ADVANCES IN CRISPR-CAS9 GENOME ENGINEERING TECHNOLOGIES IN SACCHAROMYCES CEREVISIAE AND TESTS FOR CIS-REGULATORY EVOLUTION

A DISSERTATION SUBMITTED TO
THE DEPARTMENT OF GENETICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Justin Daniel Smith

January 2016
© 2016 by Justin Daniel Smith. All Rights Reserved.
Re-distributed by Stanford University under license with the author.

This work is licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License. 
http://creativecommons.org/licenses/by-nc/3.0/us/

This dissertation is online at: http://purl.stanford.edu/dd819yf1057

Includes supplemental files:
1. Supplemental files for Chapter 2 (Chapter 2 Supplemental.zip)
2. Supplemental files for Chapter 3 (Chapter 3 Supplemental.zip)
3. Supplemental files for Chapter 4 (Chapter 4 Supplemental.zip)
4. Supplemental files for Chapter 5 (Chapter 5 Supplemental.zip)
I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Ronald Davis, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Andrew Fire

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Gavin Sherlock

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Lars Steinmetz

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.
Abstract
This thesis is organized into two primary areas of research. I will first discuss work investigating cis-regulatory evolution. I will then discuss work on advancing genome engineering and parsing oligonucleotide libraries.

It is now known that noncoding regions comprise the vast majority of genomic regions under selective constraint in mammals. There is evidence to suggest that one such class of these noncoding regions, cis-regulatory regions (promoters and enhancers), are likely the major source of evolutionary adaptation. Despite this, the ability to detect natural selection on noncoding regions has lagged behind the ability to detect selection on coding regions. In Chapter 1, we introduce a new test to detect selection on cis-regulatory elements, and demonstrate its utility on three mammalian transcriptional enhancers. In Chapter 2, we investigate the evolution of resistance to the mycotoxin citrinin by comparing two closely related species of budding yeast, Saccharomyces paradoxus and Saccharomyces cerevisiae. Applying a genome-wide test for selection on cis-regulation, we identified five genes involved in resistance in S. paradoxus, four of which are necessary for resistance and increase resistance in S cerevisiae when over-expressed. This work demonstrates how multiple cis-regulatory changes can be combined to produce complex adaptation.

In the second half of this dissertation, I discuss my work on CRISPR-Cas9. Over the past 3 years, CRISPR has dramatically improved the ease of genome engineering. In Chapter 2, we build an improved CRISPR activator (CRISPRa) to simultaneously induce overexpression (in S cerevisiae) of the four genes we identified to confirm their role in citrinin resistance. In Chapter 3, we build an inducible CRISPR interference (CRISPRi) system and then create a library of guide RNAs (gRNAs) to 20 different drug responsive genes to determine design rules for the creation of effective gRNAs for CRISPRi in S cerevisiae. We determined that gRNAs targeted to a region with low nucleosome occupancy and high chromatin accessibility within a window of 0 to 200bp upstream of the transcription start site (TSS) are more likely to be effective. In Chapter 4 we confirm and refine these rules in a much larger
library consisting of ~9000 unique strains. Additionally, we present a novel method for parsing complex oligonucleotide libraries into single, sequence verified DNA sequences using high throughput sequencing and yeast synthetic biology. We utilize this technology to create and characterize a collection of individual inducible CRISPRi strains to the vast majority of essential and aerobic essential genes in *S. cerevisiae*. In Chapter 5, we compare mismatch tolerance for Cas9 *in vitro* and *in vivo* (*S. cerevisiae*) and test a variety of truncated and full-length gRNAs (with 17, 18, and 20 nucleotides of complementarity sequence). We observed notable differences between *in vitro* and *in vivo* Cas9 cleavage specificity profiles, with *in vivo* cleavage being more sensitive/less tolerant to mismatches. Both length of truncated gRNAs showed an increase in specificity *in vivo* but, not *in vitro* with most targets. Interestingly, in Chapter 3 we found catalytically dead Cas9 (dCas9) showed no reduction in mismatch tolerance with truncated gRNAs, suggesting this improvement in specificity may have to do with Cas9's ability to cleave DNA and not its ability to bind its target sequence.
Acknowledgements

I would like to first thank by advisor Ron Davis, for giving me the intellectual freedom to pursue my interests and for how much he cares about his students.

I would also like to thank my committee, Lars Steinmetz, Hunter Fraser, Gavin Sherlock and Andy Fire, all of whom have mentored me and been key to my success in graduate school.

I also want to thank the members of the Stanford Genome Technology Center, who have enriched both my science and my experience. I especially want to thank Maureen Hillenmeyer, who first invited me to rotate in the group. Additionally Sundari Suresh, Uli Schlecht, and Leopold Parts - I appreciate their many contributions to my projects and my development as a scientist. I also want to give a special thanks to Bob St. Onge. Bob mentored and advised me throughout my time at the Genome Center, and has also been instrumental to my development as a scientist.

Additionally, I want to thank my classmates. I am lucky to have been a student with so many bright and intelligent people. I want to especially thank to Becky Fu, who worked extra hard on a short time scale to help me finish Chapter 5 of this thesis, and who has been a great collaborator during my PhD.

I also want to thank my friends in the Stanford Kayak Club, who provided me an escape on weekends, and many great adventures and conversations.

I want to give a special thanks to Carlos Barbas. If it were not for Carlos’s willingness to have me as an undergraduate in his lab, I may never have become interested in genome engineering. I also want to thank Russell Gordley, who spent an extensive amount of time mentoring during my time at Scripps.

I want to thank Eva, who has shared so many adventures with me during graduate school, and taught me the value of being ridiculously silly, and helped me get through the tough parts of graduate school.

And last, I want to thank my sister Robyn, and my parents, Ron and Diane, who raised me to have an inquisitive mind. I appreciate that my family rarely pressured me, but rather engage me by sharing a love of nature and science, and for putting up with my many terrariums, aquariums, ponds, pets, and gardens, all of which helped cultivate my interests in science.
Table of Contents

Abstract iv
Acknowledgements vi
Introduction 1
  Cis-Regulatory Evolution 2
  Write: Oligonucleotide Library Synthesis 3
  Manipulate: Genome Editing and Genome Engineering Technologies 4
  Outline of The Thesis 7

Chapter 1: A novel test for selection on cis-regulatory elements reveals positive and negative selection acting on mammalian transcriptional enhancers 8
  Authors 8
  Abstract 8
  Introduction 9
  Results 12
  Discussion 22
  Subsequent Research 25
  Materials and Methods 25
    Obtaining present-day enhancer sequences 25
    Ancestral reconstructions 26
    Calculating $K_u/K_n$ and $K_d/K_n$ 27
  Acknowledgements 28

Chapter 2: Dissecting the genetic basis of a complex cis-regulatory adaptation 30
  Authors 30
  Abstract 30
  Author Summary 31
  Introduction 31
  Results 33
    A sign test reveals selection on the cis-regulation of citrinin-induced genes 33
    Sp has recently evolved resistance to citrinin 35
    Candidate genes revealed by RNA-seq in hybrids 37
    Characterization of the candidate genes 39
    Individual promoters contribute to citrinin resistance 42
    Discussion 44
  Methods 49
    Sign test for selection on cis-regulation 49
    Fitness assays 49
    RNA-seq and ASE analysis 50
    Gene knockouts 51
    Over-expression with CRISPR/Cas9 51
    Strain barcoding 53
    Competitive growth assays 53
    Promoter replacement strain construction 54
    Quantitative PCR 55
  Acknowledgements 55

Chapter 3: Quantitative CRISPR interference screens in yeast identify chemical-genetic interactions and new rules for guide RNA design 56

vii
Chapter 5: Distinct patterns of Cas9 mismatch tolerance in vitro and in vivo and effects of truncated guide RNAs

Authors 113
Abstract 113
Introduction 114
Results and Discussion 115
  Effects of single base variants on Cas9 in vivo and in vitro cleavage specificity 119
  Effects of adjacent double and deletions variants on Cas9 in vivo and in vitro cleavage specificity 123
Conclusions 133
Methods 134
  Plasmid and Strain construction 134
  Cas9 in vitro cleavage specificity assay 136
  Cas9 in vivo cleavage specificity assays 136
  Formation of gRNA 5′ ends 137
  Cas9 in vitro and in vivo sequence retention calculation 137

Appendices 139
  Chapter 2 139
  Chapter 3 141
  Chapter 4 143
  Chapter 5 144

Bibliography 186
List of Tables

Table 1.1 Ku/Kn and Kd/Kn at varied p-value cutoff at which we classified SNVs as silent or functional. 19
Table 1.2 Overall phylogeny wide p-values and Ku/Kn, Kd/Kn ratios for the three enhancers. 20
Supplemental Table 1.3. Outgroup species used to aid in ancestral reconstruction of each enhancer. 27
Table 4.1 Library counts in different conditions. 96
Table 4.2. gRNAs targeting more than one ORF. 99
Table 4.3 Overlap with DNase I Hypersensitivity Sites (DHSs). 101
Table 5.1: Table of all gRNAs used in experiments and their corresponding targets and sequences. 116
Table S2.1. Read counts from RNA-seq in the Sc/Sp hybrid. 140
Table S2.2. All primers/oligos used in this work. 140
Additional file 3.1. Raw growth data measurements for 3. 1C. 141
Additional file 3.2. Raw data for qPCR 141
Additional file 3.3. The small molecule inhibitor/gene combinations constituting the reference chemical-genetic interactions used for this study. 141
Additional file 3.4. Counts and analysis for 8 replicates in no drug (DMSO 1%) 141
Additional file 3.5. Counts and analysis of three replicates in 20µM fluconazole. 141
Additional file 3.6. Lists the gRNAs that were synthesized, their target location, and aimed strand (template or non-template) 141
Additional file 3.7. List of sequenced samples, their respective barcodes, MiSeq run information, and other metadata. 141
Additional file 3.8. Sequence read information and raw counts from MiSeq data. 141
Additional file 3.9. ATc fitness effects (A), drug effects (D), and no-drug ATc effects (A0) for each gRNA in every tested condition. 141
Additional file 3.10. Drug-specific effects (D) for all drugs tested for each gRNA. 141
Additional file 3.11. Average drug-specific effects for each guide set (gRNAs directed against the same gene). 141
Additional file 3.12. Raw growth data for Figure 3.3. 141
Additional file 3.13. GCMS measurements of ergosterol intermediates in yeast treated with 1181-0519 141
Additional file 3.14. Nucleosome occupancy scores of gRNA regions based on tiling array analysis 141
Additional file 3.15. A range of additional potential determinants of guide effect. 141
Additional file 3.16. Number and significance of transcription factor site overlaps with positive control gRNA target sequences. 141
Additional file 3.17. Primers, Strains, and Plasmids used in this study. 141
Additional file 3.18. Analysis of potential off-target binding sites. 141
Table S4.1 Custom Array synthesized oligos used to create CRISPRi collection 143
Table S4.2 Sequences of all gRNAs contained in the CRISPRi collection. 143
Table S4.3 CRISPRi YPD growth assay and annotation data 143
Table 4.4 CRISPRi YPEG collection data 143
Table 4.5 Primers, strains and plasmids. 143
Table S5.1 Primers used to build vectors in this study 144
Table S5.2: List of Cas9 cleavage experiments and corresponding sequence datasets. 182
Table S5.3 All target oligos synthesized for the Type II library. 185
List of Figures

Chapter 1: A novel test for selection on *cis*-regulatory elements reveals positive and negative selection acting on mammalian transcriptional enhancers

Figure 1.1 Outline of our approach.
Figure 1.2 Selection on LTV1.
Figure 1.3 Selection on ALDOB.
Figure 1.4 Selection on ECR11.
Figure 1.5 Effect of excluding random subsets of sites within each enhancer from analysis

Chapter 2: Dissecting the genetic basis of a complex *cis*-regulatory adaptation

Fig 2.1. Results of the sign test for selection on *cis*-regulation.
Fig 2.2. *S. p*

Chapter 3: Quantitative CRISPR interference screens in yeast identify chemical-genetic interactions and new rules for guide RNA design

Figure 3.1. Overview of regulatable CRISPRi and individual strain analysis.
Figure 3.2. Parallel analysis of CRISPRi-induced fitness defects in pooled cultures.
Figure 3.3. Erg25 regulates fluconazole sensitivity.
Figure 3.4. Quantitative comparison of full-length and truncated gRNAs.
Figure 3.5. Effect of target location and accessibility on gRNA efficacy.

Chapter 4: A scalable method for parsing complex oligonucleotide libraries and creation of a CRISPRi *Saccharomyces cerevisiae* strain collection

Figure 4.1 Oligonucleotide library parsing strategy.
Figure 4.2 Molecular probe oligo set purity and abundance.
Figure 4.3 CRISPRi Recipient Strain.
Figure 4.4 CRISPRi collection experimental workflow.
Figure 4.6 Histograms for gRNA Library
Figure 4.7 Chromatin and Positional Effects in YPEG.
Figure 4.8. Oligo design and PCR strategy.

Chapter 5: Distinct patterns of Cas9 mismatch tolerance *in vitro* and *in vivo* and effects of truncated guide RNAs

Figure 5.1 Target Sequences and Workflow
Figure 5.2 In vivo cleavage effects of single base variants
Figure 5.3 In vitro cleavage effects of single base variants
Figure 5.4 In vivo cleavage effects of adjacent double variants
Figure 5.5 In vitro cleavage effects of adjacent double variants
Figure 5.6 In vivo cleavage effects of deletion variants
Figure 5.7: In vitro cleavage effects of deletion variants

Appendices

Chapter 2

Figure S2.1. Reproducibility of biological replicates.
Figure S2.2. Agreement between RNA-seq and microarray data.

Chapter 3

Figure S3.1. Characterization of single-plasmid system for CRISPRi in yeast.
Figure S3.2. Schematic of the workflow for parallel analysis of CRIPSRi-induced fitness defects in pooled cultures.

Figure S3.3. Replicate Plots

Figure S3.4. Parallel analysis of CRIPSRi-induced fitness defects in pooled cultures.

Figure S3.5. Effect of 1181-0519 on ergosterol metabolite levels in yeast.

Figure S3.6. Quantitative comparison of full-length and truncated gRNAs.

Figure S3.7. The effect of nucleosome occupancy and chromatin accessibility on gRNA efficacy.

Figure S3.8. Effect of sequence context on guide efficacy.

Figure S3.9. The effects of guide RNA secondary structure and DNA/RNA duplex formation strength (x-axis) on guide efficacy (y-axis).

Chapter 5

Figure S5.1 Yeast random variant library cloning vector map.

Plasmid genbank file can be found in Supplemental files.

Figure S5.2 Cleavage results for in vivo variant libraries for type I with 18 and 20 nt complementarity gRNAs.

Figure S5.3 Plots of Transition variant effects.

Figure S5.4 Cleavage results for in vitro unc-22A random variant libraries with 17, 18, and 20 nt complementarity gRNAs (type I library).

Figure S5.5 In vitro cleavage profiles for 180 minutes incubation for target library type II with full-length and truncated gRNAs (EGFP-1, EGFP-2, rol-6, and unc-22A).

Figure S5.6 Single base in vitro transversion effects of the 20 nt complementarity gRNA with NEB Cas9 with the unc-22A random variant library-2 at 30 degrees Celsius.

Figure S5.7) Fu et al. 2014 Cas9 (ref.[72]) in vitro cleavage experiments with type I unc-22A random variant library-2.

Figure S5.8 In vitro cleavage effects of adjacent double variants.

Figure S5.9 In vitro cleavage effects of deletion variants.

Figure S5.10 This Figure shows the full detailed single base in vitro variant effects shown in Figure 5.3 for the 180 minute time point.

Figure S5.11 EGFP-1-Reverese (R) and EGFP-1 Off Set (OS).

Figure S5.12 Cas9 in vivo experiments on yeast random type I variant library-2 with extended generation times.

Figure S5.13 Single base in vivo transversion effects on the unc-22A gRNA for a physically distinct type I library, yeast random variant library-1.

Figure S5.14 Single base in vivo variant effects of 20 nt complementarity for KU70 and BY4741 yeast strains.
Introduction

Our ability to understand biology has always been defined by the technologies available to us. It was the invention of the microscope in the 17th century that led to the discovery that biological organisms are made of cells. It took many more years for the discovery of the basic foundations of genetics due to technological limitations. It was less than 75 years ago that we first learned that DNA transmitted the hereditary material of life [1, 2], and not until 1953 with advances in x-ray crystallography that the structure of DNA was determined[3]. It took another two decades for DNA sequencing technologies to be developed [4, 5], and it was not until 1977 that the first simple genome, that of the phage phi X174, was sequenced [6]. During the 1990s and early 2000s, the ambitious Human Genome Project led to a technology development race that resulted in the development of new sequencing technologies that made sequencing the entire human genome possible [7, 8]. The next-generation sequencing technologies employed in modern genetics studies are a result of this technological revolution.

Upon reflection on this history, it becomes clear that technology development is an essential driver in improving our understanding of the central questions of biology. Most recently, the advent of three enabling technologies has dramatically improved our ability to address fundamental biological questions. These technologies can be broken into three categories: read (next-generation sequencing), write (oligonucleotide and gene synthesis), and manipulate (genome engineering technologies) genetics.

In this dissertation, I will discuss how we utilized these technologies to study cis-regulatory evolution. I will then discuss my work on advancing technologies for writing and manipulating DNA. But first, I will provide some background and motivations for these projects.
Cis-Regulatory Evolution

The idea that non-coding regulatory sequences could lead to evolutionary changes is nearly as old as the discovery of such regulatory sequences. Soon after the publication of the first cis-regulatory sequence characterized, the lac operon in 1961, it was speculated that mutations in such regulatory sequences could lead to evolutionary changes [9–11]. The argument was based on the recognition that for a gene to function properly, it is not only important that the gene product is functional, but that it is produced under the right circumstance. Further evidence for this idea came in 1975 with the realization that homologous proteins in humans and chimpanzees are nearly identical, and yet these species show dramatic phenotypic differences[12]. It was argued that such dramatic changes could not be the result of coding differences alone, and that non-coding regulatory variations must be primarily responsible[12].

It is now known that noncoding regions comprise the vast majority of genomic regions under selective constraint in mammals[13], suggesting that these regions have an important functions. There is evidence to suggest that these sequences are likely the major source of evolutionary adaptations in general, across a wide diversity of organisms [11, 12, 14–18]. Despite this, our ability to detect natural selection in noncoding regions has lagged far behind our ability to detect selection in protein coding regions [19]. Non-coding regions can show highly accelerated evolutionary change in the absence of positive selection do to variation in mutation rates, biased gene conversion, and transmission distortion [20]. Even when a genomic region can be demonstrated to be under selection, it can be difficult to determine if that region represents a cis-regulatory element or how that element affects gene expression [20].

An alternative approach to detect regulatory evolution is to look for its effect in genome-wide gene expression data to detect genes under positive selection. However, this can be difficult due to the lack of a null model of neutrality[20]. It can be quite difficult to distinguish adaptive expression changes driven by positive selection from neutral divergence resulting from mutation and genetic drift[20, 21].

Chapters 1 and 2 of this dissertation will discuss some of these limitations in more detail and our efforts to address them. Specifically, in Chapter 1, we develop a novel test for cis-regulatory evolution that was facilitated by the synthesis of libraries.
of enhancers with all possible single nucleotide variants. In Chapter 2, we investigate differences in drug resistance between two closely related species of budding yeast, and how subtle effects of individual regulatory elements can be combined, via natural selection, into a complex adaptation.

Write: Oligonucleotide Library Synthesis

The synthesis of DNA chemically was first accomplished in 1955. It would take another 25 years and several iterations of different chemistries until the more modern method of chemical DNA synthesis, the phosphoramidite method, was developed [22, 23]. Many additional later improvements to the chemistry and manufacturing process have made phosphoramidite synthesis a very reliable and fast method for synthesizing short single stranded stretches of DNA, known as oligonucleotides (or oligos for short)[22, 24]. Our ability to synthesize oligos has been essential for the development of many molecular biology techniques such as polymerase chain reaction (PCR), oligonucleotide probes, and DNA sequencing. Today this process is largely automated and conventional synthesis techniques yield individual oligos of lengths of ~100nt at a cost of around $0.05-0.15 per base with error rates of around 1 in 200 bases [24]. Still, this presents some limitations if one wants to synthesize thousands of different oligos, or if one wants to synthesize entire genes affordably. In order to accomplish these goals, an alternative approach was needed.

During the early 1990s, Affymetrix developed some of the first methods for spatially localized polymer synthesis on arrays [25, 26]. Since then several different technologies have been developed to synthesize oligos on arrays including those of Agilent’s ink jet printing of nucleotides with phosphamadite chemistry[27–29], and Custom Array’s semiconductor-based electrochemical acid production to selectively de-protect nucleosides[30]. After synthesis on arrays, oligos can then be released en masse to create oligonucleotide libraries of high complexity (~10,000 – 250,000 oligos in each pool). If priced at a cost per base of unique sequence generated, oligos from microarrays are 2 to 4 orders of magnitude cheaper than oligos synthesized individually. Current cost per base ranges from $0.00001-0.0001, depending on the
size of the library, the length of the oligos, and the platform that is used [24]. These oligo libraries can be used to assemble genes [24, 31], create shRNA [32] and CRISPR-gRNA libraries [33–35], create molecular probes [36], interrogate cis-regulatory regions [37, 38], and for other applications [24]. These oligos libraries were a key enabling technology for chapters 1, 3, 4, and 5 of this dissertation.

However there are limitations to microarray oligo library synthesis. Oligo synthesis has an error rate of about 1/200 nt [24], which can be a problem for some applications. For example, most genes assembled for oligo libraries will have at least a few synthesis errors [24]. To deal with this, assembled genes are often cloned into plasmids in Escherichia coli, and multiple individual clones are sequenced by Sanger sequencing until a correct clone is found. This step is time consuming and expensive. Additionally, the distribution of individual oligos in the pool of oligos is not even, with some sequences occurring at frequencies up to several fold greater than others. In chapter 4 we will discuss these limitations in detail, and present a strategy to economically obtain sequence verified DNA from oligos integrated into individual clones of yeast using next-generation sequencing.

**Manipulate: Genome Editing and Genome Engineering Technologies**

Genome editing has historically been very difficult. Most approaches have utilized the homologous recombination machinery of cells to target DNA to a desired site. But simply providing DNA with homology is relatively inefficient in eukaryotes. The absolute frequency of homologous recombination between donor and target sequences ranges from 1 event per $10^4$ to $10^7$ cells in yeast and in mouse ES cells [39]. While this can be useful when a powerful selection can be applied, with other experimental organisms and cell types where screening and selection procedures are not feasible, low recombination rates are a significant barrier to the feasibility of gene editing [39]. This motivated researchers to seek a way to increase recombination efficiency. Inspired by the discovery that natural recombination events are initiated by single double-strand breaks (DSBs), researchers hypothesized that inducing a double strand break at a target site could improve integration efficiency. To test this, early experiments introduced a unique recognition site for a site-specific DNA endonuclease at a single site in the genome and then cut by introduction of the corresponding
enzyme [40–43]. In both yeast and mammalian cells, this technique improved the efficiency of homologous recombination at the target site by several orders of magnitude [40–43].

While this demonstrated potential to dramatically improve recombination efficiency, this technique was limited by the requirement to introduce a specific target site for the endonuclease. This motivated research to create programmable site-specific endonucleases. The first class of such enzymes to be developed was the zinc-finger nucleases. These chimeric nucleases consist of a FokI nuclease domain devoid of its DNA binding domain (DBD), fused to Cys2His2 zinc fingers (ZFs) DBDs[39, 44]. Each finger domain consists of ~30 amino acids and recognizes 3 bases of DNA [45] and can be combined in a modular fashion to recognize different length targets[39]. To facilitate the targeting of new sites, zinc-finger domains were developed to recognize nearly all of the 64 possible nucleotide triplets [46]. Once zinc-finger libraries were created, they were also used to create other programmable DNA binding enzymes, such as synthetic transcription factors [47, 48], recombinases [49], and synthetic methyltransferases [50, 51].

Despite all these advantages, the difficulty and expense of creating new zinc finger DBDs and other inherent limitations of zinc finger technology limited their use [52]. An alternative to zinc-fingers emerged in 2009 with the discovery of the code of transcription activator like effectors (TALEs) from the plant pathogenic bacteria Xanthomonas [53]. The DBD of TALEs contain several repeats of 33-35-amino acids, with each repeat recognizing one base of DNA [53]. These TALEs were fused to nucleases and transcription factors, building off previous work in zinc fingers [46]. While less expensive and more predictable targeting can be achieved with TALEs than with zinc fingers[52], TALEs are still limited by the requirement of building an entirely new protein for each new target.

The past three years have seen a revolution in gene editing technologies thanks to the discovery of the mechanism of action of the bacterial type II clustered regularly-interspaced short palindromic repeats (CRISPR) adaptive immune system. This immune system is modulated by the CRISPR locus, which store information in the form of short sequences homologous to phage or plasmids that the bacteria was
previously exposed to, and a RNA guided nuclease referred to as CRISPR associated protein 9 (Cas9), which is targeted to invading phage or plasmid and cleaves the pathogenic DNA [54]. Cas9 depends on a protospacer adjacent motif (PAM, which consists of a GG dinucleotide for Streptococcus pyogenes Cas9, the most commonly used Cas9) and two RNAs, the CRISPR RNA (crRNA) expressed from the CRISPR locus which contains a sequence specific to a 20 nucleotide target, and a constant trans-activating crRNA (tracrRNA) required for binding of the crRNA to Cas9 [54]. To simplify this system for genome engineering, these two RNAs can be combined into a single RNA referred to as a guide RNA (gRNA)[54]. Because the targeting of a new site only requires the creation of a new gRNA, CRISPR-facilitated genome engineering is much simpler than previous technologies. This has enabled genome editing in a wide variety of organisms[54–57]. Additionally, the Cas9 protein can be further modified to act as a programmable DBD effector. Two point mutations can yield a catalytically dead Cas9 (dCas9)[54]. Like zinc-fingers and TALEs, dCas9 can be made to function as a transcriptional activator or repressor (also known as CRISPR interference, or CRISPRi) capable of modulating gene expression in eukaryotes [58–62].

CRISPR offers several advantages over TALEs and ZFs. Most importantly, new gRNAs can be made quickly and economically from short oligos – there is no requirement to build new enzymatic DBDs. As the specificity-determining region of the gRNA is short (generally 20 bases in length), this offers compatibility with highly-parallel array-based oligonucleotide synthesis and the potential to create libraries of gRNAs to thousands of targets. Several groups have taken advantage of this, and generated genome-wide libraries for knocking out [63–66], silencing[34], and activating genes[34, 67].

Despite all the advantages of CRISPR, some limitations still exist. Not all gRNAs work equally well [34, 68]. Additionally, several groups have demonstrated that Cas9 can bind to imperfect targets and have off-target effects [62, 69–72]. We will discuss these concerns in detail in the introductions to chapters 3 and 5 of this dissertation. In chapters 3 and 4 we will discuss work we did to help understand rules for design of effective gRNAs for CRISPRi in yeast. In chapters 3 and 5 we will
discuss work we have done on understanding the specificity of Cas9 DNA binding and nuclease activity.

Outline of The Thesis

This thesis will cover work from 5 separate projects for which I was a primary contributor during my graduate career. In this thesis, I will discuss how we leveraged data from a study using oligo library synthesis to characterize all point mutations in three mammalian enhancers to develop a novel test to detect selection on these cis-regulatory elements (Chapter 1). I will also discuss how we were able to detect an example of complex multi-locus cis-regulatory evolution in budding yeast and how we experimentally validated this result, in part by using an improved CRISPR-activator I created (Chapter 2). I will then discuss my work determining rules for efficient CRISPR interference in *Saccharomyces cerevisiae*, including a novel finding on Cas9’s preference for nucleosome free open chromatin regions (Chapter 3). Next, I will discuss a new technology we developed to parse oligo libraries into individual, sequence verified clones using *S. cerevisiae*. We applied this technology to create a CRISPRi strain collection with gRNAs to most of the essential and aerobic essential genes in *S. cerevisiae* (Chapter 4). Finally, I will discuss our work on understanding of the specificity profile of *Streptococcus pyogenes* Cas9 nuclease, both *in vitro* and *in vivo* in *S cerevisiae*, and how truncated gRNAs allow for improved specificity in *S cerevisiae* (Chapter 5).
Chapter 1: A novel test for selection on cis-regulatory elements reveals positive and negative selection acting on mammalian transcriptional enhancers

Authors
Justin D. Smith¹, Kimberly F. McManus², and Hunter B. Fraser²*

¹Department of Genetics, and ²Department of Biology, Stanford University, Stanford CA 94305.
*Correspondence: hbfraser@stanford.edu

This chapter is a reproduction of work published in Molecular Biology and Evolution (2013) 30 (11): 2509-2518. doi: 10.1093/molbev/mst134 with the exception of section titled ‘Subsequent Research’. The dissertation author [JDS] performed the majority of the analyses, created all the figures, and co-wrote the paper with the other authors. The section ‘Subsequent Research’ is a summary of the research of others that built off the test for selection developed in this chapter.

Abstract
Measuring natural selection on genomic elements involved in the cis-regulation of gene expression—such as transcriptional enhancers and promoters—is critical for understanding the evolution of genomes, yet it remains a major challenge. Many studies have attempted to detect positive or negative selection in these noncoding elements by searching for those with the fastest or slowest rates of evolution, but this can be problematic. Here we introduce a new approach to this issue, and demonstrate its utility on three mammalian transcriptional enhancers. Using results from saturation mutagenesis studies of these enhancers, we classified all possible point mutations as up-regulating, down-regulating, or silent, and determined which of these mutations have occurred on each branch of a phylogeny. Applying a framework analogous to $K_a/K_s$ in protein-coding genes, we measured the strength of selection on up-regulating and down-regulating mutations, in specific branches as well as entire phylogenies. We discovered distinct modes of selection acting on different enhancers: while all three have experienced negative selection against down-regulating mutations, the selection pressures on up-regulating mutations vary. In one
case we detected positive selection for up-regulation, while the other two had no detectable selection on up-regulating mutations. Our methodology is applicable to the growing number of saturation mutagenesis data sets, and provides a detailed picture of the mode and strength of natural selection acting on cis-regulatory elements.

**Introduction**

Noncoding regions comprise the vast majority of genomic regions under selective constraint in mammals [13], harbor most common genetic variants influencing human disease [73], and may be the source of most evolutionary adaptations [12, 14, 15]. Yet our ability to measure natural selection in noncoding regions has lagged far behind our ability to do so in the small fraction of the genome that codes for protein [19].

In order to infer natural selection, one must be able to reject a null model of neutral evolution [74]. In protein-coding regions, a convenient proxy for neutral mutations is the synonymous mutations that do not alter the amino acid sequence. By comparing the rate of accumulating potentially functional nonsynonymous mutations (abbreviated as $K_a$) to the synonymous rate ($K_s$), selection can be inferred [75]: a slower nonsynonymous rate ($K_a/K_s < 1$) reflects negative (purifying) selection against changing protein sequence; a faster rate ($K_a/K_s > 1$) reflects positive selection for changing protein sequence; and approximately equal rates ($K_a/K_s \approx 1$) means that selection cannot be inferred. Because synonymous and nonsynonymous sites are interdigitated within every protein-coding gene, their mutation rates should not differ greatly, facilitating a direct comparison of the two. This framework has proven useful for studying protein-coding regions, but $K_a/K_s$ cannot be calculated for noncoding regions.

Many studies of noncoding evolution have therefore taken alternative approaches, such as scanning multiple genomes for noncoding regions with unusually rapid evolutionary rates [76–79]. For example, one of the first studies of rapid evolution in the human lineage discovered 49 “human accelerated regions” (HARs), which are enriched near genes involved in transcriptional regulation [76]. However this study illustrates a major caveat for such analyses: although accelerated divergence is typically attributed to positive selection, many HARs were subsequently shown to
likely result from biased gene conversion [80, 81], a process that can lead to rapid divergence in the complete absence of selection. This highlights the need for comparisons to neutrally evolving sites that are matched in mutation rate to each noncoding region of interest. Although many candidate sources of neutral sites have been proposed (e.g. nearby intronic sites, transposons, or synonymous sites), none of these is entirely neutral, and because they are all located outside of the noncoding elements being tested, regional differences in mutation rates or recombination rates are possible [82]. As a result, there is still no consensus as to what constitutes a suitable reference for noncoding elements [19].

Another issue with studies scanning genomes for regions of rapid divergence is that even when positive selection is acting on a cis-regulatory element (e.g. for higher transcriptional activity), there is likely to be negative selection simultaneously acting on the same element (e.g. purging mutations that disrupt the element’s function) [19]. When negative selection dominates—as is likely to be required for almost any element to maintain its function—scanning the genome for rapid evolution will not detect the positive selection, because the rate of the entire element will not be faster than neutral. In other words, “averaging” across sites in a noncoding element that are subject to different selection pressures will reduce power to detect selection.

One potential solution is to separate sites within a cis-regulatory element into multiple classes, and compare evolutionary rates between classes. This has the potential to solve both of the issues discussed above: mutation rates should be similar if sites of different classes are interspersed, and power to detect selection will be maximized if sites within each class experience similar selection pressures. An example of this approach is a metric called $K_b/K_i$, which compares the evolutionary rate within known transcription factor binding sites (TFBS) ($K_b$) with that outside TFBS ($K_i$), with an excess of substitutions in TFBS ($K_b/K_i > 1$) suggesting positive selection [83]. Although this has advantages over measuring the overall evolutionary rate across an entire enhancer or promoter, it has several drawbacks. Perhaps the most important is that most naturally occurring genetic variants affecting transcription factor binding fall outside of any recognizable TFBS [84, 85]. Not only are these effects not captured by $K_b$, they also can render $K_i$ an underestimate of the neutral
evolutionary rate (when mutations outside of TFBS are under negative selection), leading to inflated $K_u/K_s$ and possibly false inference of positive selection. This same caveat applies to more recent studies of selection on TFBS as well [86]. Another method that takes into account the effects of substitutions on TFBS motif strengths, while not assuming that changes outside of TFBS are neutral, still does not account for the effects of functional substitutions outside of known TFBS [87].

Our goal here was to design a robust framework for detecting natural selection in noncoding regions. To achieve this, we utilized data from saturation mutagenesis studies, which measure the effect of every possible single nucleotide variant (SNV) within specific cis-regulatory elements [37, 88–90]. In these studies, sequence constructs containing every SNV within a promoter or enhancer are produced, either by large-scale oligonucleotide synthesis [88–90] or by traditional oligonucleotide synthesis with some degeneracy introducing random SNVs [37]. Their transcriptional outputs are then measured, most often using high-throughput RNA sequencing of short transcribed “barcodes” that uniquely identify each construct. This type of data allows the classification of all possible SNVs into one of three classes: up-regulating (i.e. increasing transcription over the reference enhancer sequence), down-regulating, or silent. We can then compare the rate of fixation of (for example) up-regulating SNVs with the silent ones that are likely under little or no selection. Because the three classes of sites are interdigitated with one another, differences in regional mutation rates are unlikely. Directly analogous to $K_a/K_s$, we call these metrics $K_u/K_n$ and $K_d/K_n$ for measuring selection on up-regulating and down-regulating substitutions, respectively. Perhaps the most important novel aspect of these metrics is that because the up- and down-regulating SNVs are assessed for selection separately, we can potentially detect positive selection on one class, even when negative selection on the other class would have masked any signal when considering the element as a whole.

As an initial proof-of-principle, we applied our approach to the three liver enhancers studied by Patwardhan et al. (2012) [37]. The three enhancers, named ALDOB, ECR11, and LTV1, are located within or proximal to the ALDOB, DHRS9, and Zfp36 genes respectively [91, 92]. ALDOB is a glycolytic enzyme, fructose-1,6-bisphosphate aldolase; DHRS9 is a short chain alcohol dehydrogenase/reductase; and
Zfp36 is a zinc-finger RNA-binding protein that binds and degrades cytokine mRNAs. These three enhancers were dissected by measuring the transcriptional output \textit{in vivo} of over 640,000 distinct mutant enhancers, differing on average from the wild-type by SNVs at 2.1\%-3.1\% of sites; each nucleotide within each enhancer was mutated, on average, in over 4,000 distinct constructs. This saturation mutagenesis allowed the robust empirical estimation of each SNV’s effect on transcription, providing the input for our approach.

Results

Our methods for detecting natural selection on noncoding elements, $K_u/K_n$ and $K_d/K_n$, are outlined in Figure 1.1. Briefly, they compare the rate of up- or down-regulating substitution with the rate of silent substitution from the same enhancer element (see Materials and Methods). Values significantly greater than one indicate likely positive selection, whereas those less than one imply negative selection, directly analogous to the commonly used protein-coding metric $K_a/K_s$ [75]. Because the selective regimes are inferred for up- and down-regulation separately, they may reflect entirely distinct modes of selection.

