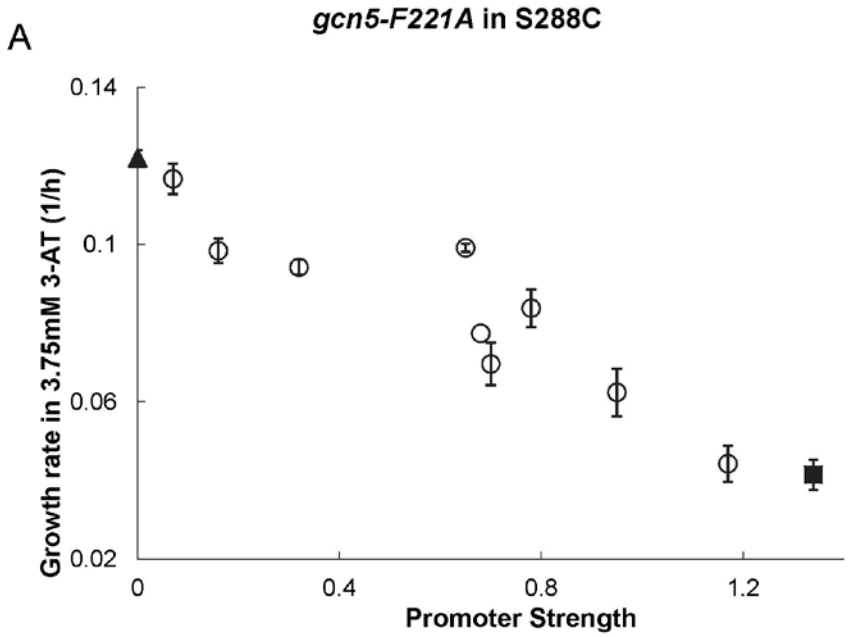
Wild-type and mutant *GCN5* mRNA levels were measured using RT-PCR and whole cell mRNA extracted from S288C wild-type, *gcn5Δ* and *gcn5-F221A* (with 5 different promoter strengths) strains. Average Ct values and standard deviation were calculated from triplicates, and mRNA levels were normalized relative to the wild-type sample. At the lowest promoter strength of .07, mutant *gcn5-F221A* is expressed at levels 2.5 fold higher than wild-type, and at the highest promoter levels of .95 and 1.17, *gcn5-F221A* is expressed at levels 8-10 fold higher than wild-type *GCN5*.

6.3.2 *gcn5-F221A* competitively inhibits the catalytic function of Gcn5p in a dose-responsive manner

After identifying candidate dominant mutants for Gcn5p, three genetic tests were used to establish and validate the gradation and competition of catalytic activity. The first test involved the HIS3 locus, a known acetylation target for Gcn5p[224]. Gene activation of HIS3 by Gcn5p-based acetylation enables higher tolerance to a histidine analogue, 3-aminotriazole (3-AT). In a *gcn5Δ* strain, HIS3 expression is decreased, leading to amino acid starvation in the presence of 3-AT and decreased cell growth. Each expression cassette controlling *gcn5-F221A* was transformed into S. cerevisiae S288C and growth rate was evaluated in the presence of 3-AT (Figure 6.3) to determine the

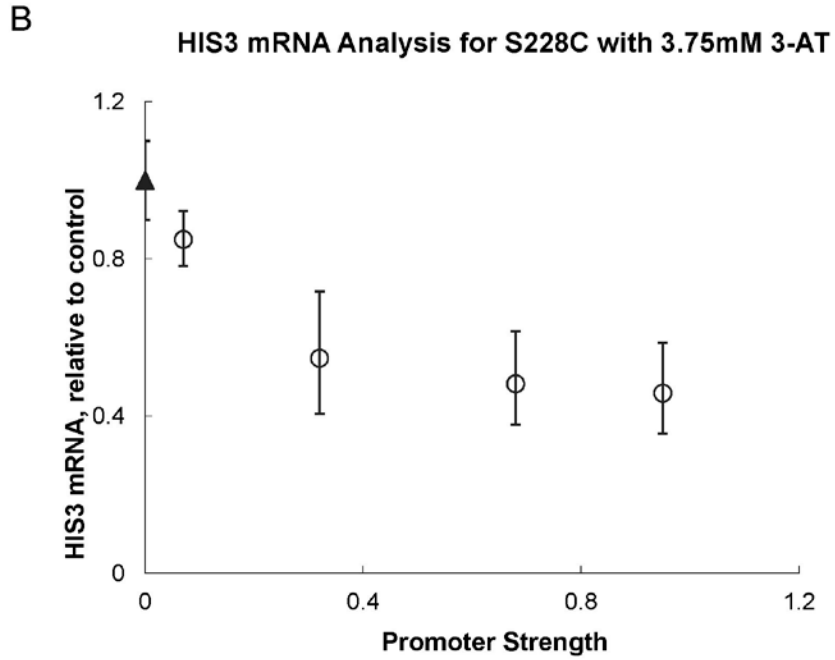
impact of mutant expression on the HIS3 locus. Strains with low expression of the mutant allele most-closely resembled the wild-type strain, whereas at higher expression levels, strains resembled that of the gcn5 null strain. Transcription of HIS3 was found to decrease in a manner that followed a competitive inhibition curve (Figure 6.3). This data provides strong evidence that the *gcn5-F221A* allele competes for Gcn5p acetylation sites in the HIS3 promoter region and effectively decreases HIS3 transcription.

Figure 6.3: *gcn5-F221A* can impart a graded phenotype as detected by starvation assays



The *gcn5-F221A* mutant was expressed in S288C with varying promoter strengths, and starvation response measured via growth rate in minimal media supplemented with 3.75mM 3-aminotriazole. **A.** Growth rates of strains harboring *gcn5-F211A* (○) were compared to wild-type (▲) and *gcn5Δ* strains (■). Error bars represent the standard deviation of biological triplicates. Increasing the expression level of *gcn5-F221A* (through progressively stronger promoters) results in a decrease in growth rate approaching the value of the knockout strain.

Figure 6.3 (continued)



B. *HIS3* mRNA levels were measured for select promoter strengths (.07, .32, .68 and .95) using RT-PCR. As *gcn5-F221A* promoter strength increases, *HIS3* expression decreases following a competitive inhibition pattern. These results demonstrate that *gcn5-F221A* can exhibit a graded, competitive phenotype at *HIS3* as measured by starvation response.

While the growth rate trend (Figure 6.3) shows a clear correlation between mutant expression and growth rate, the trend is linear rather than an inhibition curve. We believe this arises from the more indirect measurement of growth rate, which is impacted by many cellular and environmental factors, and therefore integrates multiple signals, not just HIS3 expression levels. By comparison, the measurement of HIS3 mRNA levels (Figure 6.3) is a more direct measurement and thus presents the more expected competitive inhibition curve.

A second test involved inhibiting the well-characterized Gcn5p-based regulation of the Pho5 promoter[225-227]. This test was conducted in two *pho80* knockout strains of yeast, a haploid (BY4741) and diploid (BY4743), as this mutation results in a constitutively active Pho5 promoter[226]. Both hosts contained an episomal synthetic gene

circuit with the Pho5 promoter regulating expression of the fluorescent protein yECitrine. The activation of the Pho5 promoter was assayed in yeast strains harboring the collection of plasmids with graded *gcn5-F221A* expression (Figure 6.4). In response to increased expression of *gcn5-F221A*, the fluorescent signal was observed to decrease. This is consistent with the hypothesis that this mutant allele directly competes with the native Gcn5p protein. We found that mean fluorescence followed a competitive inhibition model, where increased expression of the dominant mutant decreased the mean fluorescence.

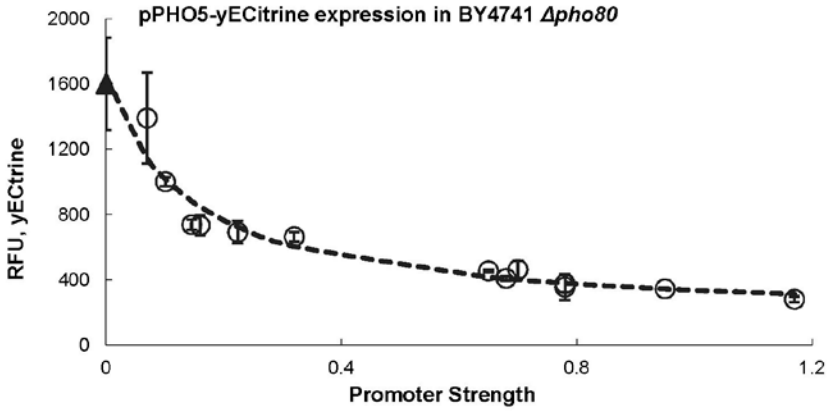
We were able to fit our data to the Hill-slope competitive inhibition model:

$$signal = Bottom + \frac{Top - Bottom}{1 + 10^{PS - \log(IC50)}}$$

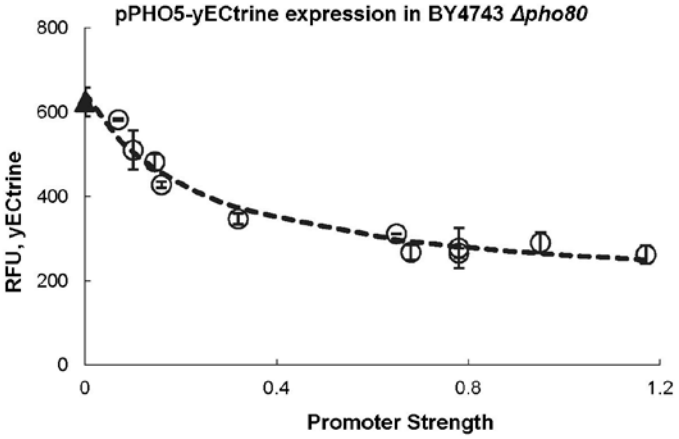
where Top is the signal strength in the absence of competition, Bottom is the signal strength of a competitively saturated system and IC_{50}, or 50% effective concentration, occurs when the signal strength is reduced to the value halfway between the upper and lower bounds. The signal is a measure of average fluorescence in RFU and PS is the relative promoter strength of the dominant mutant. A best-fit was determined using a sum of least squares regression with the experimentally determined values. These models demonstrate that increased promoter strength is required to inhibit the two chromosomal copies of *GCN5* in a diploid strain (curve shown in Figure 6.4). Results of the IC_{50} values are found in Figure 6.4**.** Finally, yECitrine mRNA levels decreased as a function of *gcn5-F221A* expression (Figure 6.4). This test demonstrates the ability of a dominant mutant allele approach to make a direct measurement relating the grading of acetyltransferase activity to downstream gene expression (in this case, Gcn5p acetyltransferase activity and Pho5 promoter activity).

Figure 6.4: Evaluating competitive inhibition by *gcn5-F221A* using a synthetic construct

A



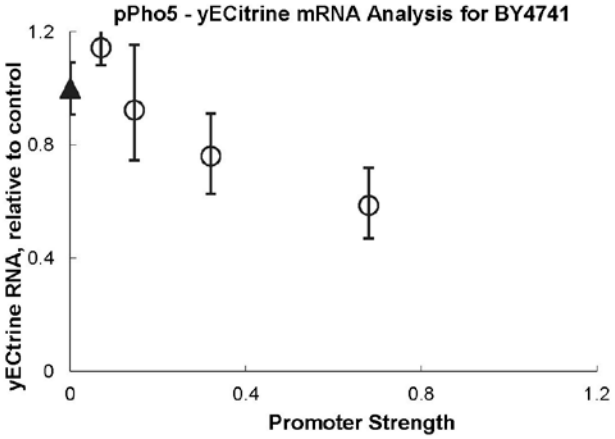
B



C

Cell Line	Curve Top (RFU)	Curve Bottom (RFU)	IC_{50} (Rel. promoter strength)
BY4741 (haploid)	1649	153	13.8
BY4743 (diploid)	645	167	24.3

D



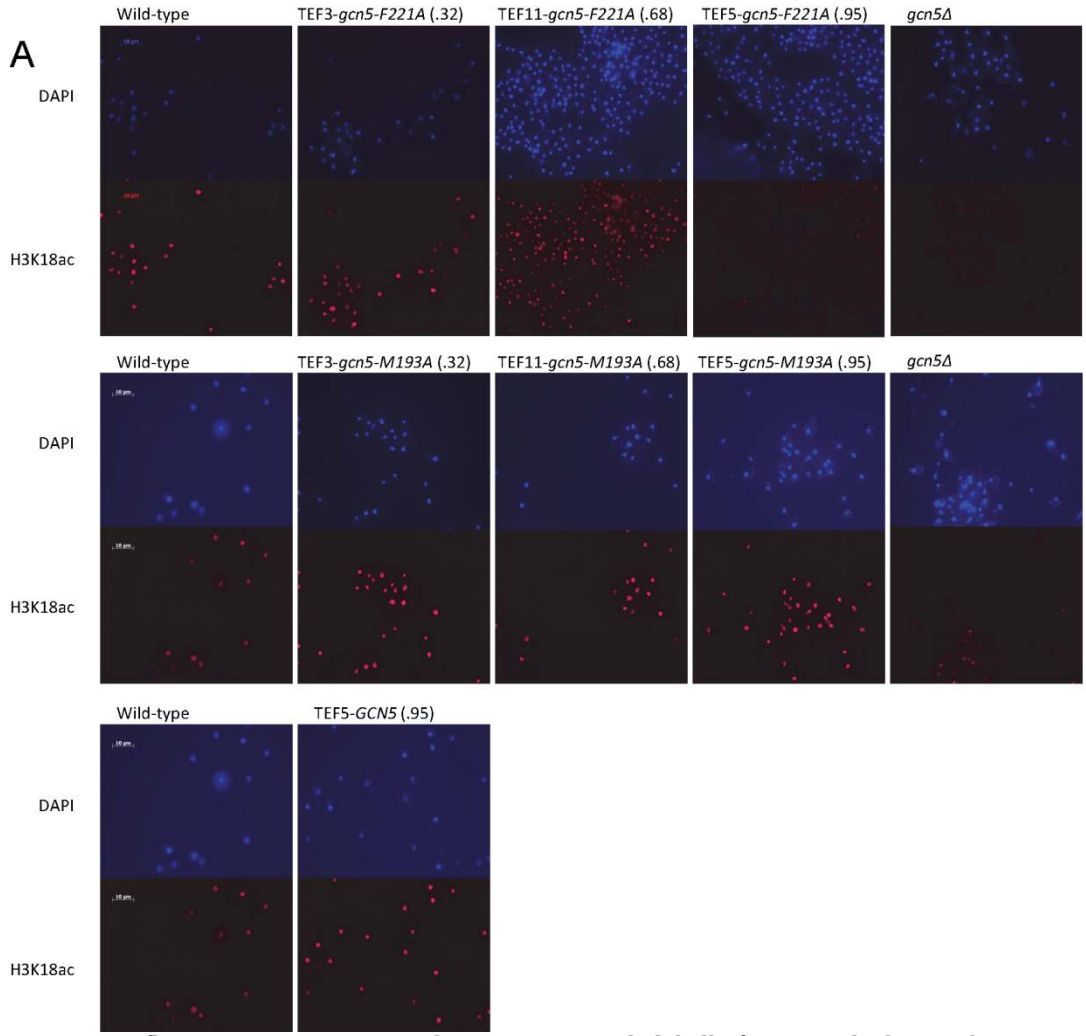
The *gcn5-F221A* mutant was expressed with varying promoter strengths in **A.** the haploid BY4741 *pho80Δ* and **B.** the diploid BY4743 *pho80Δ*, and co-expressed with a second plasmid, containing the yECitrine gene driven by pPho5. Average fluorescence in mid-exponential phase is reported and error bars represent standard deviations of biological triplicates. The data was fit to a Hill-slope competitive inhibition model (dashed line) and IC_{50} values were extracted (**C**), indicating the relative promoter strength of *gcn5-F221A* resulting in half-maximal inhibition. **D.** yEcitrine mRNA levels were measured using RT-PCR for select promoter strengths (.07, .16, .32, and .70).

6.3.3: Gcn5p graded dominant mutants competitively inhibit global histone acetylation at H3K18

In a third test, we sought to demonstrate that the graded dominant mutant, *gcn5-F221A*, was directly impacting histone acetylation. In *S. cerevisiae*, lysine 18 of histone 3 is primarily acetylated by Gcn5p, with very little acetylation occurring in a *gcn5Δ* strain[228]. An immunofluorescence assay for acetylated H3K18 residues was conducted using mid-exponential phase, fixed yeast cells. Three promoter strengths (0.32, 0.68, and 0.95 relative to wild-type TEF) were used to drive the expression of *gcn5-F221A* and these strains were compared to wild-type and *gcn5Δ* strains. Additionally, two further controls (over-expression of wild type *GCN5* and graded expression of a catalytically active mutant allele, *gcn5-M193A*) were used to demonstrate the specific acetylation inhibition only afforded by *gcn5-F221A*. Neither the wild-type *GCN5* nor the catalytically active mutant *gcn5-M193A* showed a change in global H3K18ac. By comparison, expression of the inactive, dominant mutants, *gcn5-F221A*, resulted in a dose-dependent decrease of H3K18 acetylation. These results, illustrated and quantified in Figure 6.5, demonstrate that *gcn5-F221A* competes directly with native Gcn5p, resulting in reduced histone acetylation.

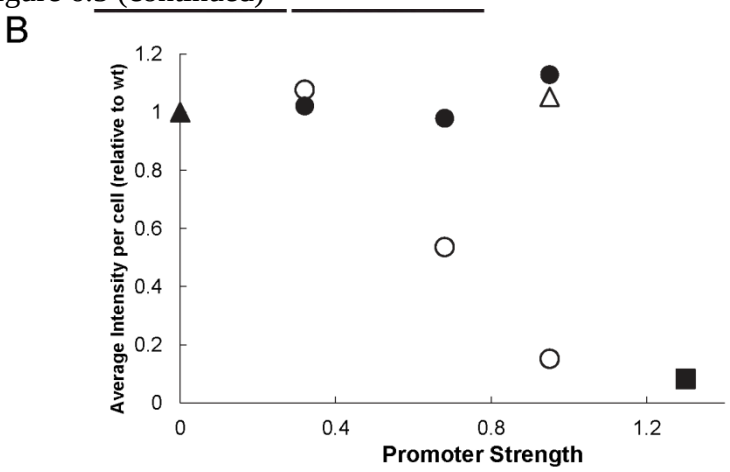
This immunofluorescence test was repeated with an additional catalytically inactive, dominant mutant (*gcn5-E173A*). As shown in Figure 6.6, the *gcn5-E173A* mutant also results in global attenuation of H3K18 acetylation. Using a high strength promoter, acetylation levels are very similar to that of the *gcn5Δ* strain. These results indicate that like *gcn5-F221A*, this second, graded dominant mutant is able to competitively inhibit Gcn5p and directly interfere with acetylation activity.

Figure 6.5: Global acetylation at H3K18 is attenuated by expression of mutant *gcn5-F221A*



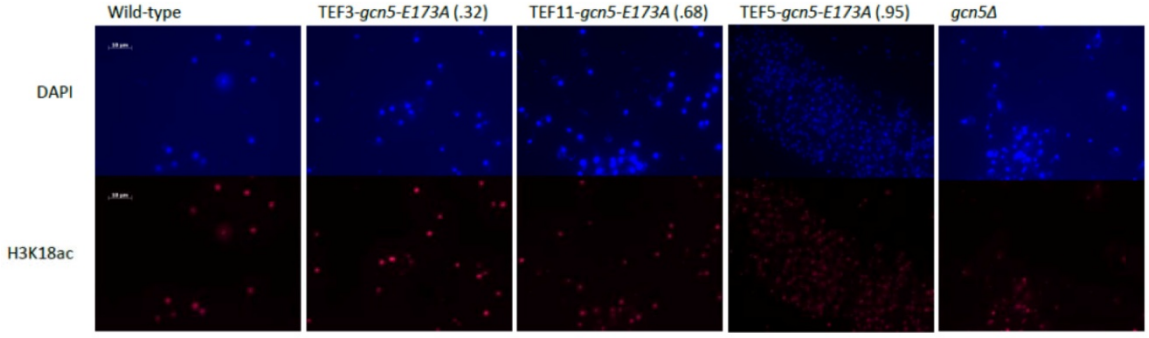
Using immunofluorescence, H3K18 acetylation was assayed globally for strains harboring the *gcn5-F221A* mutant (○) expressed with varying promoter strengths (.32, .68, and .95). For comparison, *gcn5-M193A* mutant (●) (fully functional) with the same promoter strengths, along with wild-type (▲), *gcn5Δ* (■) and wild-type with *GCN5* (Δ) over-expressed, were also examined. The primary antibody targets H3K18ac and the secondary antibody is an IgG tagged with DyLight 649. Cells were stained with DAPI to visualize nuclear material. **A.** Cells were imaged with both DAPI and Cy5 filters. The *gcn5-F221A* mutant results in global attenuation of H3K18ac and approaches *gcn5Δ* strain at high strength promoters. By comparison, the *gcn5-M193A* mutant and wild-type with *GCN5* result in no change to acetylation levels.

Figure 6.5 (continued)



B. Average cell intensity was quantified using Metamorph software and normalized relative to the wild-type.

Figure 6.6: Global H3K18 acetylation is attenuated by expression of mutant *gcn5-E173A*



Using immunofluorescence, H3K18 acetylation was assayed globally for strains harboring the *gcn5-E173* mutant expressed with varying promoter strengths (0.32, 0.68, and 0.95), along with wild-type and *gcn5Δ* cells. The primary antibody, raised in rabbit, targets H3K18ac, and the secondary antibody is an anti-rabbit IgG tagged with DyLight 649. All cells were also stained with DAPI to visualize nuclear material. Cells were imaged with both a DAPI and Cy5 filter. The *gcn5-E173A* mutant results in global attenuation of H3K18 acetylation. Average cell intensity quantification, using Metamorph software, confirms that increased *gcn5-E173A* expression decreased acetylation (average cell intensity from left to right: 133050, 89607, 48178, 37252, 32128).

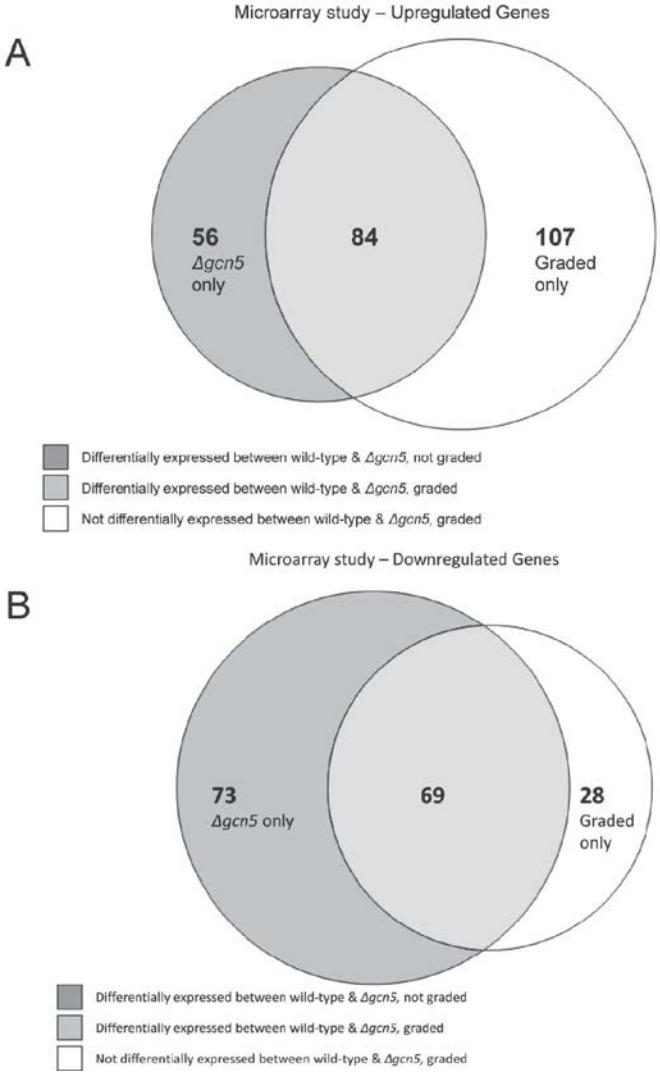
6.3.4: Combining expression profiling with a graded dominant mutant approach reveals novel Gcn5p targets and function

Next, we sought to evaluate the global influence of Gcn5p acetyltransferase activity on yeast gene expression. By using our approach, genes whose expression

changes as a function of *gcn5-F221A* level are changing as a result of decreased acetyltransferase activity. Using microarrays, we identified and classified differentially expressed genes between *S. cerevisiae* (S288C) wild-type, the *gcn5* null strain, and mutant *gcn5-F221A* expressed at three different promoter strengths (0.32, 0.68 and 0.95)[229]. A total of 282 genes were found to be differentially expressed (p-value<0.05, abs($\log_2$)>1) between the wild-type and knockout strain. This dataset overlaps a previously reported gene expression study for *gcn5Δ* with 98% coverage[230]. A total of 288 genes were found to be differentially graded in response to *gcn5-F221A* (i.e. genes whose expression changes monotonically in response *to gcn5-F221A* and all of which had p-values<0.05). Despite these similar numbers, only 153 genes (53%) found in the knockout data set overlap with the graded dominant mutant dataset (Figure 6.7c). This initial analysis indicates that, for multifunctional proteins, classifying genes and regulation based exclusively on knockout data is misleading. This data set augments our knowledge of gene targets regulated by Gcn5p-acetyltransferase activity.

Despite the common conception that Gcn5p-based acetylation is gene activating, we posit that Gcn5p-based acetylation serves a dominant role in maintaining global gene repression in yeast. We found that over-expression of a catalytically inactive dominant mutant led to up regulation of 66% of affected genes (Figure 6.7a), whereas the *gcn5* null strain significantly overestimates the number of under-expressed genes (Figure 6.7b). This finding is unexpected and not evident from traditional knockout experiments, as gene expression changes in the knockout strain were equally distributed between over and under expression. This is the first time that Gcn5p-based acetylation has been implicated with global gene repression, and may be a direct function of Gcn5p or an indirect result of additional gene regulators that are controlled by Gcn5p.

Figure 6.7: Expression analysis comparing a graded dominant mutant of *gcn5-F221A* to *gcn5Δ*.



Microarray analysis was conducted for strains expressing the *gcn5-F221A* mutant at varying promoter strengths (.32, .68, and .95) along with wild-type and *gcn5Δ* cells. **A.** Of the genes found to be up-regulated compared to the wild-type, only 84 were commonly identified by both the dominant mutant and knockout, and 107 were only identified by the dominant mutant. **B.** Fewer genes were found to be down-regulated, of which only 69 were commonly identified and 28 were only identified by the dominant mutant.

