Massively parallel analysis of the functional effects of mutations

Matthew Saul Rich

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Stanley Fields, Chair

Maitreya Dunham

Robert Waterston

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Massively Parallel Analysis of the Functional Effect of Mutations

Matthew Saul Rich

Chair of the Supervisory Committee:
Stanley Fields, PhD
Genome Sciences

Massively parallel assays can dramatically advance our understanding of biological processes. Coupling them with modern mutagenesis techniques allows for fine mapping of the link between genotype and phenotype. In this dissertation, I first discuss the current state of the field of using massively parallel assays and analysis to study the functional effects of mutations on both coding and non-coding sequences, and I describe many methods for creating large libraries of variants for using these assays. I then describe two studies applying massively parallel assays to address questions in evolution and cancer. In Chapter 2 I measured the effects of mutations in the promoter of the yeast gene *SUL1,* and showed that no mutation (or combination of high-fitness single mutations) was capable of increasing fitness to the extent of amplification of the gene. In Chapter 3 I measured the effects of synonymous mutations in an exon of the tumor suppressor *TP53*, and showed that multiple synonymous mutation had detrimental effects on protein expression due to changes in splicing patterns. In Chapter 4, I discuss initial work developing a

new method for linking genotypes found in the same cell by using *trans*-splicing ribozymes, a technology that could enable massively parallel all-by-all screening. In Chapter 5 I developed a fine-mapping approach based on creating large libraries of chimeric sequences using DNA shuffling, and applied the method to yeast flocculation. Finally, in Chapter 6 I discuss some of the outstanding questions for massively parallel assays, including research directions that I believe will be important in the future.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This work presented in this dissertation would have been incredibly difficult without the support of many people. I would like thank the members of the Department of Genome Sciences for their help and encouragement throughout my graduate work.

I thank Stan Fields. I certainly could not have found a more thoughtful and rigorous advisor for my studies, and it was through his guidance that I became the scientist that I am today. Stan has managed to form a wonderful lab full of engaging and imaginative people, all of whom deserve my thanks. I thank Lea Starita and Doug Fowler, who besides being fantastic and motivating colleagues became two of my closest friends.

To Alan, Ben, Matthew, Max, Vijay, and all the other Genome Sciences students: without your friendship, these seven long years may have felt much longer (if that's even possible).

I thank my parents, who gave me everything I could ever want or need.

I especially thank Cindi, who will always show me the best way to be a spouse, an academic, and a comedian, and whom I will always admire.

And finally, I thank the Mariners, because without them I wouldn't know the joy that is watching Felix Hernandez throw a perfect game, or Brad Miller throw a baseball into the crowd.

Chapter 1. Introduction

1.1 Massively parallel assays for measuring the functional effects of mutations

Classically, methods to analyze the constraints on sequences and the effects of mutations relied on the characterization of individual mutants, created either randomly (e.g., by chemical mutagenesis), by site-directed mutagenesis, or by deletion (for a discussion of mutagenesis methods, see section 1.2). The phenotyping of mutants provided a great deal of information about the functional constraints on a sequence, such as the location and identity of binding sites in a promoter, or the important residues in a protein. However, these experiments were laborious, requiring significant effort both to create and to phenotype variants, and therefore were rarely exhaustive.

With the advent of high throughput sequencing came the ability to rapidly monitor the frequencies of many sequences in a population in parallel (*1*). This advance allows the fitness of thousands to millions of sequences to be measured in parallel, greatly exceeding the throughput of classical mutagenesis studies and providing high resolution information on the functional constraints on sequences. Many massively parallel methods have been developed that take advantage of high-throughput sequencing technology. Here, I will review many of these methods, describing how they are used to study the effects of mutations on gene, protein, and non-coding sequence function.

Another important requirement for carrying out massively parallel experiments is the creation of large libraries of variants to assay. In section 1.2, I will review many currently used mutagenesis

methods, and drawing on my experience and that of others, provide guidelines and considerations for aspects of variant library design.

1.1.1 *Gene knockout technologies: deletion collections and insertional mutagenesis*

For roughly the last 15 years, massively parallel assays have been used to track the fitness of gene deletion and overexpression libraries. This methodology has been especially pervasive in *S. cerevisiae*, in which every open reading frame has been deleted and replaced with a uniquely-barcoded cassette (*2*). This collection of deletion strains can be competed and assayed in parallel by measuring – either by microarray hybridization (*3*) or high-throughput sequencing (*4*) – the frequencies of deletion cassette barcodes remaining in a population after a selection has been imposed on it. In addition to deletion collections in haploid, heterozygous diploid, and homozygous diploid backgrounds, collections of barcoded plasmid-borne open reading frames (*5*, *6*) are available for nearly the entire yeast proteome. Each collection has unique barcodes, allowing them to be pooled and assayed in parallel to study the fitness effects of gene copy number in various conditions (*7*).

Transposons have been used in many organisms to study gene function, because their insertion into a coding sequence can interrupt and knock out a gene. The payload carried by many transposons can also be altered, making them a customizable tool for probing the function of the genome. Transposon insertion sites can be rapidly identified in parallel by digesting genomic DNA with a frequent-cutting restriction enzyme that cuts inside the transposon, circularizing these cut fragments, and specifically amplifying transposon-genome junctions using inverse PCR. These amplicons can then be mapped to the genome, locating the insertion site of the

transposon. Transposon mutagenesis has been used to create gene knockout libraries in multiple organisms (*8-11*). Libraries containing massive numbers of transposon insertions allow for high-resolution functional characterization of genomic sequence. Methods in yeast have shown that massively-parallel characterization of transposon insertion can rapidly identify essential genes, as well as constrained domains within essential genes, and even nucleosome occupancy (*12*, *13*). Transposons have also been used to map the sites of *in vivo* protein-DNA interactions in both yeast and mammalian cells (*14*, *15*). In this method, transcription factors are tethered to a transposon, such that the transposon inserts into the genome near sites that are bound by the transcription factor. These sites can be interrogated by high-throughput sequencing, and the assay can be multiplexed to multiple transcription factors by barcoding the transposons.