![Figure 1.1 Outline of our approach.](image-url)

Using expression maps of enhancers and the effect of every single possible SNV combined with present-day sequences and ancestral reconstructions, our method estimates $K_u/K_n$ and $K_d/K_n$ to detect evidence of selection on cis-regulatory elements. $K_u/K_n$ or $K_d/K_n$ values significantly greater than one signify positive selection, whereas values significantly less than one indicate negative selection.
To avoid potential issues with distant comparisons (including less accurate ancestral reconstruction and epistatic interactions; see Discussion), we focused on species within the same phylogenetic order as the original mutagenized enhancer (rodents for LTV1, and primates for ALDOB and ECR11). We obtained orthologous sequences for the three enhancers, aligned them, and reconstructed ancestral sequences by maximum likelihood (see Materials and Methods). We then calculated $K_u/K_n$ and $K_d/K_n$ for each individual branch in the phylogeny, as well as for overall phylogenies.

For the LTV1 enhancer, we found an overall $K_u/K_n$ ratio of 1.45 (Fisher’s Exact $p = 1.5 \times 10^{-5}$ for the null model of neutrality) and $K_d/K_n$ ratio of 0.51 ($p = 1.2 \times 10^{-9}$) among rodents (Figure 1.2). No individual branches had a significant $K_u/K_n$ ratio, but three individual branches (branches D, F, and M in Figure 1.2C) did have a significantly lower $K_u/K_n$ ratio than expected under neutrality. These results suggest that within rodents there was positive selection for up-regulating mutations in this enhancer, coupled with negative selection against down-regulating mutations.
Figure 1.2 Selection on LTV1.
(A) Phylogenetic tree of the rodent species for which LTV1 enhancer sequences were analyzed. Ancestral nodes are labeled L1–L7, and branches are labeled A–M. (B) Ku/Kn and (C) Kd/Kn values are plotted for each branch of the phylogenetic tree. Below each bar is the number of up- or downregulating (top row) and neutral (bottom row) substitutions inferred for that branch. Asterisks mark branch-specific Ku/Kn or Kd/Kn values that differed significantly from neutral (Fisher’s exact \( P < 0.05 \)).

Possible Up: 206
Possible Neutral: 254
Overall \( K_u/K_n = 1.45 \)
Fisher’s Exact \( p = 1.5 \times 10^{-5} \)

Possible Down: 446
Possible Neutral: 254
Overall \( K_d/K_n = 0.51 \)
Fisher’s Exact \( p = 1.2 \times 10^{-9} \)
downregulating (top row) and neutral (bottom row) substitutions inferred for that branch. Asterisks mark branch-specific Ku/Kn or Kd/Kn values that differed significantly from neutral (Fisher’s exact P < 0.05).

For the ALDOB enhancer, we found an overall K\textsubscript{u}/K\textsubscript{n} ratio of 1.14 (p = 0.56) and a K\textsubscript{d}/K\textsubscript{n} ratio of 0.48 (p = 1.7 x 10\textsuperscript{-4}) in primates (Figure 1.3). No individual branches had a significant K\textsubscript{u}/K\textsubscript{n} ratio, while one branch (branch E in Figure 1.3C) had a significantly lower K\textsubscript{d}/K\textsubscript{n} ratio than expected under neutrality. These results suggest that within primates there was negative selection against down-regulating mutations, but no detectable selection on up-regulating mutations in this enhancer.
Figure 1.3 Selection on ALDOB.

(A) Phylogenetic tree of the primate species for which ALDOB enhancer sequences were analyzed. Note that A8 appears twice as the reconstruction did not differ between these two nodes. Ancestral nodes are labeled A1–A10, and branches are labeled A–T. (B) $K_u/K_n$ and (C) $K_d/K_n$, as in figure 1.2. “Undef” indicates an undefined value.
For the ECR11 enhancer, we found an overall $K_d/K_n$ ratio of 0.95 ($p = 0.90$) and a $K_d/K_n$ ratio of 0.67 ($p = 0.065$) (Figure 1.4), with no individual branches reaching significance. Therefore we cannot reject the null hypothesis of neutral evolution for primates as a whole. However closer examination revealed that these results were primarily driven by just two branches (D and E, the lemur clade) which accounted for 65% of all down-regulating substitutions in primates, and were evolving with no detectable selection against down-regulating SNVs ($K_d/K_n$ values of 0.90 and 0.93; Figure 1.4c). Excluding the lemur branches resulted in an overall $K_d/K_n$ of 0.44, significantly lower than expected under neutrality ($p = 0.018$). In contrast, $K_d/K_n$ was similar within vs. outside the lemur clade (values of 0.998 and 0.91, respectively). This suggests that the ECR11 enhancer may have experienced selection against down-regulating SNVs specifically in simians (monkeys and apes), although the ad hoc nature of this analysis precludes a definitive conclusion.
Figure 1. 4. Selection on ECR11.

(A) Phylogenetic tree of the primate species for which ECR11 enhancer sequences were analyzed. Note that E3 appears twice as the reconstruction did not differ between these two nodes. Ancestral nodes are labeled E1–E11, and branches are labeled A–V. (B) Ku/Kn and (C) Kd/Kn, as in figure 1.

Be quantified in these enhancers because the set of mutant enhancers tested for activity contained not only every possible SNV, but also over 99.999% of all possible pairs of variant sites, only ~0.1% of which exhibit pairwise epistasis in their effects on transcription (Patwardhan et al. 2012).

Furthermore, even among the epistatic SNVs, only the minority that alter the SNVs’ classification (as upregulating, downregulating, or silent)—as opposed to the magnitude of up- or downregulation—would have any impact on our approach. The rarity of strong epistasis implies that most...
To test the robustness of our results for all three enhancers, we repeated the overall-phylogeny $K_u/K_n$ and $K_d/K_n$ calculations for each enhancer with three variations. First, we excluded individual branches from the analysis (Table 1.2), to test whether any single branches may have disproportionate effects. Only for ECR11 $K_u/K_n$ was there a mixture of nominally significant and non-significant results (primarily due to the lemur clade described above). Second, we tested the effect of excluding random subsets of sites within each enhancer from analysis, to determine whether a small number of outlier sites may be driving the results. Even with up to 75% of sites excluded from analysis, we do not see any bias towards increased or decreased $K_u/K_n$ or $K_d/K_n$, though power to detect selection does decrease (Figure 1.5). Third, we varied the p-value cutoff at which we classified SNVs as silent or functional. At three different cutoffs, we observed very similar patterns of significance as in our original analysis (Table 1.1). Together, these analyses suggest that our results are not driven by outlier branches, outlier sites, or the exact p-value cutoff used.

<table>
<thead>
<tr>
<th></th>
<th>Functional p &lt; 0.01, silent p &gt; 0.01</th>
<th>Functional p &lt; 0.1, silent p &gt; 0.1</th>
<th>Functional p &lt; 0.01, silent p &gt; 0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTV1</td>
<td>$K_u/K_n = 1.47, p = 8.6e-06$</td>
<td>$K_u/K_n = 1.42, p = 7.9e-05$</td>
<td>$K_u/K_n = 1.42, p = 0.00013$</td>
</tr>
<tr>
<td></td>
<td>$K_d/K_n = 0.53, p = 3.8e-09$</td>
<td>$K_d/K_n = 0.53, p = 4.3e-08$</td>
<td>$K_d/K_n = 0.51, p = 2.2e-8$</td>
</tr>
<tr>
<td>ALDOB</td>
<td>$K_u/K_n = 1.58, p = 0.056$</td>
<td>$K_u/K_n = 1.37, p = 0.16$</td>
<td>$K_u/K_n = 1.48, p = 0.10$</td>
</tr>
<tr>
<td></td>
<td>$K_d/K_n = 0.49, p = 0.00077$</td>
<td>$K_d/K_n = 0.52, p = 0.0013$</td>
<td>$K_d/K_n = 0.51, p = 0.00055$</td>
</tr>
<tr>
<td>ECR11</td>
<td>$K_u/K_n = 0.83, p = 0.57$</td>
<td>$K_u/K_n = 1.10, p = 0.63$</td>
<td>$K_u/K_n = 0.82, p = 0.57$</td>
</tr>
<tr>
<td></td>
<td>$K_d/K_n = 0.75, p = 0.22$</td>
<td>$K_d/K_n = 0.67, p = 0.040$</td>
<td>$K_d/K_n = 0.74, p = 0.21$</td>
</tr>
</tbody>
</table>

Table 1.1 $K_u/K_n$ and $K_d/K_n$ at varied p-value cutoff at which we classified SNVs as silent or functional.
### Table 1.2 Overall phylogeny wide p-values and Ku/Kn, Kd/Kn ratios for the three enhancers.

Significance was called for SNVs with P < 0.05 and Neutral was called for P > 0.05

<table>
<thead>
<tr>
<th>ALDOB</th>
<th>Branch</th>
<th>Nodes</th>
<th>Ku/Kn</th>
<th>up-P-value</th>
<th>Kd/Kn</th>
<th>down-P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A2-A1</td>
<td>1.1439</td>
<td>0.5611</td>
<td>0.4797</td>
<td>1.69E-04</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>A3-A2</td>
<td>1.1439</td>
<td>0.5611</td>
<td>0.4797</td>
<td>1.69E-04</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Microcebus_murinus-A2</td>
<td>1.2710</td>
<td>0.3578</td>
<td>0.4579</td>
<td>3.42E-04</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>A4-A3</td>
<td>1.0650</td>
<td>0.7642</td>
<td>0.4579</td>
<td>2.07E-04</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Tarsius_syricha-A3</td>
<td>1.2201</td>
<td>0.5076</td>
<td>0.5087</td>
<td>3.69E-03</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>A5-A4</td>
<td>1.0469</td>
<td>0.8804</td>
<td>0.5122</td>
<td>1.05E-03</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>A10-A4</td>
<td>1.0469</td>
<td>0.8804</td>
<td>0.5122</td>
<td>1.05E-03</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>Callithrix_jacchus-A10</td>
<td>1.1251</td>
<td>0.6575</td>
<td>0.4654</td>
<td>1.87E-04</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>Saimiri_boliviensis-A10</td>
<td>1.1814</td>
<td>0.4641</td>
<td>0.4504</td>
<td>1.09E-04</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>Macaca_mulatta-A5</td>
<td>1.1632</td>
<td>0.5508</td>
<td>0.4656</td>
<td>2.39E-04</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>A6-A5</td>
<td>1.1439</td>
<td>0.5550</td>
<td>0.5037</td>
<td>7.39E-04</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Nomascus_leucogenys-A6</td>
<td>1.0894</td>
<td>0.7696</td>
<td>0.4506</td>
<td>8.49E-05</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>A7-A6</td>
<td>1.1623</td>
<td>0.5578</td>
<td>0.4726</td>
<td>1.50E-04</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>A8-A7</td>
<td>1.1814</td>
<td>0.4641</td>
<td>0.4804</td>
<td>3.65E-04</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>Pongo_abelii-A7</td>
<td>1.1439</td>
<td>0.5550</td>
<td>0.4731</td>
<td>3.28E-04</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>Gorilla_gorilla-A8</td>
<td>1.1814</td>
<td>0.4641</td>
<td>0.4954</td>
<td>2.90E-04</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>Homo_sapiens-A8</td>
<td>1.1251</td>
<td>0.6575</td>
<td>0.4804</td>
<td>2.52E-04</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>A9-A8</td>
<td>1.1623</td>
<td>0.5578</td>
<td>0.4874</td>
<td>2.14E-04</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>Pan_paniscus-A9</td>
<td>1.1439</td>
<td>0.5611</td>
<td>0.4797</td>
<td>1.76E-04</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>Pan_troglodytes-A9</td>
<td>1.1623</td>
<td>0.5578</td>
<td>0.4874</td>
<td>1.38E-04</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTV1</th>
<th>Branch</th>
<th>Nodes</th>
<th>Ku/Kn</th>
<th>up-P-value</th>
<th>Kd/Kn</th>
<th>down-P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>L2-L1</td>
<td>1.4752</td>
<td>9.34E-06</td>
<td>0.5085</td>
<td>4.44E-09</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Oryctolagus_cuniculus-L2</td>
<td>1.4303</td>
<td>3.52E-04</td>
<td>0.5069</td>
<td>5.18E-08</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>L3-L2</td>
<td>1.4294</td>
<td>5.20E-05</td>
<td>0.4838</td>
<td>4.21E-10</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Lctidmys_tridecamelineatus-L3</td>
<td>1.4550</td>
<td>1.67E-04</td>
<td>0.5183</td>
<td>9.97E-08</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>L4-L3</td>
<td>1.4385</td>
<td>3.34E-05</td>
<td>0.4996</td>
<td>1.02E-09</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>L7-L4</td>
<td>1.5413</td>
<td>1.48E-05</td>
<td>0.5339</td>
<td>5.68E-07</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>Mus_musculus-L7</td>
<td>1.3981</td>
<td>2.44E-04</td>
<td>0.5034</td>
<td>2.51E-09</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>Rattus_norvegicus-L7</td>
<td>1.4366</td>
<td>7.88E-05</td>
<td>0.5173</td>
<td>1.74E-08</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>L5-L4</td>
<td>1.4253</td>
<td>1.17E-04</td>
<td>0.4755</td>
<td>4.54E-10</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>Octodon_degus-L5</td>
<td>1.4485</td>
<td>1.19E-04</td>
<td>0.4866</td>
<td>6.97E-09</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>L6-L5</td>
<td>1.5132</td>
<td>2.43E-06</td>
<td>0.5177</td>
<td>1.19E-08</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Cavia_porcellus-L6</td>
<td>1.4039</td>
<td>7.12E-04</td>
<td>0.5018</td>
<td>3.23E-08</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>Heterocephalus_glaber-L6</td>
<td>1.4283</td>
<td>3.51E-04</td>
<td>0.5188</td>
<td>1.04E-07</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ECR11</th>
<th>Branch</th>
<th>Nodes</th>
<th>Ku/Kn</th>
<th>up-P-value</th>
<th>Kd/Kn</th>
<th>down-P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>E2-E1</td>
<td>0.9242</td>
<td>0.7951</td>
<td>0.6848</td>
<td>0.0799</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>E9-E2</td>
<td>0.9648</td>
<td>1.0000</td>
<td>0.6792</td>
<td>0.0646</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>E3-E2</td>
<td>0.8484</td>
<td>0.5876</td>
<td>0.7465</td>
<td>0.1781</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Daubentoncia_madagascariensis-E2</td>
<td>1.0594</td>
<td>0.7803</td>
<td>0.6125</td>
<td>0.0552</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Otolemur_garnettii-E9</td>
<td>0.7936</td>
<td>0.5483</td>
<td>0.5667</td>
<td>0.0425</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>E10-E3</td>
<td>0.9397</td>
<td>0.8957</td>
<td>0.6661</td>
<td>0.0608</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>Callithrix_jacchus-E10</td>
<td>0.9809</td>
<td>1.0000</td>
<td>0.6005</td>
<td>0.0202</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>Saimiri_boliviensis-E10</td>
<td>0.9724</td>
<td>1.0000</td>
<td>0.7206</td>
<td>0.1473</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>E11-E3</td>
<td>0.9975</td>
<td>1.0000</td>
<td>0.6412</td>
<td>0.0454</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>E4-E3</td>
<td>0.9242</td>
<td>0.7951</td>
<td>0.6848</td>
<td>0.0799</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>Papio_anubis-E11</td>
<td>0.9648</td>
<td>1.0000</td>
<td>0.6792</td>
<td>0.0646</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Macaca_fascicularis-E11</td>
<td>0.9728</td>
<td>1.0000</td>
<td>0.6253</td>
<td>0.0361</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>Nomascus_leucogenys-E4</td>
<td>0.9319</td>
<td>0.8958</td>
<td>0.6905</td>
<td>0.0800</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>E5-E4</td>
<td>0.9493</td>
<td>0.8977</td>
<td>0.6683</td>
<td>0.0647</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>Pongo_abelii-E5</td>
<td>0.9975</td>
<td>1.0000</td>
<td>0.7022</td>
<td>0.0985</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>E6-E5</td>
<td>0.9493</td>
<td>0.8977</td>
<td>0.6683</td>
<td>0.0647</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>Gorilla_gorilla-E6</td>
<td>0.9648</td>
<td>1.0000</td>
<td>0.6792</td>
<td>0.0646</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>E7-E6</td>
<td>0.9570</td>
<td>0.8978</td>
<td>0.6737</td>
<td>0.0643</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>Homo_sapiens-E7</td>
<td>0.9018</td>
<td>0.6988</td>
<td>0.6683</td>
<td>0.0647</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>E8-E7</td>
<td>0.9493</td>
<td>0.8977</td>
<td>0.6683</td>
<td>0.0647</td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>Pan_paniscus-E8</td>
<td>0.9648</td>
<td>1.0000</td>
<td>0.6792</td>
<td>0.0646</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>Pan_troglodytes-E8</td>
<td>0.9493</td>
<td>0.8977</td>
<td>0.6683</td>
<td>0.0647</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1.5 Effect of excluding random subsets of sites within each enhancer from analysis
Discussion

We have developed a new framework for measuring the strength of natural selection acting on cis-regulatory elements. Taking advantage of massively parallel experimental measurement of the effects of every possible point mutation, our approach can reveal either positive or negative selection acting on either up- or down-regulating mutations. The three enhancers we used as a proof-of-principle have all experienced negative selection against down-regulating mutations in at least some lineages, and one (LTV1) has also been subject to positive selection for up-regulation. Although we do not know the reason for this positive selection, it may involve an advantage of lower levels of cytokines whose mRNAs are degraded by Zfp36 (the target of LTV1) [93]. It is interesting to note that in no case was the phylogeny-wide $K_u/K_n$ significantly less than one, while $K_d/K_n$ was significantly less than one for all three enhancers (in at least part of the phylogeny). This may reflect the tendency for over-expression to be less deleterious than under-expression [94]. Re-examination of this pattern once data are available for more enhancers would be informative.

A key assumption of this approach is a minor role of epistasis, or context-dependence of SNVs. If epistasis between SNVs was widespread, we would have to measure the impact of each SNV in the precise genetic background in which it occurred, a considerably more challenging experiment than mutagenesis of a single enhancer. Epistasis can be quantified in these enhancers because the set of mutant enhancers tested for activity contained not only every possible SNV, but also over 99.999% of all possible pairs of variant sites, only ~0.1% of which exhibit pairwise epistasis in their effects on transcription [37]. Furthermore, even among the epistatic SNVs, only the minority that alter the SNVs’ classification (as up-regulating, down-regulating, or silent)—as opposed to the magnitude of up- or down-regulation—would have any impact on our approach. The rarity of strong epistasis implies that most SNVs will likely have the same effect direction (e.g. up-regulating) whether occurring in the background of (for example) a human enhancer, or the same enhancer in an ancestral primate, and thus epistasis is unlikely to have a significant effect on the overall patterns of selection that we have inferred. However more experiments would have to be performed to establish this definitively.
Although our framework is based on the widely used $K_a/K_s$ approach for protein-coding sequences, it does not suffer from several important limitations of $K_a/K_s$. For example: 1) $K_a/K_s$ reflects an “average” selection pressure across sites; if both positive and negative selection are acting on the same protein, only the more dominant one (typically negative selection) will be apparent (while it is possible to estimate $K_a/K_s$ at single codons or to partition proteins into different selection classes, these approaches require very large numbers of aligned sequences and/or nontrivial assumptions about the distribution of $K_a/K_s$ values among classes [95]). 2) Because we almost never know the functional impact of specific amino acid substitutions, a $K_a/K_s > 1$ is uninformative with respect to what trait natural selection is actually favoring (e.g. higher or lower activity of an enzyme). 3) Synonymous sites are not actually neutral; treating them as such, as is the common practice, inflates estimates of $K_a/K_s$ and can lead to spurious evidence of positive selection. This is not merely a theoretical concern, as synonymous sites have been found to be under negative selection in every species studied to date [96–100].

In contrast, our approach provides improvements in all three of these areas. Specifically: 1) Our metrics are able to detect both positive and negative selection acting simultaneously on different subsets of sites (as exemplified by LTV1), therefore avoiding much of the problem of “averaging” across sites suffered by $K_a/K_s$ applied to entire proteins. 2) Because we know the functional effects of each mutation on enhancer activity, we can infer not only the mode of selection (e.g. positive), but also what the selection is for (e.g. greater transcription). 3) Our metrics do not assume that a certain subset of sites (such as synonymous sites) are neutral, but instead rely on thousands of empirical measurements of the effect of every individual SNV.

It is also informative to compare our approach with previous studies that have scanned genomes for accelerated evolution of non-coding regions [76–79, 101]. Perhaps the two most significant limitations of these previous studies are 1) the lack of a suitable neutral reference with the same mutation rate as each noncoding region of interest [19]; and 2) only regions with the most rapid overall divergence are detected, so that any signature of positive selection can be overpowered by negative selection acting within the same region [19]. Our approach provides a solution to both of these
issues. First, by classifying SNVs as neutral only if they have no measurable effect on transcriptional output, we have a reliable neutral reference that is interspersed with non-neutral sites (and thus should be robust to regional variation in mutation rate), in much the same way as synonymous and nonsynonymous sites in protein-coding regions. Second, by distinguishing between up- and down-regulating mutations, we have the ability to detect both positive and negative selection acting simultaneously on different sites within a single enhancer, as we observed for LTV1. Importantly, this positive selection would have been missed by any approach that simply scans for an overall evolutionary rate faster than neutral, because the negative selection on down-regulating mutations (amplified by the fact that there are over twice as many possible down-regulating mutations as up-regulating mutations) actually leads to an overall rate for LTV1 that is slower than neutral (The neutral-site divergence of LTV1 across rodents is 115 substitutions / 254 possible neutral SNVs = 45.3%, while the overall divergence of LTV1 across rodents is 352 substitutions / 906 possible SNVs = 38.9%). Only by accounting for the direction of each possible mutation’s effect were we able to detect the positive selection that has occurred.

Despite its advantages, there are a number of important limitations of our test. Perhaps most important is the current dearth of saturation mutagenesis data sets that can be used as input for the test [37, 88–90]. However these studies are becoming increasingly straightforward to implement (e.g. no longer requiring access to expensive large-scale oligonucleotide synthesis technology [37]), making them accessible to any investigators. Moreover the fact that all sites within an enhancer need not be analyzed to detect selection (Figure 1.5) suggests that partial mutagenesis is a viable option. A second caveat is that because trans-acting factors can change between species, SNV effects may be species-specific. While this can certainly be an issue at long timescales (e.g. across vertebrates [102, 103]), nearly all human-mouse gene expression divergence has been found to be cis-acting [104], making this a minor concern at the even shorter timescales used here. Third, the SNVs are classified according to their transcriptional effects in the livers of mice raised in the laboratory; whether they may have other effects in different tissues/environments is unknown. Additional limitations related to the saturation mutagenesis data include the lack of
information on indels, and potential effects of SNVs not captured by the experiment (e.g. mutations that influence enhancer activity in the chromosomal but not the plasmid context, or with effect sizes too small to measure). We expect that most of these limitations will be addressed by more comprehensive saturation mutagenesis studies (perhaps targeted toward the indels or SNV combinations observed in nature) in the near future.

Many extensions to this test are possible. For example, intra-species polymorphism data could be incorporated to allow a McDonald-Kreitman framework to be applied [105], which may allow more sensitive detection of positive selection. In addition, SNV effect sizes could be incorporated to potentially increase the power to detect selection. For example, even when the number of up-regulating SNVs observed in a phylogeny is consistent with the neutral expectation, if they are shifted towards very strongly up-regulating, selection for up-regulation may still be detectable [87]. Finally, our approach could also be applied to protein-coding regions that have been subjected to saturation mutagenesis [106–108]. The framework we have introduced here will likely have many other extensions as well, as our ability to determine the effects of mutations in both coding and noncoding regions continues to evolve.

**Subsequent Research**
Since the publication of this work, Agoglia and Fraser 2015 [109] extended the methods used here to attempt to detect regulatory selection on exonic enhancers. In addition for coding for proteins, exons can also impact transcription by encoding regulatory elements. However, it is a matter of debate whether such features confer heightened selective constraint, or evolve neutrally. Angoglia and Fraser expanded on the methods developed here to disentangle the sources of selection acting on exonic enhancers. They were able to determine that the three mammalian exonic enhancers they studied were not under selective constraint for their regulatory function[109]. This is a demonstration of the general utility of the test developed in Chapter 1 of this dissertation.

**Materials and Methods**

*Obtaining present-day enhancer sequences*
Orthologous mammalian enhancer sequences were identified using the sequences from Patwardhan et al. 2012 (ALDOB (hg19:chr9:104195570-104195828), ECR11 (hg19: chr2:16993082-169939701), and LTV1 (mm9:chr7:29161443-29161744)) as BLAST query sequences. Using the NCBI Genomes (chromosome) database, we identified the genomic region from each species with the highest sequence identity to the query (with at least 70% identity), and then confirmed its genomic proximity to the putative target gene. If the enhancer sequence was not available in this database, the Whole-Genome Shotgun Contigs (wgs) database was used. In this case, to determine whether the enhancer sequence was located adjacent to the correct gene, several exons of the putative target gene were also input to BLAST to ensure that these sequences mapped to the same contig.

Ancestral reconstructions

Ancestral reconstructions were performed using the Ancestor v1.1 web server, which implements a context-dependent maximum likelihood substitution inference algorithm [110]. We provided Ancestor v1.1 with alignments of present-day enhancer sequences, including outgroup species not shown in our trees (to improve reconstructions of the most basal nodes; species listed in Table 1.3). Six separate alignments were constructed for each enhancer using different alignment algorithms or combinations of algorithms: 1. PRANK; 2. MUSCLE MSA; 3. tCoffee; 4. ClustalW2; 5. A combination of tcoffee_msa, clustalw_msa, muscle_msa, and clustalw_pair using http://tcoffee.crg.cat/; and 6) a combination of clustalw_pair and lalign_id_pair using http://tcoffee.crg.cat/ [111–114]. Default settings were used for all alignments. Alignments were inspected by eye and a small number of poorly aligned regions were manually adjusted (or in one case, the ECR11 alignment using clustalw_pair and lalign_id_pair, excluded). In addition, the Hominoidea (apes and human) contain a LINE element in ECR11, which we removed from our reconstructions (human sequence:

GAAAAATAGATCAATTTGTTCCTCACTCATAGGTGGGAATTTGAACCAATGAG
AACACATGGACACAGGAAGGGGAACATCACACATCGGGGCCTGTGTGTTG
GTGGGGGAGGGGGAGGGTAGATGCATTTAGGAGATATATCTAACGTAAAT
GACGTGTTAATGGGAGCAGCACACCAAACATGGCACATGATACATATGTA
### Supplemental Table 1.3. Outgroup species used to aid in ancestral reconstruction of each enhancer.

<table>
<thead>
<tr>
<th>ALDOB</th>
<th>ECR11</th>
<th>LTV1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rattus norvegicus</td>
<td>Cavia porcellus</td>
<td>Macaca fascicularis</td>
</tr>
<tr>
<td>Cavia porcellus</td>
<td>Heterocephalus glaber</td>
<td>Gorilla gorilla</td>
</tr>
<tr>
<td>Heterocephalus glaber</td>
<td>Spermophilus tridecemlineatus</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>Oryctolagus cuniculus</td>
<td>Oryctolagus cuniculus</td>
<td>Pan paniscus</td>
</tr>
<tr>
<td>Bos grunniens</td>
<td>Ochotona princeps</td>
<td>Pan troglodytes</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>Bos grunniens</td>
<td>Pongo pygmaeus</td>
</tr>
<tr>
<td>Tursiops truncatus</td>
<td>Bos taurus</td>
<td>Callithrix jacchus</td>
</tr>
<tr>
<td>Equus caballus</td>
<td>Tursiops truncatus</td>
<td>Bos grunniens Bos taurus</td>
</tr>
<tr>
<td>Ailuropoda melanoleuca</td>
<td>Vicugna vicugna</td>
<td>Sus scrofa</td>
</tr>
<tr>
<td>Canis lupus</td>
<td>Equus caballus</td>
<td>Mustela putorius</td>
</tr>
<tr>
<td>Pteropus vampyrus</td>
<td>Pteropus vampyrus</td>
<td>Ailuropoda melanoleuca</td>
</tr>
<tr>
<td>Myotis lucifugus</td>
<td>Myotis lucifugus</td>
<td>Canis lupus</td>
</tr>
<tr>
<td>Loxodonta africana</td>
<td>Dasyus novemcinctus</td>
<td>Pteropus vampyrus</td>
</tr>
<tr>
<td></td>
<td>Trichechus manatus</td>
<td>Myotis lucifugus</td>
</tr>
<tr>
<td></td>
<td>Loxodonta africana</td>
<td>Condylura cristata</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Trichechus manatus</td>
</tr>
</tbody>
</table>

We provided Ancestor v1.1 with an ultrametric phylogeny of all the species included in the alignments (the ‘best dates’ nexus tree from [115]). Branch lengths were rescaled to fall between 0 and 1. For two species missing from this phylogeny we used the most closely related species in the tree as its replacement for our branch length calculations (For Olive Baboon, *Papio anubis*, we substituted *Papio hamadryas*, the Hamadryas Baboon. For the Sumatran Orangutan, *Pongo abelii*, we substituted *Pongo pygmaeus*, the Bornean Orangutan).

For each ancestral node, its six alignments were used to create separate reconstructions, which were then aligned with tCoffee (using default settings). The most frequently observed base in each position was used to generate a consensus, followed by manual curation for ambiguous positions. For a small number of ambiguous positions where there was no consensus, the reconstructions derived from the tCoffee alignments (algorithm #3 above) were given priority. The final consensus reconstructions were used in all subsequent calculations.

### Calculating $K_o/K_n$ and $K_d/K_n$
$K_u/K_n$ and $K_d/K_n$ were determined by applying expression differences due to single nucleotide variants (SNVs) [37] to our ancestral reconstructions to detect evidence of selection (Figure 1.1). Insertions and deletions (indels) were not included in these calculations, as their impact on transcription was not tested [37].

We calculated $K_u$ as the ratio of observed significant ($p < 0.05$, quantifying the probability of the SNV having no effect on transcription [37]) up-regulating SNVs divided by the total possible number of significant up-regulating SNVs in each enhancer (Figure 1.1). An equivalent calculation was performed for $K_d$ and $K_n$ (replacing up-regulating mutations by down-regulating [$p < 0.05$] or neutral [$p > 0.05$]). Varying this $p$-value threshold (e.g. to 0.01 or 0.1) had little effect on our results (Supplemental Table 3). Patwardhan et al. (2012) tested two independent libraries for LTV1, so we used Fisher’s Method to combine the two sets of $p$-values for use in our calculations.

Fisher’s Exact Test was applied to a 2x2 contingency table to determine whether each $K_u/K_n$ or $K_d/K_n$ was significantly different from one (e.g. for $K_u/K_n$, the table columns were up-regulating or neutral, and the rows were observed or not observed in a given branch or phylogeny). For calculations at the level of phylogenies, the counts of each SNV class in each branch were summed.

If our three classes of SNVs (up-regulating, down-regulating, and silent) had different mutation rates, this could affect the estimation of $K_u/K_n$ and $K_d/K_n$. For example, higher mutation rates of neutral SNVs might lead to underestimates of both metrics. To test if this was an issue for the three enhancers studied here, we tabulated the number of transitions and transversions among each class of SNVs within each enhancer (because transitions have a higher mutation rate than transversions). We found no significant difference for any of them ($p = 0.20, 0.47, \text{and} 0.88$ for ALDOB, LTV1, and ECR11 respectively). For enhancers where there is a difference in mutation rates between classes, this could be easily incorporated by adjusting $K_u$, $K_d$, and $K_n$ appropriately.

**Acknowledgements**
We would like to thank the members of the Fraser Lab for helpful discussions. This work was supported by the National Institute of General Medical Sciences (grant
number 1R01GM097171-01A1). JDS is supported by a Stanford Graduate Fellowship and a Genentech Graduate Fellowship. HBF is an Alfred P. Sloan Fellow and a Pew Scholar in the Biomedical Sciences.

References (See bibliography).
Chapter 2: Dissecting the genetic basis of a complex cis-regulatory adaptation

Authors
Santiago Naranjo1*, Justin D. Smith2,3*, Carlo G. Artieri1, Mian Zhang1, Yiqi Zhou1, Michael E. Palmer1, Hunter B. Fraser1#
1Department of Biology, 2Department of Genetics, and 3Stanford Genome Technology Center, Stanford University, Stanford CA.
*These authors contributed equally.
#Corresponding author: hbfraser@stanford.edu

This chapter is an exact reproduction of work published in PLoS Genetics 11(12): e1005751. doi: 10.1371/journal.pgen.1005751. It was a collaborative effort between the dissertation author [JDS] and several members of Hunter Fraser’s group. HBF SN JDS conceived and designed the experiments. SN JDS YZ MZ CGA performed the experiments. SN JDS MZ YZ MEP CGA HBF analyzed the data. HBF and JDS wrote the paper. The dissertation author [JDS] specifically conceived, created and validated the CRISPR:Cas9-activator and gRNA plasmids and performed experiments using them. Additionally the dissertation author [JDS] designed and performed the competitive growth experiments.

Abstract
Although single genes underlying several evolutionary adaptations have been identified, the genetic basis of complex, polygenic adaptations has been far more challenging to pinpoint. Here we report that the budding yeast Saccharomyces paradoxus has recently evolved resistance to citrinin, a naturally occurring mycotoxin. Applying a genome-wide test for selection on cis-regulation, we identified five genes involved in the citrinin response that are constitutively up-regulated in S. paradoxus. Four of these genes are necessary for resistance, and are also sufficient to increase the resistance of a sensitive strain when over-expressed. Moreover, cis-regulatory divergence in the promoters of these genes contributes to resistance, while exacting a cost in the absence of citrinin. Our results demonstrate how the subtle effects of
individual regulatory elements can be combined, via natural selection, into a complex adaptation. Our approach can be applied to dissect the genetic basis of polygenic adaptations in a wide range of species.

**Author Summary**

Adaptation via natural selection has been a subject of great interest for well over a century, yet we still have little understanding of its molecular basis. What are the genetic changes that are actually being selected? While single genes underlying several adaptations have been identified, the genetic basis of complex, polygenic adaptations has been far more challenging to pinpoint. The complex trait that we study here is the resistance of *Saccharomyces* yeast to a mycotoxin called citrinin, which is produced by many other species of fungi, and is a common food contaminant for both humans and livestock. We found that sequence changes in the promoters of at least three genes have contributed to citrinin resistance, by up-regulating their transcription even in the absence of citrinin. Higher expression of these genes confers a fitness advantage in the presence of citrinin, while exacting a cost in its absence—a fitness tradeoff. Our results provide a detailed view of a complex adaptation, and our approach can be applied to polygenic adaptations in a wide range of species.

**Introduction**

Historically, most studies pinpointing the genetic basis of polymorphic traits have focused on protein sequence changes of large effect, because these have been the most amenable to identification. For example, thousands of coding region mutations have been implicated in human diseases with Mendelian inheritance [116]. In contrast, non-coding mutations and polygenic traits have traditionally received far less attention.

However this situation has radically changed with the advent of genome-wide association studies (GWAS). In these studies, millions of SNPs can be tested for statistical association to any trait of interest. Two clear patterns have emerged from hundreds of human GWAS: most traits are highly polygenic, and most associations are
in non-coding regions that are not in linkage disequilibrium with protein-coding changes (and thus cannot be acting via changes in protein sequence) [73, 117, 118]. For example, out of 697 loci associated with height, 86% are non-coding [119]. This suggests that natural selection will overwhelmingly result in polygenic *cis*-regulatory adaptations, composed of variants with very small individual effects, since selection acts on whatever heritable variation is available. Indeed, thousands of loci are involved in selection on height in Europeans, a clear example of polygenic adaptation [120]. Moreover, recent genome-wide comparisons of the proportions of adaptations in coding vs. non-coding regions in humans and sticklebacks support the prevalence of non-coding adaptations [14–16].

Two methods have successfully identified the genes underlying regulatory adaptations. One is QTL/association mapping, which has led to several beautiful examples of single-locus adaptations [121]. However QTL mapping in most species is only practical for loci of large effect, and thus is not well-suited for studying the evolution of complex traits, which are by definition polygenic.