Figure 6.7 (continued)

C

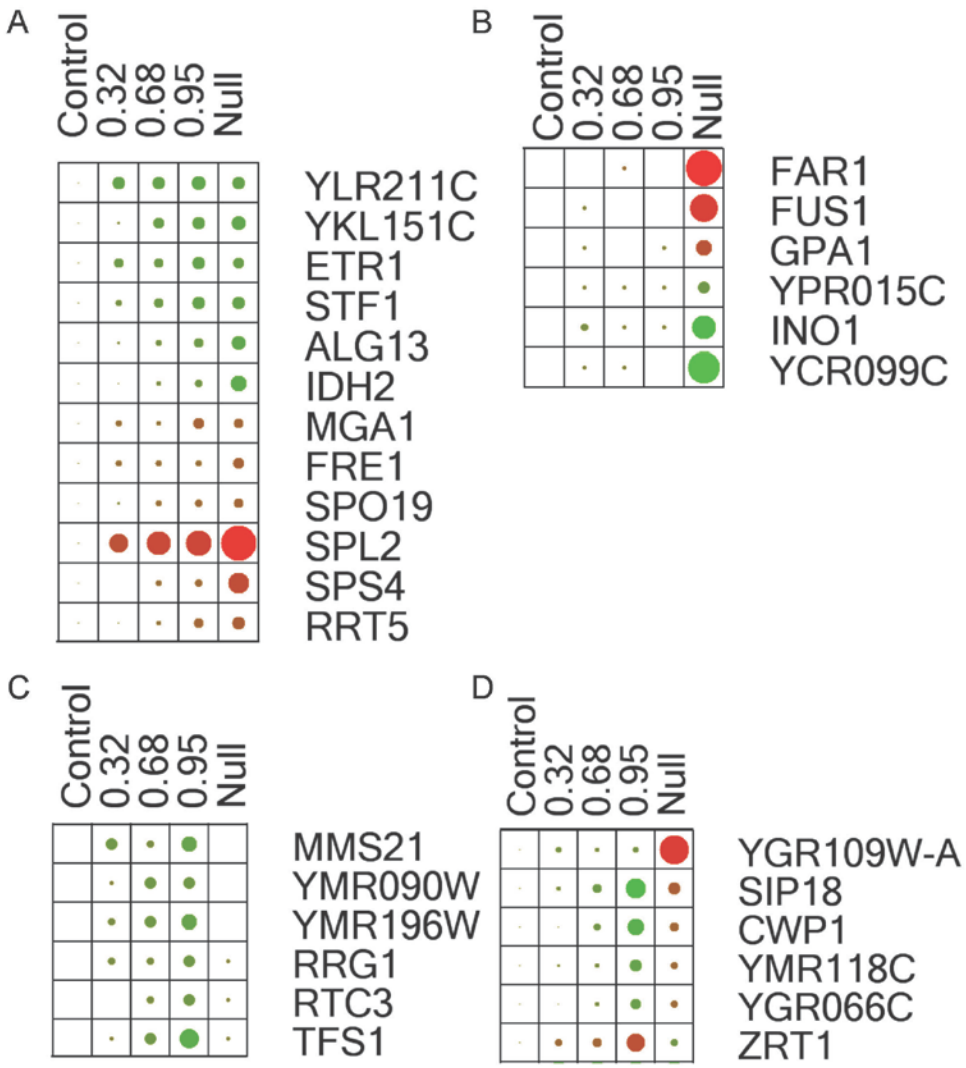
Gcn5-categorization	Number of genes	% of genes exhibiting a graded response
Genes graded/impacted by *gcn5-F221A*	288	100
Correctly determined by KO	153	53
Up-regulated	191	66
Down-regulated	97	34
False negatives	36	12.5
Opposite trends	44	15
Putative genes	64	22

C. Characterization of the 288 genes observed to exhibit a graded response with respect to increasing levels of *gcn5-F221A*.

Four non-mutually exclusive classifications of gene expression were used to characterize the targets found in this study—catalytically associated, non-catalytically associated, false negatives (compared to a knockout), and opposites—by comparing these gene targets to data obtained using the traditional knockout approach. A subset of Gcn5p-impacted genes illustrates these trends (Figure 6.8). The 288 genes identified through *gcn5-F221A* inhibition of native Gcn5p acetylation display a 'graded' response and are therefore associated with Gcn5p catalytic activity. Within this classification, variations in the response to level of gradation exist. Some genes (such as ETR1 and YLR211C, Figure 6.8a) achieve maximal gradation (a plateaued response matching that of a knockout condition) at low levels of *gcn5-F221A*. We posit that similarly responding genes (with low grading thresholds) are strongly impacted by Gcn5p acetylation and potentially have the fewest redundant epigenetic modification mechanisms in yeast. In contrast, genes that require higher levels of the dominant mutant to achieve maximal gradation (such as IDH2 and RRT5, Figure 6.8a) are less sensitive to acetylation by Gcn5p or have more redundant regulation mechanisms. This approach

allows for an evaluation of gene thresholding responses, an important concept in systems biology modeling.

Figure 6.8: Gene expression heat maps for select genes illustrating unique traits.



Four key (non-mutually exclusive) trends in gene expression were observed in this study. In the heat map, red indicates underexpression, and green overexpression relative to the control, and the size of the dot is proportional to magnitude of expression. A select sample of genes was used to display the four trends. **A.** Catalytically associated genes have expression that changes (either up or down compared to control) as a function of *gcn5-F221A*. **B.** 129 non-catalytically associated genes (changed in *gcn5Δ*, but not graded by *gcn5-F221A*) were identified. **C.** 36 false negative genes were identified, in which expression is graded by the dominant mutant, but are unchanged in the knockout strain. **D.** An additional set of 44 genes demonstrate an opposite effect in the presence of the dominant mutant compared to the *gcn5Δ*.

We observed 129 'non-catalytically associated' genes in this data set (genes differentially expressed in the knockout strain, but not significantly impacted in response to the graded dominant mutant) (Figure 6.8b). Since *gcn5-F221A* only inhibits acetyltransferease activity, we hypothesize that these non-catalytically associated genes are not influenced by Gcn5p acetylation activity, but instead are influenced by another indirect effect of Gcn5p, such as protein complex association. Furthermore, we observed that 36 (12.5%) of those genes impacted by the dominant mutant showed no change in expression between the wild-type and knockout strains (Figure 6.7c**,** Figure 6.8c). These 'false negatives' are clearly impacted by Gcn5p activity, and we hypothesize that these genomic loci are directly acetylated by Gcn5p, but in its absence, another HAT with redundant functionality steps in.

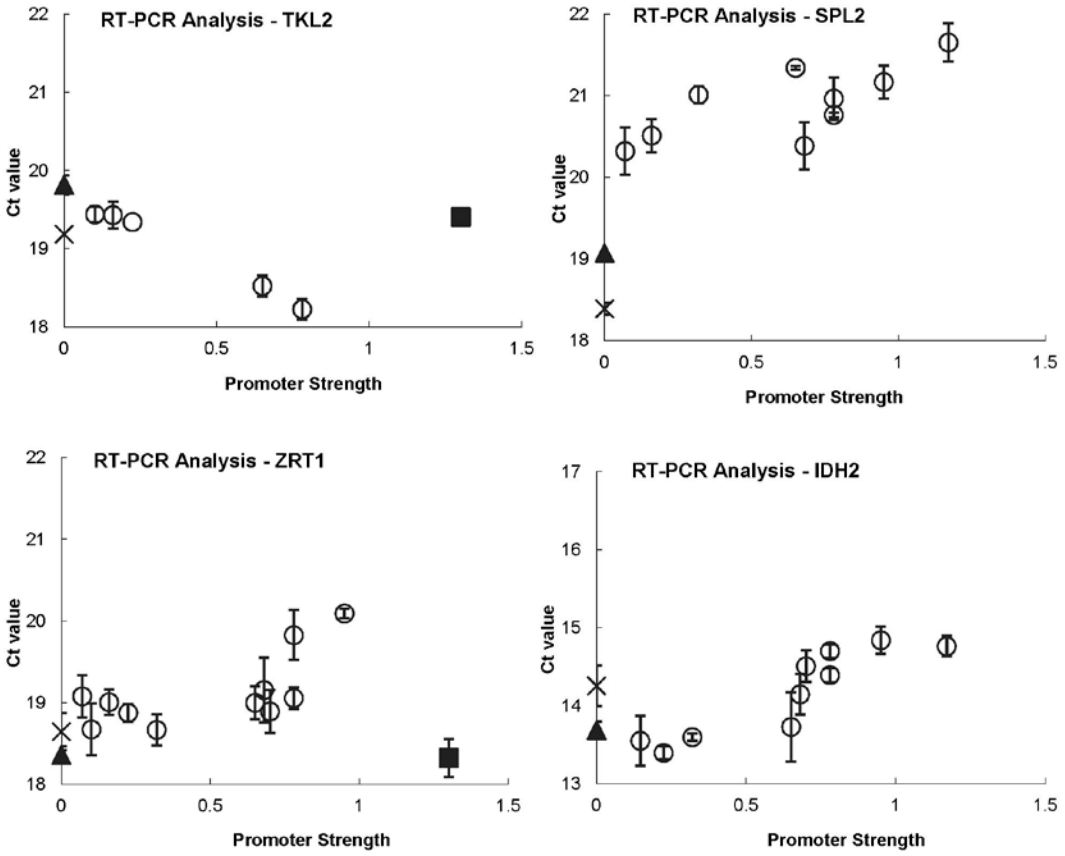
Finally, we observed that 44 genes (15%) impacted by the dominant mutant display an 'opposite' effect in expression as predicted by the gene knockout (Figure 6.7c**,** Figure 6.8d). In the majority of these cases, these genes were over-expressed in response to *gcn5-F221A* but significantly decreased in expression in the *gcn5Δ* strain. Further analysis of this set of genes indicates that nearly half are shown to be associated with the SAGA complex in an independent study[230]. It is likely that this 'opposite' phenomena is due to the partitioning of Gcn5p function and targeting across the domains (potentially the catalytic and bromodomains). In the case of the graded dominant, targeting of the SAGA complex can still occur and thus transcription is increased at these genes. However, in a gene knockout, the entire Gcn5p transcriptional coactivator is missing and thus transcription is impeded significantly. Underacetylated H4 histone proteins have also been shown to have a biased association with SAGA-regulated genes[230], further solidifying the SAGA-complex link to these opposite genes. These results provides another example that removing a globally functioning protein like Gcn5p results in an

artificial genetic background with misleading observations regarding true protein-DNA interactions. Moreover, these results highlight how novel hypotheses of function can be deduced from this approach.

6.3.5: Perturbation control experiments for *gcn5-F221A* mutant

The impact of these various gene classes was further evaluated using RT-PCR for select genes exhibiting a graded response. We used additional promoter strengths to allow for a more quantitative, high-resolution measurement of the impact of Gcn5p catalytic activity. Our global microarray study identified 288 genes whose expression levels were impacted by the *gcn5-F221A* dominant mutant. We selected four of these graded genes; *TKL2*, *SPL2*, *IDH2* and *ZRT1*, and quantified the impact of dominant mutant expression on mRNA levels using real-time PCR. These tests confirmed the results of the microarray study, while adding better resolution by including additional promoter strengths. These results are depicted in Figure 6.9. As an additional complementation control, we included a p416-TEF5-*GCN5* wild-type plasmid transformed into *gcn5Δ*. The cycle threshold values for this complementation control match very closely with that of the wild-type yeast, indicating that the observed differences in gene expression in both this study and the microarray study are not an artifact of replicative plasmids expressing high levels of Gcn5p.

Figure 6.9: RT-PCR of select graded genes confirms microarray findings.

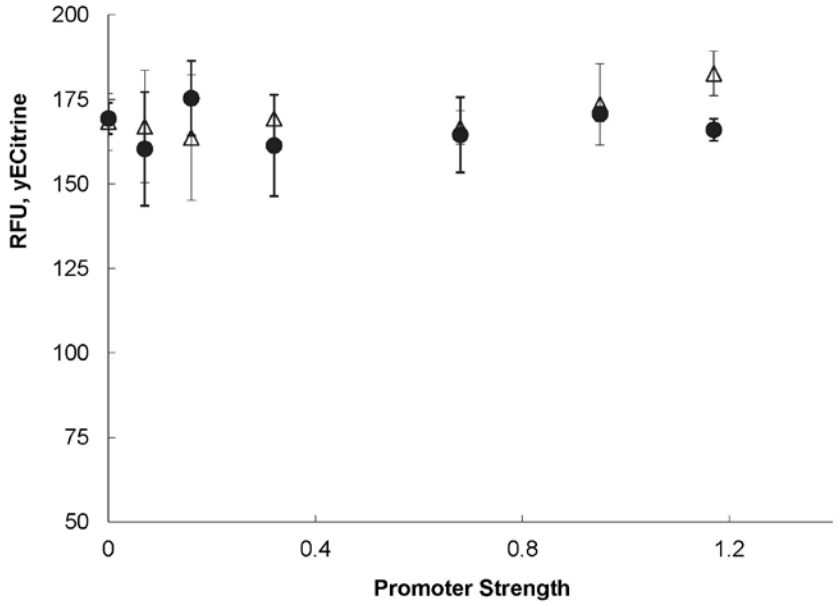


As a follow-up to the *gcn5-F221A* microarray study, 4 graded genes (TKL2, SPL2, ZRT1, IDH2) were selected for higher resolution RT-PCR analysis. RNA was extracted from S288C wild-type cells (▲), S288C *Δgcn5* (■), S288C *Δgcn5* with p416-TEF$_5$-GCN5 (**X**), and S288C with p416-TEF$_x$-*gcn5 F221A* (○). Real-time PCR was performed as previously described using primers 25 to 32 for *TKL2, SPL2, ZRT1* and *IDH2* respectively. Based on the microarray study, *TKL2* was categorized as graded and up-regulated by the dominant mutant with *Δgcn5* displaying false negative behavior. Both trends are reflected in the real time PCR data for *TKL2*. *SPL2*, *ZRT1* and *IDH2* were all categorized as graded and down-regulated by the dominant mutant from the microarray data. Additionally, *Δgcn5* displayed opposite behavior for *ZRT1*. These behaviors are again reflected in the real time PCR data for these three genes. Furthermore, *ZRT1* and *IDH2* show significant gradation at a much higher promoter strength compared to *TKL2* and *SPL2*.

Finally, we sought to determine what, if any, impact varying expression of mutant Gcn5p had on the expression of native Gcn5p. A p415-pGcn5-yECitrine plasmid was constructed, with both a short and long *GCN5* promoter, and co-transformed with the p416-TEF$_x$-*gcn5-F221A* plasmid collection. In this system, fluorescent protein

expression is controlled by the *GCN5* promoter, thus this construct serves as a promoter-based transcription reporter. Using mid-exponential, biological triplicates and flow cytometry, we measured fluorescent levels across the full range of mutant *gcn5-F221A* expression. Regardless of promoter strength driving mutant *gcn5-F221A*, we observed no change in fluorescent expression (Figure 6.10). This result indicates that expression of *gcn5-F221A* does not create artificial feedback or perturbations of native *GCN5* expression. Thus, these results demonstrate the clear link between the data we observe and the lack of catalytic function inherent in *gcn5-F221A*.

Figure 6.10: Expression of *gcn5-F221A* does not influence the native *GCN5* promoter



We sought to determine what impact, varying expression of mutant Gcn5p had on the expression of native Gcn5p. A p415-pGcn5-yECitrine plasmid, with both a short (●) and long (Δ) *GCN5* promoter, is co-expressed with the p416-TEF_x-*gcn5-F221A* plasmid collection. In this system, fluorescent protein expression is controlled by the *GCN5* promoter. Regardless of promoter strength, we observed no change in fluorescent expression.

6.3.6: Chromatin DB analysis determines histone modifications using a graded dominant mutant approach

We next sought to see whether specific chromatin modifications can be deduced from microarray data alone when using a graded dominant mutant approach. To do so, genes identified in our microarray study were analyzed using Chromatin DB[231] (Appendix E). Using genes identified by a *gcn5Δ* knockout (including subclasses of up-regulated, down-regulated, and differentially expressed), no significant enrichment or depletion of chromatin lysine acetylation is evident. This same lack of enrichment or depletion is observed using the microarray data obtained by a separate and independent *gcn5Δ* study[230]. However, by examining the graded up genes with low grading thresholds identified in this study, significant depletion is seen in H2BK11ac, H2BK16ac, H3K18ac, H3K14ac, and H3K23ac with p-values of less than 10^{-3} to 10^{-4}. This profile of histone modifications mimics those observed in a Gcn5p binding study[213]. Furthermore, the 'false negative' gene set exhibits acetylation depletions for the same lysine residues as genes that are graded up. This clearly demonstrates that 'false negative' genes are indeed direct targets of Gcn5p acetylation and explains why the vast majority of these 'false negative' genes increase in expression in response to *gcn5-F221A*. In contrast, the non-catalytically associated data set exhibits no enrichment or depletion of chromatin lysine modifications. In the case of those genes exhibiting an 'opposite' response, the primary histone modification that is observed is a depletion of H4K16ac (p-value $<10^{-3}$). It is well known that Sir2p and Esa1p are responsible for targeting H4K16[232], which implicate the actions of these proteins as potential compensators for Gcn5p. Collectively, these results demonstrate that the graded dominant mutant approach can identify the canonical acetylation targets of Gcn5p[233].

6.3.7: Gene ontology analysis reveals new cellular processes impacted by Gcn5p acetylation

Gene ontology and network analysis tools were used to further classify the genes influenced by *gcn5-F221A* activity and evaluate the dataset (Appendix F). Three functional classes (nucleolus, ribosome biogenesis, and RNA metabolic processes) were significantly enriched in the set of genes exhibiting an under-expression graded response. Nearly 70% of the genes exhibiting under-expression were associated with these functional classes. Furthermore, one gene ontology class (oxidoreductase activity) was overrepresented in genes exhibiting an over-expression in a graded fashion, and is thus a target for Gcn5p-based gene repression. This analysis expands the role of Gcn5p activity to other fundamental cellular processes.

6.3.8: Graded dominant mutant approach can interface with phenotypic and genetic assays

Finally, we sought to demonstrate how the graded dominant mutant approach can be used in conjunction with phenotypic and genetic assays. Prior to this work, it was unclear whether acetylation or protein-protein interaction is the root cause of *gcn5Δ* synthetic lethal genes. To address this issue, we paired a gene deletion with various promoter strengths driving *gcn5-F221A* to simulate the lethal double knockout strain in the haploid yeast BY4741. Twenty two of these synthetic lethal genes were selected for this study and evaluated (Table 6.1). Only three gene knockouts, *Δccr4*, *Δrsc2* and *Δrtt109*, were highly impacted in a graded fashion by the dominant mutant. *Δrtt109*, a HAT known to acetylate H3K56 and H3K9[234], demonstrated the most significant impact. Fifteen of the gene deletions were moderately impacted and four showed almost no change. Collectively, these results implicate the relative importance of Gcn5p's catalytic activity versus its protein and DNA interactions. As a comparison, ten BY4741 null

strains were selected at random to serve as a control group. None of these strains showed a growth-rate dependent response to the *gcn5-F221A* mutant, indicating the significance of the results described above.

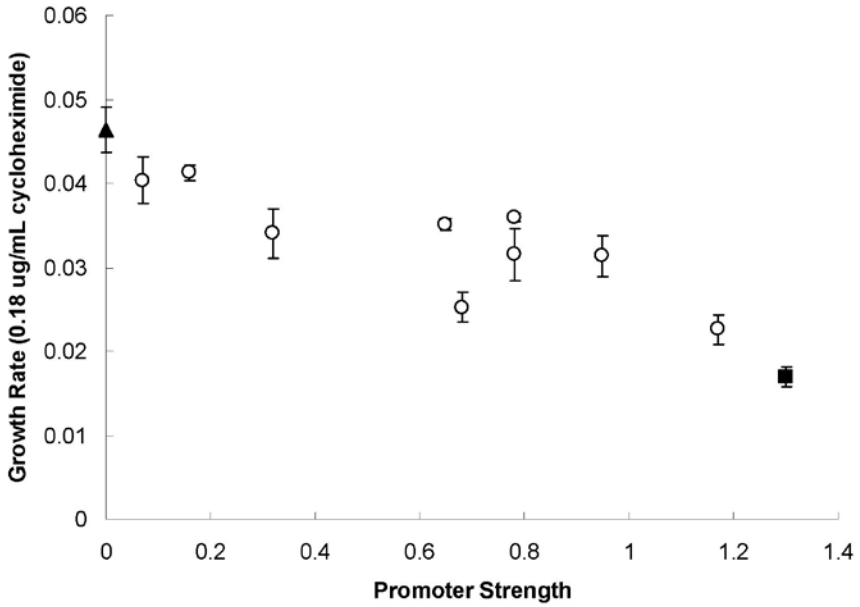
Finally, we sought to investigate the global impact of Gcn5p acetyltransferase activity on cellular phenotypes. To do so, we evaluated the impact that grading this activity has on the basis of documented large-scale chemical tolerance assays of null mutants. Yeast strains with a *gcn5* null allele have previously been shown to have increased sensitivity to cycloheximide[235], ethanol[235], 5-fluorouracil[236], KCl[237], $MnCl_2$[237], $CaCl_2$[237], and sulfanilamide[238]. Growth inhibition assays were performed using strains containing gradations of the dominant mutant, as well as a wild-type control and a *Δgcn5* control, as described in the Materials and Methods.

Table 6.1: Impact of *gcn5-F221A* on the growth rate of *GCN5* synthetic lethal genes.

***gcn5* synthetic**	**Control**	**.16 growth**	**.32 growth**	**.68 growth**	**.95 growth**
ccr4	.225 ± .003	.218 ± .007	.192 ± .000	.180 ± .003	.168 ± .001
eaf7	.252 ± .003	.225 ± .001	.232 ± .002	.225 ± .001	.208 ± .003
elp3	.229 ± .001	.220 ± .002	.212 ± .002	.208 ± .001	.208 ± .002
hhf2	.249 ± .003	.234 ± .004	.222 ± .001	.209 ± .002	.206 ± .002
hht2	.253 ± .006	.227 ± .022	.238 ± .004	.209 ± .006	.221 ± .009
hsl1	.260 ± .006	.250 ± .004	.244 ± .002	.228 ± .009	.222 ± .013
hsl7	.244 ± .001	.232 ± .004	.243 ± .005	.216 ± .007	.197 ± .003
iki3	.225 ± .002	.217 ± .002	.219 ± .003	.208 ± .002	.206 ± .003
leu2	.255 ± .004	.237 ± .004	.232 ± .002	.227 ± .002	.219 ± .001
mot2	.271 ± .008	.248 ± .005	.258 ± .006	.234 ± .003	.230 ± .003
nam2	.233 ± .001	.228 ± .003	.235 ± .000	.217 ± .005	.219 ± .000
not5	.259 ± .002	.246 ± .003	.244 ± .001	.230 ± .001	.221 ± .002
paa1	.245 ± .005	.236 ± .004	.195 ± .004	.188 ± .005	.197 ± .010
pap2	.263 ± .005	.250 ± .002	.246 ± .002	.236 ± .001	.232 ± .006
pho23	.231 ± .003	.228 ± .001	.218 ± .004	.219 ± .004	.210 ± .003
rad6	.193 ± .008	.180 ± .002	.176 ± .005	.163 ± .004	.159 ± .002
rpd3	.260 ± .003	.247 ± .005	.240 ± .002	.232 ± .002	.222 ± .003
rtt109	.204 ± .003	.167 ± .010	.154 ± .007	.131 ± .006	.131 ± .004
rsc2	.209 ± .013	.194 ± .001	.194 ± .004	.167 ± .006	.158 ± .000
sin3	.195 ± .001	.183 ± .002	.186 ± .008	.162 ± .022	.173 ± .005
snf2	.176 ± .003	.161 ± .012	.139 ± .020	.146 ± .002	.152 ± .019
spt20	0.149 ± .001	0.144 ± .002	0.151 ± .002	0.142 ± .002	0.140 ± .004
Random	**Control**	**.16 growth**	**.32 growth**	**.68 growth**	**.95 growth**
cad1	.155 ± .022	.128 ± .070	.190 ± .018	.163 ± .028	.175 ± .024
hpa2	.124 ± .006	.161 ± .024	.139 ± .006	.138 ± .024	.132 ± .007
lsb3	.136 ± .039	.131 ± .042	.130 ± .007	.131 ± .006	.132 ± .003
nma2	.163 ± .031	.141 ± .015	.147 ± .022	.114 ± .014	.160 ± .007
nup170	.125 ± .014	.135 ± .023	.151 ± .023	.110 ± .014	.159 ± .020
rpl20b	.120 ± .034	.113 ± .019	.128 ± .063	.125 ± .011	.152 ± .018
vps28	.147 ± .017	.138 ± .023	0.139 ± .018	.145 ± .002	0.126 ± .043
YBR287W	.154 ± .021	.172 ± .022	.144 ± .033	.144 ± .002	.138 ± .034
YML131W	.138 ± .009	.120 ± .007	.144 ± .031	.166 ± .016	.126 ± .022
YNL234W	.141 ± .019	0.164 ± .017	0.165 ± .040	0.154 ± .020	0.154 ± .039

We examined 22 gene knockouts with known synthetic lethal interactions to *gcn5* null. The *gcn5-F221A* dominant mutant was expressed at varying levels in the background of a knockout strain and growth rate was measured. Three gene knockouts (shown in bold), are highly impacted by the *gcn5-F221A* mutant and exhibited a more than a 20% reduction in growth rate when the mutant was highly expressed. The majority of the genes show a moderate decrease in growth rate (shown in underline) as mutant expression is increased while several showed no growth rate changes. Ten gene deletions were randomly selected as a control and none exhibit any response.

Figure 6.11: Decreased growth rate caused by cycloheximide treatment of S288C is linked with Gcn5p acetylation activity.



Using an S288C wild-type strain expressing *gcn5-F221A* at varying promoter strengths (○), we measured growth rate in the presence of 0.18 μg/mL cycloheximide. As mutant expression increased, growth rate decreased and approached that of the *gcn5Δ* strain (■). At low mutant expression levels, growth rate resembled that of the wild-type strain (▲). This demonstrates that the cellular response to cycloheximide is linked with Gcn5p acetylation. A similar impact was observed at the *HIS3* locus (Figure 6.3), a known Gcn5p gene target.

When treating the strains with cycloheximide, we observed a graded, linear decrease in growth rate that coincided with increasing *gcn5-F221A* expression, showing that cycloheximide tolerance is controlled by Gcn5p acetyltransferase activity (Figure 6.11). Assays performed using ethanol, 5-fluorouracil, KCl, 4mM $MnCl_2$, and 8mM $MnCl_2$ as a growth inhibitor did not exhibit this trend (Table 6.2). These growth inhibitors are akin to the non-catalytically associated gene expression data set, and we hypothesize that increased sensitivity to these growth inhibitors is not a result of decreasing cellular Gcn5p acetylation, but by a separate, indirect effect. Despite prior reports, sulfanilamide, $CaCl_2$, and 40mM $MnCl_2$ inhibitors did not impact growth rate for any of the strains in our liquid-culture based experiment.

Table 6.2: Putative *GCN5*-Dependant growth inhibitors tested for an impact with mutant

Putative *GCN5*-Dependent	Test Culture	Previous Study	Previous
Ethanol	6% w/v	High throughput	Solid
Cycloheximide	0.18μg/mL	High throughput	Solid
Sulfanilamide	200μg/mL	High throughput	Liquid
5-Fluorouracil	15μg/mL	High throughput	Solid
KCl	1M	Individual	Solid
$CaCl_2$	0.25M	Individual	Solid
$MnCl_2$	4mM	Individual	Solid
$MnCl_2$	8mM	Individual	Solid
$MnCl_2$	40mM	Individual	Solid
$MnCl_2$	200mM	Individual	Solid

Previous studies have identified the following compounds and concentrations which inhibited growth of a *Δgcn5* strain compared to wild-type yeast. We tested the growth of S288C strains expressing *gcn5-F221A* in the presence of these compounds in liquid media, as described in the Materials and Methods.

6.4 CONCLUDING REMARKS

Using a graded dominant mutant approach and Gcn5p as a case study, we are able to determine global gene targets and impacts, and to extract the causative linkage between the catalytic domain of Gcn5p and gene regulation. In particular, we (1) find evidence that Gcn5p is more involved in cell-wide gene repression, instead of the accepted gene activation associated with HATs, (2) identify previously unknown gene targets and interactions for Gcn5p-based acetylation, (3) quantify the strength of some Gcn5p-DNA associations, (4) demonstrate that this approach can be used to correctly identify canonical chromatin modifications, (5) establish the role of acetyltransferase activity on synthetic lethal interactions, and (6) identify new functional classes of genes regulated by Gcn5p acetyltransferase activity—all six of these major conclusions were unattainable by using standard gene knockout studies alone. These results demonstrate the power of the graded dominant mutant approach, which unlike traditional methods, only impacts one particular facet of the querying protein (in this case, acetyltransferase

activity) and is therefore especially useful for studying multifunctional proteins and global regulators.

Based on the results presented here, we would recommend that a graded dominant mutant approach be utilized in conjunction with a traditional gene knockout to study gene regulatory proteins, especially those that serve multiple functions. The resulting data is higher-resolution and more accurately defines protein domain function and influence. While demonstrated here for the case of acetyltransferase activity of the yeast protein Gcn5p, this approach can theoretically be extended to other proteins and domains of interest. This approach uniquely enables a systems biology view of the cell while at the same time leveraging synthetic biology tools[239]. The identification of dominant mutations that can remove single functions are either well-documented for many proteins of interest or can be identified with the proper genetic screens. Additionally, promoter libraries with documented expression capacity are available for most major model systems[222,240,241]. Thus, this approach is generalizable for other proteins in classes such as epigenetic modification, signaling cascades, and transcriptional regulation as well as for essential genes, which cannot be deleted, and this method is not necessarily restricted to the yeast system studied here. In addition, this approach can be combined with any cell state assay including, but not limited to, gene expression analysis, phenotypic assays, genetic screens, ChIP analysis, and metabolomics. In conclusion, the graded dominant mutant approach is able to circumvent the problems seen in standard genetic approaches and can provide a causative linkage between specific protein function and phenotype.