1.1.2 *CRISPR-based methods for studying gene and sequence function*

Clustered regularly interspaced short palindromic repeats (CRISPR)-Cas is a bacterial and archaeal system for defense against infecting sequences, in which Cas nuclease proteins are targeted to DNA in a sequence-specific manner by a short targeting guide RNA (gRNA, (*16*)). Targeting requires a short sequence, called a protospacer adjacent motif, at the 3' end of the target, allowing for high-densities (though not full genome coverage) of targets. Cas proteins (most commonly, Cas9, which is a monomer) cleave DNA, causing double-stranded breaks. These breaks can create deletions if they are repaired by non-homologous end joining (NHEJ) often causing frameshifts in coding sequences. CRISPR was adapted for use in mammalian cells (*17-19*) and many other models, including bacteria (*20*), yeast (*21*, *22*), nematodes (*23*), fruit flies (*24*), mice (*25*), and plants (*26*). CRISPR assays are well-suited for massively-parallel analysis, since the sequence of the guide RNA can be used as a proxy for the location of the

Cas9. In these assays, gRNA libraries targeting tens of thousands of coding sequences are integrated via lentivirus into cell lines; mutations (putatively loss-of-function) are allowed to accumulate; and the frequencies of gRNAs are measured over the course of a selection using high-throughput sequencing, yielding fitness estimates for gene knockouts in that selection (*27*, *28*).

Guide RNAs libraries can also be designed to target loci at high density, and the diversity of NHEJ-derived mutations can be used to finely map functional elements (*29-31*). Similarly, variant libraries made *in vitro* can be used as repair donors after cleavage at a specific site, creating many point mutations *in situ* (*32*). These assays require a functional selection, like the expression of a downstream gene in the case of mutagenesis to regulatory DNA. With additional development, it may be possible to assay the regulatory effects of mutations on many targets. I will discuss potential experiments in Chapter 6.

Cas9 has two nuclease domains responsible for DNA cleavage (*16*). Mutating the active sites of either domain yields a nuclease that cuts only a single DNA strand, which can be used with pairs of guides to make deletions with low off-target effects (*33*). Mutating both active sites creates a nuclease-dead version of Cas9 (dCas9) that still has DNA-binding activity, creating an RNA-guided protein module capable of bringing effectors to specific genomic loci. Various effector domains have been fused to dCas9 or tethered to the guide RNA through the use of protein-aptamer pairs (like the MS2 RNA hairpin and MS2 coat protein), such as transcriptional activation domains (*33-37*), repressive domains (*34*, *38*), or chromatin regulators (*39-41*). Methods have also been developed to tether RNA cargo (like long noncoding RNAs) to Cas9

(*42*). These methods are all suitable for massively parallel analysis by sequencing gRNAs after a selection, allowing for rapid, sequence-specific perturbation and analysis of many molecular mechanisms in the cell.

1.1.3 *Deep mutational scanning*

Site-directed mutagenesis has been widely used to identify functional residues in proteins. While many studies have mutagenized proteins based on prior knowledge, approaches like alanine scanning (*43*) provide a general method to probe the functional constraints at many residues with no prior knowledge about a protein. Alanine is a small amino acid with only a methyl group side chain, and therefore mutation to alanine would change other side chains from residues, potentially disrupting protein structure and function.[1] While the initial alanine scanning methodology (*43*) required site-directed mutagenesis to create mutants, combinatorial alanine mutagenesis libraries have been used to more efficiently scan proteins (*45*, *46*).

Because alanine provides only a single physiochemical change to a protein, mutagenizing proteins more broadly should yield richer data about the true constraints of residues. Creating comprehensive mutagenesis libraries by site-directed mutagenesis is impractical, and without high-throughput sequencing methods, experiments necessarily require a screen for only functional variants, which are then Sanger sequenced individually. As such, it is impossible, or at least extremely labor-intensive, to fully measure the activity of all variants to a sequence.

[1] As an interesting side note, analyses of many deep mutational scanning datasets has shown that alanine may not be the best single-residue mutation for the purposes of functionally scanning a protein (*44*). Instead, histidine and asparagine are better correlated with the functional effects of all amino acid mutations, perhaps making them a better option for single amino acid mutagenesis.

Deep mutational scanning (*47*) leverages advances in high-throughput sequencing and oligonucleotide synthesis to track the frequencies of thousands of protein sequence variants in a population in parallel. A general deep mutational scanning workflow consists of creating a library of many thousands of variants of a sequence (see section 1.2 for further information), selecting that library for function, and high-throughput sequencing the initial and selected populations of variants. A DNA variant's change in frequency in the population is concordant with the activity of the protein, and as such a decrease in frequency (often as compared to the wild type sequence) implies that the variant codes for a less active (or less stable) protein, and an increase in frequency implies that the variant codes for a more active protein. Assaying the population of variants at multiple points during a selection provides a quantitative measurement of the activity of all variants in a population, rather than just those remaining after selection, as is the case for an alanine scan or forward genetic screen.

Deep mutational scanning requires an assay that is amenable to a direct linkage of genotype (*i.e.,* variant DNA sequence) and phenotype (*i.e.,* protein activity). Generally, this linkage is accomplished using cells and coupling growth or fluorescence to the activity of the mutagenized protein. In practice, this linkage has been accomplished in a number ways, using a number of different model systems.

Growth selections provide a simple method for measuring variant activity or fitness. Libraries have been selected for resistance to drugs (*48*, *49*), metabolic capacity (*50*), adaptive capacity (*51-53*), and oncogenic gain- or loss-of-function (*54*). Genes from influenza (*55-58*) and HIV (*59*) have been assayed based on cell culture infectivity using deep mutational scanning. Deep

mutational scanning of essential genes can also be performed using temperature-sensitive (*60*) or inducible promoter alleles (*61*) of the endogenous gene.

Organismal phenotypes other than growth are also accessible to deep mutational scanning, as long as variants can be selected for. In one example, researchers mutagenized the DNA-binding domain of a yeast transcription factor, Ste12, which is necessary for both mating and invasive growth (*62*). Two selections were performed, one for mating ability and one for invasive growth into agar, in which invasive cells were extracted from plates with a blender!

Additionally, methods that have linked selectable markers to *in vivo* molecular assays are applicable to deep mutational scanning. For instance, the yeast two-hybrid assay, which can select for protein-protein interaction *in vivo* (*63*) can be used to measure protein binding to a library of protein variants, as was the case for *BARD1* and *BRCA1* (*64*). The yeast three-hybrid assay, which detects protein-RNA interactions, has also been applied to deep mutational scanning to study the binding of the HIV protein Tat to the TAR RNA element (M.S.R and Daniel Melamed, unpublished) and to engineer yeast Pumilio domains to bind RNA sequences of interest (Daniel Melamed, unpublished).