The second method is the “sign test” framework that we and others have developed [20, 122]. The goal of this approach is to identify cases where selection has led to up- or down-regulation of multiple genes via independent mutations. First the *cis*-regulatory divergence between two species is quantified genome-wide via allele-specific expression (ASE) analysis in an F1 hybrid [20]. This results in directionality information for every gene (e.g. for gene X, the species A allele is up-regulated compared to the species B allele). Any group of genes whose expression is evolving under the same selection pressure in the A and B lineages should have a similar frequency of A alleles up-regulated as in the entire genome. For example if 50% of genes with ASE have A allele up-regulation, then any random subset of these should have roughly 50% as well. If a strong deviation from 50% is detected for some gene set—e.g. all 20 genes in a pathway have the A alleles up-regulated—then this indicates the action of lineage-specific selection. Importantly, this approach does not make many assumptions of other methods for detecting selection such as constant population size, lack of epistasis, or neutrality of synonymous sites [20]. And unlike approaches that can rank genes but cannot determine which (if any) of them are
inconsistent with neutrality (such as $F_{ST}$, iHS, and the ratio of expression divergence between species to expression diversity within species), the sign test’s null model allows for confident identification of gene sets under lineage-specific selection [20]. Finally, the sign test is most powerful when many genes are involved, making it uniquely well-suited for studying complex traits.

In the current work, we applied the sign test to ASE data from a hybrid between two species of budding yeast, *Saccharomyces cerevisiae* and *S. paradoxus* (specifically the reference strains S288c and CBS432, hereafter abbreviated *Sc* and *Sp*; note that these abbreviations refer specifically to the two reference strains, and not the two species as a whole). The results led us to focus on a specific mycotoxin (a toxin produced by fungi) called citrinin. Citrinin is produced by a number of Ascomycota fungi, including several species in the *Aspergillus*, *Penicillium*, and *Monascus* genera. It increases mitochondrial membrane permeability and causes oxidative damage via an unknown mechanism, and is a potent nephrotoxin in mammals [123, 124]. Because citrinin is toxic to yeast growing on fermentable carbon sources, which do not require mitochondria, toxicity is likely caused by oxidative damage to other cellular components [123]. The transcriptional response may be an important means to mitigate citrinin toxicity, and in *Sc*, hundreds of genes are induced or repressed in response to even a low level of citrinin [125]. Although citrinin is a common food contaminant, making it a major health concern for both humans and livestock [123, 124], and has been the subject of hundreds of publications, the evolution of citrinin resistance has not been previously investigated.

**Results**

*A sign test reveals selection on the cis-regulation of citrinin-induced genes*

In an effort to identify groups of genes subject to lineage-specific selection, we applied a sign test (described above) to allele-specific gene expression levels from an *Sc/Sp* hybrid grown in YPD (rich glucose) media [126]. Our objective was to identify any gene sets (such as pathways or other functionally related groups) with a biased directionality of ASE, implying the action of lineage-specific natural selection [20]. Testing a collection of publicly available gene sets (see Methods), we found only one
set with a highly significant bias: genes induced by exposure to citrinin, a naturally occurring mycotoxin, were over-represented among genes with Sp-biased ASE. Of 11 genes that were reported to be induced at least 10-fold in response to citrinin (in \textit{Sc}) [125], four were among the top 1% of Sp-biased ASE genes (Fig. 2.1; hypergeometric \( p = 2.7 \times 10^{-6} \)). In fact, these four genes included the first and third most Sp-biased genes in the entire genome. In other words, Sp alleles have markedly higher expression of several genes that are induced in \textit{Sc}’s citrinin response, even though ASE was measured in the absence of citrinin [126].

\begin{figure}
\centering
\includegraphics[width=0.8\textwidth]{fig2_1.png}
\caption{Results of the sign test for selection on \textit{cis}-regulation.}
\end{figure}

For a description of the test and its results, see the main text and Methods. Note that only two citrinin-induced genes are shown, but four were present in the top 1% of Sp-biased genes, and ten in the top 25%.

doi:10.1371/journal.pgen.1005751.g001

In contrast, none of the 11 citrinin-induced genes had even slightly Sc-biased ASE in the \textit{Sc}/Sp hybrid. All 11 had some degree of Sp-biased ASE, and 10/11 were in the top 25% of Sp-biased genes (Fig. 2.1; \( p = 7.8 \times 10^{-6} \)). The complete absence of
citrinin-induced genes among Sc-biased genes, taken together with their 40-fold enrichment among the most strongly Sp-biased genes, is not consistent with a neutral model of gene expression evolution leading to random directionality of ASE [20]; instead, it suggests that lineage-specific selection shaped the cis-regulation of these genes.

*Sp has recently evolved resistance to citrinin*

The results of our sign test (Fig. 2.1) suggested that Sp’s constitutive up-regulation of Sc’s citrinin-induced genes may confer a greater resistance to citrinin. To test this, we measured the growth of Sc and Sp in three concentrations of citrinin: 0 ppm (parts per million), 300 ppm, and 600 ppm. At 300 ppm, Sc showed a clear shift towards slower growth, while Sp was essentially unperturbed (Fig. 2.2A-B). At 600 ppm, Sp was affected, but not nearly as strongly as Sc; the effect on Sp at 600 ppm was roughly similar to that on Sc at 300 ppm. These results are consistent with our hypothesis that Sp would show increased resistance.
Fig 2.2. *Sp* recently evolved resistance to citrinin.
A. Sc growth at different concentrations of citrinin. Three replicates are shown for each condition. B. Sp growth at different concentrations of citrinin. Three replicates are shown for each condition. C. Two metrics of resistance, maximum growth rate and maximum cell density, are shown for 25 *S. paradoxus* strains and Sc. Results are plotted as log-ratios relative to Sc. D. Citrinin resistance treated as a binary trait and plotted on a phylogenetic tree [127]. The most parsimonious evolutionary scenario is shown in red text. Branch lengths are not to scale. doi:10.1371/journal.pgen.1005751.g002

To investigate the evolutionary history of the resistance phenotype, we measured growth rates and cell densities at saturation (“max OD”) for a panel of 25 diverse *S. paradoxus* strains. These strains belong to four major clades, defined on the basis of partial genome sequences [128], which also reflect their geographic origins: European (including Western Russia), East Asian (Eastern Russia and Japan), Hawaiian, and American. We also included Sc as an outgroup. Across all strains, the effect of citrinin on growth rate and on max OD was highly correlated (Pearson $r = 0.91$; Spearman $r = 0.96$), and strains fell into distinct clusters based on their sensitivities (Fig. 2.2C). The most sensitive group included Sc, the Hawaiian strain,
both American strains, and 3/18 European strains. In contrast, the resistant group consisted of all four East Asian strains, and 15/18 European strains (including the two most resistant strains, $Sp$ and Q62.5).

Representing resistance as a binary trait on the $S.\ paradoxus$ phylogeny, the most parsimonious explanation is that resistance evolved after the split between the Hawaiian/American clade and the European/East Asian clade, and was then lost in a subset of the European strains (Fig. 2.2D) (this loss could have occurred via new mutations, or admixture from a sensitive $S.\ paradoxus$ strain; see Discussion). Any other scenario would require multiple independent gains and/or losses. These results indicate the difference between $Sc$ and $Sp$ (Fig. 2.2A-B) is likely due to resistance being gained in a recent ancestor of $Sp$, as opposed to being lost in the $Sc$ lineage.

To determine whether citrinin resistance represents a more general pleiotropic trait—such as resistance to many different toxins—we compared our growth data (Fig. 2.2C) to growth rates of the same strains in 200 diverse conditions [129]. These include several oxidative stress agents (aminotriazole, paraquat, dithiothreitol, CdCl$_2$, and CoCl$_2$) and dozens of other toxins. None of these showed a similar pattern of resistance across strains as we observed for citrinin: the maximum correlation across all 200 conditions was $r = 0.44$ ($n = 23$ strains; not significant after correction for 200 tests). In fact, $Sp$ showed the least resistance to all five oxidative stress conditions (among 22 $S.\ paradoxus$ strains and $Sc$), the opposite of our observation for citrinin. Together, these results suggest that citrinin resistance does not represent a more general resistance to toxins or oxidative stress.

**Candidate genes revealed by RNA-seq in hybrids**

To further characterize the effect of citrinin on gene expression, we performed RNA-seq on $Sc/Sp$ hybrid yeast exposed to 600 ppm citrinin, as well as in rich media lacking citrinin for comparison. This allowed us to measure the effect of citrinin on each gene’s ASE, and thus to measure the cis-regulatory contributions of each species to the induction or repression of each gene. We found strong agreement between our biological replicates ($r = 0.96$-$0.99$; Fig S2.1), and moderate concordance with published microarray data from the $Sc/Sp$ hybrid [126] and $Sc$’s response to citrinin [125] (Fig. S2.1).
To determine if the transcriptional response to citrinin was species-specific, we analyzed the responses of *Sc* and *Sp* alleles separately in our hybrid RNA-seq data. We observed highly concordant responses to citrinin (Fig. 2.3A). In fact, among 114 genes with at least 3-fold induction or repression for alleles from both species, all of them responded to citrinin in the same direction for both (*r* = 0.96 for these genes). As a result, ASE is broadly similar in rich media and in citrinin, as has been reported for the *Sc/Sp* hybrid in other conditions [126].

**Fig 2.3. RNA-seq reveals candidate genes.**

A. Citrinin has largely similar effects on *Sc* alleles and *Sp* alleles in the *Sc/Sp*hybrid; i.e. ASE is similar with or without citrinin. B. Candidate genes were identified based on having strong *Sp*-biased ASE (in YPD; FDR < 0.05) [16], and strong up-regulation in response to citrinin (of both *Sc* and *Sp* alleles; the smaller of these two in our RNA-seq data is plotted; binomial *p* < 10^{-5} for each biological replicate of each allele). doi:10.1371/journal.pgen.1005751.g003
To identify candidate genes that may be contributing to the polygenic selection that we detected (Fig. 2.1), we selected those with the strongest combination of citrinin-induction and Sp-biased ASE. To visualize this, we plotted each gene’s citrinin response against Sp/Sc ASE in the absence of citrinin (Fig. 2.3B). Requiring at least 4-fold citrinin induction and 4-fold Sp-biased ASE, we identified four genes; a fifth gene that was slightly below the ASE cutoff was included as well (Fig. 2.3B, red points). These constituted our top candidates for genes that may be involved in the evolution of citrinin resistance in Sp. Notably, no genes had both a 4-fold ASE bias and citrinin-response in any of the other three possible pairs of directions (the three additional quadrants in Fig. 2.3B), indicating the rarity of this combination.

If citrinin resistance evolved after the split of European/East Asian and American/Hawaiian strains (Fig. 2.2D), we may expect that genes involved in this adaptation would show an ASE bias towards Sp (European) alleles in hybrids between these two clades, as we observed for Sc/Sp. To test this, we performed RNA-seq in a hybrid between Sp and an American S. paradoxus strain (DBVPG6304). We could assess ASE for 4/5 candidate genes; three of these showed Sp-biased ASE (RTA1, binomial \(p = 2 \times 10^{-4}\); FRM2, \(p = 0.05\); CIS1, \(p = 8 \times 10^{-8}\)), while the fourth had strong American-biased ASE (RSB1, \(p = 3 \times 10^{-55}\)). Therefore our prediction of ASE directionality was validated for 3/4 candidate genes.

**Characterization of the candidate genes**

As an initial test of these five candidate genes, we deleted each gene individually from Sp and tested the effect on citrinin resistance (all five were nonessential in YPD). For four of the five gene deletions, Sp resistance was significantly decreased (Fig. 2.4A); the only exception was RSB1, the same gene that did not show Sp-biased ASE in the Sp/American hybrid. These four genes are involved in a variety of functions: a mitochondrial glutathione peroxidase involved in the oxidative stress response (GPX2), an oxidoreductase also involved in oxidative stress response (FRM2), a lipid-translocating exporter of the plasma membrane (RTA1), and a mitochondrial protein of unknown function (YLR346C, which we rename as CIS1, for “CItrinin Sensitive knockout”). The functions of all four genes are
consistent with a role in citrinin resistance: three are involved in mitochondria/oxidative stress, processes directly related to citrinin [123, 124]; and the fourth, RTA1, has been implicated in toxin resistance (Sc strains missing RTA1 are highly sensitive to a mycotoxin called myriocin [130], and over-expression confers resistance to multiple toxins [131]). Considering the four lines of evidence converging on these genes—induction in response to citrinin, constitutive up-regulation of their Sp alleles (in two different hybrids), the effects of their knockouts on Sp citrinin resistance, and functional annotations—we focused further efforts on these four candidates.

A complement to testing whether gene deletion leads to trait loss is to test whether over-expression leads to trait gain. To explore this, we used the CRISPR/Cas9 system, in which the Cas9 protein can be directed to a specific genomic site by use of a guide RNA (gRNA) complementary to the target DNA site [54]. When a transcriptional activation domain is fused to a nuclease-dead mutant of Cas9 (called dCas9), the resulting protein acts as a strong activator of its target genes [59]. This is an attractive system for simultaneous over-expression of multiple genes, since multiple guide RNAs can be delivered to cells on a single plasmid.
Fig 2.4. Effects of candidate gene deletion and over-expression.
A. The five candidate genes were deleted from Sp, and the effect on resistance to 300 ppm citrinin was measured. Four deletions increased sensitivity (t-test p < 0.005 for each), and were studied further. Error bars show 1 S.E. B. The effects of three different dCas9 fusion proteins targeted to four candidate genes were measured by qPCR. C. Illustration of our strategy to over-express four genes via dCas9 fusion protein. D. We used direct competition for 40 generations to measure the relative fitness of the four-gene over-expression strain (with a Gal4-dCas9-VP64 dual fusion) vs. a control strain containing the same plasmid, but lacking any gRNAs. Conditions were YPD, and YPD + 300 ppm citrinin. Error bars show 1 S.E doi:10.1371/journal.pgen.1005751.g004

To test the efficacy of this system, we compared the induction levels of our four candidate genes using three different dCas9-activator fusion genes. One was a previously published fusion to VP64 (a domain derived from a strong viral transcription factor) [59]; the second was a dCas9-VP64 fusion that we created using a dCas9 gene with an alternative codon optimization; and the third was a dual-fusion that we created with VP64 at the C terminus and the Gal4 activation domain at the N terminus. We found that all three dCas9 constructs could induce all four genes, though the level of induction varied substantially between genes, with FRM2 achieving far higher induction than the other three (Fig. 2.4B), perhaps due to its lower basal
expression level. Comparing the three dCas9 fusion proteins, we observed similar induction levels for all three, though our dual-fusion construct achieved slightly (1.2-1.8-fold) higher induction for all four genes. We then used this dCas9 dual-fusion gene to over-express all four genes in a single strain (Fig 2.4C).

To measure the effect of over-expression on fitness, we performed direct competition between the over-expression strain vs. a control (a strain with a dCas9 plasmid lacking any gRNAs), for 40 generations (Fig 2.4D). We incorporated 6-base barcodes into each plasmid, and sequenced these before and after each competition to estimate strain abundances; each strain was tagged by three different barcodes, to reduce any barcode-specific biases. In the absence of citrinin, the over-expression strain had a small but consistent (across 12 replicates) fitness disadvantage of 0.6% per generation (resulting in 20% lower abundance than the control after 40 generations). In contrast, in 300 ppm citrinin, the over-expression strain had a 0.8% fitness advantage (37% higher after 40 generations; t-test p = 3x10^{-5} comparing conditions). These results suggest that over-expression of these genes leads to a condition-specific fitness tradeoff.

**Individual promoters contribute to citrinin resistance**

Cis-regulatory divergence in mRNA levels can be caused by changes at either the transcriptional or post-transcriptional (e.g. mRNA stability) level. We hypothesized that the major causal mutations may have affected transcription via changes in promoter sequences between Sc and Sp. To test this, we replaced the promoter region (~1 kb of noncoding DNA upstream of each gene) of each of our four candidate genes in Sc with the orthologous promoter from Sp (Fig. 2.5A), using an approach that leaves no foreign DNA behind [132]. Boundaries were chosen to overlap stretches of perfect Sc/Sp conservation to ensure correct placement of the Sp sequence in the Sc genome, as well as to facilitate homologous recombination at the target site.
Fig 2.5. Promoter replacement reveals contributions of individual regulatory regions.

A. Illustration of our promoter replacement strategy. Green promoters on the left are from Sc, blue promoters on the right are from Sp. B. RNA-seq in the Sc/Sp hybrid, which measures the overall cis-regulatory divergence between these strains, is in approximate agreement with the effect of the promoter replacements, measured by qPCR ($r = 0.92$). Error bars show 1 S.E. C. After competitive growth for 40 generations, promoter replacement strains for 3/4 candidate genes show a similar pattern of fitness advantage in the presence of 300 ppm citrinin, and disadvantage in its absence. Error bars show 1 S.E. doi:10.1371/journal.pgen.1005751.g005

In each “promoter-replacement” strain we performed quantitative PCR on the downstream gene, to assess the effect on mRNA levels. We found between 1.4 and 6.9-fold up-regulation of the four Sp promoters, compared to the orthologous Sc promoters. Comparing these results to the total cis-regulatory divergence based on our RNA-seq data in the Sc/Sp hybrid (Fig. 2.2A), we found good concordance (Fig. 2.5B,
This suggests that the promoters are likely responsible for most, if not all, of the \textit{cis}-regulatory divergence for these genes between \textit{Sc} and \textit{Sp}.

To test the fitness effects of this \textit{cis}-regulatory divergence, we pooled all four promoter replacement strains together with their \textit{Sc} parent, and grew them in direct competition. We observed a similar pattern for 3/4 candidate genes, with the \textit{Sp} promoter leading to 0.2-0.3\% higher fitness in the presence of citrinin, but \~0.6\% lower fitness in its absence (Fig 2.5C; 18 replicates per condition; t-test $p < 0.01$ for each). The fourth gene, \textit{RTA1}, showed no detectable fitness difference in the two growth conditions ($p = 0.51$). These results suggest an overall similar condition-specific fitness tradeoff of the natural \textit{cis}-regulatory divergence as we observed for the 4-gene over-expression strain, with each of three promoters contributing a small amount to citrinin resistance.

\textbf{Discussion}

In this work, we have discovered a polygenic \textit{cis}-regulatory adaptation, and investigated its genetic basis and effect on fitness. Our results establish that changes in the promoters of at least three genes have contributed to \textit{Sp}'s recently evolved resistance to citrinin, which has a fitness cost in the absence of citrinin. The role of these genes in \textit{Sp}'s citrinin resistance is supported by six lines of evidence: induction in response to citrinin, constitutive up-regulation of their \textit{Sp} alleles (compared to \textit{Sc} and an American strain of \textit{S. paradoxus}), functional annotations, decreased resistance via gene deletion in \textit{Sp}, increased resistance via over-expression in \textit{Sc}, and the fitness effects of promoter-replacements.

Isolating the effects of individual promoter regions has allowed us to gain a deeper understanding of how distinct loci can contribute to polygenic adaptation. For example, the fitness cost of increased expression in the absence of citrinin may explain why these genes are induced in response to citrinin, as opposed to being expressed at constitutively high levels. Despite this cost, \textit{Sp} does express all four genes at somewhat higher levels than \textit{Sc} even without citrinin present (Fig 2.3B), which may confer a net benefit if citrinin is encountered more often by \textit{Sp} than by \textit{Sc}. 

\textit{r} = 0.92).
Alternatively, the fitness costs we observed in Sc may be reduced in Sp by compensatory changes at other loci.

Having identified this gene expression adaptation, a natural question is what selective pressure(s) caused it. Although citrinin represents a plausible candidate, since it is a widespread naturally occurring toxin, it is also possible that something else (e.g. another toxin) was the actual selective agent, with citrinin resistance being a pleiotropic side-effect. Unfortunately there is no experiment that could unambiguously identify the selective agent, since pleiotropic effects of unknown/unmeasured traits are always a possibility; however our finding that citrinin resistance does not correlate with fitness across 200 other growth conditions [129] suggests that it is not a highly pleiotropic trait.

An interesting aspect of our data is the strong correlation between maximum growth rate and maximum cell density across wild isolates (Fig. 2.2C). Similar correlations ($r = 0.85$ and $0.77$) were also observed between these same two variables among other strains of yeast [129, 133]. Ibstedt et al. [133] proposed that the correlations between fitness components of natural isolates are not due to pleiotropic effects, but rather that selection has fixed multiple variants affecting different fitness components in particular lineages. In other words, the presence of a correlation between fitness components may, by itself, be evidence of selection on condition-specific fitness [133]. In the case of citrinin resistance in Sp, this interpretation is quite consistent with our detection of polygenic selection via the sign test.

Another intriguing property of the resistance across S. paradoxus isolates is that resistance is roughly bimodal (Fig. 2.2C). This makes sense if resistance evolved once, and has not changed drastically since then in most of the strains tested here. The exception to this is the three European strains that have lost their resistance, perhaps due to new mutations and/or variants introduced via admixture with sensitive strains. Whether resistance was lost once or multiple times among European strains would be difficult to determine—although these three strains do not cluster together within the European phylogeny [128], this represents only a genome-wide average phylogeny that many loci, including those involved in the loss of citrinin resistance, may not follow.
Our inference that resistance evolved within *S. paradoxus* (Fig. 2.2D) is based on parsimony, i.e. explaining a phylogenetic pattern with the fewest possible transitions. However if loss of citrinin resistance is more likely than gain then a scenario with an ancestral state of resistance, followed by three loss events (one in *Sc* and two in *S. paradoxus*), could be more plausible. If this were the case, the lineage-specific selection that we detected with the sign test would more likely represent selection for lower expression of these genes in *Sc*, to avoid the fitness cost of their expression in the absence of citrinin (Fig. 2.4D and 2.5C). A relaxation of selection (i.e. loss of constraint) is another possible explanation for down-regulation [20], though in this case would not be consistent with either the fitness effects of promoter-replacements (which are large enough to be strongly selected), or the correlation of fitness components (Fig. 2.2C) that suggests selection is acting on citrinin resistance within *S. paradoxus* [133]. Therefore regardless of the ancestral state, the differences we have observed between *Sc* and *Sp* are likely to be adaptive.

In this work, we have only explored one facet of this polygenic adaptation, namely *cis*-regulatory divergence of mRNA levels in a specific environment (YPD media at 30°C). Many other types of evolutionary changes may also be involved, such as changes in protein sequences, *trans*-acting factors, translation, etc. Moreover, it is quite likely that the *cis*-regulatory divergence of additional genes also contributes to *Sp*’s citrinin resistance; for example in our initial analysis (Fig. 2.1), there were five genes with strong citrinin-induced upregulation and moderate *Sp*-biased ASE (in the top 25% of genes) that we did not pursue, as well as many more genes with weaker supporting evidence. Consistent with the idea that additional loci likely contribute, neither the four promoter replacements nor the 4-gene overexpression strain we tested can account for the large difference in resistance between *Sc* and *Sp* (Fig 2.2A-B) if we assume additivity of the promoter-replacement fitness effects and equivalence of the fitness measurements in these two experiments.

For regulatory changes of *RTA1*, for which we could not detect an effect on citrinin resistance, many possibilities exist—e.g. condition-specific effects (gene-by-environment interaction), or genetic background-specific effects (gene-by-gene interaction, i.e. epistasis), or effects that were too small to measure, or it may have no
effect at all. It is clear that even for this one adaptation (as well as every other polygenic adaptation), we are still far from having a complete understanding.

Looking ahead, an important question is what approaches will be most useful for understanding the genetic basis of evolutionary adaptations. QTL mapping has been very effective in localizing large-effect loci down to specific genomic regions, and has led to the identification of several genes underlying adaptations [121]. However it does have limitations. First, an adaptive phenotype must be identified before starting QTL mapping; though in many cases, such as citrinin resistance, we do not know ahead of time which traits may be adaptive. Second, QTL mapping is generally only practical for loci of large effect (with the exception of pooling-based approaches in yeast [134]), which may be rare among adaptations, given the highly polygenic nature of most traits [73, 117, 118]. Third, specific genes are not implicated; QTL peaks typically contain dozens or even hundreds of genes, of which only one may be involved in the trait. And fourth, even for loci of large effect, QTL mapping requires generating, genotyping, and phenotyping hundreds of F2 individuals to result in a reasonably sized QTL peak.

In contrast, the sign test approach we have employed allows us to circumvent these limitations. First, no adaptive phenotype must be known a priori. Second, loci of very small effect can be identified. Third, specific genes are implicated. And fourth, no F2 crosses are required; a single F1 is sufficient to measure all cis-regulatory divergence via RNA-seq. Of course, the sign test is limited to polygenic cis-regulatory adaptations in which one direction of change (up- or down-regulation) dominates; but for these, which may be quite frequent (see Introduction), the sign test can rapidly identify high-confidence candidate genes. Finally, the sign test can be applied to a wide range of species, including fungi, plants, and metazoans [135–145].

A major limitation of the sign test has been that it has previously relied on functional annotations of genes, thus limiting it to species with reasonably well-annotated genomes. However in this work, we found that gene sets and specific candidate genes can be defined on the basis of gene expression data alone (such as the citrinin-response data of ref. 15). Combined with other recent advances in genetic tools, such as CRISPR/Cas9, we believe that pinpointing specific genes and genetic
variants underlying complex, polygenic adaptations is now readily achievable across a wide range of species.
Methods

Sign test for selection on cis-regulation

Allele-specific expression data from the Sc/Sp hybrid in YPD were obtained from [126]. Two genes with strong Sp-biased ASE due to being used as auxotrophic markers in Sc (URA3 and MET17) were removed, leaving 4394 measured genes. The top 1%, 5%, and 25% of genes with the strongest ASE in each direction were submitted to FunSpec, a tool for gene set enrichment analysis [142]. The top enrichment at all three cutoffs for Sp-biased ASE (Hypergeometric $p < 0.01$ after Bonferroni correction for multiple tests) was the Gene Ontology set “response to toxic substance.” Examination of the enriched genes revealed that their annotations were all derived from a single publication, measuring the transcriptional response to citrinin [125]. Thus subsequent analysis (Fig. 2.1) was performed using these induced genes as the set of interest. Induced genes were defined as those with at least 10-fold average induction (Table 1 of ref. [125]), excluding four paralogous genes found to have cross-hybridization [125]. Out of 17 induced genes, 11 had ASE measurements [126] (In decreasing order of Sp/Sc ASE ratio from ref. 16: RTA1, CIS1, RSB1, FRM2, ECM4, FLR1, AAD4, GRE2, GTT2, YML131W, and YLL056C). None of these 11 genes were next to one another in the genome, so their ASE was likely caused by independent cis-regulatory changes, as required by the sign test [20]. No significant ASE directionality bias was observed for citrinin-repressed genes.

Fitness assays

To perform quantitative growth rate measurements (Fig 2.2A-C and 2.4A), we grew strains in 96-well plates and measured OD600 at 12-minute intervals using an automated plate reader (Tecan) until cultures reached saturation. Experiments were performed at 30°C in YPD media [145] (with or without citrinin). Citrinin (Santa Cruz Biotechnology) was dissolved in DMSO at a concentration of 20,000 ppm, prior to addition to growth media. An equal amount of DMSO was added to the no-citrinin controls, as was in each matched experiment with citrinin+DMSO. All growth assays were performed with each strain distributed across the plate (e.g. alternating
rows/columns for the different strains/conditions) to minimize any spatial variation in conditions across the plate.

Maximum growth rate was estimated by performing a linear regression on \( \log_{10}(\text{OD}) \) values vs. time, for every set of 20 consecutive time points (4 hours), and the highest slope was recorded. Maximum OD was calculated across all time points. For clarity of presentation, we show growth rates as log-ratios relative to a fixed reference (\( \text{Sc} \) in Fig. 2.2C, and \( \text{Sp} \) in Fig. 2.4A). For example in Fig. 2.4A, a strain with a 2-fold greater reduction in max growth rate than \( \text{Sp} \) after citrinin exposure would have a value of 1; a 4-fold lower reduction in max growth rate than \( \text{Sp} \) would have a value of -2.

**RNA-seq and ASE analysis**

An \( \text{Sc}/\text{Sp} \) diploid hybrid yeast strain was produced by mating \( \text{Sc} \) (BY4716: \( \text{MAT} \alpha \text{ly}2 \text{ura}3 :: \text{KAN} \)) and \( \text{Sp} \) (CBS432: \( \text{MAT}a \text{ura}3 :: \text{HYG} \)), followed by selection on plates containing hygromycin B and G418 (Sigma). Four single colony picks of the hybrid strain were grown overnight in 3 ml cultures of YPD medium at 30°C. Upon confirming that the cultures were in log phase, two of the replicates were treated with 91 \( \mu \text{l} \) of DMSO, while the other two were treated with 91 \( \mu \text{l} \) of DMSO containing citrinin such that its final concentration in the culture was 600 ppm. Cultures were allowed to continue shaking at 30°C for two additional hours followed by harvesting of the pellets by centrifugation.

RNA was immediately isolated from the pellets using the MasterPure Yeast RNA Purification Kit (Epicenter). Two \( \mu \text{g} \) of RNA from each of the samples was subsequently used to create sequencing libraries using the low-throughput protocol in the TruSeq RNA Sample Preparation Kit v2 (Illumina). Individually barcoded libraries were pooled and sequenced (resulting in a total of \(~156.6\) million single-end 101 bp reads) on a single lane of an Illumina HiSeq 2000 instrument.

Reads were trimmed to 50 bp and mapped to a concatenated reference containing both parental genomes [139] using Bowtie version 0.12, allowing no mismatches [143]. Counts of reads mapping to a high-confidence curated ortholog set [144] were determined using htseq-count with the union option [146]. ASE was
determined as the ratio of reads mapping to the $Sp$ allele divided by those mapping to the $Sc$ allele in the concatenated reference genome. Only genes with at least 20 reads mapping to each allele were retained. All reads are available in the NCBI SRA (accession PRJNA270666), and allele-specific read counts are given in Table S1.

To create the DVBPG6304/$Sp$ hybrid, we started with diploid strains of each parent, and replaced one allele of the $URA3$ gene with an antibiotic resistance gene ($hyg$ in DVBPG6304 and $kan$ in $Sp$). Transformants were selected on the corresponding antibiotic. The diploids were then sporulated to create haploids, and mixed to allow random mating. Hybrids were selected on plates containing hygromycin B and G418. RNA-seq was performed as described above, with ~27.3 million single-end 36 bp reads generated.

**Gene knockouts**

The complete coding regions of five candidate genes were replaced with the $hphMX6$ antibiotic resistance gene via PCR-mediated gene disruption [145] in $Sp$. Transformants were grown on hygromycin B (Sigma).

**Over-expression with CRISPR/Cas9**

Molecular cloning was done with Gibson Assembly [147]. $E. coli$ minipreps were performed with QIAprep Spin Miniprep Kits (Qiagen). Transformation and preparation of competent $E. coli$ DH5α were prepared using Mix & Go $E. coli$ Transformation Kit & Buffer Set and Zymo Broth (Zymo Research). Competent $S. cerevisiae$ (strain BY4741) was prepared either by standard lithium acetate transformation protocols [145] or using Frozen-EZ Yeast Transformation II Kit (Zymo Research).

Two yeast constructs were obtained from [148] via Addgene: p414-TEF1p-Cas9-CYC1t containing human codon optimized *Streptococcus pyogenes* Cas9 under control of the Tef1 $S. cerevisiae$ promoter; and p426-SNR52p-gRNA.CAN1.Y-SUP4t, expressing a guide RNA under control of the SNR52 polIII promoter.

Catalytically inactive Cas9 (dCas9) was made by by introducing the D10A and H840A mutations [54]. We next PCR amplified the Gal4 activator and linker domains
from the pDEST23 plasmid from the ProQuest Two-Hybrid System (Life Technologies), and fused it to the N-terminus of dCas9. We also obtained an alternative dCas9 activator utilizing four repeats of the minimal domain of the herpes simplex viral protein 16 (VP64) fused to the C-terminus of dCas9 [59]. We hypothesized that combining a Gal4 activator domain on the N-terminus with a C-terminal VP64 domain would produce a more potent activator than either activator fusion individually. To control for potential differences due to different codon optimizations, we cloned the VP64 activator domain from ref. [59] onto the dCas9 gene from ref. [148]. We also cloned the dCas9-VP64 activator into the pRS416 vector under control of the Tef1 promoter so that all Cas9 fusions were expressed under the same plasmid and promoter backgrounds. To build the Gal4-dCas9-VP64 activator, we linearized the pRS414-Tef1-nLS-Gal4-dCas9-Cyc1t plasmid and added VP64 to the C-terminal end via Gibson Assembly, followed by cloning into the pRS416 vector background.

We also made a gRNA plasmid from the p426-SNR52p-gRNA.CAN1.Y-SUP4t for cloning new guides by replacing the 20 base specificity sequence for the Can1.Y gRNA with a SacI restriction site. To activate the four candidate genes, we first designed two gRNAs for each gene. Because AG PAM sequences were reported to work as well as GG PAM sequences with dCas9 activators in human cell lines [62], we tested gRNAs with both type of PAM. Guides were selected to target between 50-250 bp upstream of the transcription start site (TSS) of each gene, because this window produced the best activation of other genes as well as in published data [59]. Guides were built from 60 base oligonucleotides, which contained 20 bases of overlap for Gibson cloning on either side of a 20 base unique specificity sequence. These single-stranded oligos were directly cloned into SacI digested pRS425-SacI-gRNA vector via Gibson assembly using an oligo to vector ratio of >100:1. This ratio was required because we found that Gibson Assembly less efficiently integrated single-stranded oligos than double-stranded fragments. PCR and Sanger sequencing were used to screen colonies. Confirmed clones were grown on LB carb media overnight and mini-prepped. These plasmids were then transformed into BY4741 containing the pRS416-dCas9-VP64 and grown on synthetic complete media plates lacking leucine.
and uracil for two days. From these plates, single colonies were selected and grown in liquid synthetic complete media lacking leucine and uracil. Activity of each gRNA was determined by qPCR. Interestingly, we found that none of the gRNAs designed with an AG PAM increased expression levels, suggesting that guides with an AG PAM may not be functional in *S. cerevisiae*.

Once each gRNA was tested by qPCR, we sought to create a construct expressing a combination of guides, one to each of the four target genes. From each individual gRNA plasmid, we amplified the promoter, gRNA, and terminator using primers that produce unique overlaps to assemble the gRNAs in an ordered fashion. These PCR products were then assembled via Gibson Assembly in the absence of vector, followed by a second round of Gibson Assembly into pRS425. By doing a two-step Gibson Assembly, we drastically reduced the rate of misassembled products resulting from the high sequence similarity of the fragments. All constructs were verified by sequencing.

The four-gene over-expression strain in Fig. 2.4C was created by transforming the pRS416-Gal4-dCas9-VP64 plasmid and pRS425 with 4 gRNAs into BY4741. The control strain was identical to this, except using pRS425 with a nonfunctional gRNA.

**Strain barcoding**

Strains were barcoded by transforming a pRS416 vector (either empty pRS416 or pRS416-nGal4-dCas9-cVP64 for the dCas9-activator strains in Fig 2.4D) containing a 6-base barcode. These were constructed by integrating an oligo with a random 6mer in the center flanked by sequences matching our sequencing primers, and sequences matching the vector at the NsiI site for Gibson Assembly. Assembled vectors were sequence verified and then transformed into each strain of interest. All barcodes were at least two nucleotides separated from each other. For each strain, we transformed three unique barcodes. Each barcoded strain was then grown to saturation and mixed evenly with the other strains of that background and frozen as glycerol stocks.

**Competitive growth assays**
Aliquots of the pooled barcoded strains were recovered in YPD media for 4 hours, and then diluted to OD 0.025 for the experiments. Yeast culturing and sample collection was performed using a cell-screening platform that integrates temperature-controlled absorbance plate readers, plate coolers, and a liquid handling robot. Briefly, 700 µl yeast cultures were grown in 48 well plates at 30°C with orbital shaking in either YPD media or YPD + 300ppm citrinin in Infinite plate readers (Tecan). To maintain cultures in log phase, 23 µl of the culture was removed when it reached an OD of 0.76, transferred to a well containing 700 µl of media, and then allowed to grow further. After seven such dilutions, 600 µl of the culture was collected at OD 0.76 and saved in a 4°C cooling station (Torrey Pines). This amounted to approximately 40 culture doublings from the beginning of the experiment. Fresh media transfers were triggered automatically by Pegasus Software and performed by a Freedom EVO workstation (Tecan).

After sample collection, yeast plasmids were purified using the Zymoprep Yeast Plasmid Miniprep II kit (Zymo Research). Purified plasmids were used as a template for PCR with barcoded sequencing primers that produce a double index to uniquely identify each sample. PCR products were confirmed by gel electrophoresis. After PCR, samples were combined and bead cleaned with Sera-Mag Speed Beads Carboxylate-Modified particles (Thermo Scientific). Sequencing was performed using Illumina MiSeq. Reads were counted only if they were a perfect match to the expected 6 bp barcode and flanking sequences.