Chapter 7: Conclusions and Major Findings

Compared to other cellular hosts, our ability to genetically manipulate and engineer human cell lines (and other mammalian hosts such as CHO) is limited. In particular, robust, stable, high expression of heterologous DNA is difficult and standard approaches are both unreliable and variable in their results. Resolving this issue through the development of tools that enable predictable gene expression would be greatly advantageous. Not only would such tools have great traction in mammalian cell research, they could enable medical progress in the areas of gene therapy and disease states. Furthermore, these tools would be directly applied to the production of protein therapeutics and could result in significant time and cost savings to current practices.

We evaluated the influence of a selection marker, a nearly ubiquitous element of mammalian expression cassettes, on the cell line development process[172]. This is the first study to compare four common antibiotics, hygromycin, neomycin, puromycin and Zeocin, in the same context. We evaluated the selection agents on the quality of selected populations, the stability of those populations, and the expression levels of clonal populations. Across all metrics and two human cell lines (HT1080 and HEK293), Zeocin outperformed the other antibiotics, and is therefore recommended for human cell line development.

We sought to identify sites across the human genome that support stable, high level expression of transgenic DNA. It is well established that integration locus strongly influences foreign gene expression levels, yet little effort has been made to catalogue or identify the quality of sites across the genome. Although several sites have been used for transgenic integration, and in some cases the performance of those sites have been evaluated[66], the majority of these studies look only at protein coding regions, which

represents less than 3% of the human genome. Alternatively, we applied an unbiased, random integration approach, followed by isolation of productive, single copy clonal cell lines, and retrieval of the integration locus. In doing so, we identify eight sites in the genome that support heterologous gene expression. These sites are distributed in both intronic regions and non-protein coding regions, which demonstrates that searching only exonic regions is limiting. Furthermore, we find transgenic integration at the majority of the loci caused negligible perturbation to expression of surrounding genes, indicating these loci could be used in gene therapy applications[65]. These results are profound, and can be coupled with site-specific targeting methods to reliably deliver transgenic DNA to a hospitable genomic locus.

Site-specific integration and genome editing techniques are critical for genetic manipulations including controlled integrations and endogenous gene deletions. In recent years, this component of mammalian cell engineering research has seen significant technological advancement[167,242]. Nonetheless, there remains room for improvement. We sought to optimize a Cre recombinase method for recombination-mediated cassette exchange[173]. We used a dual fluorescent assay to measure recombination efficiencies without selection, and examined parameters including mutant targeting sequences, delayed introduction of the recombinase, and varying the ratio of the recombinase to exchange cassette. Pairing a highly mutated targeting sequence with a native sequence, we observed a more than two-fold average increase in recombination efficiency. By delaying the introduction of Cre recombinase 18 to 20 hours, we achieved recombination efficiencies of more than an order of magnitude greater than previous reports. These optimizations greatly improve the efficiency of Cre based recombination-mediated cassette exchange and can be applied to related enzymes such as Flp recombinase and ΦC31 integrase.

Another major component of this dissertation utilizes a systems biology approach to develop generic cell engineering techniques. As our ability to collect global information about cellular hosts has increased through transcriptomics, proteomics, metabolomics and other approaches, we are able to generate vast amounts of data under a variety of environmental conditions. This data is continually being generated, is often publicly available and enables a systems level view of the cell. In some cases, this information is incorporated and used to evolve and update the state of the art, including *in silico* metabolism models and gene interactions databases. However, there are many common cell engineering approaches that have not been reevaluated on a systems level, which could be used to guide the development of better recombinant DNA tools[239].

Codon optimization, which is a widely used technique for heterologous gene expression, is one such tool not previously been examined from a systems biology perspective. Currently, synonymous codon optimization is determined through the codon usage of all protein coding genes in the host's genome. Although this approach improves gene expression in many cases, it fails inexplicably in other cases. We have developed an alternative, condition-specific codon optimization method. This method determines synonymous codon optimization using the genes known to be highly expressed under relevant growth conditions and uses codon context to stochastically generate optimized gene variants. We implemented this approach in *S. cerevisiae* and successfully improved expression of three heterologous genes (eGFP, CatA and Lac1) under three different conditions (constitutive high expression, stationary phase growth and xylose and glucose as carbon sources). Condition-specific codon optimization is a useful tool for metabolic engineering and can be generically applied to a variety of environmental conditions and can be implemented in any genome-sequences cellular host.

Multifunctional proteins perform important functions in eukaryotic cells including gene regulation and epigenetic modification. These proteins are typically studied through gene deletion, which is problematic because gene and protein targets, as well as phenotypic changes, cannot be directly linked to any one domain. We developed an alternative Graded Dominant Mutant Approach that can be used to study multi-functional proteins[229]. This approach uses a dominant mutant allele able to competitively inhibit the domain of interest expressed in the presence of the wild-type allele. Combining the dominant mutant with a promoter library allows for modulated mutant expression and discrete levels of competitive inhibition. A case study with *S. cerevisiae* histone acetyltransferase, Gcn5p, was performed. Using a Graded Dominant Mutant Approach and global microarray, we were able to identify over 289 catalytically associated Gcn5p gene targets and in many cases, determine the strength of their interactions with Gcn5p. Through gene ontology, we identified several biological functions associated with Gcn5p acetylation and determined for the first time that this modification is largely repressive in *S. cerevisiae.* The quality of this information far exceeds that which can be achieved with a Δ*gcn5* knockout strain. Furthermore, this approach is generic and can be coupled with many global techniques and implemented in a variety of cellular hosts.

Collectively, these studies constitute significant contributions to the genetic tools available for eukaryotic hosts. I have developed and characterized novel approaches that enable precise, stable gene expression in mammalian cells. Furthermore, I applied a systems level approach to develop methodologies that enable complex phenotypes in *S. cerevisiae*. These methods are generic and can be adapted to higher eukaryotes. The development of these tools represents a novel addition to the field of cellular engineering and has many applications in both biotechnology and medicine.

Chapter 8: Proposals for Future Work

This work is composed of studies in the areas of predicted gene expression for mammalian cellular hosts and cell engineering through systems biology. Collectively, these experiments have resulted in significant findings that will be impactful in basic biology, cell engineering, biotechnology and medicine. Furthermore, this work can be used to guide and outline the future direction of studies in these areas. Here, I outline unanswered questions as well as several follow-up studies that can be carried out to expand upon these findings.

We evaluated four common antibiotic selection systems to determine their impact on human cell line development. In doing so, we identify that Zeocin performs better in three critical areas; recombinant population identification, stable gene expression, and candidate cell line identification. These trends were consistently demonstrated across two industrial human cell lines, HT1080 and HEK293. There is interest in extending this type of study to other mammalian cell types, including Chinese hamster ovary (CHO) cells, because of their importance in the biopharmaceutical industry and protein production[14]. In addition to examining the antibiotic markers studied here, it would be advantageous to extend the study in CHO to auxotrophic marker systems (dihydrofolate reductase and glutamine synthetase) because of their popularity in industrial cell line development. One concern in addressing the influence of selection markers in CHO is that transgene duplication is commonplace[38], as well as chromosomal aberrations[243,244], which could confound results of the study. For this reason, transgene copy number must be carefully evaluated.

Additionally, it would be interesting and valuable to try to understand why Zeocin identifies better recombinant cell populations compared to other common antibiotics.

Zeocin's mode of action involves DNA cleavage, while the other antibiotics evaluated here either interfere with ribosome function or translation. Although it is clear that Zeocin functions through a different mechanism compared to other selection systems, the exact details of this mechanism have not been previously explored. By gaining a better understanding of Zeocin's intercellular mechanism, it may be possible to identify other antibiotics or Zeocin analogues that act in a similar manner and are likewise advantageous and robust for cell line development applications. Regardless, increased options for mammalian selection markers would increase flexibility for synthetic circuit design.

We have identified eight stable, high transcription loci in the human genome using a GFP reporter system. The initial identification of these loci was done in HT1080. As a follow-up, these sites are being retargeted in a second human cell line, HEK293, and compared to illegitimate integration to validate that the loci are advantageous and not cell-line specific. In addition to fluorescent proteins, these sites should be evaluated with secreted proteins, such as SEAP and a model IgG, which better represent target products for these cells. This will also allow for productivity measurements on a per cell basis. Finally, bioreactor scale-up experiments can be conducted on the 1-2 most productive cell lines to demonstrate the industrial applicability of this technology and determine maximum titers.

Furthermore, it would be advantageous to identify a larger set of stable, high transcription loci. Although we were able to establish sufficient populations, both clonal and heterogeneous, exhibiting high and stable expression, we were limited by our ability to determine the corresponding loci of integration. The low-throughput methods we employed were both slow and not sufficiently robust. Alternatively, hundreds of integration loci could be more easily identified using high-throughput deep sequencing

techniques. This approach would eliminate the necessity of single clonal populations, as well as PCR amplification of the integration locus. Identifying a larger set of advantageous integration sites would provide increased flexibility for cell line engineering applications and expand our understanding of Safe Harbor criteria for gene therapy applications[65].

We examine and optimize one approach for site-specific retargeting using the Cre recombinase enzyme. In doing so, we identify two criteria that significantly impact efficiency; mutated *lox* targeting sequences and delayed introduction of Cre recombinase. These modifications result in retargeting efficiencies of 8-13%[173]. We observed that of the mutated *lox* targeting sequences we evaluated, the heavily mutated *lox*Fas site out-performed the less-mutated *lox*5171 and *lox*2272 sites. Thus, more research should be conducted to identify highly mutated *lox* sites that can further improve efficiencies. Additionally, this method has not been directly compared to the many other genome editing techniques available, including ZFNs, TALENs and CRISPR. It would be advantageous to directly compare the efficacy of several site-specific targeting methods at several integration loci. This would not only help evaluate the robustness of the approaches, but also identify if some approaches are less variable than others with regards to locus.

The condition-specific codon optimization approach has been demonstrated as an improvement over conventional codon optimization in *S. cerevisiae* using two heterologous genes. This approach is currently being extended to a third case. We are demonstrating that carbon-source codon optimization (glucose vs. xylose) is advantageous in the expression of a laccase gene native to *M. albomyces*. Laccases are a useful class of enzymes with applications in industrial biosciences because of their ability to oxidize phenolic compounds. Furthermore, the utilization of alternative carbon

sources is useful in applied biosciences. Additionally, other useful classes of enzymes and relevant process conditions can be evaluated.

Condition-specific codon optimization is a generic approach that combines systems level information and codon context to optimize gene sequence. To demonstrate that this approach is generic, it should be extended beyond *S. cerevisiae* to other industrially relevant cellular hosts including *Pichia pastoris, Bacillus subtilis,* and *Yarrowia lipolytica.*

The graded dominant mutant approach is a methodology for understanding the intercellular interactions of a single domain of a multifunctional protein. This method was applied to the acetylation domain of Gcn5p in *S. cerevisiae*, which enabled the identification of previously unknown gene targets. This methodology can be expanded to the other Gcn5p domains, the Ada2 binding domain and the bromodomain. After identifying dominant mutants corresponding to these other loci, this approach can be used globally to map all of the gene targets for each of Gcn5p's three domains. Furthermore, the mutants can be co-expressed to get a complete picture of Gcn5p's targets and how they relate to each functionality.

Additionally, this graded dominant mutant approach is generic and can be expanded to other enzyme classes and organisms. It would be advantageous to use this approach, in conjunction with site-specific genome editing techniques, to study epigenetic factors in human cell lines. Because gene knockout is so difficult in mammalian cell lines, this graded dominant mutant approach would be a useful alternative to understand *in vivo* interactions of regulatory proteins. Many epigenetic classes, including methylases and histone deacetylases have been implicated in cancer disease states[245-249] and would be good candidates for this work.

Collectively, the experimental findings described herein resulted in significant findings that will be impactful in basic biology, cell engineering, biotechnology and medicine and have laid the groundwork for both precise and complex cellular engineering of eukaryotic organisms. These findings can be used to guide future experiments and used as a framework for the extension of these approaches to other cellular hosts.

Chapter 9: Materials and Methods

9.1 Common Materials and Methods

9.1.1 Conditions and media for human cell growth

The suspension-adapted and serum free cell lines were provided by Shire Pharmaceuticals. HT1080[250] was established from ATCC #CCL-121 and HEK293[251] from ATCC#CRL-1573. HT1080 cells were grown in a defined media with an added 4mM glutamine, 1x penicillin-streptomycin, and pH adjustment to 7.20. HEK293 cells were grown in HyClone media (ThermoFisher) supplemented with 4mM glutamine and 1x penicillin-streptomycin. All cells were passaged every 48 to 72 hours and seeded between 2e5 and 3e5 viable cells/mL. Cell viability, concentration, and size were measured using a Beckman Coulter ViCell. Shake flasks were maintained at 37ºC, 5% CO_2, humidity above 80% and 125 rpm.

9.1.2 Conditions and media for microbial growth

Bacteria were grown in lysogeny broth with ampicillin at 37ºC. Yeast were grown at 30ºC. YPD media contained 20g/L yeast extract, 10g/L peptone and 10g/L glucose. Minimal media contained yeast nitrogen base and 20g/L glucose and was supplemented with amino acids; 0.77 g/L of CSM –Ura (MP Biomedicals) for p426/416 vectors, 0.77 g/L of CSM –His for p413 vectors and 0.79 g/L of CSM for p41K vectors. For gentamycin resistant strains, media was supplemented with 200 μg/L of G418. Agar plates were grown in standing incubators and cultures in shakers operating at 225 rpm. Passage numbers for yeast cultures were kept low (2–3) for all experiments.

9.1.3 Transfection and selection of human cell populations

Prior to transfection, plasmid DNA was extracted from 150mL of DH10β culture using the Qiagen High Speed Maxi Prep kit. If linearization was necessary, DNA was digested overnight with appropriate restriction enzymes at 37ºC. Linearized DNA was purified using a phenol-chloroform extraction. To establish recombinant cells, batches of 12 million viable cells were re-suspended in RPMI media (0.75 mL per cuvette) and transfected with 30 to 50μg of plasmid DNA using a 4mm electroporation cuvette, 950μF of capacitance and either 350V (HT1080) or 160V (HEK293). Surviving cells were then transferred to the respective growth media and allowed to recover for 48-72 hours before the addition of selection pressure. Selective pressure was maintained until culture viability was above 90%.

9.1.4 Isolation of human cell clonal populations

To establish single cell clones, resistant cultures were diluted to permit the addition of one cell per well in EX-CELL CHO cloning media (Sigma), supplemented with 4mM glutamine, 1x penicillin-streptomycin, and 1x Insulin-Transferrin-Selenium (Invitrogen). 150μL were plated per well in 96 well plates. After one week, 100μL of cloning media was added. Upon the appearance of a cell mass, the contents of wells were transferred to a six well plate and split when confluent. The single cell clones were expanded until freezer stocks were established. Freezer stocks consisted of 12 million cells in 1mL of 10% DMSO preconditioned media.

9.1.5 Flow cytometry for human cell populations

GFP or mStrawberry fluorescence profiles were examined using either a FACS Fortessa or FACS Calibur. In the case of the FACS Fortessa, the 488nm laser was used to detect GFP fluorescence (505 LP, 530/30). Forward scattering had a voltage setting of

82, side scattering of 181, GFP fluorescence of 381. Forward and side scattering data were linear and fluorescence was collected on a logarithmic scale. In the case of the FACS Calibur, forward scattering had a voltage setting of E00, side scattering 371, GFP fluorescence 381 and Strawberry fluorescence 375 and ampgain was 1.00. Compensation for GFP and Strawberry fluorescence was set to 0.8% and 33%, respectively. Analysis of flow cytometry data was performed using FlowJo version 7.6.

9.1.6 Flow cytometry for yeast populations

Flow cytometry was used to determine eGFP and yECitrine expression. Stationary phase culture was used to inoculate 6mL of appropriate media at an OD600 of 0.005. Biological triplicates were grown for approximately 12 hours, allowing cultures to reach mid-log phase. Cells were pelleted and re-suspended in cold water. Fluorescent expression profiles were determined using both a FACS Calibur and Fortessa and compared to a control population. For the FACS Calibur, forward scattering had a voltage setting of E00 and ampgain of 2.96, side scattering a voltage of 505 and ampgain of 1.00 and fluorescence a voltage of 551 and ampgain of 1.00. Forward and side scattering data were linear and fluorescence was collected on a logarithmic scale. Threshold was set to a forward scattering value of 52. For the FACS Fortessa, forward scattering had a voltage setting of 209 and ampgain of 1.00, side scattering a voltage of 209 and ampgain of 1.00 and fluorescence a voltage of 308 and ampgain of 1.00. Forward and side scattering data were linear and fluorescence was collected on a logarithmic scale. Threshold was set to a forward scattering value of 5000 with an OrOperator and area scaling of 0.71. Gating and statistical analysis of the data was performed using FlowJo 7.6.

9.2 MATERIALS AND METHODS FOR CHAPTER 2

9.2.1 Plasmid Construction

The pAML-hrGFP plasmids (Figure 2.1) were constructed through modification of the pCI-neo vector (Clonetech) by introducing a human optimized GFP gene (hrGFP), an EMVC-based IRES site, and one of four distinct selection marker genes. The plasmid backbone (including the cytomegalovirus (CMV) immediate-early enhancer promoter, polyA tail and bacterial replication elements) were amplified from the pCI-neo vector using primers 1 and 2 (Table A.1), which include BamHI restriction sites. Using primers 3 and 4, the hrGFP gene was amplified from pIRES-hrGFP-1a (Stratagene) and cloned directly after the CMV promoter using a HindIII restriction site. The IRES site, amplified from pIRES-hrGFP-1a (Stratagene) using primers 5 and 6, was cloned directly after the hrGFP gene using an XbaI restriction site. Finally, using a ClaI restriction site, each of the four selection marker resistance genes were added directly after the IRES site. The puromycin resistance gene (600 bp) was amplified using primers 7 and 8 from the pLKO.1-puro plasmid (Addgene). The hygromycin resistance gene (1029 bp) was amplified from pAG26[252] using primers 9 and 10. The Zeocin resistance gene (375 bp) was amplified from the pSV40-zeo2 plasmid (Invitrogen) using primers 11 and 12 and. The neomycin resistance gene (795 bp) was amplified from the pCI-neo vector using primers 13 and 14. The total plasmid size (excluding selection marker) was 4475 bp. The largest size difference between two markers is 654 base pairs (hygromycin and Zeocin), which represents 13% of total plasmid size. All plasmids were transformed and propagated in *E. coli* DH10β. All primers for this work can be found in Appendix A, Table A.1.

9.2.2 MIC_{75} Measurements

In order to fairly compare the four selection markers, pools were established using a selection concentration corresponding to an MIC_{75} for both HT1080 and HEK293, as determined empirically through experiments. At time zero, 30 mL of healthy HT1080 or HEK293 wild-type cells (viability above 90%, density of 3e5 cells/mL) were treated with four different selection agents at six different concentrations and monitored for cell viability over the course of 8-10 days. For both cell lines, concentrations of 10, 25, 50, 75, 100 and 200 µg/mL of Zeocin and hygromycin were used, 25, 50, 75, 100, 200 and 350 µg/mL for neomycin and 5, 2.5, 1, 0.75, 0.5, 0.25 and 0.1 µg/mL for puromycin. Based on viability data, an MIC_{75} was calculated for each selection agent; In HT1080, 50 µg/mL for hygromycin and neomycin, 70 µg/mL for Zeocin and 0.7 µg/mL for puromycin. In HEK293, 85 µg/mL for hygromycin, 170 µg/mL for neomycin, 60 µg/mL for Zeocin and 1.0 µg/mL for puromycin. These values are summarized in Table 2.1.

9.2.3 Cell Line Development

For each antibiotic, 3 batches of 12 million viable cells were transfected with the appropriate DNA construct, as previously described. Surviving cells were transferred to the respective growth media and allowed to recover for 72 hours before the addition of selection pressure at the measured MIC_{75} levels. Excluding neomycin, these selection conditions were sufficient to establish stable libraries and employed for selection. Neomycin selection at the MIC_{75} levels resulted in significant cell debris, especially in the HEK293 library. In HEK293 cells, the neomycin concentration had to be lowered, first to 85 µg/mL and eventually to 42.75 µg/mL, in order to establish a stable pool with viability above 90%. Because we were unable to establish a neomycin-resistant HEK293 population at the MIC_{75} concentration, we did not attempt a duplicate, nor include

neomycin in the mRNA analysis. Although plasmid size can sometimes influence transfection efficiency, we observed no such trend with the plasmids used in this study.

9.2.4 Real-Time PCR Measurements

Relative GFP expression at the RNA level was evaluated for hygromycin, puromycin and Zeocin stable HT1080 populations. Five million cells were collected by centrifugation from each population and RNA was extracted using the RiboPure RNA kit (Ambion). 1100 ng of RNA were then converted in a 40μL total reaction volume to cDNA using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). Real-time PCR for both the hrGFP gene and the RPS11 housekeeping gene were performed in technical triplicate using primers 15 and 16, and 17 and 18 (Table A.1), respectively. Relative GFP expression was determined using a comparative calculation and the observed RPS11 values.

9.3 Materials and Methods for Chapter 3

9.3.1 Plasmid Construction

The pIRES-hrGFP plasmid (Figure 1a) was constructed through modification of pIRES-hrGFP-1a (Stratagene). The Zeocin resistance gene was amplified from pSV40-zeo2 (Invitrogen) using primers 57 and 58 (Table A.2) and cloned into pIRES-hrGFP-1a using BamHI and XhoI. Mutant *lox* sites were included for retargeting purposes and constructed as previously described[173]. An XbaI site was introduced using the QuikChange II Site-Directed Mutagenesis Kit (Stratagene) with primers 59 and 60. The pHL-GFP plasmid (Figure 3.1b) was provided by Shire Pharmaceuticals.

9.3.2 Sterile FACS Sorting

After stable cell selection was completed (approximately 15-25 days), cells were prepared for flow cytometry sorting and analysis. For each sort, 300,000 cells of the top 10-15% of the population (based on GFP expression) were isolated using a FACS Aria. This population was transferred to a six well plate and split every 24-48 hours, expanding the population until another sort was feasible. This process was iterated twice to ensure stringent selection and sustained expression.

9.3.3 Methods for identifying integration loci

Low throughput methodologies for identifying integration loci rely on approaches that both isolate and amplify genomic DNA adjacent to the transgene. We utilized three primary approaches to identify the integration sites in our high expression clones: TAIL PCR, inverse PCR and plasmid recovery. TAIL PCR utilizes three interlaced PCR reactions to amplify genomic fragments adjacent to the integrated transgene. Long primers, specific to the integrated sequence flanking the gDNA, along with an arbitrary, degenerate primer(s) of 12-16 base pairs in length are used in each PCR reaction. Based on previous reports, we adapted a methodology that uses three interlaced PCR reactions to enrich the flanking genomic DNA fragment[253-256]. The transgene was linearized prior to transfection and the long, specific primers were designed to cover approximately the last 200 bp of the linearized cassette. This methodology was used to successfully identify the integration loci for clones A, B, C, H, I and J, and further details for each clone can be found in the Appendix B.

A second approach, inverse PCR, was adapted from previous reports [39,50,63,64,257]. Genomic DNA was first fragmented, either with a restriction enzyme located close to the end of the transgene or in a non-biased fashion by shearing. DNA was sheared using a HydroShear instrument between SC7 and SC12 for 20 cycles to obtain fragments

between 3 and 6 kb in size. A low density ligation reaction is then used to circularize the genomic fragments, after which the ligation mixture is then subjected to a PCR reaction using inverted primers specific to the transgene. In the even that a circular fragment containing both the transgene and the adjacent DNA was formed, the entire fragment can be amplified and then sequenced. This approach was used to successfully identify the integration locus for clone E and further details are discussed in the Appendix B.

The third approach, plasmid recovery, provided better capture and recovery of genomic fragments, thereby increasing our coverage of the human genome. In this method, genomic DNA was fragmented, either with a specific restriction enzyme or by shearing to achieve fragments between 3 and 6 kb in size. These fragments were then ligated into a bacterial expression vector to create a genomic fragment library. This library was then transformed into DH10β using electroporation and plated on LB plates supplemented with antibiotics. After overnight growth, the plates were scraped and the colonies were pooled and mixed. Plasmid DNA was extracted and subjected to an inverse PCR reaction (as described above) using primers specific to the GFP gene. The amplified DNA was then sequenced. This approach was used to successfully identify the integration loci for clones D, F and G, and further details are discussed in the Appendix B.

9.3.4 Real-Time PCR

Whole cell RNA was extracted using the RiboPure kit (Ambion) and 5e6 cells per sample. RNA was converted to cDNA using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). Relative mRNA expression for genes of interest was measured and compared to a common housekeeping gene, RPS11. The primer pairs for each gene can be found in Table A.2, (1-56) and were designed using PrimerExpress

software. Roche SYBR Green 2x master mix was used to prepare samples in triplicate. The Viia7 Applied Biosystems instrument and software was used to run RT-PCR and analyze results. The comparative Ct method was used to normalize measurements relative to RPS11.

9.3.5 Site-Specific Retargeting

The hCas9 plasmid (41815), previously constructed by the Church group[89], was obtained from Addgene. Two gRNA constructs were designed, as recommended previously[89], each containing a unique, 23 base-pair sequence homologous to a portion of the 5th intron of the Grik1 gene on chromosome 21 and ordered as a single gBlock from IDT. These gRNA sequences, Grik1A and Grik1B, are TGTACAAAAAAGCAGGCTTTAAAGGAACCAATTCAGTCGACTGGATCCGGTA CCAAGGTCGGGCAGGAAGAGGGCCTATTTCCCATGATTCCTTCATATTTGCAT ATACGATACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACA CAAAGATATTAGTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTA GTTTGCAGTTTTAAAATTATGTTTTAAAATGGACTATCATATGCTTACCGTAA CTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAAC ACCGCTATTTTAGATATATAGCAGTTTTAGAGCTAGAAATAGCAAGTTAAAA TAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTT TCTAGACCCAGCTTTCTTGTACAAAGTTGGCATTA and TGTACAAAAAAGCAGGCTTTAAAGGAACCAATTCAGTCGACTGGATCCGGTA CCAAGGTCGGGCAGGAAGAGGGCCTATTTCCCATGATTCCTTCATATTTGCAT ATACGATACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACA CAAAGATATTAGTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTA GTTTGCAGTTTTAAAATTATGTTTTAAAATGGACTATCATATGCTTACCGTAA

CTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAAC ACCGTGGGGGTTATACCACTCGTGTTTTAGAGCTAGAAATAGCAAGTTAAAA TAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTT TCTAGACCCAGCTTTCTTGTACAAAGTTGGCATTA respectively. The hrGFP-Zeocin construct (Figure 2.1) was linearized and used for genomic integration.

HT1080 and HEK293 cells were transfected as previously described. Control samples included 20µg of hCas9 and 20µg of the GFP construct only. The targeting samples included 20µg of hCas9, 20µg of the GFP construct and 20µg of either Grik1A or Grik1B. Each transfection was re-suspended in 30mL of fresh media and grown for 72 hours prior to Zeocin selection at MIC_{75} levels. Following the recovery of the cultures, GFP expression profiles were determined using flow cytometry. Whole cell RNA was extracted and mRNA expression levels of RPS11 and Zeocin were determined are previously described.

9.4 Materials and Methods for Chapter 4

9.4.1 Plasmid Construction

The pIRES-hrGFP plasmid (Figure 4.1) was constructed through modification of pIRES-hrGFP-1a (Stratagene). The Zeocin resistance gene was amplified from pSV40-zeo2 (Invitrogen) using primers 1 and 2 (Table A.3) and cloned into pIRES-hrGFP-1a using BamHI and XhoI. Mutant lox sites (FAS, 2272 and 5171) were constructed by annealing two primers (Table A.3**:** 3, 4 for loxFAS; 5, 6 for lox2272; 7, 8 for lox5171). Primers were re-suspended at a concentration of 500mM in 1x T4 ligase buffer, denatured at 95ºC for 2.5 minutes and slowly cooled to 25ºC. DNA was purified using the MERmaid spin kit (MP Biologicals) and *lox* sites were cloned in front of the CMV

promoter using NsiI. An XbaI site was introduced using the QuikChange II Site-Directed Mutagenesis Kit (Stratagene) with primers 27 and 28.