Fluorescence can be used to measure protein abundance, though unlike the continuous fitness measurements in growth selections, these measurements are discrete, arising from cells being sorted into discrete fluorescence bins. Fluorescent reporter abundance has been used for measuring transcription rates (*65*), protein stability (*66*), and RNA splicing (Chapter 3), among other molecular phenotypes.

In vitro phenotypes can also be assayed by deep mutational scanning, provided that the necessary genotype-phenotype linkage is achieved. For instance, phage and yeast surface display methods can be used *in vitro* to select for proteins that bind to a ligand (*47*) or ubiquitin ligase activity (*64*, *67*). Yeast surface display has also been used to select for protein-ligand binding affinity (*68*) and protein solubility (*69*). Surface display and deep mutational scanning have been used to study antibody affinity and stability (*70*-*72*). Sorting of microfluidic droplets has also been used for deep mutational scanning assays to measure enzyme activity using a fluorogenic substrate (*73*).

Deep mutational scanning is a general assay that can be applied to sequences that do not code for proteins. tRNAs have been the target of multiple deep mutational scanning studies (*74*, *75*), as tRNA suppression is an easily-selectable phenotype. Similarly, the protein-RNA interaction between the MS2 coat protein and its RNA hairpin ligand has been analyzed by deep mutational scanning *in situ* on an Illumina flow-cell (*76*).

Deep mutational scans have also be applied to regulatory sequences, if the target of those sequences is a sortable or selectable marker; for instance, bacterial promoters (*65*) and yeast origins of replication (*77*). In Chapter 2, I present a deep mutational scan of a yeast promoter, in which we assay for function based on fitness during chemostat growth.

1.1.4 *Massively parallel reporter assays for regulatory sequence function*

Identifying functional sites in regulatory sequences like promoters and enhancers has been a important pursuit for many years. Early work to define these elements involved creating various mutations to a sequence, for example, by truncation (*78*), internal deletion (*79*), activating sequence replacement (*80*), or random mutagenesis (*81*). Expression can be measured directly by northern blotting or reverse-transcription and PCR, or indirectly by the expression of a reporter like beta-galactosidase (lacZ), luciferase, or a fluorescent protein. Similar to the cases described above for early protein mutagenesis experiments, regulatory mutagenesis experiments generally required a significant amount of effort both to create mutations and to assay expression. In addition, the use of many standard reporters limits the ability to multiplex experiments, as lacZ or luciferase can be excreted, confounding pooled experiments.

Massively parallel reporter assays (MPRA) leverage high-throughput RNA sequencing methodologies to measure the expression level of many sequences in a single experiment. In these experiments, libraries of regulatory sequences (either variants of a specific sequence or genomic sequence of a putative regulatory region) are cloned such that they drive the expression of a transcript containing a barcode. Variants and their expressed barcodes can be linked by high-throughput sequencing. The library is then expressed in a cell, and the barcodes, both as DNA and RNA, are sequenced. By comparing the frequencies of barcodes as DNA (of which there should be one plasmid) and as RNA, an expression level for each variant can be calculated.

Initial MPRA experiments measured the expression of array-synthesized promoter variants (*82-84*). Array synthesis limits the length of sequences to be assayed, and this method has been

extended by tiling synthesized oligonucleotides across putatively regulatory loci, which allows for fine mapping of long regions (*85*), or using polymerase cycling assembly to create longer, combinatorially mutated variant libraries (*86*). MPRA has been also used to map regulatory regions in multiple organisms by cloning fragmented genomic DNA upstream of a barcode (*87-90*). While most MPRA experiments are performed in cell culture, the method has been extended *in vivo*, by measuring the expression of enhancer variants in mouse livers using a hydrodynamic tail vein assay (*86*).

Many MPRA are performed in cell culture by transient transfection, which allows for massive population sizes (*e.g.,* 10^8 fragments assayed in (*90*)). A recent study has shown, though, that episomally-expressed variants are less reproducible and less predictable based on known functional annotations compared to variants that are genomically integrated using lentivirus (*91*).

By definition, MPRA is only able to assay RNA expression, but organismal phenotypes often manifest at the protein level. As such, other methods have been developed to extend MPRA-based methodology to also assay protein expression simultaneously with gene expression. In these methods (*92*, *93*), promoter variants are barcoded in the 3' untranslated region of fluorescent reporter gene. Protein abundance is measured by fluorescence-activated cell sorting and RNA expression is calculated as in MPRA, by comparing the read counts of barcodes in both DNA and RNA. This methodology has also been applied to measure the protein-expression effects of libraries of promoters (*94*, *95*), 5' UTRs (*96*), and 3' UTRs (*97*). Experiments like these are then able to make conclusions about the effects of variants on expression at both the RNA and protein level.

1.1.5 *Massively parallel cell barcoding methods*

Many biological questions can be addressed by monitoring single cells. Cell barcoding has been used in many cases to uniquely identify single cells or cell trajectories. Barcodes have been used to study the clonal dynamics of cells in yeast culture (*98*), cancer (*99*), and hematopoiesis (*100-102*). Methods for directly measuring cell lineages have also been developed (*103-106*), allowing for rapid study of many differentiation processes. Cellular barcoding and high throughput sequencing has also been employed in mapping neuronal connections (*107*, *108*). Another need for cellular barcoding is assaying combinations of genotypes (*e.g.,* variant libraries encoded on multiple plasmids per cell, or multiple gene deletions per cell). Some methods accomplish this by recombining DNA barcodes between library loci (*109-111*). In Chapter 4, I discuss initial success in assaying combinations of multiple barcodes using *trans*-splicing ribozymes.

1.2 Techniques for creating variant libraries for massively parallel assays

Massively parallel assays require the generation of a library of sequences. In this section, I will review some of the methods currently used to create these libraries. I will focus on libraries of targeted sequences for use in assays (like DMS, MPRA, and others I described in the previous section) that attempt to determine the function of specific bases or residues in those sequences.