**Promoter replacement strain construction**

*In vivo* site-directed mutagenesis, known as delitto perfetto [132], was performed as described [62]. Briefly, the Kluyveromyces lactis URA3 gene was amplified using primers containing ~70 bp of homology to each Sc promoter. This PCR product was transformed into Sc, and correct incorporation into each promoter was verified by PCR. Each URA3 gene was then removed by transforming a PCR-amplified orthologous Sp promoter, containing enough flanking DNA sequence (identical between Sc and Sp) to allow specific targeting of the PCR product. Because the efficiency of delitto perfetto is maximized when transforming longer DNA
molecules, as well as double-stranded DNA [33], transforming long PCR products (as opposed to shorter, single-stranded synthetic oligonucleotides) is a useful modification. Counter-selection of the resulting transformants on 5-FOA (Sigma) allowed isolation of successfully engineered strains that had replaced each URA3 gene with the correct Sp promoter, which were then sequence-verified.

Quantitative PCR

To measure the expression levels of our four candidate genes (Figs. 2.4C and 2.5B), we followed our previously published methods for qPCR [17]. We first grew each strain in YPD, and harvested them in log-phase (OD$_{600}$~1) by centrifugation. We extracted total RNA with the MasterPure Yeast RNA Purification Kit (Epicentre), and quantified them with a NanoDrop2000 spectrophotometer. Total RNA samples were diluted to a concentration of 500 ng/µL and then reverse transcribed into cDNA with SuperScript III RT (Invitrogen), following manufacturer protocols. cDNA was diluted 1:100 prior to qPCR. qPCR was performed on an Eco Real-Time PCR machine (Illumina) following manufacturer’s protocols. To quantify changes in mRNA abundance, six control genes previously noted for their stability across conditions [149] were measured in each experiment: ACT1, TDH3, ALG9, TAF10, TFC1, and UBC6. All measurements were performed in triplicate. Data were analyzed using qBase Plus software (Biogazelle) [150].

Acknowledgements

We thank members of the Fraser Lab and the Stanford Genome Technology Center for discussions and advice.

References:
See Bibliography

Supporting Materials.
See Appendices
Chapter 3: Quantitative CRISPR interference screens in yeast identify chemical-genetic interactions and new rules for guide RNA design

Authors
Justin D. Smith, Sundari Suresh, Ulrich Schlecht, Manhong Wu, Omar Wagih, Gary Peltz, Ronald W. Davis, Lars M. Steinmetz, Leopold Parts, and Robert P. St.Ongé

1 Stanford Genome Technology Center, Department of Biochemistry, Stanford University, 3165 Porter Drive, Palo Alto, CA 94304
2 Department of Genetics, Stanford University School of Medicine, Stanford, California, USA.
3 Department of Anesthesia, Stanford University School of Medicine, Stanford University, Stanford, California 94305, USA.
4 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Genome Campus, CB101SD, United Kingdom
5 European Molecular Biology Laboratory (EMBL), Genome Biology Unit, 69117, Heidelberg, Germany
6 Corresponding authors: Leopold Parts, Genome Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117, Heidelberg, Germany
Phone: +44 1223 834244
Email: leopold.parts@sanger.ac.uk
Robert P. St.Ongé, Stanford Genome Technology Center, Department of Biochemistry, Stanford University, 3165 Porter Drive, Palo Alto, CA 94304
Phone: 650-721-2976
Email: bstonge@stanford.edu

7 Current address: Wellcome Trust Sanger Institute, Hinxton, CB101SA, United Kingdom
Phone: +44 1223 834244

This chapter is a reproduction of work that is under a final round of revision with Genome Biology. The dissertation author [J.D.S.] along with, L.P., & R.P.S. conceived the experiments, analyzed the data, and wrote the paper. J.D.S., S.S., U.S., & R.P.S. performed experiments. M.W. & G.P. conceived and performed the metabolite profiling experiments. O.W. built the webtool. L.M.S. & R.W.D. provided valuable insight and advice.
Abstract

Background:

The CRISPR/Cas9 toolbox for genome editing has recently been expanded to include approaches for modulating gene expression. Genome-scale CRISPR interference (CRISPRi) has been used in human cell lines; however, the features of effective guide RNAs (gRNAs) in different organisms have not been well characterized. We sought to define rules that determine gRNA effectiveness for transcriptional repression in *Saccharomyces cerevisiae*.

Results:

We created an inducible single plasmid CRISPRi system for gene repression in yeast, and used it to analyze fitness effects of gRNAs under 18 small molecule treatments. Our approach correctly identified previously described chemical-genetic interactions, as well as a new mechanism of suppressing fluconazole toxicity by repression of the *ERG25* gene. Assessment of multiple target loci across treatments using gRNA libraries allowed us to determine generalizable features associated with gRNA efficacy. Guides that targeted regions with low nucleosome occupancy and high chromatin accessibility were clearly more effective. We also found that the best region to target gRNAs was between the transcription start site (TSS) and 200bp upstream of the TSS. Finally, unlike nuclease-proficient Cas9 in human cells, the specificity of truncated gRNAs (18 nt of complementarity to the target) was not clearly superior to full-length gRNAs (20 nt of complementarity), as truncated gRNAs were generally less potent against both mismatched and perfectly matched targets.

Conclusions:

Our results establish a powerful functional and chemical genomics screening method and provide guidelines for designing effective gRNAs, which consider chromatin state and position relative to the target gene TSS. These findings will enable
effective library design and genome-wide programmable gene repression in many genetic backgrounds.

**Keywords**
CRISPRi, dCas9, yeast, *Saccharomyces cerevisiae*, nucleosome, chemical-genetic interactions

**Background**

The bacterial type II CRISPR (clustered regularly interspaced palindromic repeats) associated Cas9 nuclease can be targeted to DNA using an engineered guide RNA (gRNA), enabling genome editing in a variety of organisms[54–57]. The Cas9 protein can be further modified to act as a programmable effector. Two point mutations can yield a catalytically dead Cas9 (dCas9)[54] which alone can serve as an effective programmable transcriptional repressor in bacteria[151]. With further modification, dCas9 can be made to function as a transcription activator or repressor (aka CRISPR interference, or CRISPRi) capable of modulating gene expression in eukaryotes[58–62], including in *Saccharomyces cerevisiae*[58, 59]. One of the advantages of CRISPR/Cas9 over previous methods of genome engineering such as Transcription Activator Like Effector Nucleases (TALENs) and Zinc Fingers is the compatibility of the specificity-determining region of the gRNA (generally 20 bases in length) with highly-parallel array-based oligonucleotide synthesis. Thus, large libraries of gRNAs can be readily synthesized and cloned for functional genomic or genome editing applications. Several groups have taken advantage of this, and generated genome-wide libraries for knocking out [63–66], silencing[34], and activating genes[34, 67].

The tremendous potential of the CRISPR/Cas9 system has motivated efforts to better understand factors that influence its efficacy. Gilbert *et al.* [34] characterized
the ideal genomic region to target gRNAs for effective repression in K562 human myeloid leukemia cells. They found CRISPRi worked best using gRNAs that direct dCas9-KRAB to a window of -50bp to +300 bp relative to the transcription start site (TSS) of a gene, with a maximum effect observed in the 50-100bp region just downstream of the TSS[34]. It is currently not known if these rules for guide positioning apply to other cell lines or organisms. Further, not all gRNAs targeted to this window functioned equally well, and therefore additional factors likely influence efficacy.

Understanding and limiting the off-target activity of CRISPR/Cas9 is also important for most applications of the system. Several groups have demonstrated that CRISPR/Cas9 can tolerate some mismatches between the gRNA and the target, indicating potential to cut or bind unintended sites[62, 71, 72, 152–154]. One strategy that has proven effective in preventing off-target cutting in human HEK293 and U2OS cells is to truncate the gRNA’s region of target site complementarity from 20 nt to 17 nt or 18 nt [71, 155]. The specificity of these truncated gRNAs has only been tested in human cells, however, and only with nuclease-proficient Cas9.

Here, we present a versatile platform for high-throughput characterization of CRISPR/Cas9 gRNA libraries in *Saccharomyces cerevisiae*. Informed by existing chemical-genomic data, we designed and tested gRNAs directed to twenty genes whose expression was predicted to influence sensitivity to specific small molecule inhibitors of growth. Repression of these genes by dCas9-Mxi1 indeed produced quantifiable and drug-specific growth defects, which we then used to assess a variety of factors potentially influencing efficacy and specificity. We evaluated the effect of genome position, chromatin accessibility, nucleosome and transcription factor occupancy of the target site, as well as the length, sequence, and secondary structure of the gRNA. While our major goal was to determine rules predictive of CRISPR/Cas9 function in yeast, our experiments also revealed surprising biological insights, including a novel cellular mechanism for resistance to the antifungal drug
fluconazole. Collectively, our results advance the development of CRISPRi as a powerful approach for functional genomics.

Results

Single plasmid system for CRISPR interference in yeast

We designed and constructed a plasmid for regulatable CRISPRi in yeast (Figure 1A). The plasmid is a derivative of pRS416[156], which contains a yeast centromeric origin of replication and a *URA3* selectable marker. To this backbone, we added the complete open reading frame (ORF) for catalytically inactive *Streptococcus pyogenes* Cas9 (dCas9) to which the Mxi1 transcriptional repressor was fused at the C-terminus[58]. We also added the tetracycline repressor (tetR) ORF, a tetO-modified RPR1 RNA polymerase III promoter[59, 157], a NotI restriction site, and common gRNA sequence. The NotI site enables rapid cloning of short oligonucleotides encoding the target complementarity region of the RNA guide. In this system, TetR and dCas9-Mxi1 are constitutively expressed from the *GPM1* and *TEF1* promoters respectively, whereas the gRNA is inducibly expressed by addition of anhydrotetracycline (ATc) to the growth medium (Figure 1B).

To validate this system as a rapid and versatile approach for transcriptional silencing in yeast, we designed gRNAs targeting the *ERG11, ERG25, CRG1*, and *SEC14* genes. Previous work has demonstrated that these four genes are haploinsufficient in the presence of the small molecule inhibitors fluconazole, 1181-0519, cantharidin, and 4130-1276[158, 159], respectively. Thus, we reasoned that transcriptional repression by dCas9-Mxi1 should produce a growth defect in the presence of the appropriate chemical compound. Based on previous small scale studies[59], guides targeting regions near the TSS of each gene were synthesized, and inserted into the NotI site of our expression construct (Methods). The growth rates of transformants were then measured in increasing concentrations of the appropriate compounds, in both the presence and absence of ATc (Methods). In all four cases, and as expected, ATc-induced expression of the gRNA resulted in increased small
molecule sensitivity relative to the empty-vector control (Figure 1C, Additional file 3.1).

Figure 3.1. Overview of regulatable CRISPRi and individual strain analysis.
A) Schematic of expression construct for regulatable CRISPRi in yeast. Key features include ORFs expressing dCas9-Mxi1 and the tetracycline repressor (TetR), as well as a tetracycline inducible gRNA locus containing the RPR1 promoter with a TetO site, a NotI site for cloning new gRNA sequences encoding target complementarity, and the constant part of the gRNA. B) In the absence of
anhydrotetracycline (ATc) TetR binds the gRNA promoter and prevents PolIII from binding and transcribing the gRNA. This in turn prevents dCas9-Mxi1 from binding the target site. In the presence of ATc, TetR dissociates and gRNA is expressed, allowing dCas9-Mxi1 to bind its target locus, and repress gene expression. C) CRISPRi-induced drug-sensitivity. Transformants expressing gRNAs directed against CRG1, ERG11, ERG25, and SEC14 (indicated above each panel), were grown in the presence of a specific small molecule (i.e. cantharidin, fluconazole, 1181-0519, and 4130-1276, respectively). Growth of the gRNA-expressing strain and the empty vector control, in the presence and absence of ATc (see legend), is plotted in each panel. Growth relative to the ‘no-drug’ control is indicated on the y-axis (see Methods), and was measured at increasing concentrations of each small molecule (x-axis).

We characterized the system further and showed that ATc-dependent repression was titratable by addition of increasing concentrations of ATc to the culture (Figure S3.1). Quantitative PCR (qPCR) analysis of transcript levels revealed rapid repression within ~2.5 hours following ATc addition, but slow reversibility. Repression levels varied among the gRNAs assayed with the most effective gRNA repressing transcription roughly 10 fold. Even though we observed modest ATc-independent small molecule sensitivity for one of four gRNAs (CRG1) in Figure 3.1C (possibly indicating leaky expression), qPCR analysis of CRG1 transcript levels did not reveal significant gene repression in the non-induced (-ATc) condition (Figure S3.1, Additional file 3.2). Thus, the collective data are consistent with strong transcriptional control of the guide.

**High-throughput CRISPRi via gRNA Library Screening**

Akin to the DNA barcodes used for the yeast deletion collection[160, 161], the short specificity-determining regions of gRNAs (i.e. the sequence complementary to the target) can act as unique identifiers of individual strains. Like barcodes, these can be readily quantified using next-generation sequencing[162, 163], thereby enabling high-throughput strain phenotyping following competitive growth in pooled cultures. By taking advantage of this, and inexpensive array-based oligonucleotide DNA synthesis, we sought to establish a quantitative assay for guide efficacy, with the goal of uncovering generalizable rules for effective use of CRISPRi in yeast (Figure S3.2). In total, we created and tested five gRNA libraries comprised of a total of 989 unique gRNAs (Additional file 3.6), in the presence of various small molecule inhibitors of growth (Additional file 3.7). The guide counts following competitive growth were
highly reproducible between biological replicates, indicating that the assay is robust (Figure S3.3A-B, Additional file 3.4, Additional file 3.5).

Figure 3.2. Parallel analysis of CRIPSRi-induced fitness defects in pooled cultures.
A) Effect of gRNA expression on growth. gRNA sequencing counts following growth in induced (+ATc) vs uninduced (-ATc) conditions were used to calculate the ATc-effect ($A_0$) for each gRNA, which were median-centered and plotted on the y-axis. Each point represents a gRNAs directed against one of 20 different target genes (gene_tiling20 library). gRNAs are color-coded and arranged
alphabetically on the x-axis by target gene. In the plot, \( A0 \) was constrained between 4 and -4. B) Effect of small molecules on detecting gRNA-induced growth defects. For each gene target (x-axis), the number of gRNAs inducing a growth defect (median-centered \( A < -1 \)) in standard conditions, and in the presence of its paired reference small molecule is plotted on the y-axis (see legend). C) Fluconazole-specific growth defects (y-axis) are plotted for each gRNA (x-axis), which are color-coded and arranged alphabetically by target gene. The drug/gene pair representing the reference chemical-genetic interaction is highlighted in grey. D) Drug-specific effects for the \( ERG11 \) gRNA set in 25 different drug conditions (x-axis). Points are color-coded by condition. Large black dots represent the mean in each drug condition, and are colored red if >1 or if <-1. In C and D, drug-specific effect (\( D \)) was constrained between 4 and -4. E) Heatmap illustrating the average drug-specific effect for each guide set (y-axis), in each condition (x-axis). A guide set refers to the group of guides directed against the same gene. Drug-sensitivity is indicated in red, drug-resistance in blue. Previously defined chemical-genetic interactions are arranged on the diagonal and are outlined in green. Triangles above indicate cases where the same compound was assayed at increasing concentrations.

We first tested a library (i.e. the ‘gene_tiling20’ library) of 238 guides targeting protospacer adjacent motif (PAM)-containing positions (on both the template and non-template strands) between 150bp upstream of the TSS and +50bp relative to the ORF start of 20 different genes. These 20 genes included the four described in Figure 3.1C, and 16 others that have a specific small molecule partner that when added to a culture at the correct dose, will render that gene haploinsufficient[158]. Repression of a target gene is expected to increase sensitivity to its specific small molecule partner, but in general, not to other compounds under study. We leveraged this “reference set” of chemical genetic interactions (Additional file 3.3) to benchmark the assay and assess gRNA effects.

All but one (\( CRG1 \)) of the genes targeted by the ‘gene_tiling20’ library are essential for viability. Consistent with effective repression of an essential gene, several guides exhibited fewer sequencing reads following growth in the presence of ATc, compared to that in the absence of ATc (Figure 3.2A). Notably, however, the majority of gRNAs targeting the 19 essential genes did not elicit a growth phenotype. Raw sequence read information, and the ATc-induced fold change (\( A \); see Methods) of each gRNA, are listed in Additional file 3.8 and Additional file 3.9, respectively.

We challenged this library with 18 different small molecules from our reference set (Additional file 3.3), and consistently observed that addition of a small molecule to the culture increased the number of gRNAs causing growth defects. Specifically, in addition to those gRNAs that inhibit growth by virtue of their potent repression of an essential gene, additional gRNAs targeting genes known to be
haploinsufficient for the added compound became depleted following competitive growth (Figure S3.4A). For example, although no guides directed against the essential gene *ERG11* inhibited growth (median-adjusted $A < -1$) under standard conditions, six *ERG11* guides produced growth defects when yeast were cultured in 20µM fluconazole, an antifungal drug that inhibits the Erg11 protein (Figure 3.2B). These results were representative of the other compounds tested. In each case, the addition of compound increased assay sensitivity, allowing the effects of guides that only weakly modulate transcription to be detected (Figure 3.2B). Nonetheless, not all guides produced a growth defect even when induced in the presence of their partner small molecule. Factors influencing guide efficacy are explored in detail later in the manuscript.

**Exploring small molecule mechanism-of-action**

To specifically explore small molecule mechanism-of-action (MoA), we calculated the “drug-specific effect” ($D$; see Methods) on each strain by comparing induced (+ATc) cultures grown in the presence of a small molecule, to those grown without the small molecule. This drug-induced fold change metric identifies only those genes that are dosage sensitive to the test compound. These genes are powerful descriptors of a compound’s MoA. For example, comparing relative guide counts following growth +/- 20µM fluconazole primarily identifies strains in which *ERG11* is repressed as sensitive to fluconazole (Figure 3.2C). Moreover, *ERG11*-repressor strains were, for the most part, not affected by the other compounds assayed (Figure 3.2D). These results were representative of the other, previously-defined, chemical-genetic interactions that comprised our reference set (Additional file 3.10, Figure S3.4B-C).

Collective analysis of strains expressing gRNAs directed against the same gene further verified that the small molecules tested specifically affected strains predicted by our reference set (Figure 3.2E, Additional file 3.11). Interestingly however, we also observed several examples where a small molecule affected the growth of a strain not predicted beforehand (off-diagonal red signal in Figure 3.2E). As this figure
reports the average drug-specific effect ($D$) on strains in a set, off-target binding by gRNAs is an unlikely explanation for the unexpected signal. Indeed, several lines of evidence suggest many represent *bona fide* chemical-genetic interactions. In cases where the same compound was tested at multiple concentrations, these interactions were reproducible and dose-dependent. The compound 9125678 was particularly interesting, as it inhibited growth of strains repressing *ERG11, ERG13*, and *ERG25*. All three genes encode components of the yeast ergosterol pathway, suggesting a mechanism of action to be tested in future experiments.
Figure 3.3. Erg25 regulates fluconazole sensitivity.
A) Strains containing gRNA constructs directed against ERG3, ERG11, ERG25, and the empty vector control (indicated above each panel), were grown in 63.1 μM fluconazole, in either the presence (blue), or absence of 250 ng/μl ATc (red). In each panel, optical density (OD) is plotted on the y-axis, as a function of time on the x-axis. Loss of ERG3 function was previously shown to confer fluconazole resistance[164],[165] and served as a positive control. B) Similar to (A). The parental BY4741 strain was grown in 60 μM fluconazole (red), 60 μM fluconazole + 40 μM 1181-0519 (blue), or no drug (black).
We also found that the growth-inhibitory effects of fluconazole were reduced in multiple ERG25-repressor strains, suggesting that ERG25 repression confers resistance to fluconazole. Growth assays in isogenic cultures confirmed this observation (Figure 3.3A, Additional file 3.12). Similar results were obtained via chemical inhibition of the Erg25 protein with 1181-0519. This compound, predicted by metabolomic profiling to inhibit Erg25 (Figure S3.5 and Additional file 3.13), increased growth of the control strain (BY4741) in the presence of fluconazole (Figure 3.3B). Interestingly, CRISPRi-mediated repression of ERG25 caused a ~2.5 fold increase in ERG11 transcript levels, thus providing a likely explanation for the observed fluconazole resistance (Figure S3.1B and S1D, Additional file 3.2).

**Efficacy and specificity of full-length and truncated gRNAs**

Having validated our overall approach and reference set, we next evaluated factors influencing gRNA performance. To this end, we focused on the growth inhibitory effects of gRNAs specifically in the presence of their partner chemical compound. Experiments in human cell lines have demonstrated that off-target cutting by Cas9 can be mitigated by reducing the length of the gRNA’s target complementarity from 20 nt, to either 17 or 18 nt [71, 155]. To assess effects of gRNA length on CRISPRi in yeast, we created an 18 nt version of our gene_tiling20 library described above (gene_tiling18), and assayed it under the same conditions. The growth effects resulting from both versions of each gRNA were generally consistent, with many full-length and truncated versions of a gRNA exhibiting similar effects (Figure 3.4A). We found, however, that full-length gRNAs tended to produce stronger phenotypes more often: for example, 94 of 182 full-length but only 73 of 182 truncated gRNAs resulted in growth defects (gRNA effect < -2, solid grey lines in Figure 3.4A).
Figure 3.4. Quantitative comparison of full-length and truncated gRNAs.
A) gRNA effects (see methods) of 182 full-length (20 nt of complementarity to the target) gRNAs (y-axis), and their truncated (18 nt of complementarity) versions (x-axis). Each point represents a gRNA-expressing strain which was grown with the appropriate reference small molecule. Dotted and solid grey lines demarcate gRNA effects of 0 and -2, respectively. B) Heatmaps illustrating growth defects induced by gRNAs containing different mismatches to the target sequence. Full-length and truncated gRNAs are arranged by target gene on the y-axis. The reference small molecule is labeled on the right.
The mismatch position of each gRNA relative to the PAM is indicated on the x-axis (gRNAs matching the target sequence perfectly are on the far left). Missing values are indicated with an X. C) As in A, only the mismatched gRNAs described in B are plotted. Points are color-coded based on target gene. Large points represent the ‘perfect’ gRNAs, all other points represent gRNAs containing mismatches.

To compare the specificity of gRNAs with 18 nt and 20 nt of target complementarity, we selected a single functional guide for ERG11, ERG25, CRG1, and SEC14, and designed a series of derivatives containing one, two, or three mismatches to the target sequence (24 in total for each target gene, for both truncated and full-length gRNAs). gRNA-induced sensitivity to the appropriate small molecule was assayed and, as expected, expression of the ‘perfect’ gRNA resulted in sensitivity (Figure 3.4B, Figure S3.6). As previously reported[72, 152, 154], we found that mismatches located in the seed region (i.e. positions 1 – 10 relative to the PAM) were poorly tolerated by both full-length and truncated gRNAs (Figure 3.4B, Figure S3.6). In general, gRNAs containing mismatches in this region had reduced efficacy (i.e. they did not yield growth defects), while mismatches in the distal region (positions 11-20) had little influence on efficacy. Plotting the effects of full-length and truncated gRNAs against each other (Figure 3.4C) revealed that gRNAs with 20 nt of complementarity tended to be more effective repressors than those with 18 nt. Importantly however, this was true of both the perfect and mismatched gRNAs. Thus, considering their reduced efficacy against perfectly matching target sequences, truncated gRNAs did not exhibit a marked improvement in specificity compared to their full-length counterparts.

gRNA efficacy depends on accessibility and location of the target region

As illustrated above, different gRNAs directed against the same gene can have a range of efficacies (Figure 2A-D). We tested whether the effective target window reported for CRISPRi in human cell lines contributes to this variability[34]. To do so, we created a library of 383 full-length guides targeting -500bp to +500bp of the TSS region of five genes (broad_tiling library), and challenged this library with four chemical compounds. Combining these data with those from the gene_tiling20 library above, we found that the median guide effect was maximal in the window of -200bp to TSS, while guides downstream of the TSS, or further than 300bp upstream of the TSS
were less effective (Figure 3.5A). Effective repression outside the -200bp to TSS window did occur, but less frequently.

**Figure 3.5. Effect of target location and accessibility on gRNA efficacy.**

In all plots gRNA efficacy was measured in the presence of the appropriate reference small molecule. A) gRNA effect magnitude (absolute value of gRNA effects that were censored to have a maximum of 0) is plotted on the y-axis, against target position relative to the TSS on the x-axis. The median of gRNA effects in 50bp windows (solid line, big markers), overlapping by 25bp, indicates a region of 200bp immediately upstream the TSS as effective. B) gRNA effect magnitude is plotted on the y-axis, against target position (gRNA midpoint) relative to the TSS on the x-axis for four loci (indicated above each plot). gRNAs targeting template and non-template strands are indicated with "o" and "x", respectively. Standard deviation estimates are indicated with blue lines, and maximum estimated gRNA effect magnitude for each target locus is given as a grey dashed line. Nucleosome occupancy (red line), and smoothed ATAC-seq read density (green line) relative to the region maximum are scaled to the maximum effect magnitude. C) gRNA effects (y-axis) are plotted (black dots) against nucleosome occupancy score (x-axis, left) and ATAC-seq read density (x-axis, right). The median of gRNA effects in windows of 0.25, overlapping by 0.125, is indicated by the circular blue markers. The blue bars show the first and third quartiles. The Spearman correlation for the relationship with nucleosome density is 0.34, p-value = 9.6x10^{-12}. The Spearman correlation for the relationship with normalized ATAC-Seq is -0.35, p-value = 2.2x10^{-12}
The asymmetry of guide effectiveness around the TSS, and the variability between closely positioned guides indicate that absolute distance to the TSS is not the only determinant of efficacy. As yeast promoters are known to be nucleosome-free, with strictly positioned nucleosomes following the TSS[166–168], we hypothesized that chromatin accessibility and nucleosome occupancy play a role in guide efficacy. We extracted nucleosome occupancy and average chromatin accessibility scores from yeast ATAC-seq data[169], and plotted these data with guide effects in Figure 3.5B. A positive relationship between chromatin accessibility and gRNA efficacy was most apparent for the SEC14 and SSL2 loci. We next systematically quantified the influence of accessibility on guide efficacy. In the TSS -400bp to TSS+400bp window, guides targeting nucleosome-free and ATAC-seq accessible regions were more effective (Figure 3.5C). The relationship with ATAC-seq read density persists in the typically nucleosome-occupied region of TSS to TSS+400bp, suggesting that accessibility influences efficacy independently of positioning relative to the TSS (Figure S3.7A). Similar results were obtained when our data were compared to other genome-wide nucleosome position data[168] (Figure S3.7B and Additional file 3.14).

We tested a range of additional potential determinants of guide efficacy (Additional file 3.15). We first considered the sequence context of the target, and found no specific base pairs that were significantly correlated with gRNA efficacy (Figure S3.8). Next, we used data from Reimand et al[170], to seek transcription factors whose known or inferred presence in the target region is correlated with guide potency. We found a small number of cases where overlap with a transcriptional activator binding site correlated with stronger guide effects (Additional file 3.16). Finally, we did not observe a strong effect of RNA secondary structure or melting temperature on gRNA efficacy (Figure S3.9).

Our results primarily identify position relative to the TSS and chromatin state as important determinants of whether a gRNA will enable robust transcriptional repression by dCas9. For example, 39% (171/442) of full-length gRNAs (which targeted regions +/-500bp from the TSS) exhibited effective repression (gRNA effect
< -2) in our assay. On the other hand, gRNAs that target the 200bp region immediately upstream of the TSS and a nucleosome-depleted region, were effective 76% (59/78) of the time. Even though additional factors may determine whether a specific gRNA will be a strong transcriptional modulator, applying these two criteria will likely improve performance of future gRNA libraries. We have created a webtool (http://lp2.github.io/yeast-crispr/) to enable rapid design of gRNAs for effective CRISPRi in yeast.

**Discussion**

We demonstrated that CRISPRi with inducible gRNA expression is a useful and effective tool to repress genes in yeast. In particular, CRISPRi provides a good alternative to other approaches for studying essential genes[171–174]. CRISPRi constructs can be readily transformed into existing knock-out, GFP-tagged, or other collections, thus enabling genome-wide effects of repressing a particular gene to be characterized. We further showed that CRISPRi, paired with complex gRNA libraries, can be used in competitive growth assays for functional and/or chemical genomic screens. Additionally, CRISPRi plasmid libraries can easily be transformed into any number of different strain backgrounds.

Consistent with CRISPRi being specific for the intended target, increased small molecule sensitivity of gene knockdown strains was largely confined to the predicted gene/drug combinations (Figure 3.2C,3.2D,S3.4B,S3.4C). We also analyzed our gRNAs for possible sites of off-target binding using ECRISP [175], and found very few sites that could potentially result in a growth defect (by repression of an essential gene), and none in the TSS region of the 20 genes we focused on (Additional file 3.18). Nevertheless, it is difficult to know if, or to what extent, off-target binding occurred in our experiments. By employing multiple guides directed against the same target however, one can be more confident that phenotypes observed with multiple independent guides are due to repression of the intended target and not off-target repression. Using this strategy, we uncovered and confirmed a novel chemical-genetic interaction, where Erg25 repression results in resistance to the common antifungal drug, fluconazole.
In our yeast data, truncated gRNAs do not greatly reduce mismatch tolerance when used with the dCas9-Mxi1 repressor. This result is inconsistent with the findings from human cell lines using nuclease-proficient Cas9[71, 155], and implies one of several possibilities. First, truncated gRNAs could be effective in reducing mismatch tolerance in human cells, but not in yeast. Alternatively, mismatched truncated guides may reduce Cas9’s ability to cleave when compared to equivalent mismatched full length guides, but not its ability to bind to target sequences. dCas9-Mxi1 may only need to bind to the target site to induce transcriptional repression. It is therefore possible that nuclease-proficient Cas9 is still recruited to mismatched target sites by truncated gRNAs but is no longer able to cleave its target. Further studies are required to test these hypotheses. Additionally, we observed that truncated gRNAs that are a perfect match to their target are generally less potent than their full-length counterparts. Thus, we found no clear advantage in using truncated guides for CRISPRi in S. cerevisiae.

Our results on ideal guide positioning also differ from those found in human cell lines in which the optimal window for CRISPRi was found to be downstream of the TSS in the 5’UTR[34]. In yeast, we found the optimal window to be a 200bp region immediately upstream of the TSS. While this difference could be due to the different repressors used (Mxi1 vs KRAB), it could also reflect differences in chromatin structure between yeast and mammalian cells[176, 177]. We observed strong and statistically significant links between guide efficacy and nucleosome occupancy, as well as chromatin accessibility. Nucleosome positioning will likely affect gRNA function in other organisms, and thus successful gRNA design is likely to be species- and even locus-specific. Indeed, ChIP-seq analysis of dCas9 binding in mammalian cells has shown that dCas9 is more likely to bind off target sites in open chromatin regions than in closed chromatin [154]. Our study thus defines a simple design rule taking these and other correlates into account that will increase the likelihood of gRNAs having a potent repressive effect.
Conclusions

We have established a powerful functional and chemical genomics screening platform using the CRISPR/Cas9 system for targeted transcriptional repression in *S. cerevisiae*. A reference set of chemical-genetic interactions enabled sensitive measurement of gRNA efficacy at multiple loci. Most notably, we found that truncated gRNAs generally exhibited reduced efficacy towards both mismatched and perfectly matched target sequences compared to their full-length counterparts. In addition, we identify nucleosome occupancy as a major determinant of gRNA performance. gRNAs directed to a region between the TSS and 200bp upstream of the TSS were more likely to be effective. These findings will directly enable library design and genome-wide screening in yeast, and may also inform the application of CRISPRi in other organisms.

Methods

*Plasmid and Strain construction*

All primers, strains, and plasmids used in this study are listed in Additional file 3.17. All chemical compounds used in this study are listed in Additional file 3.3. Molecular cloning was done with Gibson Assembly as outlined in Gibson et al. 2009[147]. *E. coli* minipreps were performed with QIAprep Spin Minipreps (Qiagen). Preparation of competent *E. coli* DH5α and transformation used Zymo Mix & Go *E. coli* Transformation reagents and Zymo Broth. Hifi Hotstart (Kapa Biosystems), Q5 (NEB) and Phusion Hot Start Flex (Thermo Scientific) high fidelity polymerases were used for PCRs. Primers and single gRNA oligonucleotides were ordered from IDT. gRNA oligo libraries were ordered from Custom Array. DpnI treatment was used to remove template plasmids in PCRs that were followed by Gibson Assembly. Benchling.com DNA editing software was used for plasmid design. Individual constructs (not libraries) were sequenced by Sanger Sequencing (Sequetech).

To build the dCas9 repressor, we first modified pRS414-Tef1-Cas9-Cyc1t obtained from addgene[55] to introduce the D10A and H840A mutations to produce
dCas9. We also fused a nuclear localization signal to the N terminus of dCas9. The human Mxi1 domain and linker from[58] was then fused to the C terminal of dCas9.

We built our single plasmid system in the yeast pRS414 and pRS416 Cen/ARS plasmids containing the Trp1 and Ura3 markers respectively. First, we introduced an engineered Tet inducible pRPR1 PolIII promoter[157, 178], NotI site, and gRNA sequence, as well as the Tet repressor (TetR) gene under control of the GPM1 promoter and terminator into pRS414-Tef1-NLS-dCas9-Mxi1-Cyc1 at the PciI site adjacent to the ori using Gibson Assembly. These vectors are referred to as pRS41XgT. We then PCRed the gRNA and TetR and cloned them into pRS416 digested with PciI along with a bridging oligo to correct the PciI site cut in Ura3. Next we PCRed and cloned the Tef1-NLS-dCas9-Mxi1-Cyc1 section of the plasmid into this vector. We are providing our tet-inducible CRISPRi plasmid on AddGene for other investigators to study their questions of interest.

gRNA oligos were amplified with extender oligos that produced 40 bp overlaps on either side of the region of target complementarity, and then cloned into the NotI site with Gibson Assembly. The same protocol was applied both to individual oligos and libraries of oligos. These were then transformed into DH5α cells and plated on LB-agar containing carbenicillin. For individual clones, single colonies were obtained and screened by colony PCR and Sanger sequencing. Correct colonies were cultured and plasmids extracted. For libraries, all colonies were washed off plates with LB-Carb liquid and then minipreped.

Competent S. cerevisiae (strain BY4741) were prepared either by standard lithium acetate transformation protocols or using Frozen-EZ Yeast Transformation II Kit™ (Zymo Research). Transformed cells were selected on synthetic complete media (SCM) – Ura agar plates. For individual strains, single colonies were selected for additional experiments. For library preparation, all colonies were washed off plates with SCM-Ura liquid media, vortexed, and aliquoted into 25% glycerol stocks of 3.0 ODs of cells each for later use.

_Growth assays of individual strains._
Strains were grown overnight in synthetic complete media lacking uracil (SCM-Ura). Growth assays were performed in 96 well NUNC flat bottom plates in 100µl SCM–Ura cultures. Starting OD$_{600}$ was either OD$_{600}$ 0.01 or 0.03, but was consistent within individual experiments. Growth rates were determined by measuring the OD$_{600}$ every ~15minutes for at least 80 cycles at 30°C in TECAN sunrise or GENios plate readers. Drugs were dissolved in DMSO and dispensed to plates using an HP D300 Digital Dispenser (Tecan). The growth rate of a strain was calculated as follows: 1) the first 10 OD readings were averaged and subtracted from all OD readings of the corresponding curve in order to set the baseline of the growth curve to zero, 2) the area under the curve (AUC) was then calculated as the sum of all OD readings. “Relative growth” was calculated as previously described[179], and as follows: (AUC$_{\text{condition}}$ – AUC$_{\text{control}}$) / AUC$_{\text{control}}$; where AUC$_{\text{control}}$ represents the growth rate of the reference condition that was assayed on the same microtiter plate.

**qPCRs**

For qPCR experiments, strains were typically cultured in SCM–Ura media overnight, diluted to an OD/ml of 0.15 in the presence (or absence) of 250ng/ml ATC, grown further, and samples collected at the times indicated. For the ATc removal time course, cells were washed 5 times with sterile water to remove any residual ATc. RNA was extracted from samples using the Ambion RiboPure™ RNA Purification Kit for yeast (Life Technologies), or the Quick RNA Kit (Zymo Research). RNA was converted to cDNA using the High-Capacity RNA-to-cDNA™ Kit (Life Technologies). This cDNA was diluted 1:10 and then used for SYBR qPCR. qPCR primers were designed using primer3 to give products of ~75-150 bases in length (Additional file 3.17). Real time/qPCR was performed using SYBR® Green PCR Master Mix (Life Technologies) and the Applied Biosystems 7900HT Fast Real-Time PCR System running SDS V2.3. The gRNAs used for these experiments are listed in Additional file 3.17. Log2 fold change relative to a reference condition was calculated as the negative delta delta Ct (-DDCt) as follows: DDCt = [(average Ct)$_{\text{gene}}$–(average Ct)$_{\text{control gene}}$] in test condition – [(average Ct)$_{\text{gene}}$ – (average Ct)$_{\text{control gene}}$] in reference condition. Average Ct values were typically calculated from four replicates. Standard
Deviation (StdDev) was calculated as the square root of \[
(\text{StdDev of } C_{t\text{gene}})^2 + (\text{StdDev of } C_{t\text{control gene}})^2
\] as measured in the test condition.