The backbone of the Strawberry construct (Figure 4.1), including the SV40 promoter, poly-A tail, bacterial origin of replication and ampicillin marker, are derived from pSV40-Zeo2 (Invitrogen). The hygromycin B resistance gene was cloned from pAG26 (primers 9 and 10, using AscI and SpeI), the IRES from pIRES-hrGFP-1a (Stratagene) and the Strawberry gene from pmStrawberry (Clonetech). Primers 11 and 12, and 13 and 14 were used to amplify the IRES and Strawberry gene respectively. The IRES site was added using a SpeI restriction site, and SpeI and FseI restriction sites were used to add the Strawberry gene. The loxP site was added using a HindIII restriction site (primers 15 and 16). A second mutant site was added using an EcoRV site. The loxFAS site was created by annealing primers 17 and 18, lox2272 primers 19 and 20, and lox5171 primers 21 and 22. These plasmids were linearized (Figure 4.1) using SspI.

The pCI-Cre plasmid was derived from pCI-neo (Clonetech). A SwaI restriction site was added after the CMV promoter by annealing primers 23 and 24 and digesting with SacI. The wild-type Cre gene was amplified from pET28a-Cre (gift of the Jayaram Lab, UT Austin) using primers 25 and 26 and introduced using SwaI and XhoI.

9.4.2 Cell Line Development

To establish stable, GFP expressing cells, 3 batches of cells were transfected with 50μg of pIRES-hrGFP DNA as previously described. Cells were allowed to recover for 48 hours before adding 50μg/mL of Zeocin. Single cell cloning was conducted as previously described.

9.4.3 Copy Number Assay

The copy number of GFP integrants was determined for each clone as previously described[173,258]. RPPH1, a previously purported two copy human housekeeping gene, was cloned into the pIRES-hrGFP vector. RT-PCR primers were designed for RPPH1 (Table A.3, 31 and 32) and hrGFP (Table A.3, 29 and 30) and primer efficiency was calculated using a standard curve method (99% for hrGFP and 102% for RPPH1). All real-time PCR measurements used the SYBR Green Master Mix (Roche) and were run on the 7900HT model using 96-well plates (Applied Biosystems). RT-PCR was conducted with both primer sets in triplicate, using 50 to 500 ng of genomic DNA. The hrGFP copy number was determined by the ratio of hrGFP to RPPH1. Of the eighteen cell lines we measured, twelve contain a single copy and no cell line has a copy number greater than three (Table 9.1).

Table 9.1: Copy Number for GFP-expressing clones

***Lox* Pairing**	**Clone**	**Copy Number**
*lox*2272-*loxP*	A	1
	B	1
	C	1
	D	1
	E	2
	F	1
	G	1
	H	1
*lox*5171-*loxP*	I	3
	J	1
	K	2
	L	1
	M	2
*lox*FAS-*loxP*	N	2
	O	1
	P	1
	Q	1
	R	2

Eighteen clones were used to evaluate the impact of lox pairing on recombination. The GFP copy number of these clones was determined using a peer-reviewed, RT-PCR based method, as described in the Materials and Methods. Of the eighteen clones, thirteen carry a single copy of the transgene. The two loxFAS-loxP cell lines used for further studies (Q and R) are both single copy integrants.

9.4.4 Measuring Cre recombinase performance

Cre recombinase function was measured using a dual-fluorescence screening method. HT1080 cell lines stably expressing GFP were transfected with either the Strawberry construct (linear), pCI-Cre (circular), or both. In cases where both the Strawberry construct and pCI-Cre were introduced, DNA was either co-transfected (i.e. added directly to cell mixture prior to a single electroporation) or sequentially transfected. In the latter case, Strawberry DNA was transfected first and some period later, Cre DNA was introduced in a separate electroporation event. Approximately two days after electroporation, samples were analyzed using flow-cytometry as previously described.

Fluorescent quadrants were established using a wild-type HT1080 sample (Figure 4.2), with positive Strawberry fluorescence above 5 RFU and positive GFP fluorescence above 8 RFU. Excision activity due to Cre was measured by subtracting the non-fluorescent population of GFP positive HT1080 cells transfected with the Strawberry construct (control) from the same cells transfected with Strawberry and Cre constructs (test). Swapping activity due to Cre was measured by subtracting the control population exhibiting Strawberry fluorescence from the test cells (Q3, Figure 4.2). Cells exhibiting both GFP and Strawberry fluorescence (Q2, Figure 4.2) were excluded from these calculations and are a result of transient expression.

9.4.5 Southern Blot

Three batches of a GFP positive cell line with a loxP-loxFAS pairing (clone R) were transfected with 30µg of Strawberry construct per batch. Twenty-three hours post transfection, surviving cells were divided evenly into three batches and transfected with 5µg of pCI-Cre. Forty-eight hours later, using FACSAria, the cell population was sorted for red fluorescent cells (expressing Strawberry only) and dual fluorescent cells (expressing Strawberry and GFP). Sorted cells were expanded as previously described[173]. As a control, a batch of cells was transfected with 30 µg of Strawberry construct only and grown for 72 hours.

Genomic DNA was extracted with the Wizard kit (Promega) from 4 distinct cell populations; the GFP positive cell line (clone R) (Q1, Figure 4.2), the GFP positive, control cell line (72 hours post-transfection), the Strawberry-expressing sorted cell population (Q3, Figure 4.2), and the Strawberry and GFP-expressing sorted cell population (Q2, Figure 4.2). Thirty µg of DNA was digested overnight using 60U of SphI, concentrated to 50µL by ethanol precipitation, and loaded on a 0.7% agarose gel.

In the first lane, the 2-log ladder (NEB) was loaded. In the third lane, GFP positive digested gDNA was loaded, followed by the Strawberry-expressing sorted cell population, the Strawberry and GFP-expressing sorted cell population, and the GFP positive control cell line. Lanes 7 and 8 were empty, and the four gDNA samples were again loaded in the same pattern. Lane 13 was empty followed by the lambda BsteII ladder (New England Biolabs). Both ladders were labeled by T4 polynucleotide kinase (NEB) using gamma-^{32}P ATP (Perkin Elmer). The gel was run at 35V for 5 hours, then soaked for fifteen minutes with agitation in an alkaline transfer buffer (0.4N NaOH and 1M NaCl). A transfer stack was assembled with two layers of blotting paper, the agarose gel, a positively-charged nylon membrane (Ambion Brightstar), two additional layers of blotting paper and a stack of paper towels with weight on top. Alkaline buffer was used as a transfer medium overnight.

The nylon membrane was cut between empty lanes 7 and 8. The halves were soaked for 15 minutes in a neutralization buffer (0.5M Tris-Cl at pH 7.2 with 1M NaCl). Full-length PCR constructs complimentary to both the hrGFP and mStrawberry genes were created using primers 33, 34 and 13, 14 respectively. Twenty-five nanograms of the PCR product were radiolabeled using the High Primer random labeling kit (Roche) and alpha-^{32}P dTTP (Perkin-Elmer). Probes were cleaned using P-30 spin size exclusion columns (Bio-Rad). The membrane was pre-hybridized with Amersham Rapid-hyb buffer (GE) at 65ºC for 45 minutes. Each probe was added to the respective membrane, and hybridization was conducted for 2 hours at 65ºC. Four washes were conducted for 20 minutes each using 50mL of 2xSSC (.3M NaCl, 0.03M Na_3 citrate) with 0.1% (w/v) SDS for the first and second washes at room temperature, followed by 1xSSC and 0.8xSSC at 65ºC. The membranes were exposed to a phosphor screen overnight and imaged using a Typhoon Trio (GE Healthcare). Contrast was adjusted such that 0.1% of pixels were

over-exposed using ImageJ (NIH). Each half of the membrane image was aligned based on large fragment DNA that had not migrated from the top of the wells.

9.5 Materials and Methods for Chapter 5

9.5.1 Microarray Data Analysis

Codon usage profiles were assembled using publicly available microarray data, downloaded from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). Data pre-processing and normalization was performed using the Robust Multichip Average algorithm[259-261], and Bioconductor's Affy package in R version 2.15.1. Differentially expressed genes were identified using the Linear Models for Microarray Data (LIMMA) package. Probe sets were matched with *S. cerevisiae* genes using information included in Affymetrix's Expression Console Software. Genes with an adjusted p-value less than 0.05 and a log-fold change greater than one or less than negative one were considered differentially expressed. A subset of differentially expressed genes (typically 50) was used to generate a condition-specific codon usage table and matrix, as previously described in Chapter 5.

9.5.2: Plasmid construction

Yeast expression vectors were propagated in *E. coli*. All experiments were carried out in *S. cerevisiae*, with parent strains including BY4741, BY4743 and YSX3. The sequences of all genes used in this study are available in Appendix C**.** The wild-type and Blue Heron optimized CatA variants were taken from a previous study[131]. All other CatA variants were assembled using IDT's gBlocks. The eight CatA sequences were assembled in the p413-TEF vector[262]. The wild-type eGFP gene was amplified from the pZE-eGFP plasmid using primers

TAAAACACCAGAACTTAGTTTCGACGGATTCTAGAATGCGTAAAGGAGAAGA ACTTTTCA and AGGTCGACGGTATCGATAAGCTTGATATCGAATTCTTAAACTGCTGCAGCGT AGTTTTCG. The other eight eGFP variants were assembled using IDT's gBlocks. The eGFP genes were cloned into the p41K-GPD and p426-CYC plasmids using yeast homologous recombination and overlapping sequences. The expression plasmids were constructed using yeast homologous recombination and a high efficiency, lithium-acetate transformation. The formation of correct plasmids was confirmed using DNA sequencing. For each variant, three biological replicates were isolated and stored.

9.5.3: CatA Activity Assay

Yeast minimal media was inoculated at an OD600 of 0.1 using stationary phase cultures of the CatA variants. Flasks contained 200mL, 100mL and 50mL of media for the 6, 18 and 24 hour growth experiments respectively. After the designated time period, cells were pelleted and protein was extracted as previously described[131]. Total protein was determined using a Bradford assay. V_{maxes} were measured on a per microgram of protein basis using a kinetic assay measuring the conversion of added catechol to muconic acid, which can be detected at 288nm. All biological replicates were included and measurements were done in technical triplicate. Catechol was mixed with protein extract at four concentrations, 0.1, 0.2, 0.3 and 0.4 mM, and Lineweaver-Burke plots were used to calculate V_{max} in units of mM/min*μg protein. A higher V_{max} corresponds to more CatA enzyme in the protein extract.

9.5.4: Muconic acid production

High pressure liquid chromatography (HPLC) was used to measure the intercellular conversion of catechol to muconic acid in S. cerevisiae cultures as

previously described[131]. Triplicate yeast cultures expressing each CatA variant were grown in 30 mL of media for 18 hours with a starting OD600nm of 0.1. After 18 hours, cultures were spiked with 1 mg/mL of catechol and grown for an additional 24 hours. At this point, 1 mL of supernatant was filtered and analyzed using a Zorbax SB-Aq column (Agilent Technologies). The injection volume was 2.0 µL and the mobile phase was 84% 25 mM potassium phosphate buffer (pH=2.0) and 16% acetonitrile with a flow rate of 1.0 mL/min. The column was maintained at 30ºC and the UV-Vis absorption was measured at 280nm. Muconic acid production levels were calculated using a standard curve. *Cis,cis*-muconic acid standards were purchased from Sigma-Aldrich and *cis,trans*-muconic acid was provided by Draths Corporation.

9.5.5: Matrix Drift Analysis

The Frobenius matrix norm is defined as the square root of the product of the trace of the conjugate transpose of the matrix and the matrix itself. This is defined in Figure 9.1. The drift between any two codon usage matrices was determined by taking the difference between the matrices (excluding stop codon usage) and the Frobenius matrix norm of that resultant matrix of differences, or $\|\mathbf{A-B}\|_\mathbf{F}$ where **A** and **B** represent two distinct matrices of identical size.

Figure 9.1: Mathematical Definition of Frobenius Matrix Norm

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{\operatorname{trace}(A^*A)} = \sqrt{\sum_{i=1}^{\min\{m,\,n\}}\sigma_i^2}$$

The genetic interaction targets for sixteen S. cerevisiae transcription factors were identified using yeastgenome.org. Using those corresponding gene target sequences,

codon usage matrices were constructed for each transcription factor. Frobenius matrix norms were calculated for all matrix pairs, including the control matrix, using MATLAB. The Frobenius norms create the edges in the map between the nodes, as shown in Figure 5.6a. The map was constructed using the Map_Draw script (Appendix D) which uses the networkz and pygraphviz python packages and Graphviz 2.28.

9.6 Materials and Methods for Chapter 6

9.6.1: Strain and Plasmid Construction

Yeast expression vectors were propagated in *E. coli* DH10β. All experiments were carried out in *S. cerevisiae*. The genotype of the parent strains including BY4741, BY4743, and S288C and their derivatives have been previously described[229]. The BY4741 knockout strains were provided by the Marcotte laboratory (University of Texas at Austin, ICMB). S288C and BY4743 homozygous *Δpho80/Δpho80* strains were purchased from OpenBiosystems. The S288C Δ*gcn5* strain was made by replacing the wild-type *GCN5* gene with a hygromycin-B resistance gene amplified from plasmid pAG32 using primers 1 and 2 (Table A.4) and extended using primers 3 and 4, for a final fragment with 80 base pairs of genomic homology both upstream and downstream. Using a high efficiency yeast transformation protocol[263], 1 μg of fragment was transformed into competent S288C cells and were plated on YPD supplemented with 100 μg/mL hygromcyin-B. The genotype of the S288C Δ*gcn5* strain was confirmed by extracting genomic DNA and performing both a positive PCR control (primers 5-8) and a negative PCR control (primers 9 and 10).

The wild-type *GCN5* gene was amplified from BY4741 gDNA using primers 9 and 10 and cloned into the pUC19 vector using restrictions enzymes XbaI and SalI. After

confirming the accuracy of the *GCN5* sequence, mutations M193A, F221A and E173A were introduced using the Stratagene Quikchange mutagenesis kit and primers 11 to 14, and 39 and 40. The mutant GCN5 genes, as well as the wild-type gene, were cloned into the library of p416-TEF_{mutant} vectors[222] using the XbaI and SalI restriction enzymes. The *gcn5-M193A*, *F221A*, and *E173A* plasmid collections were transformed into BY4741 *Δgcn5* using a Gietz lithium acetate protocol[263] and selecting on drop out media deficient in uracil.

To allow for expression in amino acid free media, the p416-TEF_{mutant}-*gcn5-F221A* plasmid collection was modified to include a G418 resistance gene. Using primers 15 and 16, the gene was amplified from the pUG6 plasmid. The p416-TEF_{mutant}-*gcn5-F221A* plasmid collection and the resistance gene were digested with StuI and EcoRV. The new p416-TEF_{mutant}-*gcn5-F221A*-G418 plasmid collection was transformed into S288C using a Gietz lithium acetate protocol and selected on YPD plates supplemented with 200 μg/mL G418. This process was repeated to create p416-TEF_{mutant}-*gcn5-E173A*-G418 and p416-TEF_{mutant}-*gcn5-M193A*-G418 plasmid collections.

The p415-pPho5-yECitrine plasmid was constructed for fluorescence assays. The *PHO5* promoter, shown to be contained in the thousand base pairs upstream of *PHO*[225], was amplified using primers 17 and 18. Using restriction enzymes SacI and XbaI, pPho5 was cloned into the p416-TEF-yECitrine vector, replacing the TEF promoter in front of the yECitrine fluorescence gene. Using SacI and KpnI, the pPho5-yECitrine fragment was moved to the p415 plasmid, which contains a leucine auxotrophic marker. Along with the p416-TEF_{mutant}-*GCN5* plasmid collections, the p415-pPho5-yECitrine plasmid was transformed into BY4741 Δ*pho80* and BY4743 Δ*pho80* and selected on drop out media deficient in both uracil and leucine. The p415-pGcn5-yECitrine plasmid was constructed for fluorescence assays. The *GCN5* short promoter (350 bp) was amplified

using primers 41 and 42, and the long (640 bp) promoter with primers 42 and 43 directly upstream from the *GCN5* gene. Using SacI and XbaI, the p415-pPho5-yECitrine plasmid was replaced with p415-pGcn5-yECitrine, and then along with the p416-TEF_{mutant}-*GCN5* plasmid collection, transformed into BY4741 Δ*pho80* and selected on drop out media deficient in both uracil and leucine.

Twenty-two BY4741 single gene knockouts, corresponding to genes that form a synthetic lethal phenotype with *gcn5Δ*, were transformed with a p416-TEF control plasmid, and p416-TEF_{mutant}-*gcn5-F221A* plasmids with promoter strengths of 0.16, 0.32, 0.68 and 0.95. Colonies were selected in triplicate from drop out media deficient in uracil. An additional ten BY4741 single gene knockouts, selected at random, were transformed under identical conditions to serve as an experimental control.

All strains were selected and tested in biological triplicate at minimum. Some strains and assays were tested with up to 6 biological replicates.

9.6.2: Growth Experiments

Complementation studies were conducted using *gcn5*Δ strains expressing each of the *gcn5* mutant plasmid libraries. From stationary phase culture, a honey-comb plate was inoculated in triplicate with a starting OD of 0.1. Minimal media lacking uracil was supplemented with 3-aminotriazole. Using a Bioscreen C Growth Curve Analysis System, optical density measurements were taken every ten minutes for 24 hours. Temperature was maintained at 30ºC and continuous, high shaking was used. Growth rate was calculated as the slope of the natural log of optical density versus time during the exponential growth phase. The histidine starvation assay was conducted using the p416-$TEFmutant_x$-*gcn5-F221A*-G418 strains. From stationary phase culture, a honey-comb plate was inoculated with 4-6 biological replicates with a starting OD of 0.1. A minimal

media composed of glucose, yeast nitrogen base without amino acids, 3.75 mM 3-aminotriazole and 200 ug/mL G418 was used. S288C wild-type and S288C Δ*gcn5*, both transformed with an empty p416-TEF-G418 plasmid, served as controls. Optical density measurements were collected over a period of 30 hours.

Ten additional growth inhibition assays were conducted, in which a total of 50 strains were assayed, including the above controls, and 4-5 biological replicates of the p416-TEFmutant$_x$-*gcn5-F221A*-G418 strains. The 50 strains were grown to stationary phase in 3mL of YPD media supplemented with G418 and then a honey-comb plate was inoculated with a starting OD_{600} of 0.1 in 250µl of fresh media either with or without (control cultures) a putative Gcn5p-dependent growth inhibition additive (Table 6.2). Optical density measurements were collected for the 100 cultures over a period of 60 hours using the Bioscreen C.

9.6.3: Yeast Immunofluorescence

Global histone acetylation at H3K18 was measured using yeast immunofluorescence. S288C strains containing p416-TEFmutant$_x$-*gcn5-M193A*-G418, p416-TEFmutant$_x$-*gcn5-E173A*-G418, and p416-TEFmutant$_x$-*gcn5-F221A*-G418 (promoter strength 0.32, 0.68 and 0.95), as well as wild-type and knockout controls, were grown to mid-exponential phase and fixed by adding a 10^{th} volume of 37% formaldehyde for 2 hours. Cells were washed twice with PBS and resuspended in 500 µl of a spheroplasting buffer (1.2M sorbitol and 0.1M KH_2PO_4 at pH of 7.5). Cells were stored for 1-2 days at 4ºC. Spheroplasts were made by incubating 200 µl of fixed cells with 1.2 µl of zymolase (Zymo Research) and 3.2 µl of β-mercaptoethanol for 30 minutes at 30ºC. Spheroplasts were washed once with 1 mL of PBS+0.05% Tween 20 and resuspended in 100 µl of PBS+0.05% Tween 20. Slides were treated with 50 µl of 1mg/mL poly-L-

lysine (>400,000 MW) for 15 minutes, followed by 3 water washes. After the slides were completely dry, 20 µL of spheroplasts were added to each well for 5 minutes, followed by 3 PBS washes. The slide was immersed in ice cold methanol for 5 minutes and ice cold acetone for 30 seconds. After drying, the slide was rehydrated by adding 50 µl of PBS for 5 minutes, followed by a PBS wash. A blocking solution composed of PBS and 1mg/mL BSA was added (20 µL) to each slide followed by 30 minutes in a humid chamber. The slide was then washed 3 times with PBS. 20 µl of H3K18ac primary Rabbit antibody (Abcam) diluted 500-fold in blocking solution was added to each slide and incubated for 90 minutes. The slide was washed 3 times with PBS. 20 µl of anti-Rabbit Goat IgG DyLight 649 secondary antibody (Abcam) diluted 200-fold in blocking solution was added to each slide and incubated for 90 minutes in the dark. The slide was washed 3 times with PBS. 20 µl of 1 µg/mL DAPI in PBS was added to each well for 5 minutes, followed by 3 washes with PBS. A drop of fluorescent mounting medium (KPL) was added to each slide along with cover glass (#1.5 thickness) before sealing with nail polish. Slides were imaged using the Zeiss Axiovert instrument and a 100x magnifying lens. The DAPI and Cy5 filters were used respectively to image DAPI and DyLight 649 staining. Average intensity per cell was determined using Metamorph software.

9.6.4: Gene Expression Microarrays

Global mRNA analysis was conducted using whole cell RNA taken from S288C cell lines grown in minimal media. In addition to the control plasmid (no *gcn5-F221A*) and a gcn5 null strain, mutant promoter strength 0.32, 0.68 and 0.95 were tested. Cell lines were grown in biological triplicate with a starting optical density of 0.0045 and harvested at a density between 0.4 and 0.5. Whole cell RNA was extracted using the Ambion Ribo-Pure kit for yeast. cRNA synthesis and fragmentation was conducted by

the Genome Sequencing and Analysis Facility at the University of Texas using the Ambion MessageAmp Premier kit. Hybridization and scanning was performed by Asuragen in Austin, TX using Affymetrix Yeast 2.0 arrays. Data pre-processing and normalization was performed using the Robust Multichip Average algorithm[259-261] and Bioconductor's Affy package. Differentially expressed genes were identified using the Linear Models for Microarray Data (LIMMA) package, which resulted in 529 probe sets. Probe sets were matched with *S. cerevisiae* genes using information included in Affymetrix's Expression Console Software, resulting in 504 unique genes. The $\log_2$ expression data for differentially expressed probe sets are available online[229] and were deposited to Gene Expression Omnibus under accession number GSE26923.

9.6.5: Real Time PCR

Relative transcription levels were quantified using real time PCR from whole cell RNA extracts. Cell lines were grown in minimal media with a starting optical density between 0.004 and 0.005 until they reached a density between 0.4 and 0.5, at which point whole cell RNA was extracted using Ambion's Ribo-Pure kit for yeast. RNA quantification was performed with a Nanodrop 2000.

RT-PCR was conducted using whole cell RNA extracted from S288C to determine the level of mutant Gcn5p expression relative to native Gcn5p. Two control strains, wild-type S288C and S288C *gcn5Δ* carried empty vectors. Additionally, S288C with *gcn5-F221A* expressed from varying promoter strengths (.07, .32, .68, .95, and 1.17) were used. Primers were designed such that both wild-type and mutant *GCN5* would be detected. Average Ct values were normalized with respect to the wild-type sample. Since the sequences are similar between the wild-type and mutant, it was necessary to deduce the expression level.

The S288C cell lines used for the analysis of *HIS3* mRNA levels were grown in media supplemented with 3.75mM 3-aminotriazole. In addition to the control plasmid (not containing the *gcn5-F221A*), promoter strengths of .07, .32, .68 and .95 were tested. cDNA synthesis and quantitative PCR were performed simultaneously using the iScriptTM One-Step RT-PCR Kit with SYBR Green (Bio-Rad). We followed the manufacturer's instructions, with the following modifications: 100ng of whole cell RNA per 25 μL reaction, an extended, 15 minute reverse transcription time, and a 56ºC annealing temperature. For the analysis of yECitrine mRNA levels, BY4741 p415-pPho5-yECitrine, p416-TEF_{mutant}-*gcn5-F221A* cell lines were grown in minimal media. In addition to the control strain (no *gcn5-F221A*), promoter strengths of .07, .16, .32 and .68 percent were tested. We determined relative RNA concentration by comparing the cycle thresholds to *ALG9*, which has shown to be an ideal housekeeping gene for yeast[264]. Primers 19 and 20 were used to amplify *HIS3*, whereas 21 and 22 were used for *ALG9*. Primers 22 and 23 were used to amplify yECitrine.

Real-time PCR confirmation of microarray findings was conducted on a small scale using whole cell RNA taken from S288C cell lines grown in minimal media (Figure 6.9). In addition to the control plasmid (no *gcn5-F221A*) and a *gcn5* null strain, a range of promoter strengths were tested. cDNA synthesis was performed using Invitrogen's High Capacity cDNA reverse synthesis kit. For quantitative PCR, we used Roche's SYBR Green Master Mix, following the manufacturer's instructions with an annealing temperature of 58ºC. We concentrated on four gene targets; *TKL2*, *SPL2*, *IDH2* and *ZRT1*. Primers 25 and 26 were used for *TKL2*, 27 and 28 for *SPL2*, 29 and 30 for *ZRT1*, and 31 and 32 for *IDH2*. Additionally, *GCN5* mRNA levels were measured with primers 37 and 38, and mRNA extracted from AML31, 32, 34, 35 and 39. All primer sequences can be found in Appendix A, Table A.4.

9.6.6: Growth Analysis for Synthetic Lethal Genes

The impact of the *gcn5* dominant mutant on synthetic lethal genes was assessed using a growth based assay and 22 BY4741 gene knockout strains. Synthetic lethals were selected from yGcn5 interaction data, available on yeastgenome.org, and corresponding single knockout strains were transformed with p416-TEFmutant$_x$-*gcn5-F221A* plasmids (promoter strengths of 0.32, 0.68, 0.95 and control). The resulting strains were grown in minimal media for 2 days prior to inoculating a honey-comb plate with a starting OD of 0.1. Minimal media lacking uracil was used. Optical density was measured using a Bioscreen C, as previously described. An average growth rate and standard deviation were calculated from the biological replicates. Values (Table 6.1) are reported for each synthetic lethal gene and control strains.

9.6.7: TEF Promoter Engineering

Additional variants of a weak TEF promoter (TEFpmut7)[222] were generated via error-prone PCR using the Genemorph II Random Mutagenesis Kit from Stratagene and primers 33 and 34, Table A.4. Six reactions containing differing template concentrations were combined to create 2 libraries of differing error rates (Table 7.1). Libraries were cleaned using the QIAquick PCR Purification Kit and cut with SacI and XbaI restriction enzymes (New England Biolabs). Fragments were ligated into a yeast expression vector upstream of the yECitrine fluorescent gene. 150 ng of each ligation was transformed into competent *E. coli* and plated onto LB agar plates containing 100 µg/mL ampicillin. For each library, approximately 17,500 colonies were scraped and collected in liquid culture and diluted to an optical density of 6 using LB media, and plasmid DNA was extracted using a Qiagen miniprep kit. From each plasmid library, 50ng DNA was transformed[265] into *S. cerevisiae* BY4741 and plated on drop out media deficient in uracil.

Table 9.2: Reaction conditions for error-prone PCR

Reaction	Library	Plasmid Template (ng)	Mutation Rate (1/kb)
1	Low	30.18	7.5 +/- 3.5
2		9.62	
3		3.07	
4	High	0.98	9.2 +/- 4.3
5		0.31	
6		0.10	

Three new low strength mutant TEF promoters were developed for this study (Figure 9.2) using error-prone PCR and a fluorescence based screen. The error-prone PCR conditions (shown above) resulted in a mutation rate between 4 and 13.5 per kilobase.

110 yeast colonies were isolated from the libraries and grown in minimal media deficient in uracil to an optical density of 0.5 and analyzed by FACS Calibur, compared to a control strain. Mutants displaying fluorescence between 30% and 2% of the control population and low cell-to-cell variability were isolated, and plasmid DNA was extracted using the Zymoprep Yeast Plasmid Miniprep I. These plasmids were then sequenced (primers 35 and 36, Table A.4), and retransformed into yeast to confirm promoter strength. The sequences are shown in Figure 9.2. The selected promoters (Tef32, 51 and 77) have strengths of 0.10 ± 0.01, 0.15 ± 0.01, and 0.22 ± 0.02 relative to a native TEF promoter.