1.2.1 *Methods based on site-directed mutagenesis*

Many library mutagenesis methods conceptually are extensions of classical, low-throughput site-directed mutagenesis methods (*112*), but with the aim to create all possible variants of a sequence. In general, a wild type sequence is cloned onto a plasmid and it is then mutagenized

by PCR with primers containing a mutation (Figure 1.1A and (*113*, *114*)). After PCR to create mutant plasmids, the original plasmids are digested using the methylation-specific restriction enzyme *Dpn*I, and in the case of the method from Jain *et al.*, linear amplicons are circularized. Site saturation mutagenesis libraries can be created by using degenerate primers, yielding all codons at a single site. The use of tiling degenerate primers across a sequence, one base or codon per reaction, and pooling these reactions can create a saturation library of all single mutations of a sequence. These libraries require two primers per site to be mutagenized, and the necessity that the mutagenesis reactions be performed individually for each site makes these methods labor-intensive. Multiple methods have been developed to make this library preparation step massively parallel by combining the multiple mutagenesis reactions into a one-pot reaction.

One such method, Pfunkel[2] mutagenesis (Figure 1.1B and (*115*)), is an adaptation of Kunkel mutagenesis (*116*) that uses as a template a plasmid extracted from bacteria that are able to incorporate uracil into DNA. A phosphorylated primer containing an internal degenerate site is annealed and extended around the template. The new strand is then ligated with a thermostable ligase. This reaction can be cycled multiple times, linearly amplifying the template with new mutant copies. After mutagenesis, second strands are created by a single fill-in reaction with a non-mutagenic primer annealed to the plasmid backbone. The wild type template DNA can then be degraded with uracil-deglycosylase, leaving predominantly mutagenized plasmids.

Another method (Figure 1.1C and (*117*)) performs similar mutagenesis, but employs exonuclease cleavage to degrade wild type template DNA. After an initial cleavage with a nicking endonuclease and exonuclease nuclease digestion, mutagenesis is performed as in Pfunkel, with

[2] The name Pfunkel is a pun based on Kunkel and the Pfu polymerase employed in the method.

phosphorylated degenerate oligonucleotides, a high fidelity DNA polymerase, and a thermostable ligase. The second template strand can then degraded by another nicking endonuclease cleavage and exonuclease digestion, and a new second strand is synthesized from a primer on the plasmid backbone.

Both these methods can create pooled libraries in a single reaction, though each retains about 25-50% of the library as wild type. The number of mutations per sequence can be adjusted, since this number is dependent on how many mutagenesis oligonucleotides anneal to each template. Both papers demonstrating these methods (*115*, *117*) used an excess of template to bias toward more single wild type and single mutants, but multiply-mutated variants could be created by decreasing this excess. Additionally, these methods are suitable for programmed synthesis (see Section 1.2.3), rather than degenerate mutagenesis.

Finally, the choice of degenerate base or codon is important. In protein mutagenesis experiments (as for deep mutational scanning), it is important to limit the number of premature termination codons, while maximizing the spectrum of nonsynonymous mutations. Several strategies have been adopted to address this problem. The use of NNK (where K is guanine or thymine) or NNS (where S is guanine or cytosine) as a degenerate codon rather than NNN results in only a single stop codon (UAG), but limits the possible codons available for other amino acids. Algorithmic and probabilistic optimization has been used to try to determine the optimal degeneracy to use for a mutagenesis scenario (*118*-*120*).

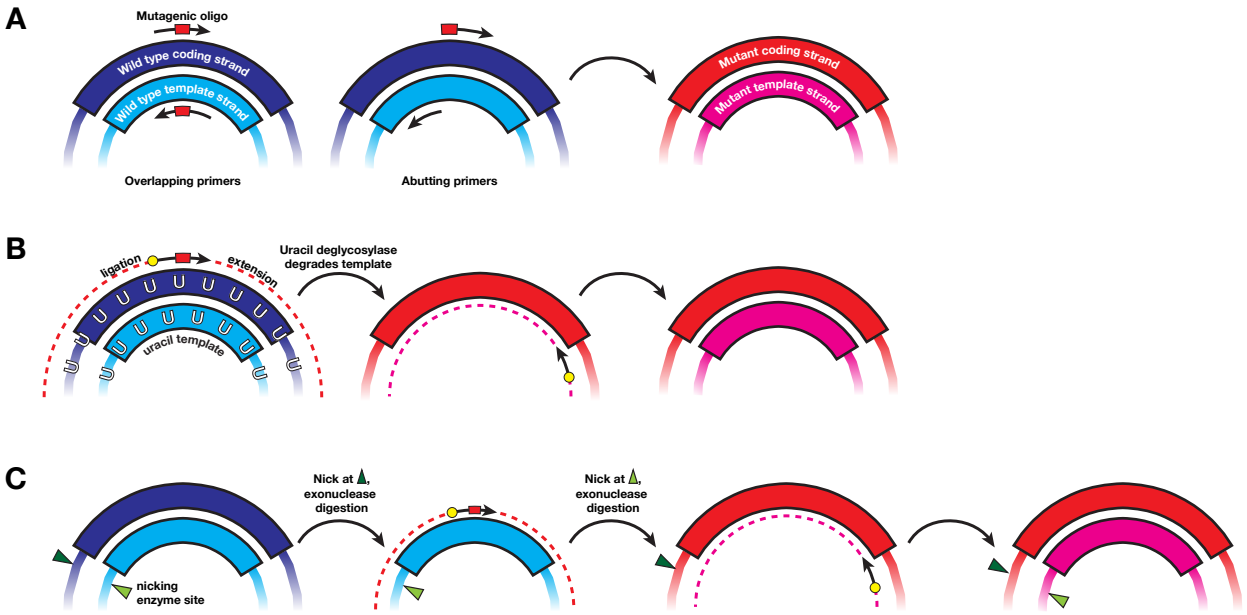

Figure 1.1. Site-directed mutagenesis methods. (A) Site-directed mutagenesis, either with overlapping or abutting mutagenic primers. (B) Pfunkel mutagenesis. (C) Nicking mutagenesis.

1.2.2 *Random mutagenesis methods*

It is often necessary or preferable to mutagenize a template randomly, since these methods are generally simpler than the site-directed mutagenesis-based methods described above. Due to the nature of random methods, the researcher is not able to control the mutations made to the template. Random mutagenesis is the simplest way to generate multiply-mutated variants, though more programmable methods for combining mutations can also be used (see Section 0).