**Library Design**

The “gene tiling” libraries were designed to a window of 150bp upstream of the TSS to 50bp into the ORF. TSS were specified as the most common transcript start position from transcript isoform profiling data [180]. A full list of genes examined are available in Additional file 3.3. Guides were designed both to the template strand and non-template strand. For each of these guides we designed versions containing 18 nt and 20 nt of target complementarity (gene_tiling18 and gene_tiling20, respectively). Even though gRNAs were designed for CDC12, ERO1, and RPO21, small molecule inhibitors specific to these genes were not assayed. For 5 of the genes (ERG11, ERG25, SEC14, CRG1, and SSL2) we designed all possible full-length guides within a window of 500bp up- and downstream of the TSS (broad_tiling). The mutant library was designed by taking the sequences for 4 gRNAs we had previously shown to be functional and making a random single base change for all positions in the seed sequence (1-10) as well as in positions 11, 12, 14, and 16. For each guide, we synthesized both an 18 and 20 nt version (mutant18 and mutant20, respectively). We used ECRISP version 4.2 to look for potential off-target binding sites in the yeast genome, allowing for up to two mismatches.

**Competitive Growth Assays**

Prior to setting up experiments, aliquots of a library were recovered in YPD media for 4 hours, and then diluted appropriately for the experiments. Yeast culturing and sample collection was performed using a cell-screening platform that integrates temperature-controlled absorbance plate readers, plate coolers, and a liquid handling robot. Briefly, 700 µl yeast cultures were grown (+/- a drug listed in Additional file 3.3, and +/- ATc) in 48 well plates at 30°C with orbital shaking in Infinite plate readers (Tecan). To maintain cultures in log phase over many doublings, 23 µls of the culture was removed when it reached an OD of 0.76, added to a well containing 700 µl of media, and then allowed to grow further. After three such dilutions, 600 µls of the
culture was collected and saved to a 4°C cooling station (Torrey Pines) when it reached an OD of 0.76. This amounted to approximately 20 culture doublings from the beginning of the experiment. Pipetting events were triggered automatically by Pegasus Software and performed by a Freedom EVO workstation (Tecan).

A key parameter in this protocol is the extent to which a drug inhibits growth of the pool. In general, drug concentrations that inhibit growth by ~20% are best for identifying chemical-genetic interactions and yielding reproducible results. If a drug was observed to inhibit the pool’s growth too strongly (e.g. by >50%), the experiment was repeated using a lower drug concentration.

After sample collection, yeast plasmids were purified using the Zymoprep Yeast Plasmid Miniprep II kit (Zymo Research). Purified plasmids were used as a template for PCR with barcoded up and down sequencing primers that produce a double index to uniquely identify each sample. PCR products were confirmed by agarose gel electrophoresis. After PCR, samples were combined and bead cleaned with Thermo Scientific™ Sera-Mag Speed Beads Carboxylate-Modified particles. Sequencing was performed using Illumina MiSeq.

**Metabolite Extraction and GCMS Analysis**

Our previously described methods were used for metabolite measurement[181]. In brief, yeast pellets of six biological replicates were homogenized in 1x PBS buffer with 0.5mg of 0.5mm glass beads/tube by vortexing for a total of 6 minutes. Every 2 minutes between vortexing, the tubes were returned back to ice. The homogenized mixture was extracted by Folch method[182]. The lower phase of the chloroform:methanol:water mixture, containing the sterol metabolites extracted from the yeast cell pellet, was collected and dried in a Speedvac. The samples were derivatized by MSTFA + 1 % TMCS and analyzed by Agilent 7200 series GC/Q-TOF. The sterols were separated on HP5-MS UI column (30 m, 0.25 mm i.d, 0.25 μm film thickness) at split ratio 20:1 using helium as carrier gas at 1 mL/min. The oven temperature program was as follows: 60 °C held for 1 min, then oven temperature was ramped at 10 °C/min to 325 °C where it was held for 3.5 min. Data were collected at acquisition rate of 5 Hz in both profile and centroid modes.
Qualitative and quantitative analysis was performed using Agilent MassHunter Workstation.

Ergosterol, lansterol and methoxyamine hydrochloride were purchased from Sigma. HPLC grade methanol, chloroform and water were from Honeywell Burdick and Jackson. The derivatization reagent MSTFA (N-methyl-N-trimethylsilyl trifluoroacetamide) with 1% TMCS (trimethylchlorosilane) was from Thermo Fisher Scientific (Bellefonte, PA).

**Sequence data analysis**

We used a combination of established tools and custom python pipelines to quantify gRNA effects from the sequence data. First, we created a synthetic reference chromosome sequence for each of the expected amplicons. The synthetic reference included Illumina adaptors, library barcode, PCR amplification priming region, PCR barcode, and the gRNA region complementary to the target. As the forward and reverse reads were expected to overlap given the 250bp read length, and average 190 bp amplicon length, we merged the reads using PEAR [183] version 0.9.4 with default parameters. The resulting FastQ file was mapped against the created synthetic reference using BWA [184] version 0.6.1-r104 with command line “bwa index [reference]; bwa aln -n3 -o3 -e1 -l22 [reference] [fastq] > [aln]; bwa samse -f [out] [reference] [aln] [fastq]”. This parameter setting allows for 3 mismatches, 3 gaps, one large gap, and short 22bp seed sequences. For each of the expected amplicons, we counted the number of perfect matches (flag NM:i:0) from the resulting alignments (Additional file 3.8, column “Count_perfect”) that were used in subsequent analyses.

**Guide fitness calculation**

We quantified fitness $f$ for guide $i$ in pool $j$, condition $k$ as the relative growth rate $f_{ijk}$. For a guide with $c_{ijk}$ counts in condition $k$, we calculated the log2-scale median centered counts $l_{ijk} = \log_2(c_{ijk} + 0.5) - \text{median}_{i' \in \text{pool}_j}(\log_2(c_{i'jk} + 0.5))$. As any such count statistics are highly variable at low abundances, it is important to also record the confidences of this value. We calculated the empirical variance estimate of $l_{ijk}$ by
resampling reads given the total library size of \(C_{jk}\) reads, and \(N_j\) guides in pool \(j\) condition \(k\). To do so, we inferred the posterior read frequency in the pool as 
\[
\text{Gamma}(c_{ijk} + 0.5, C_{jk} + 0.5N_j),
\]
 sampled 1000 observations of read counts, calculated the log-scale median centered count for each, and used the variance \(s_{ijk}^2\) of the simulated values as a variance estimate for the log scale counts. In the following, we thus model guide fitness as \(f_{ijk} \sim \text{Normal}(l_{ijk}, s_{ijk}^2)\).

**ATc-induced fold change**

To estimate the guide effect on growth, we calculated the ATc-induced fold change \((A)\). For conditions \(k+\) and \(k-\) (with and without ATc, respectively), we infer the ATc-induced fold change \(A_{ijk}\) on guide \(i\) in pool \(j\) as the difference in the fitnesses between the cultures with and without ATc, \(A_{ijk} = f_{ijk+} - f_{ijk-} \sim \text{Normal}(l_{ijk+} - l_{ijk-}, s_{ijk+}^2 + s_{ijk-}^2)\). For the control condition (1% DMSO), we had eight replicate experiments for the tiling pools, and three replicate experiments for 20µM fluconazole. We combined the \(R\) replicates \(k_1, ..., k_R\) in a natural way to obtain the variance \(\text{var}(A_{ijk}) = 1/(1/\text{var}(A_{ijk1}) + ... + 1/\text{var}(A_{ijkR}))\), and mean \(<A_{ijk}> = \text{var}(A_{ijk}) * (<A_{ijk1}>/\text{var}(A_{ijk1}) + ...) + <A_{ijkR}>/\text{var}(A_{ijkR}))\). The combined estimate across replicates is then \(A_{ijk} \sim \text{Normal}(<A_{ijk}>, \text{var}(A_{ijk}))\). gRNAs with fewer than 30 reads following growth in the minus ATc control condition were excluded from this analysis, as their effect size estimates had large variance across conditions.

The ATc-induced fold change \((A)\) of a gRNA, calculated following growth in the presence of its specific partner reference compound (Additional file 3.3), was defined as the “gRNA effect” or “guide effect” and used in Figure 3.4, 3.5C, S3.6, and S3.7. In cases where we tested multiple concentrations of a reference compound, we selected data from one concentration (these are indicated in Additional file 3.3). For the plots in Figure 3.5A, 3.5B, and S3.9, we first restrict \(A\) values to have a maximum of 0, and then take the absolute value of this. We define this as the “gRNA effect magnitude”. This calculation is based on the reasonable assumption that repression of the target gene can only result in drug-sensitivity (i.e. a negative ATc-induced fold change), and that any apparent resistance (i.e. positive ATc-induced fold change) is a result of technical noise in a poorly functional gRNA.
**Drug-specific effect estimation**

To estimate the drug-specific effect $D$ for guide $i$ in pool $j$ and drug $k$, we contrasted the guide fitness with and without drug when the system was induced with ATc, i.e.

$$D_{ijk} = f_{ijk} - f_{ij0} \sim \text{Normal}(l_{ijk} - l_{ij0}, s_{ijk}^2 + s_{ij0}^2),$$

where condition $k=0$ is the 1% DMSO control, and its parameters are for the distribution inferred from the eight replicates. gRNAs having fewer than 30 reads following growth in the presence of ATc, but without additional chemical compound, were excluded from this analysis, as their effect size estimates had large variance across conditions.

**Guide melting temperature calculation**

We used ViennaRNA package 2.0 (RNAfold version 2.1.9 ViennaRNA Package 2.0)\[185\] to calculate RNA folding and duplex formation energies for the entire targeting region, as well as the eight nucleotide seed, and the oligotm\[186\] library, both with default parameters.

**ATAC-seq and nucleosome data**

We downloaded the nucleosome occupancy and ATAC-seq insertion data from Schep et al 2015\[169\] in bigWig format, and converted it to wig using UCSC utilities (`./bigWigToWig <input> <output> -chrom=<chromosome> -start=<TSS-1000> -end=<TSS+1000>`). For each nucleotide, we used the per-base .wig output as the measure of nucleosome occupancy, and the average ATAC-seq insert count in 51 base window centered on the base as the ATAC-seq signal. Spearman correlation was calculated using the `spearmanr` function in the scipy.stats python package.

We also compared the efficacy of gRNAs targeting $ERG11$, $ERG25$, $SEC14$, and $SSL2$ to yeast nucleosome occupancy measurements reported by Lee et al\[168\]. Based on previous exonuclease footprinting experiments\[187\], we first defined a genomic region predicted to be occupied by dCas9 upon successful base pairing with the gRNA. This region (or ‘window’) consisted of the genomic target sequence recognized by the gRNA, plus 3 bases on either end. Window coordinates were defined based on the February 2006 SGD genome build as these data were used in the
analysis by Lee et al[168]. We downloaded log2 ratios representing the relative hybridization signal of total genomic DNA to nucleosomal DNA. These measurements were made using a 4bp-resolution tiling array of the yeast genome, and therefore, each gRNA has 6 or 7 log2 values within its window. These values were averaged to generate the “Nucleosome Occupancy” values plotted for each gRNA in Figure S3.7B, which were compared to the effects of gRNAs for ERG11, ERG25, SEC14, and SSL2; R and p-values (Spearman correlation) were calculated in Spotfire (Perkin Elmer).

**Target sequence context**

We considered the region of 20bp upstream of the PAM to 20bp downstream of the end of the target sequence. For each site in this region, and each of A,C,G,T bases, we calculated the relative gRNA effect in control condition for all the guides whose target sequence has that base at the considered site. Relative gRNA effect was calculated by dividing the gRNA effect for each guide by the maximum gRNA effect constrained to be between 0 and -6 for each control drug/gene set. We calculated the p-value of the median effect as the fraction of 10,000 random samples of same number of overlapping guides that have at least as large median effect. We also report p-values Bonferroni-corrected for the number of tests (4 bases x 63 sites = 252 tests).

**Overlap with transcription factor binding sites**

We downloaded data from Reimand et al 2010[170], as used in Zaugg and Luscombe 2012[166]. We considered three levels of overlap - sites overlapping middle of the specificity-determining sequence, middle 10 bases of the sequence, and any part of the sequence. For each factor, and each level of overlap, we calculated the mean guide effect magnitude in control condition of all the overlapping guides in gene_tiling20 and broad_tiling guide sets. For each level, we only considered factors that overlapped at least 10 guides. We calculated the p-value of the median effect as the fraction of 10,000 random samples of same number of overlapping guides that have at least as large mean effect. We also report p-values Bonferroni-corrected for the number of tests.
Availability of supporting data
The data sets supporting the results of this article are available in the Gene Expression Omnibus (GEO) repository, GSE71490;

List of abbreviations
Anhydrotetracycline (ATc), nucleotide (nt), guide RNA (gRNA), transcription start site (TSS), open reading frame (ORF), clustered regularly interspaced palindromic repeats (CRISPR), CRISPR interference (CRISPRi), area under the curve (AUC), mechanism-of-action (MoA), dimethyl sulfoxide (DMSO), protospacer adjacent motif (PAM)

Competing Interests
The authors declare that they have no competing interests

Author Contributions

Acknowledgements
The authors are grateful to Alicia Schep, Jason Buenrostro, and William Greenleaf for help with the ATAC-seq data, to Fahim Farzadfard, Ky Lowenhaupt, and Timothy Lu for sharing plasmids, and to Raeka Aiyar for critically reading this manuscript. This work was supported by grants from the US National Institutes of Health (P01HG000205 to L.M.S. and R.W.D. and U01GM110706-02 to R.W.D.). The research leading to these results has also received funding from the European Research Council under the European Union's Seventh Framework Programme
Supplemental Figures and Additional Files

Supplemental figures and additional data files are in the Chapter 3 supplemental file. See Chapter 3 appendices for supplemental figure and additional file legends.

Chapter 4: A scalable method for parsing complex oligonucleotide libraries and creation of a CRISPRi *Saccharomyces cerevisiae* strain collection

Authors
Justin D. Smith\(^1,2\), Ulrich Schlecht\(^1\), Joe Horecka\(^1\), Sundari Suresh\(^1\), Angela M Chu\(^1\), Richard Bennett\(^3\), Weihong Xu\(^1\), Richard Hyman\(^1\), Michael J. Proctor\(^1\), Ronald W. Davis\(^1\), Sasha F Levy\(^3\), Robert P. St. Onge\(^1,4\)

\(^1\) Stanford Genome Technology Center, Department of Biochemistry, Stanford University, 3165 Porter Drive, Palo Alto, CA 94304

\(^2\) Department of Genetics, Stanford University School of Medicine, Stanford, California, USA.

\(^3\) The Department of Biochemistry & Cell Biology Stony Brook University

\(^4\) Corresponding author: Robert P. St.Onge, Stanford Genome Technology Center, Department of Biochemistry, Stanford University, 3165 Porter Drive, Palo Alto, CA 94304 Phone: 650-721-2976 Email: bstonge@stanford.edu

* These authors contributed equally to this work.

This work represents a draft of an ongoing project we eventually intend to submit to Nature Biotechnology. The dissertation author, [JDS] was responsible for designing and building the CRISPRi recipient strain and strain collection and analyzing the data for the collection. JDS, US, and RPS analyzed the data, produced all the figures, and wrote the paper. JDS, US, RPS and SS conducted the experiments.
AMC and JH performed mitochondrial genome stability repairs the recipient strain. JDS, US, SFL, and RPS designed the experiments. JH developed the high-efficiency transformation/DNA integration protocol. RB and SFL constructed the barcoder strains. WX and RH designed molecular probes. MJP developed robotics and automated protocols. SFL and RWD provided useful insight.

Abstract

Array-based DNA synthesis can produce short DNA oligonucleotides (oligos) at larger scales and lower costs than column-based DNA synthesis platforms. Individual oligos, however, are provided as complex mixtures, in which they are often unevenly represented and contain a significant number of synthesis-derived errors. In addition, isolation of individual oligos from these mixtures typically requires unique common priming sites to be included in their design so they can be selectively amplified out of the mixture. These can limit the utility of array-synthesized oligos for some biological applications. Here, we describe a high-throughput method of retrieving sequence-verified oligos from complex mixtures. The method involves introduction of a complex oligo library into the yeast genome, and then uses yeast mating in high-density ordered arrays and site-specific recombination to barcode the exogenous DNA of oligo transformants. Next-generation sequencing is then used to identify clones containing the DNA of interest, which can be readily retrieved and archived. We demonstrate the utility of the method by creating >3000 molecular probes for culture-independent detection of bacteria, and a collection of ~9,000 strains for CRISPR interference-mediated gene repression in yeast. We go on to functionally characterize this collection, which targets the majority of essential and aerobic essential genes in yeast, as well as confirm and expand on the rules for gRNA design discussed in Chapter 3.

Introduction

The ability to synthesize DNA oligonucleotides has been essential for the development of many molecular biology techniques such as polymerase chain reaction (PCR), oligonucleotide probes, and DNA sequencing. Today many researchers
routinely order individual oligos of around 100 nucleotides in length priced at around $0.05-0.15 per base [24]. Still, oligos synthesized individually in this way become cost prohibitive for the synthesis of thousands of different oligos and for the synthesis of entire genes. One alternative is to synthesize pools of oligos on microarrays. Microarray oligos are 2 to 4 orders of magnitude cheaper than oligos synthesized individually. The cost per base ranges from $0.00001-0.0001, depending on the size of the library, the length of the oligos, and the platform that is used [24]. However, there are limitations to microarray oligo library synthesis. First of all, with microarray oligos one does not have easy access to individual oligos but rather has to deal with a complex mixture. Second, since oligo synthesis has an error rate of about 1 error per 200 bases synthesized, the majority of long oligos and genes assembled from these oligos will have errors which can limit their applications[24]. Finally, the distribution of individual oligos in oligo pools is often uneven, which can limit the utility of these oligos even when individual oligos are not needed. Here we present a method to overcome all three of these disadvantages and give researchers access to inexpensive, sequence validated synthetic DNA by leveraging the power of yeast genetics and high throughput sequencing.

Results
Our method for isolating sequence-verified oligonucleotides from complex libraries required the construction of two types of yeast strains: ‘recipient’ strains and ‘barcoder’ strains (see methods for details). Recipient strains are strains into which our oligonucleotide libraries are introduced by transformation. The genomic integration site in these strains consists of a counter-selectable marker (FCY1, which confers toxicity to the selective agent 5-fluorocytosine (5-FC) ([188])) flanked by two SceI restriction sites. Transient expression of the SceI endonuclease during transformation, results in double-strand break formation, high-efficiency replacement of FCY1 by the transforming DNA, and growth on media containing 5-FC. Our transformation protocol typically produces roughly 1 million transformants from 5ml of cells and 1pmol of transforming DNA (Figure 4.1). The integration site is genetically linked to a crippled loxP recombination site (lox66), as well as an artificial intron and one half of the URA3 selectable marker. These features are essential for barcoding of the
exogenous DNA by yeast mating in high-density ordered array format to barcoder strains.

We constructed a barcoder strain collection consisting of 1536 strains, each containing a unique 26bp barcode identifier. Similar to the integration site of the recipient strains, these barcodes are genetically linked to a crippled loxP site (lox71), and an artificial intron and the other half of the *URA3* selectable marker. Barcode sequences in each strain were verified using a combination of Sanger Sequencing and Illumina Sequencing (see methods). Each barcoder strain also contains the CRE-recombinase open reading frame (ORF) under control of a galactose inducible promoter. Thus, following mating of a transformed recipient strain to a barcoder strain, selection of diploids on media containing galactose induces site-specific recombination between loxP sites, which in turn results in the physically linkage of the barcode to the exogenous DNA. This recombination event is selectable by growth on media lacking uracil, as a functional URA3 will be reconstituted by intron splicing of the transcribed locus (**Figure 4.1**).

Yeast mating is an efficient process that has enabled high-throughput genetics through replica plating of ordered arrays [189]. Our collection of barcoder stains allows for systematic barcoding of the exogenous DNA incorporated into our recipient strains is 1536-format. Following selection of diploid loxP-recombinants, single-pot amplification of the resulting barcode-exogenous DNA hybrids is achieved by PCR with common primers. As the amplicons are relatively short, paired-end Illumina sequencing can be used to identify the exogenous DNA linked to each barcode. Importantly, employing indexed sequencing primers allows multiple sets of 1536 colonies to be interrogated in a single sequencing run. As the position of each barcoder stain on the ordered array is known, the sequencing results can be used to identify strains containing DNA of interest. These strains are then cherry-picked (from either haploid, or replicated diploid arrays), archived, combined as desired, and DNA amplified using PCR to create a new library (**Figure 4.1**). Additionally, if single stranded DNA is desired, methods exist to convert the new library back into single stranded DNA [190].
Figure 4.1 Oligonucleotide library parsing strategy.
After oligo library synthesis, oligos are amplified by PCR with primers that add overlaps for genome integration. This DNA is transformed into a strain with a landing pad consisting of LoxP, SceI sites.
flanking FCY1, an Artificial Intron sequence (AI), and half of the URA3 gene. Scel is transiently expressed during transformation. Transformants with FCY1 successfully removed by the integration event are selected on 5-flourouracil (5-FC) plates. A colony picker robot is then used to array colonies onto a 384 well plate. This plate of arrayed unknown strains is then mated against a barcoded library where each barcode is in a known position in the array. Cre-recombinase is induced with galactose to promote recombination between the LoxP sites to bring the barcode and the unknown oligo together on the same molecule in the middle of an artificial intron. This also leads to the production of a functional URA3 gene, which can be selected for on –Ura plates. Colonies are then washed from the plate, combined, and barcoded by PCR primers to specify which plate the library represents. Paired-end sequencing is then used to determine which oligos are linked to each barcode and in turn what position on which plate they are located. Finally, a stinger-adapted ROTOR robot (Singer Instruments) is used to select individual clones of interest from the arrayed transformants onto a new plate.

Our method can improve the quality of complex libraries in several ways. First, because DNA oligonucleotides are cloned and sequenced, oligonucleotides containing errors introduced during synthesis can be filtered out (Figure 4.2). In effect, our method exchanges the DNA synthesis error rate, for the error rate of the polymerase used to amplify the DNA from yeast; the fidelity of the latter being up to 1000x greater than that of the former [24, 191]. In addition, member oligos in array-derived libraries are often not uniformly represented [24] (Figure 4.2A). By combining equal numbers of cells representing each DNA oligo, our method can in theory generate equimolar DNA libraries (Figure 4.2B). Moreover, clones can be archived for future use, accessed individually, or combined as desired, allowing for customizable libraries to be created on-demand.

Molecular Probes

High-quality oligonucleotide libraries have several important medicinal and biological applications including: diagnostics[192], functional genomic libraries [32–35, 65], gene synthesis [24, 31], and investigating regulatory elements[37, 38]. We designed 7051 molecular probes for detection of 321 bacteria (~20 probes per bacteria) related to potable water quality [192]. Each probe contained 60 bases of homology to the target genome, a 9 base ‘random’ barcode, and 36 bases of priming sequence for amplification of probes that successfully hybridize to their targets. In addition to these sequence features, each probe was flanked by 20 bases of common priming sequence (for amplification of the library following array-based synthesis) and type IIS restriction sites (for removal of these common priming sequences). These probes which were a total length of 145 base were ordered from CustomArray.
Figure 4.2 Molecular probe oligo set purity and abundance.
Sequencing analysis of the purity and relative abundance of A) oligos sequenced following PCR-amplification of custom array synthesized molecular probe oligo pool and B) of PCR-amplified molecular probe pool after parsing and cherry picking with our method.

We amplified the probe library using primers that recognized the 20 base ends and which also contained Illumina sequencing adapters. The composition of the amplified library was then analyzed using paired-end sequencing. Typical synthesis error rates of 1 in 200 [24] are predicted to result in roughly 50% of the oligonucleotides of this length containing some type of error. Indeed, our analysis of the amplified library revealed that only 49% of reads (where the forward read matched the reverse read) matched a designed sequence (Figure 4.2A). Moreover, individual oligos from amplified library exhibited a wide range of concentrations, (Figure 4.2A), that were consistent with previous reports [24].

We employed our method to improve quality of this library. We first amplified the probe library using primers that contained 40 bases of homology to the integration site in our recipient strain (#2849). Following transformation, ~35,000 transformants were robotically arrayed in 384 format, mated to barcoder strains in 1536 format using a ROTOR robot (Singer Instruments), and diploid recombinants were selected on CSM-URA+Galactose media (see online methods)(Figure 4.1). Diploid clones were collectively removed from each plate, genomic DNA was extracted, and barcoded exogenous DNA amplified with indexed primer pairs specific to each plate (Figure 4.1). A 300-cycle MiSeq reagent kit was used for sequencing: 240 cycles in the forward direction to identify the oligonucleotide insert, and 60 cycles in the reverse
direction to identify the barcode. Analysis of the sequencing data basically involved binning reads by sequencing index (identifying the 1536 agar plate) and by barcode (identifying the plate position), and then extracting the most abundant sequence found at the integration site and determining whether that sequence matched one of our designed probes. 3322 colonies, each containing a different sequence-verified molecular probe was then cherry picked from the plate using a stinger-adapted ROTOR robot (Singer Instruments).

**Figure 4.2B.** Analysis of the non-matching sequences revealed that the majority are likely due to mixed colonies (i.e. cases where the selected colony in fact consisted of two different colonies) - This problem can be corrected in future experiments by more rigorously identifying these cases when analyzing the sequencing data. Parsing also dramatically improved the relative abundance of each sequence. Before parsing there were more than 2 logs of variation in the relative abundance of oligos, but after parsing the vast majority of the sequences had less than one order of magnitude of variation in abundance (**Figure 4.2B**).

There were however, a small number of sequences that were at much lower abundance. There are two potential causes for the low abundance of these sequences. First, it is possible the strains representing these colonies were missed by the Stinger robot or very few cells were transferred. We have noticed that on rare occasions, plates are not perfectly set in their nests on this instrument which can result in the Stinger ‘missing’ some colonies on the plate. Second, these initial experiments were done in a strain derived from BY4741, which has a high frequency of petites (strain that have lost their mitochondria and the ability to grow aerobically and thus form smaller colonies). If a sequence were integrated into a petite which consisted of fewer cells, it would be expected to have fewer counts during sequencing. To reduce the likelihood of petite formation, we corrected alleles known to promote mitochondrial instability in the recipient strain that was used for future experiments, including the CRISPRi recipient strain described in the next section (See Methods).
**CRISPRi Collection**

We next sought to extend our method to allow for creating CRISPR interference libraries for functional genomic screens. While several groups have used gRNA pools created from oligo libraries in different cellular backgrounds [33–35], including yeast [Chapter 3], these strains were always pooled and there was no way to easily access individual strains. Additionally, these pools were generally created from gRNAs that were present on plasmids and not integrated into the genome. Access to individual stable integrated repressor strains have advantages for studying individual genes and confirming phenotypes found in pooled screens. Because essential genes cannot be deleted, having a collection of inducible repressible strains could be useful. Indeed other groups have attempted to create similar collections in the past, but each existing collection has certain limitations and are incomplete, not covering all essential genes [171, 173, 193].

To construct a repressible essential gene collection, we first built a CRISPRi recipient strain as described in the methods which contains an integrated dCas9-Mxi1 repressor [194], a tetracycline inducible gRNA promoter, a SceI flanked FCY1 landing pad, and the constant part of the gRNA and its terminator adjacent to the Ura3/LoxP cassette (Figure 4.3A). In this new design, the integration sits outside of the artificial intron. We changed the design in this way so that we could express a functional gRNA from a promoter without interfering with the function of the intron. We took into consideration the constraints of paired end Illumina sequencing in our design. Because we are able to use PCR primers that flank the barcode and the 20 nucleotide target complementarity sequence of the gRNA and amplify a product small enough (~800nt) for Illumina paired end sequencing, we are still able to link each barcode with a specific gRNA.

Next, we designed a library of 18,803 full-length (20 nt) gRNA target complementarity sequences to 1118 essential genes and 514 aerobic essential genes and to the non-essential ADE2 gene which can be assayed for a red phenotype [195] in individual strains (see methods). Following the rules we determined in Chapter 3, we designed these guides to fall within a window of 0 to 200nt upstream of the TSS,
provided there were sufficient gRNA target sites available in that window and a defined TSS (see methods). Following the processing of ~62,000 transformants using the above protocol above, 9059 strains were selected and arrayed as individual strains representing 48% of the gRNAs designed. This CRISPRi strain collection was then grown in YPEG (aerobic growth media) or YPD (fermentative growth media) in the presence or absence of the inducing agent, anhydrotetracycline (ATc) (Figure 4.4). After growth, cells were harvested, and the gRNA locus was PCR amplified and relative gRNA counts determined to calculate log2 fold change values. Any strain with log2 fold change less than -0.5 was considered to have a detectible growth defect. For details on how these were calculated and filtered see methods. A distribution of gRNAs per ORF for the total designed guides, gRNA per ORF post filtering (see methods) in the YPEG condition, and total functional gRNAs per ORF can be see in Figure 4.6A-C.

**Figure 4.3 CRISPRi Recipient Strain.**
The CRISPRi recipient strain contains A) a modified landing cite consisting of a SceI site flanked FCY1 inbetween a Tet inducible PRPR1 promoter, and the constant part of the gRNA and the RPR1 terminator. New gRNAs are integrated by induction of Scel which cuts out FCY1 and promotes homologous recombination into the remaining site. After integration, this locus can be mated against
barcoding strains and recombined by CRE recombinase at a LoxP site as in Figure 4.1. After recombination, primers can be used to amplify the gRNA and barcode on a fragment of ~800nt that can be used for paired end sequencing to link gRNA with barcode. B) In addition to the gRNA locus, the recipient strain contains dCas9-Mxi1 and LoxP integrated at YORWdelta17. C) The strain is inducible. In the absence of anhydrotetracycline (ATc) TetR binds the gRNA promoter and prevents PolIII from binding and transcribing the gRNA. This in turn prevents dCas9-Mxi1 from binding the target site. In the presence of ATc, TetR dissociates and gRNA is expressed, allowing dCas9-Mxi1 to bind its target locus, and repress gene expression.

**Figure 4.4 CRISPRi collection experimental workflow.**
Cells were passaged 4 times in log phase in triplicate. Half the plate contained YPEG and the other half had YPD media. At the end of the experiment, cells had grown for ~12 generation. Cells were harvested from the last columns (red dashed boxes, columns 4 and 8), and genomic DNA was isolated. The gRNA locus was then PCR amplified and relative gRNA counts were determined from which relative log2 fold changes were calculated.

Substantially more gRNAs produced growth defects in YPEG (45.8%) than in YPD (20.7%) ([Table 4.1, Figure 4.5](#)). As gRNA to aerobic essential ORFs were not expected to give a growth defect in YPD, we expected fewer strains to give a growth defect in YPD, but this alone did not explain all of the difference. Even among essential genes, a greater fraction gave a growth defect in YPEG (46.3%) than in YPD (27.5%). It is possible that aerobic growth conditions are more strenuous to the yeast cells, which generally grow faster in YPD.

As YPEG produced more measurable phenotypes, we chose to use this condition to characterize what we believed to be functional gRNAs. After filtering out gRNAs with low read counts and gRNAs to 5 dubious ORFs, the set of strains included for analysis in YPEG contained 8788 gRNAs with at least one gRNA directed against 1569 distinct ORFs, with an average of 5.6 gRNAs per ORF ([Figure 4.6B](#)). Of these ORFs, 1424 were targeted by 3 or more gRNAs. In [Table 4.1](#), we show the number of gRNAs and ORFs with at least one gRNA with raw log2 fold change less than -0.5. Of the 1569 ORFs in the collection, 1349 (86.0%) had at least one
functional gRNA as defined by a growth deficit log2 fold change less than -0.5 (Figure 4.6C). A distribution of the most inhibitory gRNA (as defined by log2 fold change) per ORF can be seen in Figure 4.6D.

Table 4.1  Library counts in different conditions.
*For filtering criteria see methods
As expected, very few (8.4%) gRNAs to aerobic essential genes produced a growth defect in YPD, while nearly half (45%) did in YPEG (Table 4.1). However, as we had no reason to expect any gRNAs to aerobic genes to show a growth defect while growing continually in log phase in fermentative (YPD) media, we sought an explanation for the apparent growth defects produced by these gRNAs (Table 4.1). One possible explanation is if there was some aerobic growth occurred at the end of the experiment despite cooling the cells to 4°C which may have produce weak effects. It is also possible that while not essential in YPD conditions, some genes in our respiratory set do produce a detectable competitive growth defect when repressed. An
additional hypothesis that could potentially explain these results is if CRISPRi were able to silence nearby promoters and hence silence more than one gene.

To test this we determined which gRNAs were within 200nt (upstream or downstream) of a TSS for another gene using previously published transcript 5′-isoform data[180]. Of note, only 7.2% (160/2215) of gRNAs targeting aerobic-essential genes and exhibiting no likely unintended TSS within 200 nt produced a phenotype in YPD, while 11.3% (102/906) of gRNAs to targets with a nearby unintended TSS did (p=0.002) Table 4.2, Figure 4.7. This enrichment for functional gRNAs to targets with a nearby unintended TSS was seen in nearly all conditions tested except for essential gRNAs in YPD. There are three potential causes for this. 1) The unintended ORF that is repressed is essential and is the cause of the growth defect. Interestingly, of the aerobic gRNAs with a growth defect in YPD that had a nearby TSS, 28.8% (30/104) were nearby one of 16 essential genes, while only 16.1% (135/837) of the nonfunctional aerobic gRNAs that had a nearby unintended TSS were near on of 27 essential gene. 2) The unintended ORF is repressed and is a co-regulated ORF that has some degree of synthetic lethality with the intended target. 3) The chromatin in a region with two or more TSSs is more open and therefore CRISPR repression is more effective. This could be useful to study genes that are naturally co-regulated by a bidirectional promoter. However, it is a potential limitation of the technology if one only wishes to repress the intended ORF. To avoid this, one could design gRNAs to an ORF such that they are as far as possible from the TSS of an adjacent ORF, while still within the previously determined optimal window of 0 to 200 bases upstream of the TSS of the intended target.
Table 4.2. gRNAs targeting more than one ORF.

A comparison of gRNAs targeting only the intended ORF, to those potentially targeting an unintended nearby ORF whose TSS is within +/- 200nt of the gRNA.

Next we wanted to see if the gRNA design rules uncovered in Chapter 3 would be observable in this larger dataset. As before we found more functional gRNAs in regions of open chromatin (greater normalized ATAC-seq read density) (Figure 4.7A-B). With more data points we may be able to further refine the window relative to the TSS. In Figure 4.7C-D it appears that gRNAs within a window of 0 to approximately -100nt upstream of the TSS are most likely to be functional. Due to the design of this library, most gRNAs fall within a window of 0 to -200nt upstream of the TSS (see methods). In YPEG, 60.7% of gRNAs within 0 to -100nt upstream of the TSS produce a growth defect whereas only 47.4% of gRNAs within a 0 to -200nt window, and only 45.8% throughout the entire window produce a growth defect. Likewise 55.9% of gRNAs produce a growth defect when with normalized ATAC-seq read density is greater than or equal to 0.75 versus 45.8% from the entire set. Strikingly, 71.3% (907/1272) of gRNAs produce a growth defect (log2 < -0.5) when targeted within a window of 0 to -100nt upstream of the TSS that is characterized by an ATAC-seq read density greater than 0.75 in YPEG.

Because of the striking relationship with ATAC-seq accessibility, we hypothesized that in addition to chromatin state, other DNA binding proteins such as transcription factors may affect dCas9's activity. This relationship can go two different directions. In Chapter 3, we found a small number of cases where overlap with a transcriptional activator binding site correlated with stronger guide effects.
While this might be due to displacing an activator, one might assume in general DNA binding proteins may interfere with Cas9 ability to bind its targets. To test for this, we looked for gRNA target site overlap with DNase I hypersensitive sites (DHSs), which comprise a structural signature of the regulatory regions of eukaryotic genes [196]. We found a small but statistically significant enrichment for increased growth defects with gRNAs not overlapping DHSs in YPEG (Table 4.3). In YPD this relationship was not seen, perhaps due to a lack of statistical power resulting from the fact that most guides did not produce a detectible growth defect in YPD. Further analysis looking at the effects of sites targeted by different classes of DNA binding proteins may be warranted.