Figure 9.2: Sequence of additional low strength TEF promoters

TEF

ATAGCTTCAAAATGTTTCTACTCCTTTTTTACTCTTCCAGATTTTCTCGGACTCCGCGCATCGCCGTACCACTTCAA
AACACCCAAGCACAGCATACTAAATTTCCCCCTCTTTCTTCCTCTAGGGTGTCGTTAACTACCCGTACTAAAGGTTTG
GAAAAGAAAAAAGAGACCGCCTCGTTTCTTTTTCTTCGTCGAAAAAGGCAATAAAAATTTTTATCACGTTTCTTTTT
CTTGAAAATTTTTTTTTTGATTTTTTTCTCTTTCGATGACCTCCCATTGATATTTAAGTTAATAAACGGTCTTCAAT
TTCTCAAGTTTCAGTTTCATTTTTCTTGTTCTATTACAACTTTTTTTACTTCTTGCTCATTAGAAAGAAAGCATAGC
AATCTAATCTAAGTTT

TEFpmut32 (.10 ± 0.01)

ATAGCTTCAAAATGTCTCTACTCCTTTTTTACTCTTCCAGATTTTCTCGGACTCCGCGCACCGCCGTACCACTTCTA
AACACCCAATCACAGCATACTAAATTCCCCCTCCTCCTTCCTCTAGGGTGCCGTTAATTACCCGTACTAAAGGTTTG
GAAAAGCAAAAAGAGACCGCCTCGTCCCTTTTTCTTCGTCGGAGAAGGCAATAAGAATTTTTATCACGTTTCTTTCT
CTTGAAAACTTTTTTTTTCGATTTTGTTCTCTTTCGACGACCTCCCATTGATATGTGAGTTAGCAACCGGTCTTCAAT
TTCTCAAGTTTCAGCTTCATTTTTCCTGTTCTATTACAACTTTTTTTACTTCTTGCTCATTGGAAAGAAAGCATAGC
AATCTAATCTAAGTTT

TEFpmut51 (.15 ± 0.01)

ATAGCTTCAAAATGTCTCTACTCCTTTTTTACTCTTCCAGATTTTCTCGGACTCCGCGCACCGCCGTACCACTTCAA
AACACCCAAGCACAGCATACTAAATTCCCCCTCCTCCTTCCTCTAGGGTGCCGTTAAATACCCGTACTAAAGATTTG
GAAAAGGAAAAAGAGACCGCCTCGTCCCTTTTTCTTCGTCGGAGAAGGCAATAAAAATTTTTATCACGTTTCTTTCT
CTTGAAAACTTATTTTTCGATTTTGTTCTCCTTCGACGACCTCCCATTGATATTTGAGTTAACAAACGGTCTTCAAT
TTCTCAAGTTTCAGCTTCATTTTTCCTGTTCTATTACAACTTTTTTTACTTCTTGCTCATTGGAAAGAAAGCATAGC
AATCTAATCTAAGTTT

TEFpmut77 (.22 ± 0.02)

ATAGCTTCAAAATGTCTCTACTCCTTTTTTACTCTTCCAGATTTTCTCGGACTCCGCGCACCGCCGTACCACTTCGA
AACACCCAAGCACAGCATACTAAATTCCCCCTCCTCCTTCCTCTAGGGTGCCGTTAATTACCCGTACTAAAGGTTTG
GAAAAGGAAAAAGAGACCGCCTCGTCCCTATTTTCTTCGTCGGAGAAGGCAATAAAAATTTTTATCACGTCTCTTTCT
CTTGAAAACTTTTTTTTTCGATTTTGTTCTCTTCCGATGACCTCCCATTGATATTTGAGTTAACAAACGGTCTTCAAT
TTCTCAAGTTTCAGCTTCATTCTTCCTGTTCTATTACAACTTTTTTTACTTCTTGCTCATTGGAAAGAAAGCATAGC
AATCTAATCTGAGTTT

Three low strength TEF promoters were constructed for this study using error-prone PCR and a fluorescence based screen. Base pair mutations compared to the native TEF promoter are shown in underlined text above.

Appendix A: Primers

Table A.1: Primers from Chapter 2

Primer	Sequence
1	GATAAGGATCCGCGTATGGTGCACTCTCAGTACAATCT
2	GTGACGGATCCGCCCGGATCGATCCTTATCGGATTTTAC
3	ACTAGAAGCTTATGGTGAGCAAGCAGATCCTGAAGA
4	CTAGTAAGCTTGAATCTAGATTGTTACACCCACTCGTGCAGGCTG
5	ACTAGTCTAGACCCCTCTCCCTCCCCCCC
6	GATTATCTAGAGGTAATCGATGATTAGCATTATCATCGTGTTTTTCAAAGGAAAACCACG
7	GGTTAATCGATATGACCGAGTACAAGCCCACGGTGCGCCTC
8	GGTTAATCGATTCAGGCACCGGGCTTGCGGGTCATGCACCA
9	CTAAGATCGATATGGGTAAAAAGCCTGAACTCACCGC
10	GGTTAATCGATTTATTCCTTTGCCCTCGGACGAGT
11	GGTTAATCGATATGGCCAAGTTGACCAGTGCC
12	GGTTAATCGATTCAGTCCTGCTCCTCGGCCA
13	GGTTAATCGATATGATTGAACAAGATGGATTGCACGCA
14	GGTTAATCGATTCAGAAGAACTCGTCAAGAAGGCGAT
15	TCAGCGACTTCTTCATCCAGAGCTTC
16	ACACGAACATCTCCTCGATCAGGTTG
17	GCCCCTGCGTAATCGATAAG
18	GTCTGAATGTCCGCCATCTTC

Table A.2: Primers from Chapter 3

Primer	Target	Sequence
1	ADAM6	CTCTCTGCAGACCTATCCAAAATATATG
2	ADAM6	ATATAAACGTTTGCGGGACATGT
3	AQPEP	GAAAAGATTCAACTTGCTTATGCAAT
4	AQPEP	GATGAAGCCACAACCTCAATTATATTT
5	ARL9	ACCCAGTACTTCCTCTGGTTGTGT
6	ARL9	GCCAAAGCTTCATGGATATCTGT
7	BACH1	ACATATGAGTCCATGTGCTTAGAGAAG
8	BACH1	ACTCTGTCAGTTCCAAATGCTTTTT
9	CLDN8	TCTGGAGTAGACGTGACTTCTTTCC
10	CLDN8	CAACGAAAAGAGCAGTAGCTACAGATAC
11	pCMV	TGCCCAGTACATGACCTTATGG
12	pCMV	TGGAAATCCCCGTGAGTCAA
13	COMMD10	TACGTGGTCTTCTATGGGTCAAGAA
14	COMMD10	CAGAGTGAGCCATCTGAAGGTTAAG
15	DCC	CCAGACTAACTGCATCATCATGAGTT
16	DCC	CAACGCCATAACCGATAATATAACCT
17	DTWD2	CTCAGAAAGTGTGTTTGTGTCCATTT
18	DTWD2	TCTGCTGGATGCTGAATTATGTACA
19	GFP	TCAGCGACTTCTTCATCCAGAGCTTC

Table A.2 (continued)

20	GFP	ACACGAACATCTCCTCGATCAGGTTG
21	GRIK1	GAAGCCCAATGGTACCAATCC

22	GRIK1	CAAGCAGGCTAAGAGCACATACAT
23	hrGFP	TCAGCGACTTCTTCATCCAGAGCTTC
24	hrGFP	ACACGAACATCTCCTCGATCAGGTTG
25	IGHA2	GTGACCTCTGTGGCTGCTACAG
26	IGHA2	GTGTTTCCGGATTTTGTGATGTT
27	IGHG2	TTCGGCACCCAGACCTACAC
28	IGHG2	CACAACATTTGCGCTCAACTG
29	LOC100287225	TTCATTTTCCAGTCTCCTTCAGATG
30	LOC100287225	CTTTTTGTGATGGTGAGTTCCTTCT
31	MBD2	CCAAAGTCACAAATCATCCTAGTAATAAA
32	MBD2	CTTGTAGCCTCTTCTCCCAGAAAAG
33	REST	TGAAGTTGCTTCTATCTGCTGTTTTG
34	REST	TGTGGCCTCTAATCAACATGAAGTA
35	RPPH1	AATGGGCGGAGGAGAGTAGTCTGAAT
36	RPPH1	AGCGAAGTGAGTTCAATGGCTGAGGT
37	RPS11	GCCCCTGCGTAATCGATAAG
38	RPS11	GTCTGAATGTCCGCCATCTTC
39	SEMA3A	CAACTATCAATGGGTGCCTTATCA
40	SEMA3A	TGTAGAGTCAAAACCACCAAATGTTT
41	SEMA3D	TCTGGCAGACTGAAATGTCCTTT
42	SEMA3D	TCAGAAGCTGTTCCAGAGTAGAGGTA
43	SWMA3E	TTTTCGAGGGCATGCTATATGTG
44	SEMA3E	AGGTCCTTCCTTATGTGCATATGGT
45	SEMA6A	GGTCAGATACCGCCTTACCAAA
46	SEMA6A	CTCTGATCCCAGAAAAACCACAGT
47	SLCO3A1	GTCTACCGATACCTGTATGTCAGCAT
48	SLCO3A1	GTTTTTGATGTAGCGTTTATAGTTTTTCCT
49	SPINK2	TGCTGCTCCTGGCAGTTACC
50	SPINK2	TGAGAGCAGTTTGGCGTTCTATATT
51	SV2B	CCCCCATAATTGTCCTGATACTTG
52	SV2B	ACAGAGCGAGGGATATAGCTCAA
53	TMEM121	AGATCGGCGTGTGCATCA
54	TMEM121	GTTCTGGAAGATGAAGTAGAGCTTGA
55	VPS33B	ACAAGCTATACAAGGTGGAGAACAAG
56	VPS33B	ACAAGACTGGCAATGTATCGCATAT
57	Zeocin	CGCGGATCCATGGCCAAGTTGACC
58	Zeocin	CCGCTCGAGTCAGTCCTGCTCCTC
59	QuikChange	CTCTTCCTCAGGTTACTCATATATACTCTAGATTGATTTAAAACTTC
60	QuikChange	GAAGTTTTAAATCAATCTAGAGTATATATGAGTAACCTGAGGAAGAG

Table A.3: Primers from Chapter 4

Primer	Sequence
1	CGCGGATCCATGGCCAAGTTGACC
2	CCGCTCGAGTCAGTCCTGCTCCTC
3	CCAATGCATACAACTTCGTATATACCTTTCTATACGAAGTTGTATGCATTGGTTCTGCAGTT
4	AACTGCAGAACCAATGCATACAACTTCGTATAGAAAGGTATATACGAAGTTGTATGCATTGG
5	TGACTATGCATATAACTTCGTATAGGATACCTTATACGAAGTTATATGCATGGCA
6	TGCCATGCATATAACTTCGTATAAGGTATCCTATACGAAGTTATATGCATAGTCA
7	TGACTATGCATATAACTTCGTATAGTACACATTATACGAAGTTATATGCATGGCA
8	TGCCATGCATATAACTTCGTATAATGTGTACTATACGAAGTTATATGCATAGTCA
9	TTGGCGCGCCTAAACAACCATGGGTAAAAAGCCTGAACTCAC
10	GGACTAGTAGTACTGATTATTCCTTTGCCCTCGGAC
11	TAATGGACTAGTTTACCCCCCTCTCCCTCCC
12	TAATGGACTAGTGGTTGTGGCCATTATCATCGTGTTTTTC
13	GGACTAGTATGGTGAGCAAGGGCGAGGA
14	ATCTAGGTGGCCGGCCCTTGTACAGCTCGTCCATGCCG
15	TGACTAAGCTTATAACTTCGTATAATGTATGCTATACGAAGTTATAAGCTTGGCA
16	TGCCAAGCTTATAACTTCGTATAGCATACATTATACGAAGTTATAAGCTTAGTCA
17	TGACTGATATCACAACTTCGTATATACCTTTCTATACGAAGTTGTGATATCGGCA
18	TGCCGATATCACAACTTCGTATAGAAAGGTATATACGAAGTTGTGATATCAGTCA
19	TGACTGATATCATAACTTCGTATAGGATACCTTATACGAAGTTATGATATCGGCA
20	TGCCGATATCATAACTTCGTATAAGGTATCCTATACGAAGTTATGATATCAGTCA
21	TGACTGATATCATAACTTCGTATAGTACACATTATACGAAGTTATGATATCGGCA
22	TGCCGATATCATAACTTCGTATAATGTGTACTATACGAAGTTATGATATCAGTCA
23	ATTCGGAGCTCGTTTAGTGAACCGTCAGATCACTATTTAAATCGGTGAGCTCATTCG
24	CGAATGAGCTCACCGATTTAAATAGTGATCTGACGGTTCACTAAACGAGCTCCGAAT
25	TTACAGATTTAAATATGGGCAGCAGCCATCATCATCA
26	TTACAGCTCGAGCTAATCGCCATCTTCCAGCAGGCG
27	CTCTTCCTCAGGTTACTCATATATACTCTAGATTGATTTAAAACTTC
28	GAAGTTTTAAATCAATCTAGAGTATATATGAGTAACCTGAGGAAGAG
29	TCAGCGACTTCTTCATCCAGAGCTTC
30	ACACGAACATCTCCTCGATCAGGTTG
31	AGCGAAGTGAGTTCAATGGCTGAGGT
32	AATGGGCGGAGGAGAGTAGTCTGAAT
33	TTACAGATTTAAATATGGTGAGCAAGCAGATCCTGAAGAAC
34	TTACAGCTCGAGTTACACCCACTCGTGCAGGCTGC

Table A.4: Primers from Chapter 6

Primer	Sequence
1	AGTCTTCAGTTAACTCAGGTTCGTATTCTACATTAGATGGGCGCGCCAGATCTGTTTAGC
2	CGAAAGGAATAGTAGCGGAAAAGCTTCTTCTACGCATTACGTTTTCGACACTGGATGGCG
3	GATTGGTAAGGGAAGACCGTGAGCCGCCCAAAAGTCTTCAGTTAACTCAGGTTCGTATTC
4	ACATCGTCTCGCCGTACTAAACATTTATTTCTTCTTCGAAAGGAATAGTAGCGGAAAAGC
5	CGGGGATTCCCAATACGAGGTCGCC
6	CCTCAATTGATCACATCGTCTCGCCGTAC
7	CAGAAACTTCTCGACAGACGTCGCGG
8	GTAGGGCGTAATGATGTTTGCTTGTCAAC
9	CTAGTCTAGAAAAATGGTCACAAAACATCAGATTGAAGAGGATC
10	CTAGCGGTCGACTTAATCAATAAGGTGAGAATATTCAGGTATTTCTTTTACTTTATTATT
11	GCAGATAATTACGCTATTGGATACGCTAAAAAGCAAGGCTTCACTAAAG
12	CTTTAGTGAAGCCTTGCTTTTTAGCGTATCCAATAGCGTAATTATCTGC
13	CGGTTATGGTGCGCATCTAGCGAATCACTTAAAAGACTATGTTAG
14	CTAACATAGTCTTTTAAGTGATTCGCTAGATGCGCACCATAACCG
15	GCTAAAAGGCCTTAGGTCTAGAGATCTGTTTAGCTTGCCTCG
16	ATTACTGATATCATTAAGGGTTCTCGAGAGCTCGTTTTCG
17	GCTAGCGAGCTCTAAATACAATGTTCCTTGGTTATCCCATCGCC
18	GCTCTAGATGGTAATCTCGAATTTGCTTGCTCTATTTGTTGT
19	ACGACCATCACACCACTGAAGACT
20	CCAAAGGCGCAAATCCTGATCCAA
21	ATCGTGAAATTGCAGGCAGCTTGG
22	CATGGCAACGGCAGAAGGCAATAA
23	TTCTGTCTCCGGTGAAGGTGAA
24	TAAGGTTGGCCATGGAACTGGCAA
25	CCAACTTTGCCGCCACTTAT
26	TTGTAAGCAACCATCCCCTACA
27	ACATCGCAGTCACAATCTCTCAGT
28	CATGGGCGAGCTTGCTTAAA
29	AGTAGCGTCTGGGTTGAAAGAAAGTA
30	CAGTGTTTCTCACAACAGTGTCTTTAAT
31	TGACATTGAGAAAAACATTTGGGTTA
32	TCAACGTTTTCGTAAGTGGTCTTAA
33	CCTCACTAAAGGGAACAAAAGCTG
34	CAGTGAATAATTCTTCACCTTTAGACATTTT
35	TGTTGTGTGGAATTGTGAGC
36	TAGCATCACCTTCACCTTCAC
37	CGGCTGGACTCCCGAGAT
38	TTGTAGCTCTGTGAGTATATTCTGTATTGC
39	CGATAAGAGAGAATTCGCAGCAATTGTTTTCTGTGCCATCA
40	TGATGGCACAGAAAACAATTGCTGCGAATTCTCTCTTATCG
41	CACACACGAGCTCAGAGCAAAGACAAAAAAAATAAGACA
42	CTAGTCTAGACTAATGTAGAATACGAACCTGAGTT
43	CACACACGAGCTCTCTTAAACACTTATGGGCAGC

Appendix B: Integration Loci Identification and Validation

A variety of low throughput methodologies were used to identify the integration loci for the ten clones (A through J) discussed in Chapter 3. Using a modified TAIL PCR method[256], clones A & B were identified in an unplaced genomic contig of the human genome. A set of 3 specific primers were used in succession (GGACTCAAGACGATAGTTACCGGA, TACAGCGTGAGCTATGAGAAAGCG and TTATAGTCCTGTCGGGTTTCGCCA) along with a series of linker primers (GTGCAGCCTTGGGTCGCCGTGT/3InvdT/, CGTTTGCTATTTACGCTCCTGCCA and TACGCTCCTGCCATGTGCCGCTGG). Following sequence confirmation, the locus was confirmed using primers CTGTGAGTTGAATGCACACATCACAAAGGA (genome specific) and TTATAGTCCTGTCGGGTTTCGCCA (transgene specific).

Clone C was identified in the 26th intron of the DCC gene on chromosome 18 using TAIL PCR. The DCC gene encodes for a netrin receptor protein. Prior to TAIL PCR, clonal gDNA was fragmented by shearing. A set of 3 specific primers (TACCGCGCCACATAGCAGAACTTTAAAAGTGCTCAT, TACCGCGCCACATAGCAGAACTTTAAAAGTGCTCAT, and ACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAG) were used in succession along with the degenerate primer WGTGNAGWANCANAGA. Following sequence confirmation, the locus was confirmed using the 3rd specific primer and primer TAACACAAGAACAGAAAACCAAACACCACATGTT (genomic specific).

Clone G was identified on chromosome 7, 31kb from the SEMA3A gene, which is a secreted neuronal protein, using a modified TAIL PCR method[256]. A set of 3 specific primers were used in succession (CACTGCATTCTAGTTGTGGTTTGTCC, CGAGATAGGGTTGAGTGTTGTTCC and CCCACTACGTGAACCATCACCCTA)

along with a series of linker primers (GTGCAGCCTTGGGTCGCCGTGT/3InvdT/, CGTTTGCTATTTACGCTCCTGCCA and TACGCTCCTGCCATGTGCCGCTGG). Following sequence confirmation, the locus was confirmed using primers GAGTTGAAATGTAAACGCAATTATTTACAATGGTA (genome specific) and CGTTTGCTATTTACGCTCCTGCCA (transgene specific).

Additionally, clone H was identified in the immunoglobulin rich region of chromosome 14 using TAIL PCR. The integration locus is within the IGHG2 gene, which is an immunoglobulin heavy constant gamma 2. A set of three specific primers (CGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAA, AGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTT, and CTGATCTTCAGCATCTTTTACTTTCACCAGCGTTT) were used in succession along with degenerate primer NTCGASTWTSGWGTT. The resulting fragment was sequenced using primer GAAGGCAAAATGCCGCAAAAAAGG. This locus was confirmed by PCR using primers CTTTATTTCCATGCTGGGTGCCTGGGAAGTATGTACA (genome specific) and GAAGGCAAAATGCCGCAAAAAAGG (transgene specific).

Finally, TAIL PCR was used to identify the 5th intron of the GRIK1 gene on chromosome 21 as the integration site for clones I and J. This gene encodes for a glutamate receptor that is typically expressed in neuronal cells. For clone I, the three transgene specific primers were TGTGTGAAATTGTTATCCGCTCACAATTC, TGCCTAATGAGTGAGCTAACTCACATTAAT, and ATCGCGAGCACTTTTCGGGGAAATGT and the degenerate primer was AGWGNAGWANCA. For clone J, the three transgene specific primers were ATAACACACAATCAACAGGGGAGTGAGCTGGAGGG, TTCCTGTGTGAAATTGTTATCCGCTCACAATTCCA, and TGAGCTAACTCACATTAATTGCGTTGCGCTCACTG and the degenerate primer

was NTCGASTWTSGWGTT. The resulting fragments were sequenced using primers GCTCATGAGACAATAACCCTGATAAATGC and TTTATTTTTCTAAATACATTCAAATATGTATCCGC for clones I and J respectively. The locus was confirmed for each clone using the transgene specific sequencing primer and AGAAGAATTCAGTAGACATAGAGCTGAGGA (genomic specific).

An inverse PCR strategy was used to identify the integration locus of clone E, 9kb downstream from the SPINK2 gene, which encodes for a serine peptidase inhibitor, Kazal type 2 on chromosome 4. First, the clonal gDNA was fragmented using an AvrII restricition enzyme digest. An overnight, low-concentration ligation reaction was used to circularize the fragmented DNA. Transgene-specific inverted primers AAGGGCGACACGGAAATGTTGAATACTCATACT and AGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTC were used. Following sequencing, the locus was confirmed using primers TTTAGTAGAGACAAGGTTTCACTATGTTGGCC (genome specific) and AAGGGCGACACGGAAATGTTGAATACTCATACT (transgene specific).

Plasmid recovery was used to identify the loci of the remaining two clones, D and F. Clone D is located near the SEMA6A gene, which encodes for a transmembrane domain on chromosome 5. The clonal gDNA was first digested with NdeI and ligated with a Kanamycin fragment with NdeI sites on either end. The resulting ligation was transformed into E. coli and selected on kanamycin plates. The resulting plasmid DNA was sequenced using primer TTGACAAACTACAGCATTCTGTCCTGGG. The locus was confirmed using primers CCCAGGACAGAATGCTGTAGTTTGTCAA (genomic specific) and GCCTGGTATCTTTATAGTCCTGTCG (plasmid specific)

Clone F was identified in the 1st intron of the SV2B gene, which encodes for a synaptic vesicle glycoprotein, on chromosome 15. First, clonal gDNA was fragmented

using a HydroShear. The pUC19 bacterial expression vector was linearized and blunted. Both were end-repaired and cleaned up prior to an overnight ligation reaction to create a plasmid library of Clone F gDNA. This library was transformed into E. coli and plated on large, LB plates supplemented with ampicillin. After overnight growth, the plates were scraped, pooled and the plasmid DNA was extracted. From the plasmid DNA, primers GATCTCCTGCAGGCCGGTGTTCTTCAGGATCTGCTTGC and CCAGCCTGGGCAAGCCCCTGGGCAGCCTGCACGAGTGG (specific to hrGFP) were used to perform an inverse PCR and the resulting fragments were TOPO cloned and sequenced using M13 forward and reverse primers. The locus was confirmed using primers GGAGAGGAAGTGACATCTGAATTAGATTTTATAGG (genomic specific) and GCAGCCTGCACGAGTGGGTGTAATA (transgene specific).

Appendix C: Codon Optimized Gene Variants

All genes were assembled from gBlock fragments ordered from IDT using homologous recombination. Gene sequence was confirmed by standard sequencing.

eGFP WT

```
atgcgtaaaggagaagaacttttcactggagttgtcccaattcttgttgaattagatggtgatgttaatgg
gcacaaatttctgtcagtggagagggtgaaggtgatgcaacatacggaaaacttacccttaaatttattt
gcactactggaaaactacctgttccatggccaacacttgtcactactttcggttatggtgttcaatgcttt
gcgagatacccagatcatatgaaacagcatgacttttcaagagtgccatgcccgaaggttatgtacagga
aagaactatattttcaaagatgacgggaactacaagacacgtgctgaagtcaagtttgaaggtgataccc
ttgttaatagaatcgagttaaaaggtattgattttaaagaagatggaaacattcttggacacaaattggaa
tacaactataactcacacaatgtatacatcatggcagacaaacaaaagaatggaatcaaagttaacttcaa
aattagacacaacattgaagatggaagcgttcaactagcagaccattatcaacaaaatactccaattggcg
atggccctgtccttttaccagacaaccattacctgtccacacaatctgccctttcgaaagatcccaacgaa
aagagagaccacatggtccttcttgagtttgtaacagctgctgggattacacatggcatggatgaactata
caaaaggcctgcagcaaacgacgaaaactacgctgcagcagtttaa
```

eGFP control table

```
ATGAGAAAAGGTGAAGAATTGTTTACTGGTGTTGTTCCAATTTTGGTTGAATTGGATGGTGATGTTAATGG
TCATAAATTTTCTGTTTCTGGTGAAGGTGAAGGTGATGCTACTTATGGTAAATTGACTTTGAAATTTATTT
GTACTACTGGTAAATTGCCAGTTCCATGGCCAACTTTGGTTACTACTTTTGGTTATGGTGTTCAATGTTTT
GCTAGATATCCAGATCATATGAAACAACATGATTTTTTTAAATCTGCTATGCCAGAAGGTTATGTTCAAGA
AAGAACTATTTTTTTTAAAGATGATGGTAATTATAAAACTAGAGCTGAAGTTAAATTTGAAGGTGATACTT
TGGTTAATAGAATTGAATTGAAAGGTATTGATTTTAAAGAAGATGGTAATATTTTGGGTCATAAATTGGAA
TATAATTATAATTCTCATAATGTTTATATTATGGCTGATAAACAAAAAAATGGTATTAAAGTTAATTTTAA
AATTAGACATAATATTGAAGATGGTTCTGTTCAATTGGCTGATCATTATCAACAAAATACTCCAATTGGTG
ATGGTCCAGTTTTGTTGCCAGATAATCATTATTTGTCTACTCAATCTGCTTTGTCTAAAGATCCAAATGAA
AAAAGAGATCATATGGTTTTGTTGGAATTTGTTACTGCTGCTGGTATTACTCATGGTATGGATGAATTGTA
TAAAAGACCAGCTGCTAATGATGAAAATTATGCTGCTGCTGTTTAA
```

eGFP high expression table

```
ATGAGAAAGGGTGAAGAATTGTTCACAGGTGTGGTGCCAATTTTGGTGGAATTGGATGGTGATGTGAATGG
TCATAAGTTCTCAGTGTCAGGTGAAGGTGAAGGTGATGCTACATACGGTAAGTTGACATTGAAGTTCATTT
GTACAACAGGTAAGTTGCCAGTGCCATGGCCAACATTGGTGACAACATTCGGTTACGGTGTGCAATGTTTC
GCTAGATACCCAGATCATATGAAGCAACATGATTTCTTCAAGTCAGCTATGCCAGAAGGTTACGTGCAAGA
AAGAACAATTTTCTTCAAGGATGATGGTAATTACAAGACAAGAGCTGAAGTGAAGTTCGAAGGTGATACAT
TGGTGAATAGAATTGAATTGAAGGGTATTGATTTCAAGGAAGATGGTAATATTTTGGGTCATAAGTTGGAA
TACAATTACAATTCACATAATGTGTACATTATGGCTGATAAGCAAAAGAATGGTATTAAGGTGAATTTCAA
GATTAGACATAATATTGAAGATGGTTCAGTGCAATTGGCTGATCATTACCAACAAAATACACCAATTGGTG
ATGGTCCAGTGTTGTTGCCAGATAATCATTACTTGTCAACACAATCAGCTTTGTCAAAGGATCCAAATGAA
AAGAGAGATCATATGGTGTTGTTGGAATTCGTGACAGCTGCTGGTATTACACATGGTATGGATGAATTGTA
CAAGAGACCAGCTGCTAATGATGAAAATTACGCTGCTGCTGTGTAA
```