Mutations can be randomly made to templates by error-prone PCR (epPCR, Figure 1.2A). epPCR relies on the use of a non-proofreading DNA polymerase (like *Taq* polymerase) as well as changes to the amplification reaction conditions that favor misincorporation of bases (*121*). Generally, these conditions include increased $MgCl_2$ (or $MnCl_2$) concentrations to stabilize mispairing and changes in the concentrations of nucleotides. Adaptations of these methods

include using inosine, which basepairs with any of the four canonical nucleotides (*122*), or other base analogs (*123*), or mutating the DNA polymerase used in the reaction to make it more error-prone (*124*). Further optimization of epPCR protocols has yielded robust methods (e.g., the Mutazyme II polymerase mix from Agilent Technologies) with relatively uniform mutation frequencies. epPCR can also be carried out to generate variable mutation rates, ranging from less than one mutation per kilobase to a maximum of about 20 mutations per kilobase. This rate is generally controlled by adjusting the amount of amplification of the template; less amplification, either by starting with more template or running fewer cycles of PCR, yields less mutagenesis, and *vice versa*. The number of mutations per variant in libraries created by epPCR are Poisson distributed (see Figure 2.3).

epPCR can easily create mutagenesis libraries of long sequences (as compared to oligonucleotide-based methods). As such, it has been widely used for the directed evolution of a protein. This benefit, though, can be a detriment for massively parallel assays, in which the long distance between mutations in a sequence makes it difficult to identify which mutations are present in the same variant. Also, because the majority of epPCR mutations are single point mutations, it can be difficult to generate the nonsynonymous mutation diversity necessary for a deep mutational scanning experiment. In Chapter 2, I use error-prone PCR to make a variant library of a 493-base promoter. Error-prone PCR in that case was one of the few methods available to create a library of that length with the complexity we desired of mostly single mutations, but with a significant proportion of multiply-mutated variants.

"Doped" or "spiked" oligonucleotides can be used for shorter mutagenesis targets (Figure 1.1B). These oligonucleotides are synthesized with the wild type base at each site mixed with a defined proportion of mutant bases (e.g., at a position that is wild type adenine, 3% doping would mix 1% each of cytosine, guanine, and thymine). Because the mutant proportion can be adjusted and because, as in epPCR, the number of mutations per variant follows a Poisson distribution, doping creates a very predictable and simple mutagenesis method. A basic doping strategy, like that used in (*47*), uses a fixed percentage of mutant bases at each site across an entire sequence. Oligonucleotide synthesis can also be programmed to vary the doping mixture proportions at each site, yielding much more complex or targeted mutagenesis libraries. For example, in Chapter 3 we create a library of synonymous variants using doped oligonucleotide synthesis. In this library, every third position in a codon was mutagenized at a total of 3%, but only to bases that would create synonymous mutations at the codon. As such, most positions (which were not the third position of a codon) were not doped, some were doped 1% for all three mutant bases, and some were doped 3% for only one base.

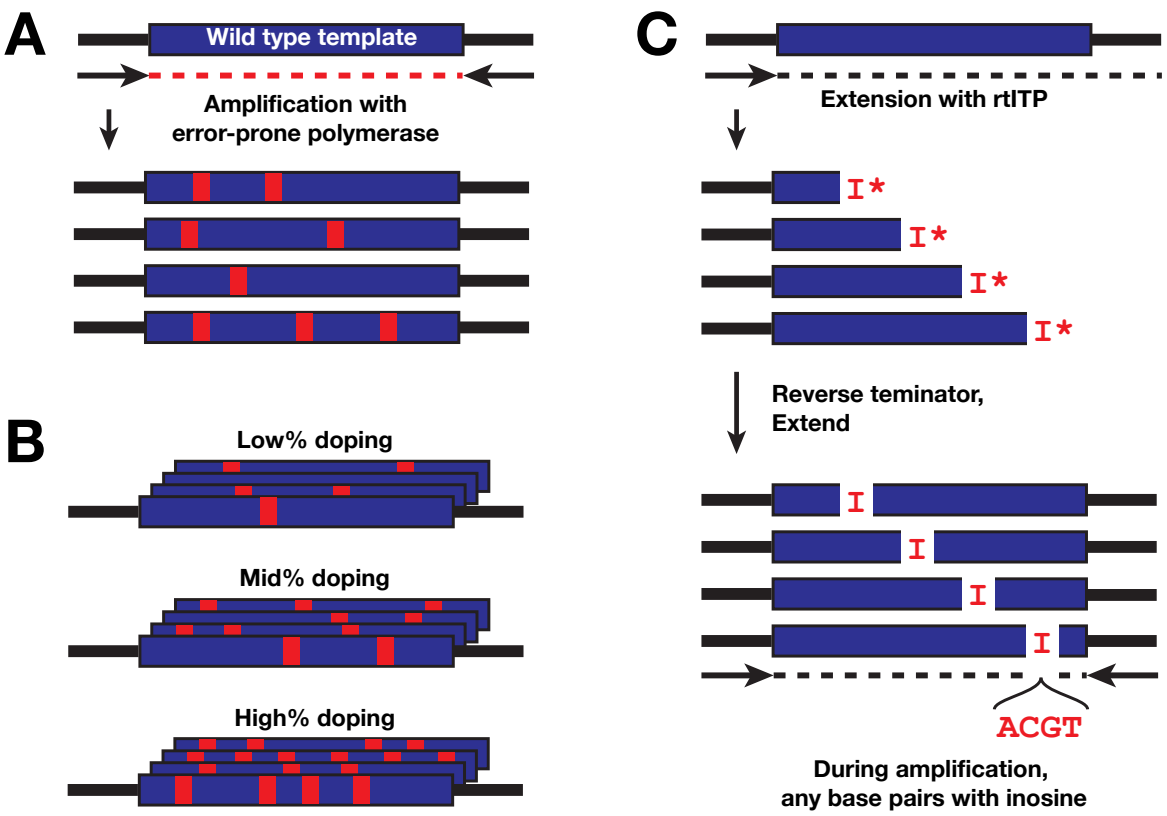

Figure 1.2. Random mutagenesis methods. (A) Error-prone PCR. Random mutations are created by an error-prone DNA polymerase. (B) Doped oligonucleotide synthesis. Three relative levels of doping are shown. (C) Random mutagenesis with reversibly-terminating inosine. The template is linearly-amplified in a reaction containing a reversibly-terminating inosine (rtITP, I*). When rtITP is incorporated, extension is blocked. The terminator can then be reversed, and the extension finished. During amplification of this product, any base will pair with inosine, making mutations at each inosine-containing sites.