**Figure 4.7 Chromatin and Positional Effects in YPEG.**
A) Log2 fold change values for each gRNA plotted against the normalized ATAC-seq read density for that gRNA. B) As in A but binned headmap. C) Log2 fold change values for each gRNA plotted against the position of the gRNAs midpoint relative to the TSS D) As in C but binned headmap.
Finally, through this collection we learned some potentially interesting new biology. While we expected genes in this collection to display a growth deficit when repressed in YPEG, we found a subset of genes that actually showed a growth increase. Many of these were represented by only a single gRNA, which could be a result of off target effects of that individual gRNA or noise. However, there were two genes, *IRA1* and *IRA2*, which both were represented by 9 gRNAs in our collection, the majority of which showed a growth increase and none of which showed a growth deficit. The mean log2 fold change for these ORFs was 1.15 and 1.89 respectively. This is unlikely to be due to random chance or off target effects due to the number of guides targeting these ORFs. Additionally, these genes are paralogs that resulted from the whole genome duplication[197], and have similar functions. The IRA1 null is listed as inviable in S288C, and IRA2 null is listed as respiratory deficient with decreased competitive fitness on the Saccharomyces Genome Database (SGD) [197], and yet when repressed, we find these strains are growing better in aerobic conditions. Thus it would appear a repression phenotype results in an opposite phenotype than that of a complete null. As both genes are negative regulators of the RAS protein, which plays a role in cell proliferation [197], it is reasonable to believe that a partial repression phenotype results in improved growth. Further tests are required to confirm these findings, but the initial results are promising.

This finding raises an interesting caveat about the percentage of gRNAs that we are characterizing as functional through growth defects. Some gRNAs may indeed be resulting in gene silencing, but are either not resulting in enough repression to give
a phenotype, or are resulting in a phenotype counter to what is expected. Thus the percentage of gRNAs that are functional in our collection may be greater than the observed number that produced a growth defect (especially considering those cases where multiple gRNAs to the same ORF show no growth defect). Because we have access to individual strains, one could isolate such strains and confirm via qPCR if gene repression is indeed occurring.

**Discussion**

Our process represents a powerful and inexpensive method to obtain individual sequence validated oligos by leveraging the power of yeast genetics and high throughput sequencing. To obtain a similar number of individual oligos as obtained in Figure 4.2B with conventional oligo synthesis, the cost would be approximately $385,000 (IDT costs are 80 cents per base for PAGE-purified ultramers), but with our method the cost is only around $3,000. This presents a huge advantage when one needs a large number of high quality, sequence verified segments of synthetic DNA.

Additionally, our process can be adapted to create functional CRISPRi repression strains. This allowed us to create a collection of inducible repression strains to the majority of essential and aerobic essential genes that can be assayed in pools for competitive fitness experiments or studied as individual strains. Unlike other collections that took a great amount of labor and expense, our collection was relatively easy and inexpensive to create. Experiments with this collection were fruitful, and we were able to confirm our finding that chromatin accessibility affects guide efficacy, and we determined that within our 0 to 200nt window upstream of the TSS, guides within 100nt of the TSS are the most functional. We can now look for additional factors that influence gRNA efficacy that may have been missed by our smaller-scale study in Chapter 3.

In addition to parsing oligo libraries, our method could be extended to other DNA libraries. With longer reads available with platforms such as Pacific Biosciences, we could parse cDNA libraries, pools of degenerate gene sequences for variant characterization, assembled full length synthetic genes for sequence validation, and other pools of longer sequences of DNA. To do this, we’d use an architecture
similar to that of the CRISPRi recipient strain, where the DNA sequence in question sits outside of the synthetic intron.

However, our method does have some limitations. We note, that retrieving DNA oligos that are poorly represented in the original array-derived library requires processing/screening many more transformants than what is required highly abundant oligos. Our method thus yields diminishing returns that are directly correlated to the composition and quality of the original library. For applications that do not require each and every one of the designed oligonucleotides to be retrieved from the mixture (such as those described here), this does not represent major limitation. On the other hand, to enable applications that require all of the designed DNA fragments to be retrieved (e.g. gene synthesis), our method will require more uniform libraries and/or improved strategies for isolating rare entities.

**Methods:**

**Construction of Barcoder Strains.**

Approximately 1536 Barcoder strains were used for high-throughput tagging of the exogenous DNA incorporated in recipient strains. These barcoder strains are derivatives of SHA345 (MATa, his3Δ, leu2Δ, met15Δ, ura3Δ, ybr209w::GalCre-NatMX, can1::MFalpha1pr-LEU2). SHA345, and the process for incorporating barcodes into this strain is described elsewhere [198]. Briefly, random barcodes were ordered as primers (IDT) and inserted into a plasmid backbone by ligation, generating a plasmid library (U3Kan66) that contains a partially crippled loxP site (lox66) [199, 200], the barcode region, the 3’ end of URA3 gene preceded by part of an artificial intron [201] and the KanMX dominant drug resistant marker. We used the DNA from this plasmid library to replace (by yeast homologous recombination) the NatMX cassette in SHA345 with lox66-Barcode-3′URA3-URA3term-KanMX. This yields SHA345+BC strains (MATa, his3Δ, leu2Δ, met15Δ, ura3Δ, ybr209w::GalCre-lox66-Barcode-3′URA3-URA3term-KanMX, can1::MFalpha1pr-URA3-MFa1pr-LEU2), where each strain contains a unique barcode. The genotype of each SHA345+BC strain was verified by assaying for growth on YPD+G418 (for KanMX), YPD+nourseothricin (for NatMX), and
CSM+galactose -uracil following mating to a “tester strain”. The barcode sequence of roughly 1100 of the SHA345+BC strains were identified by Sanger sequencing. The barcode sequence of all SHA345+BC strains were verified/identified using additional barcoder strains of the opposite mating type (see below). Mating of arrayed SHA345+BC strains to transformed recipient strains in 1536 format, followed by selection for loxP recombinants on CSM-uracil+galactose media, results in tagging (i.e. barcoding) of the exogenous DNA incorporated in recipient strains.

MATalpha barcoder strains derived from SHA349 (MATalpha, his3Δ, leu2Δ, lys2Δ, ura3Δ, ybr209w::NatMX-GalCre, can1::MFalpha1pr-HIS3-MFalpha1pr-LEU2) were also used in this study. As above, random barcodes were ordered as primers (IDT) and inserted into a plasmid backbone by ligation, generating a second plasmid library (U5Kan71), which contains a partially crippled loxP site (lox71) [199, 200], the barcode region, the promoter and 5’ end of URA3 gene followed by part of an artificial intron[201], and the KanMX dominant drug resistant marker[202]. Homologous recombination was used to replace the NatMX cassette in SHA349 with lox71-Barcode-URA3prom-5’URA3-KanMX. This produced strain #2797 (the parent for all recipient strains), as well as 768 SHA349+BC barcoder strains (MATalpha, his3Δ, leu2Δ, lys2Δ, ura3Δ, ybr209w::KanMX-URA3prom-5’URA3-Barcode-lox71-GalCre, can1::MFalpha1pr-HIS3-MFalpha1pr-LEU2). The barcode sequence in each SHA349+BC strain was verified by Sanger sequencing. We used these strains to verify/identify the barcode sequences in our SHA345+BC strains. Briefly, SHA349+BC strains were mated to SHA345+BC strains in array format, and diploids containing ‘double-barcodes’ were selected on CSM-uracil+galactose media. Collective amplification of double-barcodes, followed by Illumina sequencing, was used to verify and/or identify the barcode in each SHA345+BC strain.

**Construction of Recipient Strains**

Recipient strains are designed for high-efficiency incorporation of DNA libraries by transformation, and their subsequent tagging by mating to Barcoder strains (which are described below). All recipient strains used in this study are derivatives of strain #2797 (MATalpha, his3Δ, leu2Δ, lys2Δ, ura3Δ, ybr209w::KanMX-URA3prom-
5′URA3-Barcode-lox71-GalCre, can1::MFalpha1pr-HIS3-MFalpha1pr-LEU2), a derivative of SHA349 (MATalpha, his3Δ, leu2Δ, lys2Δ, ura3Δ, ybr209w::NatMX-GalCre, can1::MFalpha1pr-HIS3-MFalpha1pr-LEU2), that contains a partially crippled loxP site (lox71) [199, 200], a 26 bp barcode (TGCCTAAGCAGGAAGTGTGTTTGCAAC), the promoter and 5′ end of URA3 gene followed by part of an artificial intron [201], and the KanMX dominant drug resistant marker[202]. SHA349 is described elsewhere [198] and additional details on the construction of strain #2797 can be found below (Barcoder strains). To create a recipient strain, #2797 was modified as follows. First, wild-type FCY1 was replaced with the HphMX-cassette (Hygromycin B resistance cassette)[203]; the HphMX-cassette was PCR-amplified with primers 957 and 958 (yielding a 1656 bp product), and transformed into #2797. Transformants were selected on YPD+HygB. HygB-resistant clones were confirmed to grow on YNB+AS+Dex+leu+his+ura+5-fluorocytosine, and were confirmed to not grow on YNB+AS+leu+his+cytosine (further confirming loss of FCY1). The fcy1::HphMX deletion was also confirmed by PCR using primers 640 and 641 (yielding a 2081 bp product). The resulting strain was #2836. Next, the SceI-FCY1prom-FCY1-SceI cassette was inserted between the lox71 site and the 26 bp barcode of strain #2836; the 1089 bp cassette was amplified by PCR with primers 784 and 785 (which each contains one SceI site) and using the plasmid pJH143 as template, and transformed into #2836. Transformants were selected on CSM-uracil+cytosine+HygB. Successful transformants were confirmed to not grow in YPD+5-fluorocytosine (5-FC), indicating the presence of FCY1 cassette, and also by PCR using primers P45 and P40 (yielding a 1414 bp product). The resulting strain, #2849, served as the initial recipient strain, and was used in the experiments described in Figure 4.1.

**Mito-Repaired Recipient Strain**

We noted that strain #2849 was prone to spontaneous loss of mitochondrial DNA which can negatively, and unpredictably, affect growth of transformants. We therefore repaired alleles at three loci known to impact mitochondrial genome stability (SAL1, CAT5, and MIPI). First, we corrected the sal1-1 allele to wild-type SAL1 using the mega 50:50 method (and improved version of the 50:50 method described in
A correct recombinant was identified by genomic PCR and confirmed by Sanger sequencing. The resulting strain is JHY627. Next, we converted CAT5(91I) to CAT5(91M) using primers CAT5.80.1 and CAT5.80.2 together with PCR template pJH140. A correct recombinant was identified by genomic PCR and confirmed by Sanger sequencing. The resulting strain is JHY629. Finally, we converted MIP1(661A) to MIP1(661T) using a two-step allele replacement strategy using pLND44-4 as described by Dimitrov [205]. A correct recombinant was identified by Sanger sequencing of both the MIP1 QTL and a region 988 bp downstream that has a known plasmid error. The resulting strain is JHY650. DNA from the three QTLs was PCR amplified from the final strain, JHY650, and Sanger sequenced to confirm the desired alleles. All alleles were correct, and the pLND44-4 plasmid sequence error noted by Dimitrov was not present. In addition the barcoding/artificial-intron region of JHY650 strain was also confirmed by Sanger sequencing.

Construction of CRIPSri Recipient Strain

A recipient strain for guide RNAs targeting yeast genes for transcriptional repression by dCas9-Mxi1 was also created. This strain facilitates expression of fully-functional gRNAs from a tetracycline-regulatable RPR1 promoter, and also constitutively expresses a dCas9-Mxi1 fusion protein as in chapter 3. To create this strain, we first created a cassette containing SceI-FCY1prom-FCY1-SceI flanked by the RPR1-promoter and the structural part of the gRNA plus the RPR1-terminator. This integration cassette can be seen here: https://benchling.com/s/qpwEfqmX/edit and a illustration in Figure 4.3A. Briefly, SceI-FCY1prom-FCY1-SceI DNA was amplified by PCR with 128-pRPR1-FCY1-fwd and 129-gRNA-FCY1-rev from genomic DNA from the original recipient strain (#2849). The cassette was inserted into the NotI restriction site of pRS416gT-Mxi1 plasmid from chapter 3. From this plasmid a 1769 bp cassette was PCR-amplified with primers 131-pRPR1-int-fwd and 130-RPR1t-int-rev, and transformed into strain #2836. The cassette was inserted by homologous recombination adjacent to the URA3 marker, thereby replacing the KanMX marker. Transformants were selected on YNB+AS+leu+his+cytosine.
Transformants were also screened for loss of G418-resistance (*i.e.* loss of KanMX cassette), and successful integration was also confirmed by PCR using primers 894 and 895 (which yielded a 2183 bp product). The resulting strain is #2869. Next, the HphMX cassette at the FCY1 locus was replaced with KanMX by PCR amplification of KanMX cassette (with primers 142 and 143 (1282 bp), transformation into strain #2869, and then selection on YPD media containing G418. This resulted in strain #2877.

A cassette was then created to enable the incorporation of dCas9-Mxi1 into strains #2877. First, the HphMX cassette was amplified by PCR using the primers 133-Tef1-HphMX-Fwd and 132-HphMX-Rev, and inserted in pRS416gT-Mxi1 (Chapter 3) between dCas9-Mxi1 repressor and the Tet repressor with Gibson assembly. A map can be found here: [https://benchling.com/s/2Gki8et5/edit](https://benchling.com/s/2Gki8et5/edit). Primers 134-Mid-HphMX-Fwd and 137-site18dn-M13F-Rev were then used to amplify the dCas9 portion of the plasmid, and primers 135-Mid-HphMx-Rev and 136-site18up-GPM1-fwd were used to amplify the TetR portion of the plasmid. Each PCR product contained overlapping fragments of the HphMX cassette. These two PCR products were co-transformed into strain #2877, and integrated into site 18 (YORWdelta17 XV) from Flagfeldt et al., 2009 [206]. This site had been characterized as a good site to express heterologous proteins. Transformation was recovered overnight in YPD+Hyg liquid media and then further selected on YPD+Hyg agar plates. Successful integration was confirmed by PCR with primers 138-Site18-YOWdelta17-FWD and 139-Site18-YOWdelta17-REV, which sit outside the site and amplify only if the proper product is present. This strain was #2877-dCas9-Mxi1. Effective CRISPRi in this strain was confirmed by transforming it with gRNAs directed against the SEC14 and ERG25 genes, which are known to produce growth defects upon expression with ATc (as in Chapter 3). All primers, strains, and plasmids are available in Table S4.5.

**Mito-Repaired CRISPRi Recipient Strain**

We set out to improve the #2877-dCas9-Mxi1 as previously described with #2849. We used the 50:50 method[204] to alter alleles at two of the loci known to
impact mitochondrial genome stability. We corrected the sal1-1 allele to wild-type SAL1 using primers SAL1.80.1 and SAL1.80.5 and PCR template pJH136. Cassette integration was identified by growth on CSM-Ura media, and confirmed by PCR with primers SAL1.80.3 and URA3.34.6. Loss of the plasmid was determined genomic PCR with primers SAL1.3 and SAL1.4 of FOA resistant colonies, which were then Sanger sequenced for identification of the correct wild-type allele. We used the SAL1 corrected strain as the starting strain to repair the CAT5 (91I) allele to CAT5(91M) in the same manner, using CAT5.80.1 and CAT5.80.5 primers for cassette construction, CAT5.80.3 and URA3.34.6 primers for integration confirmation. FOA resistant colonies were Sanger sequenced to identify the CAT5(91M) isolates. These two RM11 allele changes reduce the petite frequency by eight-fold[205] which is sufficient improvement for our experiments. The genotype of the resulting strain, yACJ2, is: αLpha hisD1 leu2D0 lys2D0 ura3D0, ybr209w::(R)GalCre-KanMX, yorwdelta17::tef1p-dCas9-Mxi1-cyc1t HphMX pgpm1p-TetR-gpm1t can1::MFApr1-His3-MFAplα1pr-LEU2 SAL1 CAT5(91M).

Oligo Library Parsing

Oligos are amplified by PCR with primers that add 90nt overlaps for integration and are co-transformed with linear DNA (PCR product) for the Scel meganuclease into either strain #2849 or yACJ2. Each strain has a landing pad consisting of LoxP, Scel site flanked FCY1, Artificial Intron sequence (AI), and half of the URA3 gene although the orientation differs (see CRISPRi Recipient Strain Modification). Transformants were selected on 5-flourocytosine (5-FC) plates for successful removal of FCY1 signaling an integration event. Colonies were picked by a colony picker robot and arrayed onto 384 well plates. These plates of arrayed unknown strains were then mated against plates with barcoder strains arrayed in known positions using a Singer robot onto a synthetic complete minus Uracil, plus Galactose plate. Colonies were then washed from a plate, combined, an genomic prepped using YeaStar™ Genomic DNA Kit (Zymo Research). Samples from individual plates were barcoded by PCR primers to specify which plate the DNA came from. PCR products were confirmed by agarose gel electrophoresis. After PCR,
samples were combined and bead cleaned with Thermo Scientific™ Sera-Mag Speed Beads Carboxylate-Modified particles.

Paired end sequencing was performed using Illumina MiSeq. Sequence data was used to determine which oligos were linked to each barcode and in turn what position on which plate they are located. One colony for each correct unique sequence obtained was then cherry picked from the plate using a ROTOR Stinger robot (Singer Instrument) onto a new 384 well plate if it met the minimum requirements of 50 correct reads. For more details see Figure 4.1.

**Molecular Probe Oligo Set**

We designed 7051 molecular probes for detection of 321 bacteria (~20 probes per bacteria) related to potable water quality [192]. Each probe was synthesized on a 145 base oligo which contained 60 bases of homology to the target genome, a 9 base ‘random’ barcode, and 36 bases of priming sequence for amplification of probes that successfully hybridize to their targets. In addition to these sequence features, each probe was flanked by 20 bases of common priming sequence (for amplification of the library following array-based synthesis) and type IIS restriction sites (for removal of these common priming sequences). These probes were ordered from CustomArray. Before integration these oligos were amplified with primers that added 40nt of homology for integration into the FCYI site in #2849.

After parsing with the method described above, the selected probe collection was sequenced by using Illumina MiSeq sequencing. The composition of sequences where read 1 was a perfect match to read 2, and omitting singleton reads, was compared to the expected sequences for probes that were intended to be in in the set and relative abundance of each of these sequences.

**CRISPRi ORF Selection**

A list of essential gRNAs (null phenotype inviable) was downloaded from the Saccharomyces Genome Database (SGD) (yeastgenome.org)[197]. We filtered to only include essential genes for S288C. Aerobic essentials were defined as the 514 respiratory deficient genes identified by the union of respiratory deficient ORFs in Steinmetz et al [207] and Schlecht et al [208]. We also included ADE2, a non essential
gene that will give a red phenotype in null mutants[195]. We hoped to be able to test for this non-growth phenotype in individual strains isolated from the collection.

**gRNA Design and Cloning Strategy**

Guides were designed to fall with a range of 200bp upstream of the TSS of a gene as determined by transcript isoform profiling [180]. If fewer than 5 guide targets were available in this window, guides were designed to a window of 300bp upstream of the TSS. If there was not a defined TSS for an ORF, we designed guides to a window of 300bp upstream of the ORF start.

Two gRNA specificity sequences flanked with constant overlaps to the pRPR1 promoter and gRNA constant region were synthesized per oligo with a middle sequence separating them. This unique middle sequence allowed for the selective amplification of either half of the oligo during the first round of PCR (Figure 4.8).

With 5 different unique middle sequences we, were able to subsection the library into 10 different sub-libraries. 6 of these sub-libraries contained ~2100 gRNAs each from the essential set, and 4 contained ~1600 gRNAs each for the aerobic essential set.

Guides were binned into the subsets depending on the number of guides per ORF, so that more colony picking and sequencing effort could be focused on these sets so that gRNAs to ORFs with only a few gRNAs available would be more likely to be represented in our collection. Full list of oligos designed is available in Table S4.1.
Figure 4.8. Oligo design and PCR strategy.
Two gRNAs were designed per oligo separated by a unique middle sequence. Depending on the direction of the primer complementary to the middle sequence used, the left and right sequences can be amplified separately. After this initial round of PCR, a second round of PCR allows for extension of the overlaps for homologous recombination at the target site in the genome.

**CRISPRi Competitive Growth Assays**

Prior to setting up experiments, aliquots of a library were recovered in YPD media for 4 hours, and then diluted appropriately for the experiments. Yeast culturing and sample collection was performed using a cell-screening platform that integrates temperature-controlled absorbance plate readers, plate coolers, and a liquid handling robot. Briefly, 700 µl yeast cultures were grown (+/- ATc in YPD or buffered YPEG media, [Figure 4.4](#)) in 48 well plates at 30°C with orbital shaking in Infinite plate readers (Tecan). To maintain cultures in log phase over many doublings, 80 µls of the culture was removed when it reached an OD of 0.76, added to a well containing 620 µl of media, and then allowed to grow further. After three such dilutions, 600 µls of the culture was collected and saved to a 4°C cooling station (Torrey Pines) when it reached an OD of 0.76. This amounted to approximately 12 culture doublings from the beginning of the experiment. Pipetting events were triggered automatically by Pegasus Software and performed by a Freedom EVO workstation (Tecan).

After sample collection, yeast genomic DNA was purified using the YeaStar™ Genomic DNA Kit (Zymo Research). Purified genomic DNA was used as a template for PCR with barcoded up and down sequencing primers that produce a double index to uniquely identify each sample. PCR products were confirmed by agarose gel electrophoresis. After PCR, samples were combined and bead cleaned with Thermo ScientificTM Sera-Mag Speed Beads Carboxylate-Modified particles. Sequencing was performed using Illumina MiSeq. Two separate experiments were combined to give the overall strain counts used in these experiments. One experiment contained 722 strains, the other 8337 distinct strains.

**CRISPRi Data Analysis**

gRNA counts were generated by counting all perfect matches where Read 1 = Read 2 within the region, and then mapping these sequence counts to gRNAs that we had designed. Counts were then loaded into R and analyzed with DESeq2[209] to
calculate normalized log2 fold change values. In order to perform these calculations, we added one count uniformly to all counts. Additionally, we filtered out reads that had a row sum less than 200 across the samples. Sequencing counts and log2 fold change values were calculated separately for the experiment with 722 strains and the experiment with 8337 strains and merged afterwards. This is the cause of the majority of strains with greater counts in the replicates plot in Figure 4.5. All gRNAs were annotated with TSS, nucleosome, gRNA midpoint, normalized ATAC-seq, gRNA location, nearby TSS data from the webtool we created in Chapter 3 (http://lp2.github.io/yeast-crispri/). We also added data for the presence or absence of a DNAse I hypersensitive site located within 10nt in either direction of the gRNA midpoint. DNAseI hypersensitivity data was downloaded from SGD[196]. All gRNAs obtained through yeast parsing are available in Table S4.2. These annotated data sets with counts and log2 fold change values are available in Table S4.3 (YPD) and Table S4.3 (YPEG).
Chapter 5: Distinct patterns of Cas9 mismatch tolerance in vitro and in vivo

Authors
1Becky Xu Hua Fu, 3,4 Robert P. St. Onge, 1,2 Andrew Z. Fire, 1,3 Justin D. Smith

1Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA
2Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA
3Stanford Genome Technology Center, Palo Alto, CA, 94304, USA
4Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305, USA

This chapter represents a draft of a paper that has since been published in Nucleic Acids Research. The dissertation author [JDS] was responsible for proposing the idea for this project, building the plasmid libraries, and conducting all the in vivo experiments. JDS, BXHF, and AZF designed the experiments. BXHF conducted the in vitro experiments and sequencing. BXHF, AZF, and JDS analyzed the data, created the figures, and wrote the manuscript. RPS advised on the project and assisted on the automated in vivo growth experiments.

Abstract

Cas9, a CRISPR-associated RNA-guided nuclease, has been rapidly adopted as a tool for biochemical and genetic manipulation of DNA. Although Cas9 offers remarkable specificity and versatility for genome manipulation, mis-targeted events occur. To extend the understanding of Cas9 target::homology requirements, we compared mismatch tolerance for a specific Cas9::gRNA complex in vitro and in vivo (in Saccharomyces cerevisiae). A variety of truncated and full-length gRNAs (with 17, 18, and 20 nucleotides (nt) of complementarity sequence) were used with four different targets (with varying GC content and sequence features). In each case, we observed notable differences between in vitro and in vivo Cas9 cleavage specificity profiles, with a more stringent effect of mismatches on activity seen in vivo. Both lengths of truncated gRNAs showed an increase in specificity in vivo, with most targets, but not in vitro. Overall, this study highlights differences in the specificity of Cas9 cleavage between controlled in vitro conditions and chromatinized in vivo conditions.
Introduction

The CRISPR-Cas (clustered regularly interspaced short palindromic repeats) immune system evolved naturally in archaea and prokaryotes and has been adapted as a tool for genome editing and manipulation in numerous organisms. *Streptococcus pyogenes* Cas9 is the most well-characterized nuclease used in CRISPR-Cas genome editing[210–212]. Cas9 can be directed to specific genomic sites with a programmable RNA called the guide RNA (gRNA)[213], which includes 17-20 nucleotides of the target sequence complementarity next to a protospacer adjacent motif (PAM), 5’-NGG-3’. Cas9 generates a blunt double-stranded DNA break that can be repaired by nonhomologous end-joining or homologous recombination (if a template for repair is provided)[70, 211]. Due to the broad range of research utilizing Cas9 as technology and its potential in clinical applications, much effort has been put into understanding the requirements for gRNA::target complementarity for precise binding and cleavage[70, 212]. Functional studies have, in particular, identified mismatches in the PAM and ‘seed’ region (10-12 bp proximal to the PAM) as having strong effects [72, 154, 213, 214].

Despite this area of general agreement, reported requirements through the homology region provide some apparent paradoxes. For example, some studies have reported that the ‘seed’ region can tolerate mismatches [72, 215] while some report little to no tolerance[216]. These differences could be the result of varying target sequences, differences between conditions *in vitro* and *in vivo*, or other variations in the assays employed. Since Cas9 whole protein complex with gRNA and Cas9 expressed off of plasmids are popular methods of genome editing technology, we specifically wanted to assay Cas9 target mismatch tolerance under *in vitro* and *in vivo* conditions for these purposes. In this work we use previously characterized variant target libraries[72] and a newly developed variant library containing four different target sequences to assay Cas9 cleavage *in vitro* and *in vivo* with full-length and 5' truncated gRNAs to provide parallel characterizations of homology requirements for cleavage. We also sought to compare the specificity profiles of two targets characterized in human cells via fluorescence assays to the specificity profiles
determined through our previously developed *in vitro* method[72] and an extension of that method *in vivo* in *S. cerevisiae*. Finally, we sought to determine if gRNAs with a GG motif at the 3’ end of their target-specific sequences which were recently determined to dramatically increase gene editing efficiency in *Caenorhabditis elegans* would have an effect on Cas9 specificity.

**Results and Discussion**

We adapted a previous high throughput sequencing approach[72] to assess the effects of single base variants on Cas9-mediated cleavage *in vivo*. We assayed Cas9 cleavage using different gRNA species with distinct variant libraries of four distinct target sequences. The first random variant library (type I) was constructed through degenerate oligonucleotide synthesis followed by plasmid cloning. For the type I variant library, as the target (and starting point for degenerate oligonucleotide synthesis), we used a segment ('unc-22A') from the *C. elegans unc-22* gene[72] (Figure 5.1a). Assays were carried out with gRNAs of four different lengths (17 nt, 20 nt, and 20+G nt of *unc-22A* complementarity respectively Table 5.1). In order to extend the study, a second variant library (type II) was made using specified oligonucleotide synthesis for all single base variants, single base deletions, as well as a set of double adjacent variants for four targets: EGFP-1 (EGFP site 1), EGFP-2 (EGFP site 2), rol-6, and *unc-22A* (Figure 5.1a). These sequences represented a range of GC contents (80%, 50%, 45%, and 70% respectively). EGFP-1 and EGFP-2 are targets from the EGFP gene characterized previously in mammalian cells [217]. The *rol-6* target is a gRNA previously used for co-conversion in *C. elegans* genome editing[218, 219]. The *unc-22A* target was previously characterized[72]. Both the EGFP-1 and *rol-6* gRNAs contain a 3’ GG motif, but have dramatically different GC contents. A list of all sequences of the type II library can be found in Table S5.3. The results of the type II libraries are discussed in the main text while the results of type I libraries are in the supplemental.
Table 5.1)

<table>
<thead>
<tr>
<th>gRNA name</th>
<th>gRNA target sequence</th>
<th>gRNA target</th>
<th>gRNA length</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGFP-1-F&gt;</td>
<td>GGCACGGGCAAGTTCGGCGG&gt;</td>
<td>EGFP-1&gt;</td>
<td>20 nt</td>
</tr>
<tr>
<td>EGFP-1-F18&gt;</td>
<td>GACACGGGCAAGTTCGGCGG&gt;</td>
<td>EGFP-1&gt;</td>
<td>18 nt</td>
</tr>
<tr>
<td>EGFP-1-F17&gt;</td>
<td>CACCGGCAAGTTCGGCGG&gt;</td>
<td>EGFP-1&gt;</td>
<td>17 nt</td>
</tr>
<tr>
<td>EGFP-1-R&gt;</td>
<td>CCGCAAGCTGGGCCGCTGCC&gt;</td>
<td>EGFP-1&gt;</td>
<td>20 nt</td>
</tr>
<tr>
<td>EGFP-1-R18&gt;</td>
<td>GCGACCTGGCCGCTGCC&gt;</td>
<td>EGFP-1&gt;</td>
<td>18 nt</td>
</tr>
<tr>
<td>EGFP-1-R17&gt;</td>
<td>GCAAGCTGGCCGCTGCC&gt;</td>
<td>EGFP-1&gt;</td>
<td>17 nt</td>
</tr>
<tr>
<td>EGFP-2&gt;</td>
<td>GATGCGTTCTTCCTGCTTGT&gt;</td>
<td>EGFP-2&gt;</td>
<td>20 nt</td>
</tr>
<tr>
<td>EGFP-2-18&gt;</td>
<td>TGCCGTTCTTCCTGCTTGT&gt;</td>
<td>EGFP-2&gt;</td>
<td>18 nt</td>
</tr>
<tr>
<td>EGFP-2-17&gt;</td>
<td>GCGTTCTTCCTGCTTGT&gt;</td>
<td>EGFP-2&gt;</td>
<td>17 nt</td>
</tr>
<tr>
<td>Rol-6&gt;</td>
<td>GTGAGACGTCAACATATGG&gt;</td>
<td>rol-6&gt;</td>
<td>20 nt</td>
</tr>
<tr>
<td>Rol-6-18&gt;</td>
<td>GAGACGTCAACATATGG&gt;</td>
<td>rol-6&gt;</td>
<td>18 nt</td>
</tr>
<tr>
<td>Rol-6-17&gt;</td>
<td>AGACGTCAACATATGG&gt;</td>
<td>rol-6&gt;</td>
<td>17 nt</td>
</tr>
<tr>
<td>EGFP-1-OS&gt;</td>
<td>CCAGCCGAGGCGAGCTTGC&gt;</td>
<td>EGFP-1&gt;</td>
<td>20 nt</td>
</tr>
<tr>
<td>EGFP-1-OS18&gt;</td>
<td>AGGGCAAGGCAGCTTGC&gt;</td>
<td>EGFP-1&gt;</td>
<td>18 nt</td>
</tr>
<tr>
<td>EGFP-1-OS17&gt;</td>
<td>GGGCAAGGCAGCTTGC&gt;</td>
<td>EGFP-1&gt;</td>
<td>17 nt</td>
</tr>
<tr>
<td>unc-22a-20&gt;</td>
<td>GCCACCGGTCGGCCGCAATT&gt;</td>
<td>unc-22a&gt;</td>
<td>20 nt</td>
</tr>
<tr>
<td>unc-22a-18&gt;</td>
<td>ACCACCGGTCGGCCGCAATT&gt;</td>
<td>unc-22a&gt;</td>
<td>18 nt</td>
</tr>
<tr>
<td>unc-22a-17&gt;</td>
<td>CCACCGGTCGGCCGCAATT&gt;</td>
<td>unc-22a&gt;</td>
<td>17 nt</td>
</tr>
</tbody>
</table>

Table 5.1: Table of all gRNAs used in experiments and their corresponding targets and sequences.

In each experiment, we challenged a library of potential target sequences with a single Cas9::gRNA species and examined sequences that remained intact (indicating failure to cleave). Each sequence variant is represented by several distinctly "barcoded" sequence species in the initial library, with robustness of results supported by consistent results among the barcoded sub-pools[72]. As previously described[72], a log retention is calculated for each sequence. A positive retention suggests lack of cleavage while a negative retention suggests cleavage. For the yeast in vivo assay, the library of target plasmids was maintained continually in yeast, during expression of both Cas9 with a nuclear localization signal (NLS) [220] and the relevant gRNAs (Figure 5.1a-b). Following growth under inducing conditions (galactose +
anhydrotetracycline (ATc)), the resulting change in target representation was assessed as described above.

For the type I variant library, all sequences with 4–7 mismatches to the original trigger served as an ‘internal control’ as well as distinct sequences spiked in from a different target library (‘protospacer 4’ a.k.a ‘PS4’), with comparable results from both normalization methods. The type I library derives from oligonucleotides synthesized to have a baseline variation rate in gRNA homology region with a 6 nt region of random sequence flanking the gRNA and PAM homology regions[72] (Figure 5.1a). Two independently created unc-22A random variant pools from the previous study were used as a starting material in this analysis (the pools were those originally designated unc-22A RVL-1 and unc-22A RVL-2[72]). For the type II library design, the target of interest normalized against all target sequences not related to the gRNA in question. For example, if the gRNA being assayed was EGFP-1, then all the target sequences of EGFP-2, rol-6, and unc-22A were used for normalization. Type I libraries were PCR amplified from the plasmid pool from reference [72] and inserted into a yeast vector containing galactose inducible Cas9 and a tetracycline inducible unc-22A gRNA, producing a library of plasmids as in Figure 5.1b. Type II libraries were cloned into the multiple cloning site of pRS415 and gRNAs were provided on a separate plasmid (pRS416) which contained Cas9 and the tetracycline repressor as in Figure S5.1 but which lacked a target library (Figure 5.1c). Figure 5.1b-c depicts the workflow process for the massively parallel assays with these libraries.
**Figure 5.1 Target Sequences and Workflow**

a. The target sequences used for type I and II variant libraries.

b. To create a yeast target library, we first inserted the relevant unc-22A homologous sequences into the Tet-inducible guide RNA cassette in a yeast Cen/Ars vector also containing inducible Cas9. The resulting plasmids were linearized (SacII) for insertion of fragments containing target variants from the amplified target library of Fu et al. 2014\[69\]. Insertion of target variants employed Gibson Assembly\[145\], using primers giving appropriate overlaps. Following transformation into *E. coli*, we miniprepped each library and used the library for both *in vivo* and *in vitro* experiments. For the *in vivo* experiments the library was transformed into haploid *S. cerevisiae*, followed by Cas9 and gRNA induction with galactose and ATC respectively for 10 generations. Cas9/gRNA cut plasmids should then be lost, while retained plasmids were observed by PCR and sequencing (from yeast plasmid minipreps obtained after 10 generations). For the *in vitro* experiments, the *E. coli* miniprep plasmids libraries were cut *in vitro* in reactions with a mixture of purified Cas9 enzyme and *in vitro* transcribed unc-22AgRNA. For both methods the retained plasmids were isolated and amplified and sequenced and retention profiles were determined.

c. As in 1b except: Type II target libraries were created from microarray oligo pools designed to four different target sequences. These pools were PCR amplified and integrated into the multiple cloning site of a leucine selectable Cen/Ars plasmid using Gibson Assembly and *E. coli* cloning as in 1b. This library was miniprepped and used for *in vitro* experiments exactly as in 1b. For the *in vivo* experiments, the Cas9 and Tet-inducible gRNA genes were on a separate Uracil selectable Cen/Ars plasmid from the target library. First, we transformed the Cas9/gRNA plasmid into haploid *S. cerevisiae*. Next the library was transformed into the strains containing the Cas9/gRNA plasmid during a second round of transformation. During *in vivo* cutting, plasmid selection was maintained for both plasmids. Otherwise the protocols were identical.