eGFP control matrix 1

ATGCGAAAAGGCGAGGAACTGTTTACAGGTGTAGTGCCCATTTTGGTGGAGCTAGATGGTGACGTTAACGG
TCATAAATTCTCGGTATCCGGAGAAGGCGAAGGTGACGCTACCTATGGTAAACTCACCTTAAAGTTCATCT
GCACAACTGGAAAGCTTCCAGTCCCTTGGCCCACATTGGTAACGACATTTGGCTATGGTGTACAATGCTTC
GCAAGATACCCGGATCACATGAAACAACATGACTTTTTCAAGTCTGCTATGCCTGAAGGTTACGTCCAGGA
GCGGACTATTTTCTTTAAGGATGATGGTAATTACAAAACCCGTGCCGAAGTAAAGTTCGAAGGCGACACGT
TGGTAAACCGCATCGAACTTAAAGGCATAGATTTTAAGGAAGATGGTAATATCTTGGGTCATAAATTAGAA
TATAACTACAATTCTCATAACGTTTACATTATGGCTGATAAGCAAAAGAACGGAATTAAAGTGAACTTTAA
AATCAGACACAACATTGAGGATGGTTCTGTTCAATTGGCTGATCATTACCAACAGAATACACCTATCGGAG
ACGGCCCAGTTTTACTACCAGATAATCATTACTTAAGTACTCAGTCTGCATTAAGCAAGGATCCAAATGAG
AAGAGAGATCACATGGTTTTGTTGGAATTTGTAACAGCAGCCGGAATAACACATGGCATGGACGAGTTGTA
CAAAAGACCTGCGGCAAATGATGAAAACTATGCAGCTGCTGTATAA

eGFP control matrix 2

ATGAGAAAAGGTGAAGAATTGTTCACAGGAGTTGTTCCCATTTTAGTTGAATTAGACGGTGATGTAAATGG
TCATAAATTTTCTGTTTCCGGCGAAGGAGAGGGAGATGCAACGTACGGAAAGTTGACATTGAAGTTTATAT
GCACCACAGGAAAGCTTCCCGTTCCATGGCCTACCTTGGTAACGACGTTTGGTTATGGTGTTCAATGCTTT
GCTCGATATCCGGATCACATGAAGCAGCATGATTTCTTCAAGAGCGCTATGCCCGAAGGGTATGTTCAAGA
AAGAACCATTTTCTTTAAAGATGATGGCAATTATAAGACAAGAGCTGAAGTAAAATTCGAGGGAGATACAT
TGGTTAATCGAATAGAATTAAAGGGTATTGACTTTAAAGAGGATGGTAATATTCTGGGTCACAAACTTGAA
TACAATTATAATTCCCATAACGTCTACATAATGGCCGACAAACAAAAGAATGGTATTAAAGTCAATTTCAA
AATTCGTCATAACATCGAGGACGGCAGCGTCCAATTAGCGGATCACTATCAACAAAATACACCTATAGGTG
ACGGTCCCGTGCTTTTACCAGACAACCACTATCTAAGCACTCAATCTGCTCTATCGAAAGATCCTAACGAA
AAAAGAGATCATATGGTGCTGTTAGAATTTGTCACTGCGGCTGGTATTACACATGGGATGGACGAACTTTA
CAAAAGGCCCGCTGCTAATGATGAAAATTATGCAGCCGCCGTTTAA

eGFP control matrix 3

ATGAGAAAGGGCGAAGAGCTATTCACAGGTGTCGTACCAATTTTAGTAGAATTAGACGGTGATGTTAACGG
GCACAAGTTTTCTGTTTCTGGAGAAGGAGAAGGCGATGCAACTTATGGTAAATTGACTTTAAAATTTATTT
GCACGACTGGTAAATTGCCAGTTCCATGGCCAACTTTAGTTACTACATTTGGTTATGGTGTTCAGTGTTTT
GCAAGATACCCAGATCATATGAAACAACACGATTTCTTTAAATCTGCAATGCCAGAAGGTTACGTTCAAGA
AAGAACTATCTTCTTTAAAGATGATGGCAACTATAAAACAAGAGCAGAGGTTAAGTTCGAGGGTGATACTT
TGGTTAACAGGATAGAACTAAAGGGCATAGATTTCAAGGAAGATGGGAATATATTGGGTCACAAGCTAGAA
TACAATTACAATAGTCATAACGTTTATATAATGGCAGACAAGCAAAAAAATGGAATAAAGGTAAATTTTAA
AATTAGACACAATATAGAGGATGGCAGCGTTCAATTAGCTGACCATTATCAACAGAATACACCAATTGGTG
ATGGCCCAGTTCTACTGCCTGATAATCATTATTTGTCCACTCAAAGTGCCCTGTCTAAAGATCCAAATGAG
AAGCGAGATCATATGGTACTCTTAGAGTTTGTGACTGCAGCTGGTATCACACATGGGATGGATGAATTATA
CAAAAGACCAGCTGCAAATGACGAAAATTATGCCGCTGCGGTTTAA

eGFP high expression matrix 1

ATGAGAAAGGGTGAAGAACTTTTTACAGGTGTTGTTCCAATCTTGGTAGAACTGGACGGTGACGTCAACGG
TCACAAGTTCTCAGTGTCAGGGGAGGGTGAAGGTGATGCTACCTACGGTAAGCTTACTCTAAAGTTCATCT
GTACCACGGGGAAATTACCCGTGCCATGGCCAACTCTAGTTACAACTTTTGGATACGGTGTTCAATGTTTT
GCTAGATACCCAGATCATATGAAGCAACACGATTTTTTTAAATCAGCGATGCCGGAAGGTTACGTGCAAGA
AAGGACTATCTTTTTCAAAGATGACGGTAACTACAAGACCAGGGCTGAAGTTAAATTTGAAGGTGACACTC
TGGTGAACCGAATAGAATTAAAGGGTATTGATTTCAAGGAAGATGGTAACATTTTGGGTCACAAGTTGGAA
TACAATTATAACTCCCATAACGTTTACATTATGGCTGATAAACAAAAGAATGGTATCAAGGTAAATTTTAA
AATCAGACACAACATTGAAGACGGTTCCGTACAATTGGCTGATCATTATCAACAAAATACACCTATTGGTG
ACGGTCCTGTTTTACTCCCCGATAATCATTATTTGTCCACTCAATCCGCTTTGTCAAAGGATCCAAATGAA

AAGCGTGACCATATGGTGTTACTGGAATTCGTTACTGCTGCAGGGATCACGCATGGCATGGATGAATTGTA
TAAGAGACCAGCTGCCAATGACGAAAACTACGCTGCCGCCGTTTAA

eGFP high expression matrix 2

ATGCGTAAAGGTGAAGAGCTCTTCACAGGTGTCGTTCCAATCTTGGTCGAACTAGATGGTGACGTTAACGG
TCACAAGTTTTCTGTTTCCGGTGAAGGTGAAGGTGACGCTACCTATGGGAAGCTTACGCTGAAATTTATCT
GTACAACCGGTAAGTTGCCAGTCCCATGGCCAACTTTGGTAACAACATTTGGCTACGGTGTTCAATGTTTT
GCGCGCTACCCAGATCATATGAAGCAGCATGATTTCTTTAAAAGCGCCATGCCAGAGGGTTATGTCCAAGA
AAGGACGATATTCTTCAAGGACGACGGTAACTACAAGACCAGGGCTGAAGTTAAATTTGAAGGCGACACTC
TGGTGAATAGAATAGAACTGAAAGGTATCGATTTCAAGGAAGATGGTAACATTCTTGGGCATAAACTAGAA
TACAATTATAACTCCCATAACGTGTATATTATGGCTGACAAGCAAAAAAATGGAATCAAGGTTAACTTCAA
AATTCGTCATAACATCGAGGACGGTTCTGTCCAATTGGCTGATCATTATCAACAAAACACCCCAATCGGTG
ATGGTCCAGTTCTTCTGCCAGATAACCATTACTTGTCAACTCAAAGCGCACTCTCTAAGGACCCTAATGAA
AAGAGAGATCATATGGTTCTTCTCGAGTTCGTCACTGCTGCTGGTATCACTCACGGTATGGATGAACTATA
CAAGAGACCAGCTGCTAATGACGAAAATTATGCCGCTGCTGTTTAA

eGFP high expression matrix 3

ATGCGTAAAGGTGAAGAATTGTTTACAGGTGTTGTCCCAATCTTGGTCGAATTGGACGGTGACGTGAATGG
GCATAAATTTTCGGTATCTGGGGAGGGTGAAGGTGATGCTACCTACGGTAAGTTGACTCTGAAATTTATCT
GTACCACAGGCAAGTTGCCGGTACCGTGGCCCACGCTCGTTACGACGTTTGGCTACGGTGTTCAATGTTTT
GCGCGCTACCCAGACCACATGAAACAACACGATTTCTTTAAAAGCGCAATGCCGGAAGGCTACGTTCAAGA
AAGGACAATCTTTTTCAAGGACGACGGTAACTACAAAACTAGAGCTGAAGTCAAGTTTGAAGGTGACACGC
TGGTCAACCGTATTGAATTGAAAGGTATTGACTTCAAGGAAGACGGTAACATTCTTGGACATAAACTCGAA
TATAACTACAACTCCCACAACGTTTATATCATGGCCGATAAGCAAAAGAATGGTATTAAGGTGAATTTTAA
AATTCGTCACAACATTGAAGATGGTTCTGTTCAACTAGCTGACCATTACCAACAAAACACTCCAATCGGTG
ACGGTCCAGTCTTGCTGCCCGACAACCATTATCTCTCTACACAATCTGCTCTTTCTAAGGACCCAAATGAA
AAAAGGGATCATATGGTATTGTTAGAGTTTGTTACAGCAGCTGGTATCACGCATGGTATGGACGAACTGTA
CAAGAGACCAGCGGCAAACGATGAAAACTACGCTGCTGCCGTTTAA

CatA wild type

ATGGAAGTTAAAATATTCAATACTCAGGATGTGCAAGATTTTTTACGTGTTGCAAGCGGACTTGAGCAAGA
AGGTGGCAATCCGCGTGTAAAGCAGATCATCCATCGTGTGCTTTCAGATTTATATAAAGCCATTGAAGATT
TGAATATCACTTCAGATGAATACTGGGCAGGTGTGGCATATTTAAATCAGCTAGGTGCCAATCAAGAAGCT
GGTTTACTCTCGCCAGGCTTGGGTTTTGACCATTACCTCGATATGCGTATGGATGCCGAAGATGCCGCACT
AGGTATTGAAAATGCGACACCACGTACCATTGAAGGCCCGCTATACGTGGCAGGTGCGCCTGAATCGGTAG
GTTATGCGCGCATGGATGACGGAAGTGATCCAAATGGTCATACCCTGATTCTACATGGCACGATCTTTGAT
GCAGATGGAAAACCTTTACCCAATGCCAAAGTTGAAATCTGGCATGCCAATACCAAAGGCTTTTATTCACA
CTTCGACCCAACAGGCGAGCAGCAGGCGTTCAATATGCGCCGTAGTATTATTACCGATGAAAACGGTCAGT
ATCGCGTTCGTACCATTTTGCCTGCGGGTTATGGTTGCCCACCAGAAGGTCCAACGCAACAGTTGCTGAAT
CAGTTGGGCCGTCATGGTAACCGCCCTGCGCACATTCACTATTTTGTTTCTGCCGATGGACACCGCAAACT
AACTACGCAAATTAATGTGGCTGGCGATCCGTACACCTATGACGACTTTGCTTATGCAACCCGTGAAGGCT
TGGTGGTTGATGCAGTGGAACACACCGATCCTGAAGCCATTAAGGCCAATGATGTTGAAGGCCCATTCGCT
GAAATGGTTTTCGATCTAAAATTGACGCGTTTGGTTGATGGTGTAGATAACCAAGTTGTTGATCGTCCACG
TCTAGCGGTGTAA

CatA Blue Heron

ATGGAAGTTAAGATTTTTAACACTCAAGACGTACAAGATTTTTTACGTGTCGCAAGCGGATTAGAACAAGA
AGGCGGAAATCCCAGAGTAAAGCAAATAATACACAGAGTTTTATCAGATTTGTACAAAGCGATAGAAGATT
TAAATATAACTTCAGATGAATATTGGGCTGGTGTAGCATACTTAAATCAATTAGGAGCAAATCAAGAAGCA
GGATTATTATCACCCGGACTAGGTTTCGATCATTATTTAGATATGAGAATGGACGCAGAAGACGCAGCCTT
AGGTATTGAAAACGCCACGCCAAGAACAATAGAAGGACCACTTTATGTTGCAGGTGCCCCCGAATCAGTAG
GTTACGCAAGAATGGATGACGGTTCCGACCCAAATGGCCACACTTTAATTTTACACGGAACAATTTTTGAC
GCTGATGGTAAACCCCTTCCTAATGCTAAAGTTGAGATATGGCACGCAAACACTAAAGGTTTCTATTCACA
TTTTGACCCAACAGGAGAACAACAAGCATTCAACATGAGAAGATCAATTATAACAGACGAGAACGGACAAT
ACAGAGTAAGGACTATATTACCAGCAGGATACGGTTGCCCGCCAGAAGGCCCAACACAACAATTACTAAAT
CAATTAGGTAGACATGGAAATAGACCCGCTCACATTCATTATTTTGTTAGCGCAGATGGACACAGGAAATT
GACCACACAAATCAATGTTGCAGGAGATCCCTATACTTACGACGATTTTGCATACGCTACAAGAGAAGGGC
TAGTAGTAGACGCAGTAGAGCATACAGATCCAGAAGCAATAAAAGCAAATGACGTAGAAGGACCATTCGCA
GAAATGGTTTTCGACCTAAAACTTACTAGATTAGTAGATGGAGTAGATAATCAAGTTGTAGACAGACCAAG
ATTAGCAGTCTAA

CatA control 1

ATGGAAGTTAAGATTTTTAATACCCAAGATGTGCAGGATTTTTTGAGAGTTGCTTCGGGCCTAGAACAAGA
AGGTGGTAATCCTCGTGTTAAACAGATTATACACCGTGTCTTGTCCGATCTATATAAAGCAATTGAAGATT
TGAACATTACGTCAGATGAATATTGGGCTGGCGTAGCGTATTTGAACCAGTTAGGTGCTAACCAAGAGGCA
GGCTTGTTAAGTCCCGGTTTGGGCTTTGATCATTACTTGGACATGAGGATGGATGCAGAAGATGCTGCATT
AGGTATTGAAAATGCCACGCCAAGAACTATAGAAGGTCCACTTTATGTTGCAGGTGCCCCAGAAAGCGTCG
GTTACGCTCGTATGGATGATGGATCTGACCCAAATGGACACACCTTAATCTTGCACGGAACAATCTTTGAT
GCAGACGGAAAACCTCTTCCGAACGCAAAAGTGGAAATTTGGCATGCAAACACTAAGGGCTTTTACAGCCA
CTTTGACCCAACTGGTGAGCAGCAGGCATTTAACATGCGAAGAAGTATAATAACTGATGAAAACGGACAAT
ACAGAGTGAGGACCATCTTGCCAGCAGGTTACGGATGTCCTCCAGAGGGTCCCACACAACAACTACTTAAC
CAGTTAGGACGCCATGGTAATAGACCTGCTCATATTCATTACTTTGTCTCTGCGGACGGCCATAGAAAGTT
AACAACACAAATAAACGTTGCGGGTGATCCTTACACTTATGACGACTTCGCATATGCCACCCGTGAGGGCT
TAGTTGTAGATGCTGTCGAACACACTGATCCAGAAGCTATTAAGGCTAATGACGTAGAAGGTCCTTTTGCG
GAAATGGTTTTCGATTTAAAATTAACAAGATTAGTCGATGGAGTTGATAACCAAGTTGTTGATAGGCCACG
ACTTGCTGTCTAA

CatA control 2

ATGGAAGTGAAAATCTTCAACACACAGGACGTACAGGACTTTTTGAGAGTGGCATCTGGTTTGGAACAAGA
AGGGGGCAATCCTCGAGTGAAGCAGATCATTCATAGAGTGCTTTCTGATCTATACAAAGCTATCGAAGATT
TAAATATTACGTCAGACGAATATTGGGCAGGGGTCGCTTATTTGAATCAATTAGGTGCCAACCAAGAGGCA
GGCCTTTTGAGTCCAGGATTGGGATTTGACCATTACTTGGATATGCGTATGGATGCTGAGGATGCAGCATT
AGGAATAGAGAATGCAACACCCAGAACGATAGAGGGACCGTTATATGTTGCTGGTGCTCCAGAGTCAGTTG
GTTACGCCCGTATGGATGATGGCTCTGATCCAAATGGCCATACATTAATTTTGCATGGTACTATATTTGAT
GCTGATGGAAAACCACTTCCCAACGCTAAAGTTGAAATTTGGCATGCCAACACCAAGGGGTTTTATTCACA
CTTTGATCCGACAGGCGAGCAACAAGCTTTTAACATGAGACGGAGTATAATAACAGATGAAAATGGCCAGT
ATAGGGTAAGGACTATTCTACCAGCCGGTTACGGATGTCCCCCAGAAGGTCCCACACAACAATTACTAAAC
CAACTGGGACGACATGGAAATAGGCCAGCTCATATACACTACTTCGTTAGCGCTGATGGCCATAGAAAACT
AACAACACAAATAAATGTGGCAGGAGACCCTTACACATATGACGACTTTGCATATGCCACACGTGAAGGCT
TAGTTGTTGACGCTGTCGAACATACAGATCCAGAAGCTATCAAGGCTAATGACGTCGAGGGTCCTTTTGCA
GAAATGGTATTTGATTTAAAGTTAACTAGACTAGTGGATGGGGTTGACAATCAAGTAGTAGATCGCCCCAG
ATTGGCGGTGTAA

CatA control 3

ATGGAGGTAAAAATATTCAACACACAAGATGTTCAAGATTTTTTGAGAGTGGCTTCTGGCTTAGAGCAAGA
AGGTGGGAACCCAAGAGTCAAACAAATAATACACCGAGTGCTGTCAGATTTATACAAAGCTATTGAAGATC
TCAATATAACAAGCGATGAATATTGGGCCGGCGTGGCATACCTAAACCAATTAGGTGCCAATCAAGAGGCT
GGTCTTCTGAGCCCAGGCCTTGGGTTTGATCATTACTTAGACATGAGGATGGATGCTGAGGACGCAGCATT
AGGGATAGAAAATGCGACTCCAAGAACTATAGAGGGCCCACTATATGTAGCCGGCGCACCCGAAAGTGTGG
GATATGCAAGAATGGATGACGGCTCTGATCCGAACGGTCATACTTTGATACTTCACGGCACCATTTTTGAT
GCCGATGGCAAGCCATTACCGAATGCGAAGGTTGAAATTTGGCACGCTAACACTAAGGGCTTTTATTCCCA
TTTTGATCCTACAGGAGAACAACAAGCTTTTAACATGAGAAGATCAATAATCACCGACGAGAATGGCCAAT
ATAGGGTTAGAACAATATTACCAGCTGGCTACGGTTGTCCTCCTGAGGGCCCGACCCAACAGCTCCTTAAT
CAGTTAGGTCGTCATGGTAACAGACCAGCTCATATACACTATTTTGTCAGTGCAGATGGACATCGGAAATT
AACTACTCAAATAAACGTAGCAGGCGATCCGTACACATATGACGATTTCGCCTATGCAACAAGGGAAGGTC
TTGTGGTTGACGCTGTAGAGCATACCGACCCTGAAGCAATTAAAGCGAATGACGTTGAGGGTCCTTTCGCC
GAAATGGTGTTTGATTTAAAGTTAACTAGACTGGTGGATGGCGTTGATAATCAAGTTGTAGATCGCCCTAG
GCTCGCGGTGTAA

CatA stationary 1

ATGGAAGTAAAGATCTTCAACACTCAAGACGTTCAAGATTTTTTAAGAGTTGCTTCAGGACTTGAACAGGA
AGGCGGTAACCCTCGGGTAAAGCAAATTATCCATCGGGTCCTGTCTGATCTATACAAGGCAATCGAAGACC
TAAACATCACTTCTGACGAATATTGGGCGGGCGTGGCGTACCTTAACCAATTGGGAGCGAATCAAGAGGCT
GGCTTATTAAGCCCAGGCCTTGGATTCGATCACTATCTTGATATGAGAATGGACGCAGAAGATGCAGCCTT
AGGTATAGAGAACGCTACCCCAAGAACTATCGAAGGCCCATTGTATGTCGCTGGTGCCCCCGAGAGTGTCG
GTTATGCCCGTATGGATGATGGGTCGGATCCAAACGGTCATACTTTGATATTGCACGGTACTATATTCGAT
GCCGATGGTAAACCTCTGCCTAATGCAAAGGTGGAAATATGGCATGCGAATACAAAGGGATTCTACTCACA
TTTTGACCCAACGGGAGAACAACAAGCCTTCAATATGCGGCGGTCTATTATAACGGATGAGAACGGCCAAT
ACAGGGTAAGGACCATATTGCCCGCAGGGTACGGCTGCCCACCAGAAGGTCCAACTCAACAACTCTTAAAC
CAATTGGGCAGGCATGGCAACAGGCCTGCCCACATTCACTATTTCGTGTCAGCGGATGGTCACAGGAAGTT
AACAACACAAATCAACGTCGCAGGTGATCCGTACACCTACGACGATTTTGCATATGCTACCAGAGAAGGCC
TTGTAGTTGATGCTGTGGAACATACGGACCCCGAAGCGATCAAGGCCAATGATGTAGAAGGTCCTTTCGCG
GAGATGGTTTTCGATTTGAAATTGACGAGACTAGTTGATGGTGTAGATAATCAGGTTGTAGACAGACCAAG
GTTAGCAGTCTAA

CatA stationary 2

ATGGAAGTTAAAATCTTCAACACCCAGGATGTTCAAGACTTTTTGCGTGTAGCCTCCGGACTTGAACAAGA
AGGTGGTAATCCAAGAGTAAAGCAGATCATTCACAGAGTTTTATCTGATCTATACAAGGCGATCGAAGATT
TGAATATCACTTCGGACGAGTACTGGGCCGGAGTTGCTTACTTGAATCAGTTGGGTGCTAACCAAGAAGCC
GGTTTGTTGTCACCAGGCTTGGGTTTTGACCATTACCTCGACATGCGGATGGATGCTGAAGACGCGGCCCT
GGGTATTGAAAACGCTACCCCAAGGACGATAGAAGGCCCCCTTTATGTTGCAGGTGCTCCTGAGAGTGTGG
GCTATGCAAGAATGGACGACGGTTCTGACCCCAACGGTCACACATTGATTTTGCACGGTACAATCTTTGAC
GCCGATGGTAAGCCGTTGCCAAACGCTAAGGTGGAGATTTGGCATGCGAATACCAAAGGTTTTTATAGCCA
CTTCGATCCAACAGGAGAACAACAAGCTTTCAATATGAGAAGATCGATTATTACAGACGAAAACGGACAAT
ATAGAGTCAGGACTATACTGCCCGCGGGATACGGTTGTCCTCCAGAAGGTCCAACGCAGCAACTACTTAAT
CAATTAGGAAGGCATGGAAATAGACCCGCTCATATCCACTATTTCGTTTCTGCTGATGGCCATCGTAAATT
GACTACTCAAATCAACGTTGCCGGTGATCCATATACTTATGATGACTTTGCCTATGCAACACGAGAAGGCT
TAGTAGTGGACGCTGTGGAGCACACCGACCCTGAAGCTATCAAAGCTAACGACGTTGAAGGGCCTTTCGCG
GAAATGGTCTTTGACTTGAAACTAACCAGATTAGTCGACGGAGTAGATAACCAAGTTGTGGACCGGCCTCG
ATTAGCAGTCTAA

CatA stationary 3

ATGGAGGTTAAGATTTTCAACACTCAAGATGTCCAAGACTTTTTAAGAGTGGCCTCGGGGCTGGAACAAGA
AGGCGGCAATCCAAGAGTTAAACAGATCATCCATAGAGTTTTGTCCGATCTTTACAAAGCCATTGAAGATT
TAAACATCACTTCAGACGAATATTGGGCAGGAGTAGCTTACTTGAATCAGTTGGGTGCTAACCAGGAAGCC
GGTCTGCTATCTCCTGGCCTAGGTTTCGATCACTATTTGGATATGAGAATGGATGCTGAAGATGCAGCATT
AGGTATCGAGAATGCTACTCCAAGAACGATAGAAGGGCCTCTATATGTAGCAGGTGCTCCCGAGTCGGTCG
GCTACGCCCGTATGGACGACGGTTCAGATCCGAACGGACATACTCTGATTCTACATGGAACAATCTTTGAC
GCCGATGGAAAGCCCCTTCCCAACGCTAAAGTTGAAATCTGGCATGCCAATACTAAGGGATTTTATTCGCA
CTTCGATCCCACTGGTGAACAACAGGCTTTCAATATGAGGCGTAGTATCATCACTGATGAGAATGGCCAAT
ACAGAGTTAGAACAATATTACCCGCGGGATACGGGTGTCCTCCTGAAGGACCCACTCAACAATTACTCAAC
CAATTAGGTAGACATGGTAACCGCCCTGCTCATATTCACTACTTTGTGTCCGCAGACGGTCACCGTAAGTT
AACGACACAAATCAACGTTGCCGGTGACCCGTACACTTACGACGATTTCGCCTACGCTACTAGAGAGGGTT
TAGTTGTCGATGCTGTAGAACATACTGATCCGGAAGCTATTAAAGCAAATGACGTTGAAGGGCCATTTGCA
GAGATGGTTTTTGACTTGAAACTGACACGGTTAGTCGATGGGGTGGACAACCAAGTGGTTGATAGACCCAG
GCTCGCAGTCTAA