Taking doped oligonucleotides one step further leads to fully random sequences. For decades, pools of fully random oligonucleotides have been used to identify the sequences of functional genomic elements (*125-127*). While these early works were not massively parallel, recent work has demonstrated the utility of massively parallel analysis of random pools for studying splice site selection (*128*) and the effects of 5' untranslated regions on translation (*129*). Using fully

random sequences is not necessarily a "mutagenesis" method, *per se*, but analysis of these sequences can uncover contextual effects on functional elements in addition to simply their sequence constraints.

A recent mutagenesis method tried to combine the ability of site-directed techniques to make only single mutations with the ease of making random mutagenesis libraries. This method (Figure 1.2C and (*130*)) uses reversibly-terminated inosine triphosphates (rtITP), akin to chain-terminator sequencing (*131*) in that a single rtITP is incorporated during polymerase extension. The rtITP, like ddNTP in Sanger sequencing, stops extension, so one and only one rtITP is incorporated per strand. The terminator on the ITP is then chemically reversed and the extension reaction is allowed to proceed, this time without rtITP. Inosine-containing molecules are then amplified, during which the inosine is base-paired with each of the four canonical nucleotides. This method creates libraries consisting of nearly only wild type and single-mutated variants, but presently the necessity for in-house synthesis of rtITP is a large barrier for widespread use.

1.2.3 *Programmed mutagenesis*

The advent of multiplex array-based oligonucleotide synthesis (*132*, *133*) has revolutionized mutagenesis methods, as programmed arrays can produce tens of thousands to millions of high quality oligonucleotides with specific, user-defined sequences. As such, researchers can make large libraries comprising specific subsets of mutations and combinations of mutations in a sequence, or altogether novel sequences, to test for function.

Currently, array-derived oligonucleotides are relatively short (about 200 bases, at the longest), and generally are designed to contain primer binding sites on each end, allowing for near-unlimited use after synthesis (Figure 1.3A). This limitation constrains the region for mutagenesis to approximately 150 bases. Multiple methods for combining these oligonucleotides into longer sequences have been developed, though these methods either combine many oligos into relatively few completed sequences (Figure 1.3B, *134*) or only two oligos into many pairs (Figure 1.3C, *135*). To create large pools of long sequences derived from array-synthesized oligos, some methods have tiled subpools of variant oligonucleotides across wild type sequences (*136*), or combined subpools (including wild type oligos) to create multiply-mutated variants (Figure 1.3D, *48*). These methods often require removal of the flanking primer-binding sites, which has been accomplished by cleavage with type-IIS restriction enzymes (*134*) and uracil deglycosylases (*51*, *135*).

One of the first methods to use array oligonucleotides for mutagenesis across a long sequence is called Programmed Allelic Series (PALS, *51*). Oligonucleotides that mutate each codon to every other codon are annealed and extended along a uracil-containing template to create "megaprimers," which are then filled-in to create full-length sequences. By the use of uracil deglycosylase to degrade the wild type template (similar to Pfunkel mutagenesis, although uracils are incorporated into the template *in vitro* in this methodology), most of the resulting library should contain only single mutations.

Array-derived oligonucleotides can be used in many of the methods described here, in the place of individually-synthesized oligos. Methods like pFunkel (*115*) or nicking mutagenesis (*117*),

recombineering (*137*, *138*), and DNA shuffling (*139*, *140*) are well-suited for programmed mutagenesis.

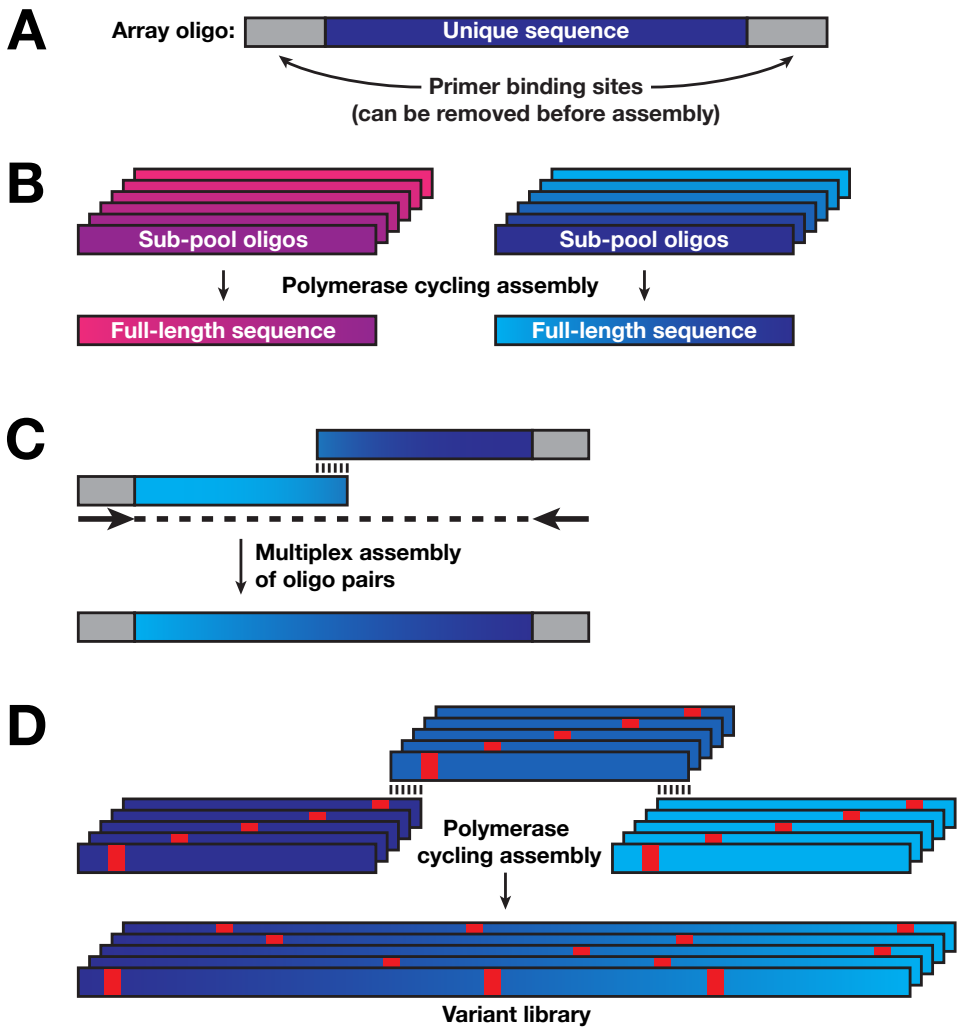

Figure 1.3. Programmed mutagenesis methods. (A) Cartoon of an array-derived oligonucleotide. Oligos are generally synthesized with flanking primer binding sites for amplification (grey). For some methods, these sites must be removed. Other methods can use them for cloning purposes. (B) Assembly of oligo pools. In order to make assemble sequences longer ~300 bases, oligos are amplified into subpools, adaptors are removed, and subpools are assembled using polymerase cycling assembly (*134*).