**Effects of single base variants on Cas9 in vivo and in vitro cleavage specificity**

From these data, we see a range of retention values, varying from sequences that are removed completely from the pool, to those that are barely affected by Cas9 nuclease activity. For type II libraries, all four targets cut with full-length gRNAs exhibited intolerance *in vivo* to variation in the ‘seed’ region (positions 11-20), but tolerated variation in the distal region (**Figure 5.2**). Results for the type I variant libraries showed similar patterns for the unc-22A gRNAs (**Figure S5.2,S5.12,S5.13**). **Figure 5.2a-d** depicts the retention values of all transversion variants following expression of the complementarity guide. For graphical simplicity only the results for transversion variants are shown in **Figure 5.2** but transition variants showed comparable retention values to transversions (**Figure S5.3**). There is an evident dip in the distal region (positions 1-10) indicating the relaxation of sequence requirements outside the seed region.
Single base transversion effects on EGFP-1 target \textit{in vivo}

A

Single base transversion effects on EGFP-2 target \textit{in vivo}

B

Single base transversion effects on \textit{rol-6} target \textit{in vivo}

C

Single base transversion effects on \textit{unc-22a} target

D

- \textbf{Time = 0 generations}
- \textcolor{orange}{Time = 10 generations, Guide=17nt}
- \textcolor{teal}{Time = 10 generations, Guide=18nt}
- \textcolor{magenta}{Time = 10 generations, Guide=20nt}
Figure 5.2 *In vivo* cleavage effects of single base variants

with a) EGFP-1, b) EGFP-2, c) *rol*-6, and d) *unc*-22A gRNAs at 10 generations. A median of effects for single base transversion variants is indicated for each position by a dot and the colors denote the length of the gRNA. This graph shows the transversion variant retention for the target region and constrained PAM nucleotides (positions 1-20 and 22-23). Only sequences with >=50 reads in the control (uncut) library were considered in calculating this median. All targets that are not complementary to the gRNA of interest are negative controls (labeled control) and wild-type targets are positive controls (labeled WT). A negative retention score indicates sequence cleavage, while a retention score of zero corresponds to samples whose cleavage is comparable to the pool of non-complementary sites. A slightly positive retention score indicates a lack of Cas9 cleavage. Over-representation is due to noise or consistent PCR effects.

Truncated gRNAs have been reported to increase the specificity of Cas9 *in vivo* in human cells[217, 221]. Examining gRNAs with 18 nt and 17nt complementarity in the yeast *in vivo* Cas9 cleavage assay, we observed most targets had a striking increase in specificity (over the gRNA with 20 nt complementarity) with at least one of either the 17 nt or 18 nt complementarities. In particular, transversions throughout both seed and distal regions led to a dramatic drop in cleavage by the 18 nt complementarity Cas9::gRNA complex for the EGFP-2, *rol*-6, and *unc*-22A targets. The 18 nt complementarity gRNA for the EGFP-1 target had a similar retention profile as the full-length while the 17nt gRNA showed intolerance to transversion variants throughout the gRNA::DNA complementarity region. For both type I and type II libraries, attempts to use a shorter gRNA (17 nt complementarity) were unsuccessful for the *unc*-22A target for reasons that were not investigated further (Figure S5.2).

The observed *in vivo* cleavage profiles of the gRNAs are notably different from previously described *in vitro* cleavage profiles with the *unc*-22A parent plasmid library and the same gRNA[72]. We tested the effects of variants *in vitro* for the yeast random *unc*-22A variant library type I and all four targets with the type II variant libraries with New England Biolabs (NEB) Cas9. Figure 5.3 shows the results of targeting EGFP-1, EGFP-2, *rol*-6, and *unc*-22A targets with full-length and truncated gRNAs (17, 18, and 20 nt complementarity) with 3 minutes of incubation. Results for the longer incubation times for the type II library can be found in Figure S5.5. The cleavage profiles for all guides show similar "horseshoe" patterns previously described in Fu et al. 2014[72]. Similar *in vitro* assays were performed on the type I yeast.
variant library for the \textit{unc-22A} gRNAs (17, 18, and 20 nt complementarity) and similar cleavage patterns were observed (Figure S5.4).

\textbf{Figure 5.3 In vitro cleavage effects of single base variants} with a) EGFP-1, b) EGFP-2, c) \textit{rol-6}, and d) \textit{unc-22A} gRNAs with 3 minutes incubation. The method of generating these graphs is identical to Figure 5.2. Type I library can be seen in S5.4 and 180 minute time points are available in Figures S5.5 and S5.10.
For all gRNAs tested in vitro, the early time points show some stringency for the seed region but the specificity is lost with longer incubation times (Figure S5.5 & S5.10). The truncated gRNAs (17 and 18 nt complementarity) and the full-length 20 nt complementarity gRNA have similar cleavage specificity profiles in vitro for all time points for all targets. We did not observe and improvement in specificity of Cas9 cleavage with the in vitro conditions tested.

To ensure temperature was not the cause of differences observed between in vivo and in vitro experiments, the in vitro assay with full-length gRNA was repeated at 30°C, the yeast in vivo incubation temperature with yeast random variant library type I (Figure S5.6). The temperature change slowed Cas9 cleavage activity in vitro but the general pattern of the cleavage profile did not vary. In addition, the cleavage assay was also performed with an independent source of Cas9 protein used previously[72]. We observed comparable results with the independent source of Cas9 nuclease assayed with gRNAs of 18 and 20 nt complementarity (Figure S5.7).

**Effects of adjacent double and deletions variants on Cas9 in vivo and in vitro cleavage specificity**

The type II variant library was constructed to have all single base deletions and consecutive double variants throughout the EGFP-1, EGFP-2, rol-6, and unc-22A targets and the effects of these deletion variants were measured in vivo and in vitro. Adjacent double variants in vivo resulted in a general decrease of cleavage efficiency for all targets and all lengths of gRNAs (17, 18, and 20 nt complementarity) (Figure 4a-d). Although the double variants in vivo caused striking decreases in cleavage with full and truncated gRNAs, some deletion variants were tolerated or increased cleavage efficiency (Figure 6a-d). For example, for the full-length guide for EGFP-1 deletions toward middle and end of the target caused cleavage comparable to the wild-type sequences (deletions denoted by 1, 2, 4, and 5) (Figure 6a).
Figure 5.4A)

**EGFP-1 Target Sequences**

1. GGGCACGGGCAGCTTGCC
2. GGGCACGGCCAGCTTGCC
3. GGGCACGGGCAGCTACCCG
4. GGGCACGGGCAGGATGCCG
5. GGGCACGGCTCTTGCCCG
6. GGGCACGGCAGCTTGCCCG
7. GGGCACGGCAGCTTGCCCG
8. GGGCTGAGGGCAGCTTGCCCG

**Retention for double variants with the EGFP-1 17nt gRNA in vivo**

**Retention for double variants with the EGFP-1 18nt gRNA in vivo**

**Retention for double variants with the EGFP-1 20nt gRNA in vivo**

---

Figure 5.4B)

**EGFP-2 Target Sequences**

1. CCTTGAGCCGTCTCTCTGCTCACGCCAT
2. CCTTGAGCCGTCTCTCTGCAAGTCGCCAT
3. CCTTGAGCCGTCTCTCTGTCGTCGCCAT
4. CCTTGAGCCGTCTCTGAGCTTGTCGCCAT
5. CCTTGAGCCGTCTCTGACTGCTTGTCGCCAT
6. CCTTGAGCCGTCTCTGACTGCTTGTCGCCAT
7. CCTTGAGCCGTCTCTGACTGCTTGTCGCCAT
8. CCTTGAGGGTCTCTCTGCTTGTCGCCAT

**Retention for double variants with the EGFP-2 17nt gRNA in vivo**

**Retention for double variants with the EGFP-2 18nt gRNA in vivo**

**Retention for double variants with the EGFP-2 20nt gRNA in vivo**

---

124
**Figure 5.4C)**

*rol-6 Target sequences*

1. GTGAGACGTCAACAATATCCAGG  
2. GTGAGACGTCAACAATTAGGAGG  
3. GTGAGACGTCAACATTAGGAGG  
4. GTGAGACGTCAAGTATAGGAGG  
5. GTGAGACGTCTTCAATATGAGG  
6. GTGAGACGAGAACAATATGAGG  
7. GTGAGACTCAACAATATGAGG  
8. GTGACTCGTCAACAATATGAGG

**Figure 5.4D)**

*unc-22a Target sequences*

1. GTAGGCACACCCTCGCAGGCTATTTG  
2. GTAGGCACACCCTCGCAGGCTATTTG  
3. GTAGGCACACCCTCGCAGGCTATTTG  
4. GTAGGCACACCCTCGCAGGCTATTTG  
5. GTAGGCACACCCTCGCAGGCTATTTG  
6. GTAGGCACACCCTCGCAGGCTATTTG  
7. GTAGGCACACCCTCGCAGGCTATTTG  
8. GTAGGCACACCCTCGCAGGCTATTTG  
9. GTAGGCACACCCTCGCAGGCTATTTG  
10. GTAGGCACACCCTCGCAGGCTATTTG

---

125
**Figure 5.4** In vivo cleavage effects of adjacent double variants with a) EGFP-1, b) EGFP-2, c) rol-6, and d) unc-22A gRNAs. This graph depicts the in vivo retention scores of each double variant assayed in the type II libraries. The key to the target variants assayed is shown in the top panel. The bases in red are the variants in the given target of interest. The method of generating the retention scores is described in Figure 5.2.

**Figure 5.5** show results of cleavage of the same gRNAs tested in **Figure 5.4** in vitro for the 3-minute time point. The double variants at the early time points have similar cleavage profile to the in vivo results and are not tolerated well. However, with longer incubation time (180 minutes) double variants in the distal region of all targets are cleaved while most variants in the seed remain intolerant (Figure S5.8). **Figure 5.7** shows the 3-minute time point in vitro results that complement the in vivo results from **Figure 5.6**. For all for all lengths of tested gRNAs, most of the deletion variants are not tolerated at the early time points. Deletion variants in the distal region of the target are more tolerated. However, with longer incubation time (180 minutes) almost all deletion variants in the distal region of targets cleave and some positions in the seed region remain intolerant (Figure S5.9).

**Figure 5.5A**)

**EGFP-1 Target Sequences**

1. GGGCACGGGCAGCTTGGCC
2. GGGCACGGGCAGCTTGGGG
3. GGGCACGGGCAGCTACCCGG
4. GGGCACGGGCAGGGATGCGCCGG
5. GGGCACGGCTCTTGGCCGG
6. GGGCACGGAGCTTGGCCGG
7. GGGCACGCAGCTTGGCCGG
8. GGGCTGGGCAGCTTGGCCGG

**Retention for double variants with the EGFP-1 17 nt gRNA in vitro**

**Retention for double variants with the EGFP-1 18 nt gRNA in vitro**

**Retention for double variants with the EGFP-1 20 nt gRNA in vitro**
Figure 5.5B)

**EGFP-2 Target Sequences**
1. CCTTGATGCCTTCTTGCTTACGG
2. CCTTGATGCGTTCTCTGCAAGTCGG
3. CCTTGATGCCTTCTTGCTTACGG
4. CCTTGATGCTTTCTCTAGCTTGG
5. CCTTGATGCCTTCTCTGAGTCGG
6. CCTTGATGCCTTCTCTGCAAGTCGG
7. CCTTGATGCCTTCTCTGAGTCGG
8. CCTTGATGCCTTCTCTGCAAGTCGG

Retention for double variants with the EGFP-2 17 nt gRNA in vitro

Retention for double variants with the EGFP-2 20 nt gRNA in vitro

Figure 5.5C)

**rol-6 Target sequences**
1. GTGAGACGTCAACAAATATTCCAGG
2. GTGACGTCAACAAATATTCCAGG
3. GTGAGACGTCAACAAATATTCCAGG
4. GTGACGTCAACAAATATTCCAGG
5. GTGAGACGTCAACAAATATTCCAGG
6. GTGAGACGTCAACAAATATTCCAGG
7. GTGAGACGTCAACAAATATTCCAGG
8. GTGACGTCAACAAATATTCCAGG

Retention for double variants with the rol-6 18 nt gRNA in vitro

Retention for double variants with the rol-6 20 nt gRNA in vitro
Figure 5.5D)

**unc-22a target sequences**
1. GTAGCCACCACCCTCAGCGCCTTCTTG
2. GTAGCCACCACCCTCAGCGCCTTCTTG
3. GTAGCCACCACCCTCAGCGCCTTCTTG
4. GTAGCCACCACCCTCAGCGCCTTCTTG
5. GTAGCCACCACCCTCAGCGCCTTCTTG
6. GTAGCCACCACCCTCAGCGCCTTCTTG
7. GTAGCCACCACCCTCAGCGCCTTCTTG
8. GTAGCCACCACCCTCAGCGCCTTCTTG
9. GTAGCCACCACCCTCAGCGCCTTCTTG
10. GTAGCCACCACCCTCAGCGCCTTCTTG

**Retention for double variants with the unc-22a 17 nt gRNA in vitro**

**Retention for double variants with the unc-22a 18 nt gRNA in vitro**

**Retention for double variants with the unc-22a 20 nt gRNA in vitro**

**Figure 5.5 In vitro cleavage effects of adjacent double variants**
with a) EGFP-1, b) EGFP-2, c) rol-6, and d) unc-22A gRNAs. This graph depicts the *in vitro* retention scores of each double variant assayed in the type II libraries. The method of data generation is described in Figure 5.2.
Figure 5.6A)  

**EGFP-1 Target Sequences**

1. GCCAGG_CACGGGCACCTCGCTGG
2. GCCAGG_ACGGCGCGCTGCTGG
3. GCCAGG_CCAGGCGACCTGCTGG
4. GCCAGG_CAAGGACCTGCTGG
5. GCCAGG_CACGGCGGACCTGCTGG
6. GCCAGG_CCAGGCGGACCTGCTGG
7. GCCAGG_CCAGGCGGACCTGCTGG
8. GCCAGG_CCAGGCGGACCTGCTGG
9. GCCAGG_CCAGGCGGACCTGCTGG
10. GCCAGG_CCAGGCGGACCTGCTGG
11. GCCAGG_CCAGGCGGACCTGCTGG
12. GCCAGG_CCAGGCGGACCTGCTGG
13. GCCAGG_CCAGGCGGACCTGCTGG

Retention for deletion variants with the EGFP-1 17nt gRNA in vivo

Retention for deletion variants with the EGFP-1 20nt gRNA in vivo

---

Figure 5.6B)  

**EGFP-2 Target Sequences**

1. CCTT_ATCGCGTCTTCTTCTTCGG
2. CCTTG_CCCTCTTCTTCGG
3. CCTTG_CCCTCTTCTTCGG
4. CCTTG_CCCTCTTCTTCGG
5. CCTTG_CCCTCTTCTTCGG
6. CCTTG_CCCTCTTCTTCGG
7. CCTTG_CCCTCTTCTTCGG
8. CCTTG_CCCTCTTCTTCGG
9. CCTTG_CCCTCTTCTTCGG
10. CCTTG_CCCTCTTCTTCGG
11. CCTTG_CCCTCTTCTTCGG
12. CCTTG_CCCTCTTCTTCGG
13. CCTTG_CCCTCTTCTTCGG
14. CCTTG_CCCTCTTCTTCGG
15. CCTTG_CCCTCTTCTTCGG
16. CCTTG_CCCTCTTCTTCGG

Retention for deletion variants with the EGFP-2 17nt gRNA in vivo

Retention for deletion variants with the EGFP-2 20nt gRNA in vivo
Figure 5.6C) **rol-6 Target sequences**

1. CGT_TGAGACGTCAACAATATGAGG
2. CGTG_GAGAGCTCAACAAATGAGG
3. CGTG_TGACGTCAACAAATGAGG
4. CGTG_TGACGTCAACAAATGAGG
5. CGTG_TGACGTCAACAAATGAGG
6. CGTG_TGAGCGTCAACAAATGAGG
7. CGTG_TGAGCGTCAACAAATGAGG
8. CGTG_TGAGCGTCAACAAATGAGG
9. CGTG_TGAGCGTCAACAAATGAGG
10. CGTG_TGAGCGTCAACAAATGAGG
11. CGTG_TGAGCGTCAACAAATGAGG
12. CGTG_TGAGCGTCAACAAATGAGG
13. CGTG_TGAGCGTCAACAAATGAGG
14. CGTG_TGAGCGTCAACAAATGAGG
15. CGTG_TGAGCGTCAACAAATGAGG
16. CGTG_TGAGCGTCAACAAATGAGG
17. CGTG_TGAGCGTCAACAAATGAGG

Retention for deletion variants with the rol-6 17nt gRNA in vivo

Retention for deletion variants with the rol-6 20nt gRNA in vivo

Figure 5.6D) **unc-22a deletion target sequences**

1. GTAG_CACCACCGTCGCAGCGACTTGG
2. GTAGC_AACCAGTCCGCGCGACTTGG
3. GTAGG_CCACCGTCGCAGCGACTTGG
4. GTAGG_CCACCGTCGCAGCGACTTGG
5. GTAGG_CCACCGTCGCAGCGACTTGG
6. GTAGG_CCACCGTCGCAGCGACTTGG
7. GTAGG_CCACCGTCGCAGCGACTTGG
8. GTAGG_CCACCGTCGCAGCGACTTGG
9. GTAGG_CCACCGTCGCAGCGACTTGG
10. GTAGG_CCACCGTCGCAGCGACTTGG
11. GTAGG_CCACCGTCGCAGCGACTTGG
12. GTAGG_CCACCGTCGCAGCGACTTGG
13. GTAGG_CCACCGTCGCAGCGACTTGG
14. GTAGG_CCACCGTCGCAGCGACTTGG
15. GTAGG_CCACCGTCGCAGCGACTTGG

Retention for double variants with the unc−22a 18 nt gRNA in vivo

Retention for double variants with the unc−22a 20 nt gRNA in vivo

Figure 5.6 In vivo cleavage effects of deletion variants

With a) EGFP-1, b) EGFP-2, c) rol-6, and d) unc-22A gRNAs. This graph depicts the in vivo retention scores of each deletion variant assayed in the type II libraries. The key to the target variants assayed is shown in the top panel. The "_" indicate the deleted base in the given target of interest. The method of generating the retention scores is described in Figure 5.2.
Figure 5.7A) EGFP-1 Target Sequences
1. GCCAGG_CACGGGCGACTTGCCTCCGCTGG
2. GCCAGGG_AGCAGGCACTTGCCTCCGCTGG
3. GCCAGGGC_GGGGCACTTGCCTCCGCTGG
4. GCCAGGGCA_GGGGCACTTGCCTCCGCTGG
5. GCCAGGGCAGGG_GACCTGCTCCGCTGG
6. GCCAGGGCAGGG_AGCAGGCACTTGCCTCCGCTGG
7. GCCAGGGCAGGG_GCTTGCTCCGCTGG
8. GCCAGGGCAGGGCACTTGCCTCCGCTGG
9. GCCAGGGCAGGGCAGGGTCTTGCTCCGCTGG
10. GCCAGGGCAGGGCAGGGTCTTGCTCCGCTGG
11. GCCAGGGCAGGGCAGGGTCTTGCTCCGCTGG
12. GCCAGGGCAGGGCAGGGTCTTGCTCCGCTGG
13. GCCAGGGCAGGGCAGGGTCTTGCTCCGCTGG

Figure 5.7B) EGFP-2 Target Sequences
1. CCTT_ATGCCTCTCTGCTTGCTCGG
2. CCTT_ATGCCTCTCTGCTTGCTCGG
3. CCTT_ATGCCTCTCTGCTTGCTCGG
4. CCTT_ATGCCTCTCTGCTTGCTCGG
5. CCTT_ATGCCTCTCTGCTTGCTCGG
6. CCTT_ATGCCTCTCTGCTTGCTCGG
7. CCTT_ATGCCTCTCTGCTTGCTCGG
8. CCTT_ATGCCTCTCTGCTTGCTCGG
9. CCTT_ATGCCTCTCTGCTTGCTCGG
10. CCTT_ATGCCTCTCTGCTTGCTCGG
11. CCTT_ATGCCTCTCTGCTTGCTCGG
12. CCTT_ATGCCTCTCTGCTTGCTCGG
13. CCTT_ATGCCTCTCTGCTTGCTCGG
14. CCTT_ATGCCTCTCTGCTTGCTCGG
15. CCTT_ATGCCTCTCTGCTTGCTCGG
16. CCTT_ATGCCTCTCTGCTTGCTCGG
Figure 5.7C) rol-6 Target sequences

1. CGT_TGAGACATCGTACAAATATGG
2. GG_TGACATCGTACAAATATGG
3. CGTGTG_GAGCTCAACTATGG
4. CGTGTG_GAGCTCAACTATGG
5. CGTGTG_AGTGCAACTATGG
6. CGTGTG_AGTGCAACTATGG
7. CGTGTG_GAGCTCAACTATGG
8. CGTGTG_GAGCTCAACTATGG
9. CGTGTG_GAGCTCAACTATGG
10. CGTGTG_GAGCTCAACTATGG
11. CGTGTG_GAGCTCAACTATGG
12. CGTGTG_GAGCTCAACTATGG
13. CGTGTG_GAGCTCAACTATGG
14. CGTGTG_GAGCTCAACTATGG
15. CGTGTG_GAGCTCAACTATGG
16. CGTGTG_GAGCTCAACTATGG
17. CGTGTG_GAGCTCAACTATGG

Retention for deletion variants with the rol-6 18nt gRNA in vitro

Retention for deletion variants with the rol-6 20nt gRNA in vitro

Figure 7d) unc-22a deletion target sequences

1. GTAG_CACACCGTCCCGCGGATTGG
2. GTAGG_ACCACCGTCCCGCGGATTGG
3. GTAGGCC_CACACCGTCCCGCGGATTGG
4. GTAGGCC_CACACCGTCCCGCGGATTGG
5. GTAGGCC_CACACCGTCCCGCGGATTGG
6. GTAGGCC_CACACCGTCCCGCGGATTGG
7. GTAGGCC_CACACCGTCCCGCGGATTGG
8. GTAGGCC_CACACCGTCCCGCGGATTGG
9. GTAGGCC_CACACCGTCCCGCGGATTGG
10. GTAGGCC_CACACCGTCCCGCGGATTGG
11. GTAGGCC_CACACCGTCCCGCGGATTGG
12. GTAGGCC_CACACCGTCCCGCGGATTGG
13. GTAGGCC_CACACCGTCCCGCGGATTGG
14. GTAGGCC_CACACCGTCCCGCGGATTGG
15. GTAGGCC_CACACCGTCCCGCGGATTGG

Retention for deletion variants with the unc-22a 18nt gRNA in vitro

Retention for deletion variants with the unc-22a 20nt gRNA in vitro

Figure 5.7: In vitro cleavage effects of deletion variants
with a) EGFP-1, b) EGFP-2, c) rol-6, and d) unc-22A gRNAs. This graph depicts the in vitro retention scores of each deletion variant assayed in the type II libraries. The method of data generation is described in Figure 5.2.
Conclusions

Cas9 has become the dominant platform for genome manipulation, thus warranting extensive research to explore the gRNA::DNA interaction of Cas9 cleavage specificity. This study provides a detailed map of Cas9 specificity and cleavage for gRNAs of various lengths in the context of *in vitro* and *in vivo* conditions. In addition, two of the four targets characterized in this study were also characterized in mammalian cells[217]. The cleavage efficiency profiles of the targets in the mammalian and yeast system show notable similarities. For example, the EGFP-1 site with the 18 nt truncated gRNA did not show a dramatic increase in specificity in the mammalian system[217] and that holds true our findings. With only four target libraries and gRNAs, it is difficult to determine a general set of rules for off target cleavage.

However it is worth noting that the EGFP-1 target is the most GC rich target and one of the two targets with a 3’ GG motif adjacent to the PAM in positions 19 and 20. It has been shown previously in *C. elegans* that gRNAs with a GG motif at the 3’ end of their target-specific sequences have a higher activity level[219]. It is possible that this, combined with a higher GC content resulted in the inability of the 18 nt guide to reduce off target cutting. When we targeted a gRNA to this guide in the reverse orientation (GFP1-R) or shifted the gRNA to use the 3’ GG as the PAM (GFP1-OffSet), the 18nt gRNA was able to improve specificity on the EGFP-1 target as with the other 3 targets (Figure S5.11). Further studies would be required for clearer understanding of Cas9 cleavage specificity.

Another interesting finding of this work is that truncated gRNAs produced a substantial improvement in specificity over full-length gRNAs in *S cerevisiae*. In Chapter 3 we found no such improvement using truncated (18nt) gRNAs with CRISPRi. CRISPRi only requires that dCas9-Mxi1 bind target loic in order to repress. This suggests that the improvement in specificity seen with truncated gRNAs is due to a reduction in the ability of Cas9 to cleave mismatched target sequences and not a reduction in its ability to bind to them.
The applications of Cas9 both as a purified protein and expressed in vivo are continuing to grow. Our results emphasize differences in Cas9 specificity in different contexts, and highlight the caution that must be taken in extrapolating findings regarding Cas9 between in vitro and in vivo conditions. The lack of specificity of Cas9 in vitro may limit many of its applications, and warrants further research aimed at improving its specificity in these conditions. A logical next step may be to test the two recently published engineered high fidelity Cas9s[222, 223] in vitro to see if they are capable of reducing in vitro mismatch tolerance. There is no guarantee these results will replicate to in vitro conditions. We were able to replicate the finding that truncated gRNAs can improve Cas9 cleavage specificity[155] in S cerevisiae, but they did yield a notable improvement in vitro cleavage assays. Our analysis of Cas9 in vitro and in vivo will be of value for understanding Cas9 cleavage specificity.

Methods
Plasmid and Strain construction

Molecular cloning was done with Gibson Assembly as outlined by Gibson et al. 2009. E. coli minipreps were performed with QIAprep Spin Minipreps (Qiagen). Preparation of competent E. coli DH5α and transformation used Zymo Mix & Go E. coli Transformation reagents and Zymo Broth or electrocompetent CloneCatcher DH5G cells purchased from Genlantis. Competent S. cerevisiae (strain BY4741 and KU70 deletion strain from MATa collection) were prepared either by standard lithium acetate transformation protocols[224] or using Zymo Frozen-EZ Yeast Transformation II. Hifi Hotstart (Kapa Biosystems), Q5 High Fidelity polymerase (NEB), and Phusion Hot Start Flex (Thermo Scientific) were used for PCRs. Primers and gRNA oligonucleotides were ordered from IDT. DpnI treatment was used to remove template plasmids in PCRs that were followed by Gibson Assembly. Benchling.com and APE DNA editing software were used for plasmid design.

Cas9-gRNA plasmids were built in the yeast pRS416 Cen/ARS plasmid containing the Ura3 marker. First we cloned an engineered Tet inducible pRPR1 Pol III promoter[157, 178], NotI site, and gRNA sequence as well as the Tet repressor gene under control of the GPM1 promoter and terminator into pRS416 at the PciI site.
adjacent to the ori using Gibson Assembly\textsuperscript{14}. This vector is referred to as pRS416gT.
Next we digested pRS416gT at the multiple cloning site with KpnI and SacI. GalL-Cas9-Cyc1t was amplified from pRS415-GalL-Cas9-Cyc1t (Addgene 43804)\textsuperscript{19} using M13 forward and reverse primers. This PCR product was Gibson assembled into cut pRS416gT and transformed into DH5α. The gRNA single stranded oligonucleotides were then cloned into the NotI cleaved site with Gibson Assembly.

For the initial experiments (type I libraries), libraries \textit{unc}-22A RVL-1 and RVL-2 from\textsuperscript{72} were amplified using Kapa HiFi Hotstart, followed by DpnI treatment to remove the template plasmid. This PCR product was then Gibson Assembled into SacII cut pRS416gT (with a specific guide sequence cloned into the gRNA locus as above) adjacent to the GPM1 terminator and transformed into DH5α. The plasmid design for the 20 nt complementarity \textit{unc}-22A plasmid is available at https://benchling.com/s/O5VobNjd or as Figure S5.1. All oligonucleotide sequences used in this work are available in Figure S5.1.

For the type II libraries, EGFP-1, EGFP-2, \textit{rol}-6, and new \textit{unc}-22A libraries, oligos were synthesized by Custom Array. These target libraries were designed to contain all possible single nucleotide changes, all possible unique single base deletions, and a set of double nucleotide changes tiling the entire target. Oligos were amplified from the Custom Array pool using 182-Fu-Library\textsubscript{fwd} and 183-Fu-Library\textsubscript{rev} which also created overlaps for Gibson Assembly. This PCR product was cloned into pRS415 cut with XhoI and SacI by Gibson Assembly. The product of assembly was then ethanol precipitated to purify and concentrate the Gibson product, and then was electroporated into CloneCatcher DH5G cells using a capacitance of 25\textmu F, resistance of 400 ohms, and voltage of 2.0 kV using an ice-cold 0.1 cm cuvette. The transformation was recovered for 90 minutes and then transferred to liquid LB carbenicillin as well as a dilution on LB carbenicillin plates to get colony counts. The resulting cells were spun down and minipreped. To create strains for each of the gRNAs, we first transformed in pRS416gT plasmids containing each of the 18 different gRNAs into BY4741. We then cloned the new pRS415 library into each of
these strains and grew in liquid synthetic complete media lacking uracil and leucine to select for both plasmids.

**Cas9 in vitro cleavage specificity assay**

Cas9 in vitro cleavage assays and gRNA transcription (17, 18, and 20 nt complementarity) were performed as described in Fu et al. 2014[72].

**Cas9 in vivo cleavage specificity assays**

For the initial type I library experiments, starting cultures were grown overnight at 30°C in synthetic complete media –Ura. These cultures were then used to inoculate experimental cultures to OD 0.1 in YP Galactose media with 250ng/ml anhydrotetracycline (ATC). To control for possible effects of non-homologous end joining (NHEJ) events, we did our initial experiments in both wild BY4741 strain and a KU70 null NHEJ deficient strain. 17, 18, 20 and 20+G nt complementarity gRNA strains were grown for 12 hours in inducing conditions. No detectible difference in the cleavage pattern was observed (Figure S5.14). Wild type BY4741 was used for all further experiments. Approximately 3000-4500 unique species were assayed in vivo. Further experiments with libraries *unc*-22A RVL-1 and RVL-2 were done with 18 and 20 nt complementarity versions of the gRNAs in BY4741. For the type II libraries (EGFP-1, EGFP-2, rol-6, and new *unc*-22A libraries), experiments were inoculated as above with 17, 18, and 20nt complementarity versions of the gRNAs, but starting cultures were grown overnight at 30°C in synthetic complete media –Ura –Leu +dextrose, and growth experiments were conducted in synthetic complete media –Ura –Leu +galactose with 250ng/ml anhydrotetracycline (ATC). For EGFP-1 target site, there were three different orientations of guide RNAs tested (EGFP-1f, EGFP-1r, EGFP-1os).

Yeast culturing and sample collection was performed using a cell-screening platform that integrates temperature-controlled absorbance plate readers, plate coolers, and a liquid handling robot. Briefly, 700 ul yeast cultures were grown in 48 well plates at 30°C with orbital shaking in Infinite plate readers (Tecan). To maintain cultures in
log phase over 10 doublings, 80 uls of the culture was removed when it reached an OD of 0.76, added to a well containing 620 ul of media, and then allowed to grow further. After two such dilutions, 600 uls of the culture was collected and saved to a 4°C cooling station (Torrey Pines) when it reached an OD of 0.76. This amounted to approximately 10 culture doublings from the beginning of the experiment. Pipetting events were triggered automatically by Pegasus Software and performed by a Freedom EVO workstation (Tecan). After sample collection, yeast plasmids were purified using the Zymoprep Yeast Plasmid Miniprep II kit (Zymo Research).

**Formation of gRNA 5' ends**

We note that all synthetic gRNAs were produced with T7 polymerase and therefore have an additional G on the 5' end. In vivo production of gRNAs utilized an engineered Tet inducible RPR1 PolIII promoter. The RPR1 promoter transcribes an 84 nt 5' leader that is cleaved in vivo in its native context (RPR1 gene) to give the mature RPR1 transcript[225]. We and others have used this promoter[226] and other Pol III promoters (SNR52) that produce leader sequences[227, 228] as a standard tool to produce effective gRNAs in vivo in yeast; nonetheless, we don't know if this 5' leader is present or has been removed at the time that the Cas9::gRNA complexes act in vivo.

**Cas9 in vitro and in vivo sequence retention calculation**

Cas9 in vitro and in vivo substrate cleavage efficiency was calculated as described in Fu et al. 2014[72]. A log retention score for each sequence in each experiment was calculated by quantifying the representation of each sequence before and after either the addition of Cas9 (in vitro) or induction of Cas9 (in vivo). A more negative retention score indicates more cleavage while a more positive retention score indicates less cleavage. Only sequences with n >= 50 counts in the non-cleaved control were considered for all experiments. A list of all Cas9 in vitro and in vivo experiments, experimental conditions, and sequencing run IDs are reported in Table S5.2.

**GEO Accession Number:** Pending
**Funding:**
R01GM37706 (A.Z.F); T32GM00779 (B.X.H.F); NSF Graduate Fellowship (B.X.H.F), P01HG000205 (R.P.S, J.D.S.)

**Acknowledgements:** We thank Michael Nonet, Ronald W Davis, Karen Artilles, and Gavin Sherlock for their help and suggestions.
Appendices

Chapter 2

Figure S2.1. Reproducibility of biological replicates.
A. Reads per gene are shown for RNA-seq replicates of the Sc/Sp hybrid grown in YPD + DMSO. B. Reads per gene are shown for RNA-seq replicates of the Sc/Sp hybrid grown in YPD + 600 ppm citrinin (dissolved in DMSO).
Figure S2.2. Agreement between RNA-seq and microarray data.

A. ASE ratios for the Sc/Sp hybrid in microarray data in YPD [126] compared to our RNA-seq data in YPD+DMSO. B. Fold-changes for Sc’s response to 300 ppm citrinin from oligonucleotide microarray data [125], compared to the Sc allele’s response (within the Sc/Sp hybrid) to 600 ppm citrinin in our RNA-seq data.

Table S2.1. Read counts from RNA-seq in the Sc/Sp hybrid.
(Supplemental Data File for Chapter 2)

Table S2.2. All primers/oligos used in this work.
(Supplemental Data File for Chapter 2)
Chapter 3
Additional files and supplemental figures are included in the supplemental file for Chapter 3.

Additional file 3.1. Raw growth data measurements for 3.1C.

Additional file 3.2. Raw data for qPCR

Additional file 3.3. The small molecule inhibitor/gene combinations constituting the reference chemical-genetic interactions used for this study.

Additional file 3.4. Counts and analysis for 8 replicates in no drug (DMSO 1%)

Additional file 3.5. Counts and analysis of three replicates in 20µM fluconazole.

Additional file 3.6. Lists the gRNAs that were synthesized, their target location, and aimed strand (template or non-template)

Additional file 3.7. List of sequenced samples, their respective barcodes, MiSeq run information, and other metadata.

Additional file 3.8. Sequence read information and raw counts from MiSeq data.

Additional file 3.9. ATc fitness effects (A), drug effects (D), and no-drug ATc effects (A0) for each gRNA in every tested condition.

Additional file 3.10. Drug-specific effects (D) for all drugs tested for each gRNA.

Additional file 3.11. Average drug-specific effects for each guide set (gRNAs directed against the same gene).

Additional file 3.12. Raw growth data for Figure 3.3.

Additional file 3.13. GCMS measurements of ergosterol intermediates in yeast treated with 1181-0519

Additional file 3.14. Nucleosome occupancy scores of gRNA regions based on tiling array analysis

Additional file 3.15. A range of additional potential determinants of guide effect.

Additional file 3.16. Number and significance of transcription factor site overlaps with positive control gRNA target sequences.

Additional file 3.17. Primers, Strains, and Plasmids used in this study.

Additional file 3.18. Analysis of potential off-target binding sites.

Supplemental Figure Legends

Figure S3.1. Characterization of single-plasmid system for CRISPRi in yeast.
A) Transformants expressing gRNAs directed against CRG1, ERG11, ERG25, and SEC14 (see legend), were grown in the presence of a specific small molecule \( \text{i.e.} \) cantharidin (10µM), fluconazole (25µM), 1181-0519(20µM), and 4130-1276(2.5µM), respectively, and in increasing concentrations of ATc (x-
axis). Growth relative to the ‘no-ATc’ control is indicated on the y-axis (see Methods). B - E) SYBR qPCR results examining the effects various gRNAs on gene expression. Each plot shows expression change (log2 fold change) relative to a control condition. All plots use ACT1 as a reference gene, with the exception of B which uses UBC6. Error bars represent the standard deviation. Raw data are available in Additional file 3.2. B) Strains expressing gRNAs targeting CRG1, ERG11, or ERG25 (plus empty vector control) were cultured in the presence or absence of ATc (see legend). Expression relative to the empty vector control in the absence of ATc (EV-Mxi1-ATc), is measured for each gene and the gRNA (x-axis). C) RBD2 expression changes following induction of the P20 strain (RBD2 gRNA) with ATc. Log2 fold change (relative to the time zero control) is plotted on y-axis, for different time points following ATc addition on the x-axis. D) Similar to C, only for P12 (ERG25 gRNA). Both ERG11 and ERG25 expression is plotted. E) CRG1 expression changes following removal of ATc from an induced culture. The P141 strain (CRG1 gRNA) was grown overnight in the presence or absence of ATc. Cells from induced culture (+ATc) were centrifuged and washed several times with water, and then resuspended in media without ATc. Samples were collected at various times following removal of ATc (x-axis). Expression relative to that of the overnight culture grown without ATc is plotted on the y-axis.