Appendix D: Python Scripts

CODONUSAGEBIAS

```
#Read a file into the program as a string, name that string Linestring
myfile = raw_input("What file should be analyzed? Remember to include .txt! ")
linestring = open(myfile, 'r').read()
#Remove all spaces, carraige returns and name string Linestring_scrub
linestring_scrub = linestring.rstrip('\r\n')
linestring_scrub = linestring_scrub.replace("\r","")
linestring_scrub = linestring_scrub.replace("\n","")
#Determine the number of codons by dividing the length of the scrubbed string by 3
CodonNumber = len(linestring_scrub)/3
#Assign each codon to a string
Ala1 = 'GCG'; Ala2 = 'GCA'; Ala3 = 'GCT'; Ala4 = 'GCC'; Cys1 = 'TGT'; Cys2 = 'TGC';
Asp1 = 'GAT'; Asp2 = 'GAC'; Glu1 = 'GAG'; Glu2 = 'GAA'; Phe1 = 'TTT'; Phe2 = 'TTC'
Gly1 = 'GGG'; Gly2 = 'GGA'; Gly3 = 'GGT'; Gly4 = 'GGC'; His1 = 'CAT'; His2 = 'CAC'
Ile1 = 'ATA'; Ile2 = 'ATT'; Ile3 = 'ATC'; Lys1 = 'AAG'; Lys2 = 'AAA'; Leu1 = 'TTG'
Leu2 = 'TTA'; Leu3 = 'CTG'; Leu4 = 'CTA'; Leu5 = 'CTT'; Leu6 = 'CTC'; Met1 = 'ATG'
Asn1 = 'AAT' Asn2 = 'AAC' Pro1 = 'CCG' Pro2 = 'CCA' Pro3 = 'CCT' Pro4 = 'CCC'
Gln1 = 'CAG' Gln2 = 'CAA' Arg1 = 'AGG' Arg2 = 'AGA' Arg3 = 'CGG' Arg4 = 'CGA'
Arg5 = 'CGT' Arg6 = 'CGC' Ser1 = 'AGT' Ser2 = 'AGC' Ser3 = 'TCG' Ser4 = 'TCA'
Ser5 = 'TCT' Ser6 = 'TCC' Thr1 = 'ACG' Thr2 = 'ACA' Thr3 = 'ACT' Thr4 = 'ACC'
Val1 = 'GTG' Val2 = 'GTA' Val3 = 'GTT' Val4 = 'GTC' Trp1 = 'TGG' Tyr1 = 'TAT'
Tyr2 = 'TAC' Stp1 = 'TGA' Stp2 = 'TAG' Stp3 = 'TAA'
#Creat list of codons, CodonList
```

```
CodonList = [Ala1, Ala2, Ala3, Ala4, Cys1, Cys2,Asp1,Asp2,
Glu1,Glu2,Phe1,Phe2,Gly1,Gly2,Gly3,Gly4,His1,His2,Ile1,Ile2,Ile3, Lys1 , Lys2, Leu1,
Leu2, Leu3, Leu4, Leu5, Leu6, Met1, Asn1, Asn2, Pro1, Pro2, Pro3,  Pro4, Gln1, Gln2,
Arg1, Arg2, Arg3, Arg4, Arg5, Arg6, Ser1, Ser2, Ser3, Ser4, Ser5, Ser6, Thr1,Thr2,
Thr3, Thr4,Val1,Val2, Val3,Val4,Trp1,Tyr1,Tyr2,Stp1,Stp2,Stp3]
#Create list CodonCount to keep track of each codon in the imported string
CodonCount = 64*[0]
#Examine each codon in the string by splicing every three characters
        for i in range(0,CodonNumber):
            codoni = linestring_scrub[i*3:(i+1)*3]
            #Compare each codon against the strings as assigned above
            for j in range(0,64):
                if codoni == CodonList[j]:
 #If a codon match is found, increment the corresponding position in CodonCount list
                    CodonCount[j]=CodonCount[j]+1
#Put codon table data into a text file named myfile
data1 = raw_input("Enter file name for simple codon table, include .txt: ")
myfile = open(data1, 'w')
#Add a header
header = "%s %s %s" % ("Codon", "Count ", " Frequency per 1000")
myfile.write(header + '\n')
#Add each codon, the count for that codon and its frequency per 1000
        for l in range(0,64):
            frequency = float(CodonCount[l]*1000.00/CodonNumber)
            line = "%s %8.2f %8.2f" % (CodonList[l], CodonCount[l], frequency)
```

```
        myfile.write(line + '\n')
        myfile.close()
#This code will now deal with creating the Markov chain associated with adjacent codons
#Define the 64x64 matrix, CodonCount_adj, which will initially contain all zeros
CodonCount_adj = [ [ 0 for i in range(64) ] for j in range(64) ]
#Examine two adjacent codons in the string by splicing every three characters
        for i in range(0,CodonNumber-1):
            codoni = linestring_scrub[i*3:(i+1)*3]
            codonk = linestring_scrub[(i+1)*3:(i+2)*3]
#Compare each codon against the strings as assigned above
            for j in range(0,64):
                if codoni == CodonList[j]:
                    #When a match is identified, save the x position for the matrix
                    xvar = j
            for l in range(0,64):
                if codonk == CodonList[l]:
                    yvar = l
#If a codon pair match is found, increment the corresponding position in CodonCount list
            CodonCount_adj[xvar][yvar]=CodonCount_adj[xvar][yvar]+1
#Reset the stop codon positions to zero, as they cannot be adjacent.
        CodonCount_adj[61]=64*[0]
        CodonCount_adj[62]=64*[0]
        CodonCount_adj[63]=64*[0]
#Put Adj. codon table data into a text file named Adjfile
data2 = raw_input("Enter file name for adjacent codon table, include .txt: ")
```

```
Adjfile = open(data2, 'w')
#Add the 64 codons across the top row
header = "%s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s %s" % (' ','GCG','GCA','GCT','GCC','TGT','TGC','GAT','GAC','GAG','GAA','TTT','TTC','GGG','GGA','GGT','GGC','CAT','CAC','ATA','ATT', 'ATC', 'AAG', 'AAA', ‘TTG', 'TTA', 'CTG', 'CTA', 'CTT', 'CTC', 'ATG', 'AAT', 'AAC', 'CCG', 'CCA', 'CCT', 'CCC', 'CAG', 'CAA', 'AGG', 'AGA', 'CGG', 'CGA', 'CGT', 'CGC', 'AGT', 'AGC', 'TCG', 'TCA', 'TCT', 'TCC', 'ACG','ACA','ACT','ACC','GTG','GTA','GTT','GTC','TGG','TAT','TAC','TGA','TAG', 'TAA')
Adjfile.write(header + '\n')
#Add each codon name, the first row represents position1
    for l in range(0,64):
        line = CodonList[l]+'\t'
#Add each codon pair count, followed by a tab
        for m in range(0,64):
            line = line+str(CodonCount_adj[l][m])+'\t'
        Adjfile.write(line + '\n')
    Adjfile.close()
print "Analysis completed, data written to files: " + data1 + " and " +data2
```

GENEDESIGNER

This script has been abbreviated in places for ease of understanding.

```
ProteinSeq=raw_input("Input protein sequence using capitalized, single letter abbreviations: ")
#Read in Protein Sequence as a string
#Define AA codes
A=["GCG","GCA","GCT","GCC"] #alanine
C = ["TGT","TGC"] #cysteine
D = ["GAT","GAC"] #aspartic acid
E = ["GAG","GAA"] #glutamic acid
F = ["TTT","TTC"] #phenylalanine
G = ["GGG","GGA","GGT","GGC"] #Glycine
H = ["CAT","CAC"]#histidine
I = ["ATA","ATT","ATC"] #isoleucine
K = ["AAG","AAA"] #Lysine
L = ["TTG","TTA","CTG","CTA","CTT","CTC"] #Leucine
M = ["ATG"] #Methionine
N = ["AAT","AAC"] #Asparagine
P = ["CCG","CCA","CCT","CCC"] #proline
Q = ["CAG","CAA"] #glutamine
R = ["AGG","AGA","CGG","CGA","CGT","CGC"] #Arginine
S = ["AGT","AGC","TCG","TCA","TCT","TCC"] #serine
T = ["ACG","ACA","ACT","ACC"] #threonine
V = ["GTG","GTA","GTT","GTC"] #valine
W = ["TGG"] #tryptophan
Y = ["TAT","TAC"] #tyrosine
#Create an 8x8 matrix in which to store the probability values, Prob_list
```

```
Prob_list=61*[0]
#Read data from the csv file and unpack values, in same order, into the Prob_list
matrix_file=raw_input("Provide the file name containing the probability distribution, remember to include .csv! ")
import csv
ifile  = open(matrix_file, "rb")
reader = csv.reader(ifile)
rownum=0
j=0
for row in reader:
   Prob_list[j]=row
   j=j+1
#Convert the strings to floats
for a in range(61):
   for b in range(61):
      Prob_list[a][b]=float(Prob_list[a][b])
#Define the GCGA list (Ala1 followed by Ala)
GCGA = [ [ 0 for i in range(2) ] for j in range(4) ]
#Fill the AA list, first element is the 3bp string followed by the probability
#Probability is read in from the Prob_list
#Strings are read in from the 'A' list at the begining of the script
#create a placeholder variable for the Prob_list
m=0
n=0
for i in range(len(A)):
```

```
    GCGA[i][0]=A[i]
    GCGA[i][1]=Prob_list[n][m]
    m=m+1
GCGC = [ [ 0 for i in range(2) ] for j in range(2) ]
#Continue in this manner for all paired amino acid combinations…
TACY = [ [ 0 for i in range(2) ] for j in range(2) ]
#Create a dictionary, Master, that pairs each list with the corresponding string
Master = {'GCGA':GCGA…'TACY':TACY}

for i in range(len(C)):
    GCGC[i][0]=C[i]
    GCGC[i][1]=Prob_list[n][m]
    m=m+1
#Continue for each element…
for i in range(len(Y)):
    TACY[i][0]=Y[i]
    TACY[i][1]=Prob_list[n][m]
    m=m+1

DNASeq='ATG' #Save DNASeq in a DNASeq string
##couplet=DNASeq+ProteinSeq[1:2] #Look at one couplet at a time
import random #Define the pick_random function
import sys
def pick_random(prob_list):
  r, s = random.random(), 0
```

```
  for num in prob_list:
    s += num[1]
    if s >= r:
      return num[0]
  print >> sys.stderr, "Error: shouldn't get here"
###Send the couplet to the pick_random function
##correct_list= Master[couplet]
##new = pick_random(correct_list)
##DNASeq=DNASeq+new
new=DNASeq
for i in range(len(ProteinSeq)-1):
    #take the last 3 bases, add the AA in ProteinSeq
    couplet = new+ProteinSeq[i+1:i+2]
    correct_list= Master[couplet]
    new = pick_random(correct_list)
    DNASeq=DNASeq+new
DNASeq=DNASeq+'TAA'
print "DNA sequence: "
print DNASeq
```

MAP_DRAW

```
#You have to install three libraries/programs before using this script
#Install networkx, pygraphviz & graphviz.
import networkx as nx
```

```
import numpy as np
import string
import pygraphviz
#Read data from the csv file and unpack values, in same order, into A
matrix_file=raw_input("Provide the file name containing the drift data, remember to include .csv! ")
size = raw_input("Input size of matrix, must be symmetric: ")
size=int(size)
#Create a symmetric matrix in which to store the probability values, A
A=np.zeros(shape=(size,size))
import csv
#Read data from the csv file into matrix A
ifile  = open(matrix_file, "rb")
reader = csv.reader(ifile)
rownum=0
j=0
for row in reader:
   A[j]=row
   j=j+1
#Convert the strings to floats
for a in range(size):
   for b in range(size):
      A[a][b]=float(A[a][b])
dt = [('len', float)]
#Resize the edge lengths by modifying the multiplier here
```

```
A = np.array(A)*8
A = A.view(dt)
G = nx.from_numpy_matrix(A)
#Create a dictionary with names for each node.  Note, names must be entered in the same order that the data is in.
mapping = {0:'CBF1', 1:'DAL82', 2:'GCN4', 3:'GLN3', 4:'HAP4', 5:'HSF1', 6:'LEU3', 7:'MBP1', 8:'MSN4', 9:'NRG1', 10:'PHO4', 11:'RTG3', 12:'SKN7', 13:'STE12', 14:'TEC1', 15:'UPC2', 16:'control'}
G = nx.relabel_nodes(G, mapping)
G = nx.to_agraph(G)
#These style attributes can be changed to meet your needs.
G.node_attr.update(color="cyan", shape='circle', fontsize='25.0', style="filled,bold")
G.edge_attr.update(color="white", width="5.0")
#Give the output a file name in the first entry
#Output format can be changed, many options including .png, .jpeg, .bmp, .gif, .pdf
G.draw('TFmap_test.pdf', format='pdf', prog='neato')
#Maps can be made in other formats, remove commenting if desired
##G.draw('testout2.png', format='png', prog='dot')
##G.draw('testout3.png', format='png', prog='twopi')
##G.draw('testout4.png', format='png', prog='circo')
##G.draw('testout5.png', format='png', prog='fdp')
```

Appendix E: Chromatin DB systems analysis of microarray data

We sought to determine which covalent chromatin modifications, if any, were enriched or depleted under the various conditions explored in our microarray experiment. Using Chromatin DB[231], which utilizes the Bonferroni Correction, we examined those genes (1) differentially expressed between the wild-type and knockout, (2) graded by *gcn5-F221A*, (3) differentially expressed between the wild-type and knockout, but not graded (non-catalytically associated), (4) graded but not differentially expressed between the wild-type and knockout (False negatives), and (5) genes with knockout expression levels opposite the expression levels in the presence of the *gcn5-F221A* mutant (opposites). We have catalogued those statistically significant covalent modifications for which each data set is enriched and depleted with corrected p values in parentheses.

1. Differentially expressed (DE) genes between wild-type & knockout

   No significant enrichment or depletion of chromatin

   DE genes between wild-type & knockout, up-regulated in knockout

   Depletion: H4Nterm ac ($<10^{-3}$)

   DE genes between wild-type & knockout, down-regulated in knockout

   Enrichment: H3K4me$_2$ ($<10^{-4}$)

   Depletion: H3 occupancy ($<10^{-4}$), H4 occupancy ($<10^{-3}$)

   DE genes between wild-type & knockout, no grading observed

   No significant enrichment or depletion of chromatin

2. Graded genes compared to wild-type

   No significant enrichment or depletion of chromatin

   Graded up compared to wild-type

   Depletion: H3K18ac ($<10^{-3}$), H3K14ac ($<10^{-3}$)

Graded up, early

No significant enrichment or depletion of chromatin

Graded up, late

Depletion: H2AK7ac ($<10^{-4}$), H2BK11ac ($<10^{-4}$), H2BK16ac ($<10^{-3}$), H3K14ac ($<10^{-4}$), H3K18ac ($<10^{-4}$), H3K23ac ($<10^{-4}$)

Graded down compared to wild-type

Depletion: H4 occupancy ($<10^{-3}$)

Graded down, early

No significant enrichment or depletion of chromatin

Graded down, late

Depletion: H4 occupancy ($<10^{-3}$)

Graded up compared to wild-type, no change in knockout

No significant enrichment or depletion of chromatin

Graded down compared to wild-type, no change in knockout

No significant enrichment or depletion of chromatin

3. Non-catalytically associated genes (differentially expressed in knockout, no gradation)

   Depletion: H2AZ occupancy ($<10^{-3}$)

4. False negative genes

   Depletion: H2BK11ac ($<10^{-3}$), H2BK16ac ($<10^{-3}$), H3K18ac ($<10^{-4}$), H3K14ac ($<10^{-4}$), H3K23ac ($<10^{-4}$),

5. Opposite genes

   Depletion: H4K16ac ($<10^{-3}$), H2AZ occupancy ($<10^{-3}$)

Appendix F: Cytoscape BiNGO systems analysis of microarray data

From our microarray study, we sought to determine which cellular processes, if any, were related to the observed cellular phenomena. Using Cytoscape BiNGO version 2.44[266], we examined the over represented cellular processes (**ORCP**) and underrepresented cellular processes (**URCP**) for subsets of genes. We examined those genes (1) differentially expressed between the wild-type and knockout, (2) catalytically associated, (3) differentially expressed between the wild-type and knockout, but not catalytically associated, (4) catalytically associated but not differentially expressed between the wild-type and knockout (False negatives), and (5) genes with knockout expression levels opposite the expression levels in the presence of the *gcn5-F221A* mutant (opposites). Statistically significant ORCPs and URCPs are reported below, with corrected p values calculated for false discovery rates in parentheses, except in the case of the false negatives and opposites for which there were no statistically significant ORCPs or URCPs.

1. Differentially expressed (DE) genes between wild-type & knockout

ORCP: cell wall (2.10e-4), conjugation (4.83e-2), extra-cellular region (4.94e-4), plasma membrane (1.57e-4)

URCP: cytoplasm (9.03e-10), DNA metabolic process (6.93e-4), triplet codon-amino acid adaptor activity (3.89e-5), endomembrane system (2.31e-2), ribosome (8.88e-4), mitochondrial envelope (3.55e-2), response to stress (1.74e-2), chromosome segregation (3.55e-2), golgi apparatus (3.55e-2), protein complex biogenesis (3.55e-2), translation (2.34e-2), structural molecule activity (3.55e-2)

DE genes between wild-type & knockout, up-regulated in knockout

ORCP: cell wall (1.80e-3)

URCP: cytoplasm (5.29e-3), DNA metabolic process (5.29e-3), triplet codon-amino acid adaptor activity (1.42e-2), translation (3.13e-2), nucleolus (2.40e-2), nucleus (8.55e-3), ribosome biogenesis (5.29e-3), RNA binding (6.51e-4), RNA metabolic process (3.49e-4)

DE genes between wild-type & knockout, down-regulated in knockout

ORCP: nucleolus (4.39e-5), ribosome biogenesis (3.05e-4), conjugation (3.14e-4), plasma membrane (3.60e-3), extra-cellular region (2.02e-2), RNA metabolic process (4.05e-2)

URCP: cytoplasm (2.65e-7), response to stress (1.04e-2), triplet codon-amino acid adaptor activity (2.17e-2), protein binding (2.17e-2), mitochondrial envelope (2.17e-2), structural molecule activity (2.17e-2), ribosome (2.17e-2), golgi apparatus (3.70e-2), protein complex biogenesis (3.70e-2)

DE genes between wild-type & knockout, no grading observed

URCP: RNA metabolic process (3.32e-2), RNA binding (3.32e-2)

2. Graded genes compared to wild-type

ORCP: oxidoreductase activity (3.00e-2)

URCP: triplet codon-amino acid adaptor activity (2.71e-5), cytoplasm (3.27e-3), translation (1.42e-5), ribosome (8.70e-4), structural molecule activity (2.25e-2)

Graded up compared to wild-type

ORCP: Oxidoreductase activity (8.86e-3), cellular protein catabolic process (3.16e-2)

URCP: RNA binding (2.28e-2), triplet codon-amino acid adaptor activity (3.11e-3), Ribosome biogenesis (3.11e-3), RNA metabolic process (4.48e-3), translation (7.53e-5), heterocycle metabolic process (1.1e-2), ribosome (2.28e-2), nucleolus (2.28e-2)

Graded up compared to wild-type, no change in knockout

ORCP: mitochondrial envelope (1.71e-2)

URCP: translation (1.10e-2)

Graded down compared to wild-type

ORCP: Nucleolus (2.53e-7), Ribosome biogenesis (5.51e-6), RNA metabolic process (6.93e-3), plasma membrane (2.52e-2)

URCP: Cytoplasm (2.87e-4), protein binding (3.09e-2), response to stress (3.09e-2), mitochondrial envelope (4.48e-2), structural molecule activity (4.48e-2)

3. Non-catalytically associated genes (differentially expressed in knockout, no gradation)

ORCP: Conjugation (1.55e-4), plasma membrane (1.55e-4), extracellular region (3.08e-3), cell wall (6.16e-3), membrane (1.83e-2)

URCP: Cytoplasm (1.00e-5), DNA metabolic process (1.97e-2), protein modification process (1.97e-2), RNA binding (3.63e-2), nucleus (2.66e-2), triplet codon-amino acid adaptor activity (3.47e-2), response to stress (3.63e-2), ribosome (3.63e-2)

4. False negative genes

No functional enrichment/depletion of cellular processes detected.

5. Opposite genes

No functional enrichment/depletion of cellular processes detected.

Table A.5: Microarray genes identified as over represented for cellular processes

Gene Category	**ORCP**	**Yeast Gene IDs**
Differentially expressed (DE) genes between wild-type & knockout	Cell wall	YDR055W YJL171C YIR039C YJR004C YNR044W YBR067C YLR042C YKL163W YLR194C YJL052W YLR040C YIL011W YOR382W YGR189C YMR008C YMR006C YDR077W
	conjugation	YCL027W YBL016W YNR044W YJR004C YNL279W YBR083W YIL037C YHR005C YKL178C YFR008W YLR452C YJL157C YGL089C
	Extra-cellular region	YDR055W YNR044W YBR067C YJR004C YKL163W YLR042C YPL123C YLR040C YIL011W YOR382W YGR189C YHR057C YMR006C YGL089C YDR077W
	Plasma membrane	YPL265W YAR033W YDR055W YBR021W YGL053W YCL027W YNL279W YBL042C YLR194C YAR027W YMR319C YPR194C YML123C YLR214W YOL020W YOR101W YJL219W YMR008C YFL051C YLR452C YLR121C YAR031W YFL041W YBR068C YIR039C YPR124W YCL048W YIR032C YPL058C YHR005C YPR192W YOL156W YKL178C YGR121C YDR508C YLR413W YOL152W
DE genes between wild-type & knockout, up-regulated in knockout	Cell Wall	YIL011W YOR382W YJL171C YDR055W YIR039C YGR189C YKL163W YLR194C YMR008C YJL052W YDR077W
DE genes between wild-type & knockout, down-regulated in knockout	Nucleolus	YLR068W YIL127C YNL175C YGR280C YBL028C YNL124W YIL096C YDR021W YMR131C YBR247C YJL050W YJL033W YBR141C YJL109C YGL029W YGR159C YKL078W YLR145W YMR128W YAL059W YCR072C

Table A.5 (continued)

	Ribosome biogenesis	YBR267W YLR068W YGR280C YNL124W YNL112W YDR101C YDR021W YMR131C YBR247C YJL050W YHR197W YJL033W YJL109C YGL029W YLR059C YNL182C YGR159C YKL078W YLR145W YCR018C YMR128W YAL059W YCR072C
	Conjugation	YKL178C YBL016W YCL027W YNR044W YJR004C YNL279W YBR083W YLR452C YIL037C YJL157C YGL089C YHR005C
	Plasma membrane	YFL041W YBR021W YPR124W YCL027W YNL279W YBL042C YIR032C YHR005C YPR192W YMR319C YKL178C YML123C YLR214W YOL020W YOR101W YGR121C YLR452C YDR508C YLR413W YOL152W
	Extracellular region	YHR057C YBR067C YNR044W YJR004C YLR042C YMR006C YGL089C YLR040C
	RNA metabolic process	YNL141W YLR068W YJL050W YJL033W YJL109C YOL124C YNL182C YGR159C YLR145W YKL078W YMR128W YCR072C YBR267W YOL125W YOL066C YIL131C YGR280C YNL040W YNL124W YNL112W YDR021W YLR298C YMR131C YBR247C YHR197W YLR059C YGL029W YCR018C YAL059W YGR129W YDR465C
Graded genes compared to wild-type	Oxidoreductase activity	YPL171C YNL274C YER069W YKL107W YLL041C YML131W YGL055W YOR136W YJL052W YOR374W YNL037C YLR214W YIR036C YGR234W YIR038C YEL024W YJR096W YAL061W YLR460C YCL026C-B YIL155C YIL111W YMR118C YGR088W YBR026C YDL085W YCR102C YOR120W YOL152W
Graded up compared to wild-type	Oxidoreductase activity	YNL274C YPL171C YER069W YLL041C YML131W YKL107W YCL026C-B YIL155C YOR136W YJL052W YIL111W YOR374W YMR118C YGR088W YNL037C YBR026C YIR036C YDL085W YIR038C YEL024W YJR096W YAL061W YOR120W
	Cellular protein catabolic process	YOR173W YFR053C YGR161C YIL155C YDR003W YPL002C YOR185C YGL156W YFR050C YLR178C YBR214W YKR098C YER054C YGR088W YJL020C YMR174C YDR358W YGL180W YIR038C YAL061W

Table A.5 (continued)

Graded up compared to wild-type, no change in knockout	Mitochondrial envelope	YIL136W YDL142C YEL039C YJL161W YIL155C YJL066C YIL111W YER004W YMR118C YPL004C YDR178W YBR147W YDL085W YDR236C YIR038C YKL087C YEL024W
Graded down compared to wild-type	Nucleolus	YLR068W YIL127C YMR290C YNL175C YGR280C YBL028C YMR269W YDR021W YDL148C YMR131C YJL050W YJL109C YGL029W YGR159C YKL078W YLR145W YPL157W YMR128W YAL059W YGR245C
	Ribosome biogenesis	YLR068W YMR290C YGR280C YNL112W YDR101C YMR269W YDR021W YLR074C YDL148C YMR131C YJL050W YHR197W YJL109C YGL029W YGR159C YKL078W YLR145W YPL157W YMR128W YAL059W YGR245C
	RNA Metabolic process	YNL141W YLR068W YMR290C YDL201W YDL148C YJL050W YJL109C YGR159C YLR145W YKL078W YMR128W YOL066C YOL125W YIL131C YGR280C YNL112W YDR021W YMR269W YLR298C YMR131C YHR197W YGL029W YPL157W YAL059W YGR129W YDR465C
	Plasma membrane	YBR294W YBR021W YOR273C YPR124W YGL255W YMR319C YHL016C YPR194C YML123C YLR214W YOL020W YGR121C YLR413W YOL152W
Non-catalytically associated genes (differentially expressed in knockout, no gradation)	Plasma membrane	YAR033W YFL041W YBR068C YGL053W YCL027W YNL279W YBL042C YCL048W YLR194C YPL058C YIR032C YHR005C YPR192W YOL156W YKL178C YOR101W YJL219W YFL051C YLR452C YLR121C YDR508C YAR031W
	Conjugation	YKL178C YBL016W YCL027W YNR044W YJR004C YNL279W YBR083W YLR452C YIL037C YJL157C YGL089C YHR005C
	Extracellular Region	YOR382W YGR189C YHR057C YBR067C YNR044W YJR004C YLR042C YGL089C YLR040C
	Cell Wall	YOR382W YJL171C YGR189C YBR067C YNR044W YJR004C YLR042C YLR194C YLR040C

Table A.5 (continued)

	Membrane	YDR492W YGL053W YBL042C YLR040C YJL082W YJL037W YLR050C YAR035W YJL157C YBR222C YAR031W YNR065C YCL048W YIR032C YFL054C YCL021W-A YER100W YDR366C YGR189C YDL072C YDR508C YNR066C YAR033W YJL171C YPL156C YCL027W YJR004C YBR067C YNL279W YLR042C YLR194C YIL037C YPL057C YEL004W YOR382W YLR145W YOR101W YFL051C YJL219W YLR121C YLR452C YHL026C YFL041W YBR068C YER060W YNR044W YDR034C-A YPL058C YHR005C YPR192W YDR275W YOL156W YKL178C YDR218C

References

1. D. A. Jackson, R. H. Symons, and P. Berg, *Proc Natl Acad Sci U S A* **69** (10), 2904 (1972).
2. L. Baldi, D. L. Hacker, M. Adam et al., *Biotechnol Lett* **29** (5), 677 (2007).
3. F. M. Wurm and C. J. Petropoulos, *Biologicals* **22** (2), 95 (1994).
4. H. Reisinger, W. Steinfellner, B. Stern et al., *Appl Microbiol Biotechnol* **81** (4), 701 (2008).
5. A. Rita Costa, M. Elisa Rodrigues, M. Henriques et al., *Eur J Pharm Biopharm* **74** (2), 127 (2009).
6. O. Kramer, S. Klausing, and T. Noll, *Appl Microbiol Biotechnol* **88** (2), 425 (2010).
7. M. Song, K. Raphaelli, M. L. Jones et al., *Journal of Chemical Technology and Biotechnology* **86** (7), 935 (2011).
8. F. M. Wurm, *Nat Biotechnol* **22** (11), 1393 (2004).
9. Pharma R&D Annual Review 2010, Available at http://www.pharmaprojects.com, (2010).
10. J. Chusainow, Y. S. Yang, J. H. Yeo et al., *Biotechnol Bioeng* (2008).
11. G. Seth, S. Charaniya, K. F. Wlaschin et al., *Curr Opin Biotechnol* **18** (6), 557 (2007).
12. Alison J Porter, Andrew J Racher, Richard Preziosi et al., *Biotechnology Progress* 26 (5), 1455 (2010).
13. The Truly Staggering Cost Of Inventing New Drugs, Available at http://www.forbes.com/sites/matthewherper/2012/02/22/the-truly-staggering-cost-of-inventing-new-drugs-the-print-version/. (2012).
14. T. Omasa, M. Onitsuka, and W. D. Kim, *Curr Pharm Biotechnol* **11** (3), 233 (2010).
15. F. Li, N. Vijayasankaran, A. Y. Shen et al., *MAbs* **2** (5), 466 (2010).
16. A. A. Shukla and J. Thommes, *Trends Biotechnol* **28** (5), 253 (2010).
17. K. Swiech, V. Picanco-Castro, and D. T. Covas, *Protein Expr Purif* **84** (1), 147 (2012).
18. Y. Durocher and M. Butler, *Curr Opin Biotechnol* **20** (6), 700 (2009).
19. R. W. Peng, C. Guetg, M. Tigges et al., *Metab Eng* **12** (1), 18 (2010).
20. S. M. Browne and M. Al-Rubeai, *Trends Biotechnol* **25** (9), 425 (2007).