(C) Multiplex assembly of oligo pairs. Pairs of synthesized oligos can be combined by PCR if overlaps are designed to be specific for each pair (*135*). (D) Assembly of full-gene variant libraries. Similar to (B), oligos containing mutations are assembled with polymerase cycling assembly to create full-length mutated genes (*48*).

1.2.4 *Combinatorial mutagenesis*

To this point, the methods discussed either aim to create single mutations across a sequence (Section 1.2.1), random mutations (and combinations of mutations) across a sequence (Section 1.2.2), or designed sets of mutations in a short sequence (Section 1.2.3). In this section, I will describe current methods for creating libraries in which only specific sites across long sequences are mutagenized.

DNA shuffling is a method by which homologous genes are fragmented, mixed, and allowed to reassemble using polymerase cycling assembly (*139*, *140*). This methodology has been widely used for directed protein evolution as a way to combine mutations from high-activity variants (*141-143*). Additionally, DNA shuffling can be used assort variants from different related sequences, and if followed by phenotyping and shotgun sequencing, *a la* bulk segregant analysis, these libraries can inform whether mutations cause phenotypes (*144*, *145* and Chapter 5).

DNA shuffling can also be used to combine single-variant libraries into multiply-mutated libraries, although to my knowledge this methodology has not been used to make variant libraries for massively parallel analyses.

Combinations of mutations can also be created *in vitro* using overlap-extension PCR (see Figure Figure 1.3D). In this method, primers that are degenerate for either wild type or a mutation at sites of interest are used to amplify fragments of a template. These fragments can then be used as "mega-primers" for subsequent rounds of PCR, in addition to other degenerate primers, until all sites are mutagenized. We used this methodology in Chapter 2 to create a small library containing 64 combinations of six mutations in five rounds of PCR. This method has been generalized (*146*) by using sets of degenerate primers for either forward or reverse amplification reactions. The amplicons from these reactions are then combined by more cycles of PCR, yielding combinations of mutations. In a similar methodology, all mutations of interest are synthesized in discrete sections by array synthesis, then combined by PCR (*48*).

These PCR-based methodologies become complicated as mutagenesis is required at more sites. An approach known as MAGE uses recombineering to create multiplexed genome edits (*137*), and the same methodology can be applied to creating multiplexed mutations to a plasmid in a recombineering-competent *E. coli* strain (*138*). Multiple recombineering rounds are necessary for high-complexity combinations, but each round is simply an electroporation of plasmid and targeting oligonucleotides and outgrowth of the electroporated cells. As one would expect, mutations at positions that are too close together (within the homology arms of a targeting oligo) are made at low efficiency, and targeting oligos containing multiple mutations recombine less efficiently than those with fewer mutations. Nonetheless, recombineering methods like this one could be used for creating combinations of mutations at sparse sites in a sequence.

CRISPR has also been used combinatorially, by creating pairs of gRNAs that are synthesized together and cloned into the same plasmid (*147-149*).

1.2.5 *Constraints for massively parallel libraries: amplicon length*

A basic requirement for massively-parallel assays is that genotypes being assayed must be able to be sequenced with high-throughput sequencers capable of millions of short reads. In this work, we employed Illumina sequencers, although the methods presented here could be adapted for other high-throughput short read sequencing technologies. While long-read sequencers, like those from Pacific Biosciences (*150*) or Oxford Nanopore (*151*), are able to phase long sequences, these technologies have both high error rates and low throughput, making them less attractive solutions for quantitatively measuring large libraries.

This requirement – sequencing many millions of amplicons in order to quantitatively measure library members – forces massively parallel assays to assay the library by sequencing short amplicons. These amplicons could be the variant library itself, though the short length of high-throughput sequencing reads (from Illumina, a maximum of 600 bases in two 300 base reads) limits the assayable length of the library, and read quality is known to degrade at the end of long reads, making genotyping each site in the library more difficult. Due to these limitations, in our experience it is currently practical to make libraries up to about 500 bases in length if they will be genotyped by direct sequencing.

In order to assay variant libraries that exceed the acceptable read length of sequencers, library members can be uniquely tagged with a short sequence, or barcode, cloned into the same

plasmid. These barcodes must first be linked with their variants, but after this initial step, only the barcodes need to be amplified and sequenced, which can be done at higher throughput and lower cost. Additionally, if variants are redundantly barcoded (which would be the case if the number of barcodes cloned into the variant library is larger than the size of the variant library itself), each barcode then becomes a unique measurement of variant activity, yielding better estimation of true activity. In Chapter 2, we employ this methodology to assay our variant library. Although the length of that library (493 bases) is shorter than 500 bases, barcoding the libraries allowed for higher-depth sequencing of many time points.

Variant barcoding also allows assaying libraries that span lengths longer than sequencer read lengths. In a method known as "subassembly" (*152*), a barcoded library is sequenced with paired-end reads such that one read sequences the barcode and another sequences different fragments of the library member tagged by the barcode. The full-length variant can then be constructed by assembling all reads paired with a single barcode. These variant fragments can be made in multiple ways, including shearing DNA and ligating sequencing adapters (*152*), Tn5 transposase-mediated fragmentation (*51*), PCR tiling across the region (*67*), or exonuclease digestion to shorten the library followed by re-circularization next to the barcode (*64*). In theory, any molecular biological methodology that creates a single amplicon containing a barcode and a library fragment short enough to sequence with a single read can be used for subassembly.