Figure S3.2. Schematic of the workflow for parallel analysis of CRIPSRi-induced fitness defects in pooled cultures.

Oligos were synthesized by oligo array synthesis by Custom Array. These were PCR-amplified and Gibson Assembled into NotI digested pRS416gT-Mxi1 and transformed into E. coli. The plasmids were purified and transformed into yeast. Once the yeast pool was built, experiments were conducted in which pools were grown in inducing (+ATc) and non-inducing conditions (-ATc), in the presence (or absence) of different small molecules. After multiple generations of growth in these conditions, yeast plasmids were extracted and the gRNA region complementary to its target was PCR-amplified and sequenced on a MiSeq. Counts of each gRNA were compared between different conditions. The table summarizes the different growth conditions compared in this study, and the information that each comparison can provide.

Figure S3.3. Replicate Plots

A) Plots comparing the gRNA sequence counts (log2-transformed) from 8 +ATc replicates and 8 – ATc replicates. B) Plots comparing the log2-transformed gRNA sequence counts for 3 +ATc replicates and 3 –ATc replicates in fluconazole.

Figure S3.4. Parallel analysis of CRIPSRi-induced fitness defects in pooled cultures.

Drug/gene pairs representing reference chemical-genetic interactions are shown in grey. (A) Effect of gRNA expression on growth in the presence of drug (indicated above each plot) as in Figure 3.2A. Median-adjusted ATc effect (A) is plotted on the y-axis. Each point represents a unique gRNA directed against one of 20 different target genes. gRNAs are color-coded and arranged alphabetically on the x-axis by target gene. B) Similar to (A), only the y-axis quantifies drug-specific effects as in Figure 3.2C. Large black dots represent the mean for each gene, and are colored red if >1 or if <-1. C) Drug-specific growth defects for each gRNA set (group of guides directed against the same gene; indicated above each plot) in 25 different drug conditions (arranged on the x-axis), as in Figure 3.2D. Points are color-coded by condition. Large black dots represent the mean in each drug condition, and are colored red if >1 or if <-1.

Figure S3.5. Effect of 1181-0519 on ergosterol metabolite levels in yeast.

A simplified schematic of the ergosterol synthesis pathway is illustrated. The abundance of six metabolites from yeast treated with 1181-0519 (blue bars), or DMSO control (black bars) is shown in the bar plots. Error bars represent SEM. Consistent with chemical inhibition of Erg25 by 1181-0519, the abundance of immediate upstream metabolite of Erg25, 4,4-dimethyl-5a-cholesta-8,24-dien-3b-ol (aka 4,4-dimethylzymosterol) increased by 6.2 fold in the presence of 1181-0519 with a p-value of 2.9x10^{-4}, while the abundance of downstream metabolite zymosterol decreased by 6.0 fold with a p-value of 1.2x10^{-5}. The abundance of ergosterol decreased by 1.3 fold with a p-value of 4.1x10^{-3}.

Figure S3.6. Quantitative comparison of full-length and truncated gRNAs.
Heatmaps illustrating the ATc effects measured for gRNAs containing different mismatches to the target sequence. Full-length (20 nt of target complementarity) and truncated (18 nt of target complementarity) are arranged by target gene on the y-axis. The mismatch position of each gRNA relative to the PAM is indicated on the x-axis (gRNAs matching the target sequence perfectly are on the far left). Missing values are indicated with an X. Each panel represents results for a different drug (indicated above).

Figure S3.7. The effect of nucleosome occupancy and chromatin accessibility on gRNA efficacy. A) As in Figure 3.5C, guide effect relative to nucleosome occupancy and chromatin accessibility, for gRNAs targeting 0 to 400 bp downstream of the TSS. The median of gRNA effects in windows of 0.25, overlapping by 0.125, is indicated by the circular blue markers. The blue bars show the first and third quartiles. The Spearman correlation for the relationship with nucleosome density is 0.06, p-value = .44. The Spearman correlation for the relationship with normalized ATAC-Seq is -0.31, p-value = 9.3x10^{-6}. B) Nucleosome occupancy based on previously published measurements[168] (see Methods) is plotted on x-axis versus guide effect (broad tiling library) on the y-axis. Negative guide effect values indicate stronger gRNA-induced growth defects. Each panel represents a different locus (indicated above each plot). Guide effect values were calculated based on growth in the drug specific to each locus. R and p-values are shown in the bottom left of each panel.

Figure S3.8. Effect of sequence context on guide efficacy. Relative gRNA effect, the ATc effect of a guide relative to the maximum ATc effect in that guide set, is plotted on the y-axis. Average (solid line, y-axis) and median (individual dots) relative gRNA effect for guides targeting sequences with A (blue), C (green), G (red), or T (cyan) at bases 20bp upstream of the PAM site to 20bp downstream of the end of the targeting region (x-axis).

Figure S3.9. The effects of guide RNA secondary structure and DNA/RNA duplex formation strength (x-axis) on guide efficacy (y-axis). Each control guide RNA is one dot, large dots with connecting solid line denote median effect in sliding windows overlapping by 50%. A) TM calculated by Oligo TM B) ΔG of gRNA/DNA sequence’s duplex predicted by RNA fold. C) TM of seed sequence (8 bases of guide nearest to PAM). D) ΔG of seed sequence gRNA/DNA duplex predicted by RNA fold. E) ΔG of the predicted secondary structure for the gRNA (including leader sequence “gtccctacgttagagatgggacacatcgctctggtctggtgcacgccgaacagaattgactgctcgcggttgcag” and the structural part of the gRNA).

Chapter 4
The Chapter 4 supplemental tables for the CRISPRi collection are available in Chapter 4 supplemental file.

Table S4.1 Custom Array synthesized oligos used to create CRISPRi collection

Table S4.2 Sequences of all gRNAs contained in the CRISPRi collection.

Table S4.3 CRISPRi YPD growth assay and annotation data

Table 4.4 CRISPRi YPEG collection data

Table 4.5 Primers, strains and plasmids.
### Table S5.1 Primers used to build vectors in this study

**Primers used to build Cas9 gRNA expression vector pRS416gT-GaL-Cas9**

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEF1_fwd</td>
<td>tcacacaggaacacagctatgaccatg</td>
</tr>
<tr>
<td>CYC1_rev</td>
<td>tggaaaaagcaggeccagtgac</td>
</tr>
<tr>
<td>57-TetR-Fwd</td>
<td>tccaaacacacacatattacaataataataaatttcttagtagataaaagtaatgtg</td>
</tr>
<tr>
<td>58-TetR-Rev</td>
<td>aaaagaatacatcaatacatctatttcgacttaagccacacttcacatattaagttg</td>
</tr>
<tr>
<td>49-pRPR1_fwd</td>
<td>ctggcccttttgctgcttttgctcggggagtctgccaattgaac</td>
</tr>
<tr>
<td>50-RPR1_term_rev</td>
<td>ctttaagctatacattgcagactaCCTGCAGGaagaacacacaggtcacaag</td>
</tr>
<tr>
<td>51-pRPR1-NotI-sgRNA-F</td>
<td>tgggagctgctccgacgacCGGCCGCgttttagagctagaaatagctaatgg</td>
</tr>
<tr>
<td>52-sgRNA-NotI-pRPR1-R</td>
<td>getatattctagetctaaaaacGCGGCCGgetgcatacgcaggcgetccacag</td>
</tr>
<tr>
<td>53-pGPM1_prom_fwd</td>
<td>gaactctttgtttctttCCTGCAGGtagctgtcaatgattgtatgtaaagattg</td>
</tr>
<tr>
<td>54-pGPM1_prom_rev</td>
<td>aaatcattctacttcgacCCCGGGgtattgtaatatgtgtttggttatttaag</td>
</tr>
<tr>
<td>55-tGPM1-term-fwd</td>
<td>cacacatatataataCCCGGGgtctgtaagaatgattgattagatttc</td>
</tr>
<tr>
<td>56-tGPM1_term_rev</td>
<td>aacacggggataacggaagaacCGCCGTTtcaactcctactccctctttg</td>
</tr>
</tbody>
</table>

**Primers used to amplify target library and add overlaps for Gibson Assembly**

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>111-pRS-Becky-Lib-Fwd</td>
<td>gaaagctgtaatgcggcgactggaacC CGCCGAGCAGTAATTCTAGAGGTACC</td>
</tr>
<tr>
<td>112-pRS-Becky-Lib-Rev</td>
<td>ACAGAATCAGGGGATAACGCAGGAAGACCCCCGCACGCTTCCCGATCT</td>
</tr>
<tr>
<td>182-Fu-Library_fwd</td>
<td>TATAGGGCGAAATTGGGTACCGGCCCCCCCTCGAGACACGACTCTTCCCGATCT</td>
</tr>
<tr>
<td>183-Fu-Library_rev</td>
<td>TTAACCCTCACTAAGGGGAACAAAGCCTGGAGCTCGCAGTAATTCTAGAGGTACC</td>
</tr>
</tbody>
</table>

**Primers used to verify KU70 deletion from deletion collection strain.**

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>114-Ku70-Fwd</td>
<td>GACTCGCCCCCGGGAGATATT</td>
</tr>
<tr>
<td>115-Ku70-Rev</td>
<td>ACAACGGTTTAGCAGACAAAG</td>
</tr>
<tr>
<td>122-Ku70-Out-REV</td>
<td>GGGCAAAACACTTGGCGTGGTT</td>
</tr>
<tr>
<td>123-Ku70-Out-Fwd</td>
<td>TGACTCTCGGTAGCACAAGTGGT</td>
</tr>
</tbody>
</table>

**Oligos used to clone Unc22 guide RNAs into vector**

<table>
<thead>
<tr>
<th>Oligo Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>P89-unc-22-20</td>
<td>ctgggagctgctcggcagGCACCACCGTCGCCGCGATTTgttttagctagaatagc</td>
</tr>
</tbody>
</table>
P90-unc-22-18  ctgggagctggtcggcagACCACCGTCGCCGCATTgttttagctagaaatagc
P91-unc-22-17  ctgggagctggtcggcagCCACCGTGCCGCCGGCATTgttttagctagaaatagc
P92-unc-22-20+G ctgggagctggtcggcagGGCACCACCGTCCGCCGGCATTgttttagctagaaatagc
P111-GFP1-F ctgggagctggtcggcagGGGCACCGCAGCTTTGCCGGGtttttagctagaaatagc
P112-GFP1-F18 ctgggagctggtcggcagGCACCGCAGCTTTGCCGGGtttttagctagaaatagc
P113-GFP1-F17 ctgggagctggtcggcagCACCGCAGCTTTGCCGGGtttttagctagaaatagc
P114-GFP1-R ctgggagctggtcggcagCCCGCAAGCTGCCCGTGCCCgttttagctagaaatagc
P115-GFP1-R18 ctgggagctggtcggcagGGCAAGCTGCCCGTGCCCgttttagctagaaatagc
P116-GFP1-R17 ctgggagctggtcggcagGCAAGCTGCCCGTGCCCgttttagctagaaatagc
P117-GFP2 ctgggagctggtcggcagGATGCCGTTCTTCTGCTTGTgttttagctagaaatagc
P118-GFP2-18 ctgggagctggtcggcagTGCCGTTCTTCTGCTTGTgttttagctagaaatagc
P119-GFP2-17 ctgggagctggtcggcagGCCGTTCTTCTGCTTGTgttttagctagaaatagc
P120-Rol6 ctgggagctggtcggcagGTGAGACGTCAACAATAGGgttttagctagaaatagc
P121-Rol6-18 ctgggagctggtcggcagGAGACGTCAACAATAGGgttttagctagaaatagc
P122-Rol6-17 ctgggagctggtcggcagAGACGTCAACAATAGGgttttagctagaaatagc
P123-GFP1-OS ctgggagctggtcggcagCCAGGGCACGGCAGCTTGCGtttttagctagaaatagc
P124-GFP1-OS18 ctgggagctggtcggcagAGGGCACGGCAGCTTGCGtttttagctagaaatagc
P125-GFP1-OS17 ctgggagctggtcggcagGGGCACGGCAGCTTGCGtttttagctagaaatagc
Figure S5.1 Yeast random variant library cloning vector map.

Plasmid genbank file can be found in Supplemental files.
Figure S5.2 Cleavage results for in vivo variant libraries for type I with 18 and 20 nt complementarity gRNAs.
(Retention data taken from experiments under the Illumina run ID: AF_SOL_588)

a. Cas9 cleavage results for the 20 nt complementarity gRNA with random variant library-2 at 10 generations. A median of effects for single base transversion variants is indicated for each position by a dot. This graph shows the transversion variant retention for the target region and constrained PAM nucleotides (positions 1-20 and 22-23). Only sequences with >=50 reads in the control (uncut) library were considered in calculating this median. Protospacer 4 sequences and sequences with 4-7 mismatches from the unc-22A target are negative controls (labeled PS4 and 4-7 mm) and wild type unc-22A targets are positive controls (labeled WT). Controls are shown at the left. The retention scores shown are from experiments done on yeast unc-22A random variant library-2. A negative retention score indicates sequence cleavage, while a retention score of zero corresponds to samples whose cleavage is comparable to the pool of highly mutated sites (with 4-7 mutations). A slightly positive retention score indicates a lack of Cas9 cleavage. Over-representation is due to noise or consistent PCR effects.

b. Results for the 18 nt complementarity gRNA with yeast random variant library-2 at 10 generations. Methods of analysis are as in Figure 5.2a.
b) Graphs depict single base in vivo variant effects on unc-22A guide. This figure uses the same data represented in Figure S5.2 but shows all variant effects. Analysis was as in Fu et al. (2014), with the analytic pipeline reproduced below for the convenience of the reader. Each variant is indicated by a dot noting differences from the canonical sequence. Retention at various time points following Cas9 addition is depicted as a dot colored according to time point. For the target region and constrained PAM nucleotides (positions 1–20 and 22–23), each dot represents median retention amongst matching library sequences carrying the variant of interest (either A,C,G or T) and distinct flanking sequences. For the flanking sequence and the N in the PAM, each dot represents median retention for library species with fully matched target/PAM regions and with the indicated flanking position constrained to the indicated base. Only species with \( \geq 50 \) reads in the control (uncut) library were considered in calculating this median. Two negative (labeled PS4) and one positive (labeled WT) controls are shown at the left. Sequences with 4–7 mismatches in the target region, and sequences from a ps4 ‘spike in’ behave in aggregate as similar ‘uncut’ pools, while a median from all sequences that match the guide and PAM sequences perfectly provides a positive control. Negative retention scores indicate increasing cleavage, while a retention score of zero corresponds to samples whose cleavage is comparable to the pool of highly mutated sites (with 4–7 mutations) and to trace amounts of unrelated (protospacer 4) DNA. Some samples showed a slightly positive retention score, indicating a lack of cleavage with either noise or consistent PCR effects leading to a slight over-representation in the post-cleavage sample.

(AF_SOL_588)
d) Single base *in vivo* effects of the 18 nt complementarity guide
e) Single base *in vivo* transversion effects with the 17 nt complementarity guide RNA in yeast *unc-22A* random variant library-2. Method of analyses identical to Figure S5.2.

Transversion effects with the 17 nt complementarity gRNA in the KU70 deletion background. (AF_SOL_572)

Transversion effects with the 17 nt complementarity gRNA in the BY4741 background. (AF_SOL_574)
Figure S5.3 Plots of Transition variant effects. In vivo cleavage single base transition effects of variants with a) EGFP-1, b) EGFP-2, c) rol-6, and d) unc-22A gRNAs. This graph depicts the in vivo retention scores of each double variant assayed in the type II libraries. The method of data generation is described in Figure 2 but for transition effects.

a) Results for EGFP-1. (AF_SOL_636)
b) Results for EGFP-2. (AF_SOL_673, AF_SOL_672)
e) Results for *rol-6*. (AF_SOL_673)
d) Results for *unc-224*. (AF_SOL_676)
Single base transition effects on unc-22A target

- Time = 0 generations
- Time = 10 generations, Guide=unc-22A L+CAG+18
- Time = 10 generations, Guide=unc-22A L+CAG+20
Figure S5.4 Cleavage results for *in vitro* unc-22A random variant libraries with 17, 18, and 20 nt complementarity gRNAs (type I library).

a. Time course Cas9 cleavage for 20 nt complementarity gRNA with yeast unc-22A random variant library. Methods of analysis are as in Figure 5.2. The dots are color coded according to the time points. (Retention data taken from experiments under the Illumina run ID: AF_SOL_594)
b. Time course Cas9 cleavage for 18 nt complementarity gRNA with yeast *unc-22A* random variant library-2. Methods of analysis are as in 5.2. (Retention data taken from experiments under the Illumina run ID: AF_SOL_596)
c.  Time course Cas9 cleavage for 17 nt complementarity gRNA with yeast *unc-22A* random variant library-2. Methods of analysis are as in Figure 5.2. (Retention data taken from experiments under the Illumina run ID: AF_SOL_594)
Figure S5.5 *In vitro* cleavage profiles for 180 minutes incubation for target library type II with full-length and truncated gRNAs (EGFP-1, EGFP-2, *rol-6*, and *unc-22A*).

a) Results for EGFP-1 gRNAs.
b) Results for EGFP-2 gRNAs.
c) Results for rol-6 gRNAs.

**Single base transversion effects on rol-6 target in vitro**

- **Time = 0 min**
- **Time = 180 min, Guide=18nt**
- **Time = 180 min, Guide=G+18nt**
- **Time = 180 min, Guide=G+20nt**
d) Results for *unc-22A* gRNAs.

**Single base transversion effects on unc-22A target**

- **Black**: Time = 0 min, Guide=17nt
- **Light Blue**: Time = 180 min, Guide=17nt
- **Purple**: Time = 180 min, Guide=18nt
- **Red**: Time = 180 min, Guide=20nt
Figure S5.6 Single base *in vitro* transversion effects of the 20 nt complementarity gRNA with NEB Cas9 with the *unc-22A* random variant library-2 at 30 degrees Celsius (AF_SOL_591).

a) Retention results for 20 nt complementarity gRNA.
Figure S5.7) Fu et al. 2014 Cas9 (ref.[72]) in vitro cleavage experiments with type I unc-22A random variant library-2 (AF_SOL_585).

a) Single base variants effects of 20 nt complementarity gRNA.
Figure S5.8 *In vitro* cleavage effects of adjacent double variants with a) EGFP-1, b) EGFP-2, c) rol-6, and d) *unc-22A* gRNAs (180 minute time point). This graph depicts the *in vitro* retention scores of each double variant assayed in the type II libraries. The method of data generation is described in Figure 5.2.

a) Results for EGFP-1.

---

**EGFP-1 Target Sequences**

1. GGGCACGGGACGCTTGGCCCGC
2. GGGCACGGGACGCTTGGCCGG
3. GGGCACGGGACGCTACCAGG
4. GGGCACGGGACGCTAGACCAGG
5. GGGCACGGGACGCTAGGCGGG
6. GGGCACGGGACGCTAGCCTCCGG
7. GGGCACGGGACGCTAGGCGGG
8. GGGCACGGGACGCTTGGCCGG

---

**Retention for double variants with the EGFP-1 17 nt gRNA in vitro**

---

**Retention for double variants with the EGFP-1 18 nt gRNA in vitro**

---

**Retention for double variants with the EGFP-1 20 nt gRNA in vitro**
b) Results for EGFP-2.

EGFP-2 Target Sequences
1. CCTGATGCCGTTTCTGCTTCAACGGCCAT
2. CCTGATGCCGTTTCTGCAAGTCGGCCAT
3. CCTGATGCCGTTTCTCTTTTTCGGCCAT
4. CCTGATGCCGTTTCTGCTTTCGGCCAT
5. CCTGATGCCGTTTCTCTTTTTCGGCCAT
6. CCTGATGCCGTTTCTCTTTTTCGGCCAT
7. CCTGATGCCGTTTCTCTTTTTCGGCCAT
8. CCTGATGCCGTTTCTCTTTTTCGGCCAT
c) Results for rol-6.

d) Results for unc-22A.
**Figure S5.9 In vitro cleavage effects of deletion variants**

with a) EGFP-1, b) EGFP-2, c) rol-6, and d) *unc-22A* gRNAs (180 minute time point). This graph depicts the *in vitro* retention scores of each deletion variant assayed in the type II libraries. The method of data generation is described in Figure 5.2.

a) Results for EGFP-1.

EGFP-1 Target Sequences

1. GCCAGG_CACGCGCGACGCTTGCCCGCTGGTGCA
2. GCCAGGACACGCGACGCTTGCCCGCTGGTGCA
3. GCCAGGACCGGCACGCTTGCCCGCTGGTGCA
4. GCCAGGGA_GCCAGCGCTTGCCCGCTGGTGCA
5. GCCAGGACACGCGCACGCTTGCCCGCTGGTGCA
6. GCCAGGACCGGCACGCTTGCCCGCTGGTGCA
7. GCCAGGACCGGCACGCTTGCCCGCTGGTGCA
8. GCCAGGACCGGCACGCTTGCCCGCTGGTGCA
9. GCCAGGACCGGCACGCTTGCCCGCTGGTGCA
10. GCCAGGACCGGCACGCTTGCCCGCTGGTGCA
11. GCCAGGACCGGCACGCTTGCCCGCTGGTGCA
12. GCCAGGACCGGCACGCTTGCCCGCTGGTGCA
13. GCCAGGACCGGCACGCTTGCCCGCTGGTGCA

Retention for deletion variants with the EGFP-1 17 nt gRNA

Retention for deletion variants with the EGFP-1 18 nt gRNA

Retention for deletion variants with the EGFP-1 20 nt gRNA
b) Results for EGFP-2.

EGFP-2 Target Sequences
1. CCGGGCGGTTCGTCGAGC
2. CCGGGCGGTTCGTCGAGC
3. CCGGGCGGTTCGTCGAGC
4. CCGGGCGGTTCGTCGAGC
5. CCGGGCGGTTCGTCGAGC
6. CCGGGCGGTTCGTCGAGC
7. CCGGGCGGTTCGTCGAGC
8. CCGGGCGGTTCGTCGAGC
9. CCGGGCGGTTCGTCGAGC
10. CCGGGCGGTTCGTCGAGC
11. CCGGGCGGTTCGTCGAGC
12. CCGGGCGGTTCGTCGAGC
13. CCGGGCGGTTCGTCGAGC
14. CCGGGCGGTTCGTCGAGC
15. CCGGGCGGTTCGTCGAGC

Retention for deletion variants with the EGFP-2 17 nt gRNA

Retention for deletion variants with the EGFP-2 18 nt gRNA

Retention for deletion variants with the EGFP-2 20 nt gRNA
c) Results for rol-6.

**rol-6 Target Sequences**

1. _TCGGACGTCAACATATGAGG
2. _GCAGACGTCAAACATATGAGG
3. CT_GGACGTCAAACATATGAGG
4. GTG_AACGTGAAACATATGAGG
5. GCGA_GGTGAGG
6. CTGGG_GTCGAGG
7. GCGAGA_GTCAACATATGAGG
8. CTGGACGTCAACATATGAGG
9. GCGGACGTGAAACATATGAGG
10. GCGGACGTCAAACATATGAGG
11. GCGGACGTCAAACATATGAGG
12. GCGGACGTCAAACATATGAGG
13. GCGGACGTCAAACATATGAGG
14. GCGGACGTCAAACATATGAGG
15. GCGGACGTCAAACATATGAGG
16. GCGGACGTCAAACATATGAGG
17. GCGGACGTCAAACATATGAGG

**Retention for deletion variants with the rol-6 18nt gRNA in vitro 180 min**

**Retention for deletion variants with the rol-6 G+18nt gRNA in vitro 180 min**

**Retention for deletion variants with the rol-6 G+20nt gRNA in vitro 180 min**
d) Results for *unc-22A*.

**unc-22a deletion target sequences**
1. GTAG\_CACCGCGGTCCGGGCA\_TTTTGCTCC
2. GTAGCG\_CCCGGTCCGGGCA\_TTTTGCTCC
3. GTAGCG\_CCCGGTCCGGGCA\_TTTTGCTCC
4. GTAGCG\_CCCGGTCCGGGCA\_TTTTGCTCC
5. GTAGCG\_CCCGGTCCGGGCA\_TTTTGCTCC
6. GTAGCG\_CCCGGTCCGGGCA\_TTTTGCTCC
7. GTAGCG\_CCCGGTCCGGGCA\_TTTTGCTCC
8. GTAGCG\_CCCGGTCCGGGCA\_TTTTGCTCC
9. GTAGCG\_CCCGGTCCGGGCA\_TTTTGCTCC
10. GTAGCG\_CCCGGTCCGGGCA\_TTTTGCTCC
11. GTAGCG\_CCCGGTCCGGGCA\_TTTTGCTCC
12. GTAGCG\_CCCGGTCCGGGCA\_TTTTGCTCC
13. GTAGCG\_CCCGGTCCGGGCA\_TTTTGCTCC
14. GTAGCG\_CCCGGTCCGGGCA\_TTTTGCTCC
15. GTAGCG\_CCCGGTCCGGGCA\_TTTTGCTCC

**Retention for deletion variants with the unc-22a 17 rIgRNA**

**Retention for deletion variants with the unc-22a 18 rIgRNA**

**Retention for deletion variants with the unc-22a 19 rIgRNA**

**Retention for deletion variants with the unc-22a 20 rIgRNA**
Single base *in vivo* transversion effects with 20+G nt complementarity gRNA with yeast random variant library. Method of analyses as in Figure 5.2. (AF_SOL_580)
Figure S5.10 This Figure shows the full detailed single base *in vitro* variant effects shown in Figure 5.3 for the 180 minute time point. Methods of analysis from Figure S5.2 were used in these experiments.

a) Single base *in vitro* variant effects on the *unc-22A* 17 nt complementarity gRNA (AF_SOL_594).
b) Single base \textit{in vitro} variant effects on the \textit{unc-22A} 18 nt complementarity gRNA (AF_SOL_596).
c) Single base *in vitro* variant effects on the *unc-22A* 20 nt complementarity gRNA (AF_SOL_594).

Figure S5.11 EGFP-1-Reverse (R) and EGFP-1 Off Set (OS)
These guides were both aimed at the same target as EGFP-1 F, but in different orientations. The EGFP-1-R used the CC motif in the constant region upstream of the EGFP-1 target site as its PAM motif. The EGFP-1-OS gRNA used the 3’GG motif as it’s PAM motif. Profiles for these gRNAs with 17nt, 18nt, and 20nt gRNAs can be seen below.

a) Specificity profile (transversions) of the EGFP-1 target cut by the EGFP-1-Reverse gRNA.
Single base transversion effects of EGFP-1 target

- Time = 0 min
- Time = 10 generations, Guide= EGFP-1-R (17nt)
- Time = 10 generations, Guide= EGFP-1-R (18nt)
- Time = 10 generations, Guide=EGFP-1-R (20nt)
b) Specificity profile (transversions) of the EGFP-1 target cut by the EGFP-1-Off Set gRNA.

**Figure S5.12** Cas9 *in vivo* experiments on yeast random type I variant library-2 with extended generation times.
a) Retention results for 20 nt complementarity gRNA (AF_SOL_579).

![Single base in vivo transversion effects on unc-22A gRNA](image)

b) Retention results for 18 nt complementarity gRNA (AF_SOL_580).

![Single base in vivo transversion effects on unc-22A 18bp gRNA](image)
Figure S5.13 Single base *in vivo* transversion effects on the unc-22A gRNA for a physically distinct type I library, yeast random variant library-1.
(AF_SOL_588)

a) Retention results for 20 nt complementarity gRNA.

b) Retention results for 18 nt complementarity gRNA.
Figure S5.14 Single base *in vivo* variant effects of 20 nt complementarity for KU70 and BY4741 yeast strains (AF_SOL_572). Methods of analysis from Figure S5.2 were used in these experiments.

a) Retention results for 20 nt complementarity gRNA in BY4741 background.

b) Retention results for 20 nt complementarity gRNA in KU70 deletion background.
Table S5.2: List of Cas9 cleavage experiments and corresponding sequence datasets.
Data for experiments can be identified by sample ID and Illumina run ID. Illumina run IDs designated as AF_SOL_###. Sample ID is the first number and letter of the filename, for example: 1a is the filename for 1a-lo-3-18bp_S1_L001_R1_001_OUTPUT_AF_SOL_596.fasta. To find the conditions and controls for experiments use the descriptions in the table (e.g. library type and gRNA length).

<table>
<thead>
<tr>
<th>Target random variant library (RVL):</th>
<th>gRNA complementarity length:</th>
<th>Sample IDs:</th>
<th>Control IDs:</th>
<th>Generations after Cas9 induction:</th>
<th>Illumina run ID:</th>
<th>Description of experiment:</th>
</tr>
</thead>
<tbody>
<tr>
<td>unc-22A RVL-2</td>
<td>20</td>
<td>3c, 3d, 3e</td>
<td>1a, 1b, 1c</td>
<td>10</td>
<td>AF_SOL_588</td>
<td>Cas9 cleavage experiment <em>in vivo</em> for 20 nt complementarity gRNA (in strain BY4741)</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>18</td>
<td>3f, 3g, 3h</td>
<td>1d, 1e, 1f</td>
<td>10</td>
<td>AF_SOL_588</td>
<td>Cas9 cleavage experiment <em>in vivo</em> for 18 nt complementarity gRNA (in strain BY4741)</td>
</tr>
<tr>
<td>unc-22A RVL-1</td>
<td>20</td>
<td>2e, 2f, 2g</td>
<td>1g, 1h, 2a</td>
<td>10</td>
<td>AF_SOL_588</td>
<td>Cas9 cleavage experiment <em>in vivo</em> for 20 nt complementarity gRNA (in strain BY4741)</td>
</tr>
<tr>
<td>unc-22A RVL-1</td>
<td>18</td>
<td>2h, 3a, 3b</td>
<td>2b, 2c, 2d</td>
<td>10</td>
<td>AF_SOL_588</td>
<td>Cas9 cleavage experiment <em>in vivo</em> for 18 nt complementarity gRNA (in strain BY4741)</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>17</td>
<td>3a, 3b</td>
<td>1g, 1h</td>
<td>12</td>
<td>AF_SOL_572</td>
<td>Cas9 cleavage experiment <em>in vivo</em> for 17 nt complementarity gRNA (in strain KU deletion)</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>17</td>
<td>2e, 2f</td>
<td>1c, 1d</td>
<td>12</td>
<td>AF_SOL_574</td>
<td>Cas9 cleavage experiment <em>in vivo</em> for 17 nt complementarity gRNA (in strain BY4741)</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>20+G</td>
<td>2b, 2c, 2d</td>
<td>3b, 3c</td>
<td>5</td>
<td>AF_SOL_580</td>
<td>Cas9 cleavage experiment <em>in vivo</em> for 20+G nt complementarity gRNA (in strain BY4741)</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>20+G</td>
<td>2e, 2f, 2g</td>
<td>3b, 3c</td>
<td>10</td>
<td>AF_SOL_580</td>
<td>Cas9 cleavage experiment <em>in vivo</em> for 20+G nt complementarity gRNA (in strain BY4741)</td>
</tr>
<tr>
<td>Target random variant library:</td>
<td>Type of Cas9:</td>
<td>gRNA concentration (ng):</td>
<td>Cas9 protein concentration:</td>
<td>Sample IDs:</td>
<td>Control IDs:</td>
<td>gRNA nt comp length:</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------</td>
<td>------------------------</td>
<td>---------------------------</td>
<td>-------------</td>
<td>--------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>la</td>
<td>1f, 1g</td>
<td>17</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>1b</td>
<td>1f, 1g</td>
<td>17</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>1c</td>
<td>1f, 1g</td>
<td>17</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>1e</td>
<td>1f, 1g</td>
<td>17</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>3a</td>
<td>3f, 3g</td>
<td>20</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>3b</td>
<td>3f, 3g</td>
<td>20</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>3c</td>
<td>3f, 3g</td>
<td>20</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>3d</td>
<td>3f, 3g</td>
<td>20</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>3e</td>
<td>3f, 3g</td>
<td>20</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>1a, 2a</td>
<td>1f, 1g</td>
<td>18</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>1b, 2b</td>
<td>1f, 1g</td>
<td>18</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>1c, 2c</td>
<td>1f, 1g</td>
<td>18</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>1d, 2d</td>
<td>1f, 1g</td>
<td>18</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>1e, 2e</td>
<td>1f, 1g</td>
<td>18</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>NEB</td>
<td>~25</td>
<td>79.5 ng</td>
<td>1g</td>
<td>2e, 2f</td>
<td>20</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>Ref. [72] Cas9</td>
<td>~25</td>
<td>2ug</td>
<td>1b</td>
<td>2g, 2h</td>
<td>18</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>Ref. [72] Cas9</td>
<td>~25</td>
<td>2ug</td>
<td>1d</td>
<td>2g, 2h</td>
<td>18</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>Ref. [72] Cas9</td>
<td>~25</td>
<td>2ug</td>
<td>1f</td>
<td>2g, 2h</td>
<td>18</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>Ref. [72] Cas9</td>
<td>~25</td>
<td>2ug</td>
<td>1h</td>
<td>2g, 2h</td>
<td>18</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>Ref. [72] Cas9</td>
<td>~25</td>
<td>2ug</td>
<td>2b</td>
<td>2g, 2h</td>
<td>18</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>Ref. [72] Cas9</td>
<td>~25</td>
<td>2ug</td>
<td>1a</td>
<td>2e, 2f</td>
<td>20</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>Ref. [72] Cas9</td>
<td>~25</td>
<td>2ug</td>
<td>1c</td>
<td>2e, 2f</td>
<td>20</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>Ref. [72] Cas9</td>
<td>~25</td>
<td>2ug</td>
<td>1e</td>
<td>2e, 2f</td>
<td>20</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>Ref. [72] Cas9</td>
<td>~25</td>
<td>2ug</td>
<td>1g</td>
<td>2e, 2f</td>
<td>20</td>
</tr>
<tr>
<td>unc-22A RVL-2</td>
<td>Ref. [72] Cas9</td>
<td>~25</td>
<td>2ug</td>
<td>2a</td>
<td>2e, 2f</td>
<td>20</td>
</tr>
</tbody>
</table>
Table S5.3 All target oligos synthesized for the Type II library.
Please see the supplemental file for Chapter 5. This file includes all oligos that were synthesized for the Type II libraries.
Bibliography


20. Fraser HB: Genome-wide approaches to the study of adaptive gene expression evolution: systematic studies of evolutionary adaptations involving gene expression will allow many fundamental questions in evolutionary biology to be addressed. Bioessays 2011, 33:469–77.


126. Tirosi I, Reikhav S, Levy AA, Barkai N: **A yeast hybrid provides insight into**
127. Boynton PJ, Greig D: The ecology and evolution of non-domesticated 
128. Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, 
Roberts IN, Burt A, Koufopanou V, Tsai IJ, Bergman CM, Bensasson D, O’Kelly 
MJT, van Oudenaarden A, Barton DBH, Bailes E, Nguyen AN, Jones M, Quail MA, 
129. Warringer J, Zörgö E, Cubillos FA, Zia A, Gjuvsland A, Simpson JT, Forshmark 
A, Durbin R, Omholt SW, Louis EJ, Liti G, Moses A, Blomberg A: Trait variation in 
yeast is defined by population history. PLoS Genet 2011, 7:e1002111.
B, Plikat U, Abraham Y, Altorfer M, Aust T, Baeriswyl L, Cerino R, Chang L, 
Estoppey D, Eichenberger J, Frederiksen M, Hartmann N, Hohendahl A, Knapp B, 
20.
D: The regulatory inputs controlling pleiotropic drug resistance and hypoxic 
response in yeast converge at the promoter of the aminocholesterol resistance 
132. Storici F, Resnick MA: The delitto perfetto approach to in vivo site-directed 
mutagenesis and chromosome rearrangements with synthetic oligonucleotides in 
133. Ibstedt S, Stenberg S, Bagés S, Gjuvsland AB, Salinas F, Kourtchenko O, Samy 
of life stage performances signals recent selection on yeast nitrogen use. Mol Biol 
AA, Kruglyak L: Dissection of genetically complex traits with extremely large 
135. Fraser HB, Moses AM, Schadt EE: Evidence for widespread adaptive 
evolution of gene expression in budding yeast. Proc Natl Acad Sci U S A 2010, 
107:2977–82.
Polygenic cis-regulatory adaptation in the evolution of yeast pathogenicity. 
137. Fraser HB, Babak T, Tsang J, Zhou Y, Zhang B, Mehrabian M, Schadt EE: 
Systematic detection of polygenic cis-regulatory evolution. PLoS Genet 2011, 
7:e1002023.
regulatory evolution across pathways in Saccharomyces. Proc Natl Acad Sci U S A 
139. Artieri CG, Fraser HB: Evolution at two levels of gene expression in yeast. 


155. Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK: *Improving CRISPR-Cas


182. FOLCH J, LEES M, SLOANE STANLEY GH: **A simple method for the