21. H. Wurtele, K. C. E. Little, and P. Chartrand, *Gene Therapy* **10** (21), 1791 (2003).

22. C. A. Gersbach, T. Gaj, R. M. Gordley et al., *Nucleic Acids Res* **39** (17), 7868 (2011).

23. S. Dietmair, L. K. Nielsen, and N. E. Timmins, *Biotechnol J* **7** (1), 75 (2011).

24. M. Wirth, J. Bode, G. Zettlmeissl et al., *Gene* **73** (2), 419 (1988).

25. F. M. Wurm, *Biologicals* **18** (3), 159 (1990).

26. V. Gurtu, G. Yan, and G. Zhang, *Biochem Biophys Res Commun* **229** (1), 295 (1996).

27. S. Rees, J. Coote, J. Stables et al., *Biotechniques* **20** (1), 102 (1996).

28. M. L. Kennard, D. L. Goosney, D. Monteith et al., *Biotechnol Bioeng* **104** (3), 540 (2009).

29. J. Y. Kim, Y. G. Kim, and G. M. Lee, *Appl Microbiol Biotechnol* **93** (3), 917 (2011).

30. A. J. Porter, A. J. Racher, R. Preziosi et al., *Biotechnol Prog* **26** (5), 1455 (2010).

31. A. Lanza, Cheng, J., Alper, H., *Current Opinion in Chemical Engineering* (2012).

32. E. Young and H. Alper, *J Biomed Biotechnol* **2010**, 130781 (2010).

33. A. S. Khalil and J. J. Collins, *Nat Rev Genet* **11** (5), 367 (2010).

34. E. Leonard, D. Nielsen, K. Solomon et al., *Trends Biotechnol* **26** (12), 674 (2008).

35. Sang Seo, Seong Kim, and Gyoo Jung, *Biotechnology and Bioprocess Engineering* **17** (1), 1 (2012).

36. M. S. Siddiqui, K. Thodey, I. Trenchard et al., *FEMS Yeast Res* **12** (2), 144 (2011).

37. A. Melnikov, A. Murugan, X. Zhang et al., *Nat Biotechnol* **30** (3), 271 (2012).

38. R. K. Koduri, J. T. Miller, and P. Thammana, *Gene* **280** (1-2), 87 (2001).

39. C. Mielke, K. Maass, M. Tummler et al., *Biochemistry* **35** (7), 2239 (1996).

40. M. G. Pallavicini, P. S. DeTeresa, C. Rosette et al., *Mol Cell Biol* **10** (1), 401 (1990).

41. R. Dekeyser, B. Claes, M. Marichal et al., *Plant Physiology* **90** (1), 217 (1989).

42. R. M. Hauptmann, V. Vasil, P. Ozias-Akins et al., *Plant Physiol* **86** (2), 602 (1988).

43. B. Miki and S. McHugh, *J Biotechnol* **107** (3), 193 (2004).

44. Ashty S. Karim, Kathleen A. Curran, and Hal S. Alper, *FEMS Yeast Research*, (2012).

45. S. A. Cheon, J. Choo, V. M. Ubiyvovk et al., *Yeast* **26** (9), 507 (2009).
46. R. Giordano-Santini and D. Dupuy, *Cell Mol Life Sci* **68** (11), 1917 (2011).
47. Simon Myers, Colin Freeman, Adam Auton et al., *Nat Genet* **40** (9), 1124 (2008).
48. R. Ikeda, C. Kokubu, K. Yusa et al., *Mol Cell Biol* **27** (5), 1665 (2007).
49. M. Esteller, *Nat Rev Genet* **8** (4), 286 (2007).
50. Christian Mielke, Karin Maass, Meike Tummler et al., *Biochemistry* **35** (7), 2239 (1996).
51. Rick S. Mitchell, Brett F. Beitzel, Astrid R. W. Schroder et al., *PLoS Biol* **2** (8), e234 (2004).
52. Astrid R. W. Schröder, Paul Shinn, Huaming Chen et al., **110** (4), 521 (2002).
53. F. M. Rosin, N. Watanabe, J. L. Cacas et al., *Plant J* **55** (3), 514 (2008).
54. F. Ozsolak, J. S. Song, X. S. Liu et al., *Nat Biotechnol* **25** (2), 244 (2007).
55. L. Feuk, A. R. Carson, and S. W. Scherer, *Nat Rev Genet* **7** (2), 85 (2006).
56. David Derse, Bruce Crise, Yuan Li et al., *J. Virol.* **81** (12), 6731 (2007).
57. Mary K. Lewinski, Masahiro Yamashita, Michael Emerman et al., *PLoS Pathog* **2** (6), e60 (2006).
58. Y. Moalic, H. Felix, Y. Takeuchi et al., *J. Virol.* **83** (4), 1920 (2009).
59. Xiaolin Wu, Yuan Li, Bruce Crise et al., *Science* **300** (5626), 1749 (2003).
60. Sanggu Kim, Yein Kim, Teresa Liang et al., *J. Virol.* **80** (22), 11313 (2006).
61. E. J. Mead, L. M. Chiverton, C. M. Smales et al., *Biotechnol Bioeng* (2008).
62. G. Dellaire and P. Chartrand, *Radiation Research* **149** (4), 325 (1998).
63. Z. Liang, A. M. Breman, B. R. Grimes et al., *Transgenic Res* **17** (5), 979 (2008).
64. H. J. Gierman, M. H. Indemans, J. Koster et al., *Genome Res* **17** (9), 1286 (2007).
65. M. Sadelain, E. P. Papapetrou, and F. D. Bushman, *Nat Rev Cancer* **12** (1), 51 (2011).
66. Justin Eyquem, Laurent Poirot, Roman Galetto et al., *Biotechnology and Bioengineering*, (2013).
67. M. V. Francia and J. M. Garcia Lobo, *J Bacteriol* **178** (3), 894 (1996).
68. D. B. Flagfeldt, V. Siewers, L. Huang et al., *Yeast* **26** (10), 545 (2009).
69. W. Y. Liu, Y. Wang, Y. Qin et al., *Mar Biotechnol (NY)* **9** (4), 420 (2007).
70. K. Smith, *Reproduction Nutrition Development* **41** (6), 465 (2001).
71. G. Silva, L. Poirot, R. Galetto et al., *Curr Gene Ther* **11** (1), 11 (2010).

72. A. Nern, B. D. Pfeiffer, K. Svoboda et al., *Proceedings of the National Academy of Sciences of the United States of America* **108** (34), 14198 (2011).

73. F. Xie, Q. Ma, S. Jiang et al., *DNA Cell Biol* (2012).

74. C. Patsch, D. Kesseler, and F. Edenhofer, *Methods* **53** (4), 386 (2011).

75. Y. Takata, S. Kondo, N. Goda et al., *Genes to Cells* **16** (7), 765 (2011).

76. S. Yamaguchi, Y. Kazuki, Y. Nakayama et al., *PLoS ONE* **6** (2) (2011).

77. B. Sauer, *Endocrine* **19** (3), 221 (2002).

78. A. Keravala, S. Lee, B. Thyagarajan et al., *Mol Ther* **17** (1), 112 (2009).

79. T. W. Chalberg, J. L. Portlock, E. C. Olivares et al., *J Mol Biol* **357** (1), 28 (2006).

80. A. L. Garcia-Otin and F. Guillou, *Front Biosci* **11**, 1108 (2006).

81. S. Grizot, J. Smith, F. Daboussi et al., *Nucleic Acids Res* **37** (16), 5405 (2009).

82. F. D. Urnov, E. J. Rebar, M. C. Holmes et al., *Nature Reviews Genetics* **11** (9), 636 (2010).

83. Pei-Qi Liu, Edmond M. Chan, Gregory J. Cost et al., *Biotechnology and Bioengineering* **106** (1), 97 (2010).

84. L. Malphettes, Y. Freyvert, J. Chang et al., *Biotechnol Bioeng* **106** (5), 774 (2010).

85. C. Mussolino and T. Cathomen, *Curr Opin Biotechnol* (2012).

86. A. W. Briggs, X. Rios, R. Chari et al., *Nucleic Acids Res* **40** (15), e117 (2012).

87. J. C. Miller, S. Tan, G. Qiao et al., *Nat Biotechnol* **29** (2), 143 (2010).

88. D. Hockemeyer, H. Wang, S. Kiani et al., *Nat Biotechnol* **29** (8), 731 (2011).

89. P. Mali, L. Yang, K. M. Esvelt et al., *Science* (2013).

90. L. Cong, F. A. Ran, D. Cox et al., *Science* **339** (6121), 819 (2013).

91. J. J. Harrington, G. Van Bokkelen, R. W. Mays et al., *Nat Genet* **15** (4), 345 (1997).

92. Z. Larin and J. E. Mejia, *Trends Genet* **18** (6), 313 (2002).

93. Y. Shibata, P. Kumar, R. Layer et al., *Science* **336** (6077), 82 (2012).

94. D. M. Camacho and J. J. Collins, *Cell* **137** (1), 24 (2009).

95. N. Ishii, K. Nakahigashi, T. Baba et al., *Science* **316** (5824), 593 (2007).

96. K. S. Lau, A. M. Juchheim, K. R. Cavaliere et al., *Sci Signal* **4** (165), (2011).

97. H. Goodarzi, B. D. Bennett, S. Amini et al., *Mol Syst Biol* **6**, 378 (2010).

98. M. A. Oberhardt, B. O. Palsson, and J. A. Papin, *Mol. Syst. Biol.* **5** (2009).

99. C. S. Henry, M. DeJongh, A. A. Best et al., *Nat. Biotechnol.* **28** (9), 977 (2010).

100. F. H. Karlsson, I. Nookaew, D. Petranovic et al., *Trends Biotechnol* **29** (6), 251 (2011).

101. K. Saito and F. Matsuda, *Annu Rev Plant Biol* **61**, 463 (2009).

102. M. F. Ciaccio, J. P. Wagner, C. P. Chuu et al., *Nat Methods* **7** (2), 148 (2010).

103. E. C. O'Shaughnessy, S. Palani, J. J. Collins et al., *Cell* **144** (1), 119 (2011).

104. K. Faust, D. Croes, and J. van Helden, *Biosystems* **105** (2), 109 (2011).

105. B. Liu and M. Pop, *BMC Proc* **5 Suppl 2**, S9 (2011).

106. J. Beal, T. Lu, and R. Weiss, *PLoS ONE* **6** (8), e22490 (2011).

107. B. Schwanhausser, D. Busse, N. Li et al., *Nature* **473** (7347), 337 (2011).

108. L. Z. Hong, J. Li, A. Schmidt-Kuntzel et al., *Genome Res* **21** (11), 1905 (2011).

109. R. Gupta, A. Bhattacharyya, F. J. Agosto-Perez et al., *Nucleic Acids Research* **39**, D92 (2011).

110. E. Portales-Casamar, S. Thongjuea, A. T. Kwon et al., *Nucleic Acids Research* **38**, D105 (2010).

111. R. Amit, H. G. Garcia, R. Phillips et al., *Cell* **146** (1), 105 (2011).

112. L. Hood, J. R. Heath, M. E. Phelps et al., *Science* **306** (5696), 640 (2004).

113. H. Y. Chuang, M. Hofree, and T. Ideker, *Annu Rev Cell Dev Biol* **26**, 721 (2010).

114. N. Gehlenborg, S. I. O'Donoghue, N. S. Baliga et al., *Nat Methods* **7** (3 Suppl), S56 (2010).

115. R. Hershberg and D. A. Petrov, *PLoS Genet* **5** (7), e1000556 (2009).

116. C. Gustafsson, S. Govindarajan, and J. Minshull, *Trends Biotechnol* **22** (7), 346 (2004).

117. J. B. Plotkin and G. Kudla, *Nat Rev Genet* **12** (1), 32 (2011).

118. J. M. Fox and I. Erill, *DNA Res* **17** (3), 185 (2010).

119. T. Tuller, Y. Y. Waldman, M. Kupiec et al., *Proc Natl Acad Sci U S A* **107** (8), 3645 (2010).

120. B. Wiedemann and E. Boles, *Appl Environ Microbiol* **74** (7), 2043 (2008).

121. L. Kotula and P. J. Curtis, *Biotechnology (N Y)* **9** (12), 1386 (1991).

122. Y. Nakamura, T. Gojobori, and T. Ikemura, *Nucleic Acids Res* **28** (1), 292 (2000).

123. R. Jansen, H. J. Bussemaker, and M. Gerstein, *Nucleic Acids Res* **31** (8), 2242 (2003).

124. P. M. Sharp and W. H. Li, *Nucleic Acids Res* **15** (3), 1281 (1987).

125. F. Wright, *Gene* **87** (1), 23 (1990).

126. P. Puigbo, E. Guzman, A. Romeu et al., *Nucleic Acids Res* **35** (Web Server issue), W126 (2007).

127. P. Puigbo, I. G. Bravo, and S. Garcia-Vallve, *BMC Bioinformatics* **9**, 65 (2008).

128. T. T. Wang, W. C. Cheng, and B. H. Lee, *Mol Biotechnol* **10** (2), 103 (1998).

129. M. F. Alexeyev and H. H. Winkler, *Biochim Biophys Acta* **1419** (2), 299 (1999).

130. Expression Optimization Service Survey Results, Available at http://www.blueheronbio.com/assets/documents/BlueHeronBioExpressionSurvey.pdf.

131. K. A. Curran, J. M. Leavitt, A. S. Karim et al., *Metab Eng* **15**, 55 (2013).

132. D. Agashe, N. C. Martinez-Gomez, D. A. Drummond et al., *Mol Biol Evol* (2012).

133. A. C. Forster, *Biotechnol J* **7** (7), 835 (2012).

134. H. Dong, L. Nilsson, and C. G. Kurland, *J Mol Biol* **260** (5), 649 (1996).

135. M. Frenkel-Morgenstern, T. Danon, T. Christian et al., *Mol Syst Biol* **8**, 572 (2012).

136. P. M. Sharp and W. H. Li, *J Mol Evol* **24** (1-2), 28 (1986).

137. M. Hirst and M. A. Marra, *Int J Biochem Cell Biol* **41** (1), 136 (2009).

138. P. M. Das, K. Ramachandran, J. vanWert et al., *Biotechniques* **37** (6), 961 (2004).

139. P. J. Park, *Nat Rev Genet* **10** (10), 669 (2009).

140. C. T. Chien, P. L. Bartel, R. Sternglanz et al., *Proc Natl Acad Sci U S A* **88** (21), 9578 (1991).

141. S. Ben-Aroya, C. Coombes, T. Kwok et al., *Mol Cell* **30** (2), 248 (2008).

142. S. Mnaimneh, A. P. Davierwala, J. Haynes et al., *Cell* **118** (1), 31 (2004).

143. R. Sopko, D. Huang, N. Preston et al., *Mol Cell* **21** (3), 319 (2006).

144. M. Schuldiner, S. R. Collins, N. J. Thompson et al., *Cell* **123** (3), 507 (2005).

145. D. K. Breslow, D. M. Cameron, S. R. Collins et al., *Nat Methods* **5** (8), 711 (2008).

146. Z. Li, F. J. Vizeacoumar, S. Bahr et al., *Nat Biotechnol* **29** (4), 361.

147. L. Sigalotti, E. Fratta, S. Coral et al., *J Cell Physiol* **212** (2), 330 (2007).

148. H. M. Byun, S. H. Choi, P. W. Laird et al., *Cancer Lett* **266** (2), 238 (2008).

149. N. Detich, V. Bovenzi, and M. Szyf, *J Biol Chem* **278** (30), 27586 (2003).

150. S. Y. Roth, J. M. Denu, and C. D. Allis, *Annu Rev Biochem* **70**, 81 (2001).

151. R. W. Johnstone, *Nat Rev Drug Discov* **1** (4), 287 (2002).

152. A. B. Bleecker, M. A. Estelle, C. Somerville et al., *Science* **241** (4869), 1086 (1988).

153. H. Alper, J. Moxley, E. Nevoigt et al., *Science* **314** (5805), 1565 (2006).

154. A. K. Wilson, F. B. Pickett, J. C. Turner et al., *Mol Gen Genet* **222** (2-3), 377 (1990).

155. A. R. Curtis, C. Fey, C. M. Morris et al., *Nat Genet* **28** (4), 350 (2001).

156. O. K. Steinlein, J. C. Mulley, P. Propping et al., *Nat Genet* **11** (2), 201 (1995).

157. P. M. Nolan, J. Peters, M. Strivens et al., *Nat Genet* **25** (4), 440 (2000).

158. P. Reddy and S. Hahn, *Cell* **65** (2), 349 (1991).

159. P. Mulsant, A. Gatignol, M. Dalens et al., *Somat Cell Mol Genet* **14** (3), 243 (1988).

160. R. J. Kaufman, *Mol Biotechnol* **16** (2), 151 (2000).

161. M. Oliva-Trastoy, M. Defais, and F. Larminat, *Mutagenesis* **20** (2), 111 (2005).

162. H. U. Bernard, G. Krammer, and W. G. Rowekamp, *Exp Cell Res* **158** (1), 237 (1985).

163. A. Valera, J. C. Perales, M. Hatzoglou et al., *Hum Gene Ther* **5** (4), 449 (1994).

164. M. E. Azzam and I. D. Algranati, *Proc Natl Acad Sci U S A* **70** (12), 3866 (1973).

165. S. Pestka, *Annual Review of Microbiology* **25**, 487 (1971).

166. J. A. Vara, A. Portela, J. Ortin et al., *Nucleic Acids Res* **14** (11), 4617 (1986).

167. Amanda M. Lanza, Joseph K. Cheng, and Hal S. Alper, *Current Opinion in Chemical Engineering* **1** (4), 403 (2012).

168. S. P. Lees-Miller and K. Meek, *Biochimie* **85** (11), 1161 (2003).

169. R. V. Merrihew, K. Marburger, S. L. Pennington et al., *Mol Cell Biol* **16** (1), 10 (1996).

170. K. Sakurai, M. Shimoji, C. G. Tahimic et al., *Nucleic Acids Res* **38** (7), e96 (2010).

171. C. J. Ramachandra, M. Shahbazi, T. W. Kwang et al., *Nucleic Acids Res* **39** (16), e107 (2011).

172. Amanda M. Lanza, Do Soon Kim, and Hal S. Alper, *Biotechnology Journal*, (2013).

173. Amanda M. Lanza, Timothy J. Dyess, and Hal S. Alper, *Biotechnology Journal*, (2012).

174. H. Wurtele, K. C. Little, and P. Chartrand, *Gene Ther* **10** (21), 1791 (2003).

175. B. Sauer and N. Henderson, *Proc Natl Acad Sci U S A* **85** (14), 5166 (1988).

176. D. A. Sorrell and A. F. Kolb, *Biotechnol Adv* **23** (7-8), 431 (2005).

177. G. D. Van Duyne, *Proc Natl Acad Sci U S A* **106** (1), 4 (2009).

178. M. Nakano, K. Odaka, M. Ishimura et al., *Nucleic Acids Res* **29** (7), E40 (2001).

179. E. Sangiorgi, Z. Shuhua, and M. R. Capecchi, *Nucleic Acids Res* **36** (20), e134 (2008).

180. E. E. Schmidt, D. S. Taylor, J. R. Prigge et al., *Proc Natl Acad Sci U S A* **97** (25), 13702 (2000).

181. Y. Q. Feng, J. Seibler, R. Alami et al., *J Mol Biol* **292** (4), 779 (1999).

182. R. W. Siegel, R. Jain, and A. Bradbury, *FEBS Lett* **505** (3), 467 (2001).

183. F. Schnutgen, N. Doerflinger, C. Calleja et al., *Nat Biotechnol* **21** (5), 562 (2003).

184. B. Bethke and B. Sauer, *Nucleic Acids Res* **25** (14), 2828 (1997).

185. K. Araki, M. Araki, and K. Yamamura, *Nucleic Acids Res* **30** (19), e103 (2002).

186. D. Esposito and J. J. Scocca, *Nucleic Acids Res* **25** (18), 3605 (1997).

187. M. Osterwalder, A. Galli, B. Rosen et al., *Nat Methods* **7** (11), 893 (2010).

188. D. R. Shimshek, J. Kim, M. R. Hubner et al., *Genesis* **32** (1), 19 (2002).

189. A. Nagy, L. Mar, and G. Watts, *Methods Mol Biol* **530**, 365 (2009).

190. R. Weng, Y. W. Chen, N. Bushati et al., *Genetics* **183** (1), 399 (2009).

191. L. Chen, L. Li, D. Pang et al., *Animal* **4** (5), 767 (2010).

192. D. A. Sorrell, C. J. Robinson, J. A. Smith et al., *Nucleic Acids Res* **38** (11), e123 (2010).

193. Y. Koresawa, S. Miyagawa, M. Ikawa et al., *Journal of Biochemistry* **127** (3), 367 (2000).

194. F. Buchholz and A. F. Stewart, *Nat Biotechnol* **19** (11), 1047 (2001).

195. E. Suzuki and M. Nakayama, *Nucleic Acids Res* **39** (8), e49 (2011).

196. P. I. Missirlis, D. E. Smailus, and R. A. Holt, *BMC Genomics* **7**, 73 (2006).

197. Y. Kameyama, Y. Kawabe, A. Ito et al., *Biotechnol Bioeng* **105** (6), 1106 (2009).

198. E. T. Wong, J. L. Kolman, Y. C. Li et al., *Nucleic Acids Res* **33** (17), e147 (2005).

199. S. Huang, Y. Kawabe, A. Ito et al., *Biotechnol Bioeng* (2010).

200. E. Will, H. Klump, N. Heffner et al., *Nucleic Acids Res* **30** (12), e59 (2002).

201. H. Gingold and Y. Pilpel, *Mol Syst Biol* **7**, 481 (2011).

202. R. Hershberg and D. A. Petrov, *Annu Rev Genet* **42**, 287 (2008).

203. M. Yarus and L. S. Folley, *J Mol Biol* **182** (4), 529 (1985).

204. B. K. Chung and D. Y. Lee, *BMC Syst Biol* **6**, 134 (2012).

205. A. Tats, T. Tenson, and M. Remm, *BMC Genomics* **9**, 463 (2008).

206. J. R. Coleman, D. Papamichail, S. Skiena et al., *Science* **320** (5884), 1784 (2008).

207. F. C. Holstege, E. G. Jennings, J. J. Wyrick et al., *Cell* **95** (5), 717 (1998).

208. J. O. Westholm, N. Nordberg, E. Muren et al., *BMC Genomics* **9**, 601 (2008).

209. P. Owrutsky and N. Khaneja, presented at the American Control Conference (ACC), 2012.

210. T. R. Hebbes, A. W. Thorne, and C. Crane-Robinson, *EMBO J* **7** (5), 1395 (1988).

211. T. I. Lee and R. A. Young, *Annu Rev Genet* **34**, 77 (2000).

212. M. H. Kuo and C. D. Allis, *Bioessays* **20** (8), 615 (1998).

213. F. Robert, D. K. Pokholok, N. M. Hannett et al., *Mol Cell* **16** (2), 199 (2004).

214. M. L. Greenberg, P. L. Myers, R. C. Skvirsky et al., *Mol Cell Biol* **6** (5), 1820 (1986).

215. R. C. Trievel, J. R. Rojas, D. E. Sterner et al., *Proc Natl Acad Sci U S A* **96** (16), 8931 (1999).

216. R. M. Imoberdorf, I. Topalidou, and M. Strubin, *Mol Cell Biol* **26** (5), 1610 (2006).

217. F. Elefant and K. B. Palter, *Mol Biol Cell* **10** (7), 2101 (1999).

218. H. Alper, J. Moxley, E. Nevoigt et al., *Science* **314** (5805), 1565 (2006).

219. G. Jansen, E. Leberer, D. Y. Thomas et al., *Methods Enzymol* **344**, 82 (2002).

220. L. Wang, L. Liu, and S. L. Berger, *Genes Dev* **12** (5), 640 (1998).

221. E. Krissinel and K. Henrick, *J Mol Biol* **372** (3), 774 (2007).

222. H. Alper, C. Fischer, E. Nevoigt et al., *Proc Natl Acad Sci U S A* **102** (36), 12678 (2005).

223. E. Nevoigt, J. Kohnke, C. R. Fischer et al., *Appl Environ Microbiol* **72** (8), 5266 (2006).

224. T. Georgakopoulos and G. Thireos, *EMBO J* **11** (11), 4145 (1992).

225. S. Barbaric, H. Reinke, and W. Horz, *Mol Cell Biol* **23** (10), 3468 (2003).

226. P. D. Gregory, A. Schmid, M. Zavari et al., *Mol Cell* **1** (4), 495 (1998).

227. R. Haguenauer-Tsapis and A. Hinnen, *Mol Cell Biol* **4** (12), 2668 (1984).

228. W. Peng, C. Togawa, K. Zhang et al., *Genetics* **179** (1), 277 (2008).

229. A. M. Lanza, J. J. Blazeck, N. C. Crook et al., *PLoS ONE* **7** (4), e36193 (2012).

230. K. L. Huisinga and B. F. Pugh, *Mol Cell* **13** (4), 573 (2004).

231. T. R. O'Connor and J. J. Wyrick, *Bioinformatics* **23** (14), 1828 (2007).

232. W. Dang, K. K. Steffen, R. Perry et al., *Nature* **459** (7248), 802 (2009).

233. S. R. Bhaumik, E. Smith, and A. Shilatifard, *Nature structural & molecular biology* **14** (11), 1008 (2007).

234. J. Feser, D. Truong, C. Das et al., *Mol Cell* **39** (5), 724.

235. A. M. Dudley, D. M. Janse, A. Tanay et al., *Mol Syst Biol* **1**, 2005 0001 (2005).

236. M. Gustavsson and H. Ronne, *Rna-a Publication of the Rna Society* **14** (4), 666 (2008).

237. Y. Xue-Franzen, A. Johnsson, D. Brodin et al., *BMC Genomics* **11**, (2010).

238. J. Botet, L. Mateos, J. L. Revuelta et al., *Eukaryotic Cell* **6** (11), 2102 (2007).

239. A. M. Lanza, N. C. Crook, and H. S. Alper, *Curr Opin Biotechnol* (2012).

240. J. Tornoe, P. Kusk, T. E. Johansen et al., *Gene* **297** (1-2), 21 (2002).

241. J. Blazeck, L. Liu, H. Redden et al., *Appl Environ Microbiol* **77** (22), 7905 (2012).

242. Ning Sun and Huimin Zhao, *Biotechnology and Bioengineering*, (2013).

243. S. C. Barranco, K. Shilkun, S. Nichols et al., *In Vitro* **17** (8), 730 (1981).

244. M. Derouazi, D. Martinet, N. Besuchet Schmutz et al., *Biochem Biophys Res Commun* **340** (4), 1069 (2006).

245. B. Barneda-Zahonero and M. Parra, *Mol Oncol* (2012).

246. K. I. Ansari, S. Kasiri, and S. S. Mandal, *Oncogene* (2012).

247. C. J. Kenyon, *Nature* **464** (7288), 504 (2010).

248. C. Sebastian, K. F. Satterstrom, M. C. Haigis et al., *J Biol Chem* (2012).

249. S. Spiegel, S. Milstien, and S. Grant, *Oncogene* **31** (5), 537 (2012).

250. S. Rasheed, W. A. Nelson-Rees, E. M. Toth et al., *Cancer* **33** (4), 1027 (1974).
251. F. L. Graham, J. Smiley, W. C. Russell et al., *J Gen Virol* **36** (1), 59 (1977).
252. A. L. Goldstein and J. H. McCusker, *Yeast* **15** (14), 1541 (1999).
253. Y. G. Liu and N. Huang, *Plant Molecular Biology Reporter* **16** (2), 175 (1998).
254. Y. G. Liu and R. F. Whittier, *Genomics* **25** (3), 674 (1995).
255. D. Rahmutula, T. Nakayama, M. Soma et al., *Endocrine* **17** (2), 85 (2002).
256. E. C. Bryda, M. Pearson, Y. Agca et al., *Biotechniques* **41** (6), 715 (2006).
257. H. Ochman, A. S. Gerber, and D. L. Hartl, *Genetics* **120** (3), 621 (1988).
258. C. Lee, J. Kim, S. G. Shin et al., *J Biotechnol* **123** (3), 273 (2006).
259. R. A. Irizarry, B. M. Bolstad, F. Collin et al., *Nucleic Acids Res* **31** (4), e15 (2003).
260. R. A. Irizarry, B. Hobbs, F. Collin et al., *Biostatistics* **4** (2), 249 (2003).
261. B. M. Bolstad, R. A. Irizarry, M. Astrand et al., *Bioinformatics* **19** (2), 185 (2003).
262. D. Mumberg, R. Muller, and M. Funk, *Gene* **156** (1), 119 (1995).
263. R. D. Gietz and R. H. Schiestl, *Nature Protocols* **2** (1), 31 (2007).
264. M. A. Teste, M. Duquenne, J. M. Francois et al., *BMC Mol Biol* **10**, 99 (2009).
265. R. D. Gietz and R. A. Woods, *Methods Enzymol* **350**, 87 (2002).
266. S. Maere, K. Heymans, and M. Kuiper, *Bioinformatics* **21** (16), 3448 (2005).