In many experiments, it may be possible to extend the library of a mutagenized library to beyond the length of a sequencing read and assay the variants in that library by either shotgun (*73*) or amplicon sequencing (*55*, *153*). By assaying the library in this way, phasing information about

variants is lost, so it becomes impossible to measure epistasis between mutations in the sequence. Similarly, without long read sequencing, it is only possible to shotgun sequence sparse combinatorial libraries, like those made by DNA shuffling. In this case, it is appropriate to apply the same analyses as bulk segregant analysis (Chapter 5 and ref. *154*).

1.2.6 *Constraints for massively parallel libraries: library diversity*

In any massively parallel experiment, it is important to consider the diversity of the library that can be assayed in a given model system, as different model systems have different population sizes. Because it is necessary to quantitatively measure the frequencies of variants in the population, expressing each library member in more than one cell (and, realistically, in at least ten cells) allows for confident measurements. For example, a detrimental variant should be found in many cells in the initial population, allowing for quantitative measurements of its low fitness.

In general, the size of the library capable of being expressed in an organism decreases as the complexity of the organism increases. *In vitro* assays like SELEX can assay over 10^{12} variants, generally randomers, which is far beyond the sequencing capacity of a high-throughput sequencer (*155*). Similarly, *in vitro* protein displays assays like phage or ribosome display can assay between 10^{6}-10^{12} variants (*156*). *In vivo* systems are capable of assaying fewer variants than *in vitro*, as each variant needs to be expressed in a cell, though not necessarily in one cell per variant, depending on the assay.

Growth- and sorting-based assays require a single variant per cell, which limits throughput. Libraries of 10^{9} bacteria or 10^{6} yeast are generally attainable. In Chapter 2, we created pools of

~$5x10^6$ yeast colony forming units transformed with library plasmids. Expressing a single variant in mammalian cell culture is more complicated, since plasmids are more unstable than they are in yeast and bacteria, but we and others have developed methods for stably expressing a single variant in mammalian cell culture with library sizes of 10^4-10^5 variants (Chapter 3 and ref. *157*).

Lentiviral infection of variants into cell culture is widely used for large scale expression of 10^4-10^5 shRNA (*158*) or CRISPR gRNA (*27*), though these experiments require infection at low multiplicity of infection to enrich for singly-infected cells, and cannot rule out multiply-infected cells. Notably, libraries in MPRA experiments, in which cells are used simply to express variants, can be transiently transfected at much higher efficiencies (up to 10^8 variants per experiment, *90*).

1.2.7 In situ *mutagenesis*

To this point, I have discussed only *in vitro* cloning of variant libraries, which requires multiple cloning steps that restrict library diversity. Recently, multiple groups have developed methods for performing mutagenesis *in situ* in the genome of both yeast and mammalian cell culture (*159*, *160*). These methods fuse cytidine deaminase (CDA) domains to Cas9. When Cas9 is targeted to a genomic locus, the cytosine bases are deaminated to uracil, which then are mutagenically repaired or base-paired with adenine during replication, creating a C>T transition. CDA efficiently mutates cytosines, but only approximately 15 bases upstream of the protospacer adjacent motif, presumably due to deformation of the DNA helix by Cas9 (*161*), limiting this method's applicability to large scale mutagenesis. One application of these proteins is specific base mutagenesis, and the method has been optimized to improve specificity and efficiency (*162*,

163). Optimizing this system to increase editing diversity, or development of other *in situ* mutagenesis techniques should enable library-scale experiments in higher organisms like nematodes or fruit flies, which are both limited by low population sizes as well as inefficient introduction of variants.

Chapter 2. Comprehensive analysis of the *SUL1* promoter of *Saccharomyces cerevisiae*

In the yeast *Saccharomyces cerevisiae*, beneficial mutations selected during sulfate limited growth are typically amplifications of the *SUL1* gene which encodes the high affinity sulfate transporter, resulting in fitness increases of >35%. *Cis*-regulatory mutations have not been observed at this locus; however, it is not clear whether this absence is due to a low mutation rate such that these mutations do not arise, or they arise but have limited fitness effects relative to those of amplification. To address this question directly, we assayed the fitness effects of nearly all possible point mutations in a 493 base segment of the gene's promoter through mutagenesis and selection. While most mutations were either neutral or detrimental during sulfate-limited growth, eight mutations increased fitness more than 5% and as much as 9.4%. Combinations of these beneficial mutations increased fitness only up to 11%. Thus, in the case of *SUL1,* promoter mutations could not induce a fitness increase similar to that of gene amplification. Using these data, we identified functionally-important regions of the *SUL1* promoter and analyzed three sites that correspond to potential binding sites for the transcription factors Met32 and Cbf1. Mutations that create new Met32 or Cbf1 binding sites also increased fitness. Some mutations in the untranslated region of the *SUL1* transcript decreased fitness, likely due to the formation of inhibitory upstream open-reading frames. Our methodology – saturation mutagenesis, chemostat selection, and DNA sequencing to track variants – should be a broadly applicable approach.

Contributions M.S.R., Celia Payen, Maitreya Dunham, and Stanley Fields developed these ideas and wrote this manuscript. MR designed and cloned all mutagenesis libraries. C.P., Monica Sanchez, and Giang Ong performed chemostat selections. M.S.R. prepared sequencing libraries

and analyzed data using custom scripts written by M.S.R. and Alan Rubin. Some strain creation was performed in Nozomu Yachie's laboratory. We thank Bill Noble and Charles Grant for help with Tomtom and FIMO and for discussions about motif matching, Kunihiro Ohta for the use of his tetrad dissection microscope. This work has been published .

2.1 Introduction

Changes in the extent or timing of gene expression can have profound effects on molecular and organismal phenotypes and thereby drive evolution . Heritable noncoding variation can alter gene expression in *cis* or in *trans* , and both have been shown to contribute significantly to gene expression variation . Two primary mechanisms by which *cis* variation can increase gene expression are increases in gene copy number and point mutations in regulatory regions. However, the relative effect size of amplification compared to point mutation is not known. It has been proposed that amplification is both quickly achieved and then reverted after fixation of fitness-increasing point mutations . This mechanism raises the question of whether amplification is simply a transitional state that provides an increased chance for a beneficial point mutation to occur. Alternatively, gene amplification may be so frequently observed because the fitness effects it confers are greater than those achievable by point mutations. Gene amplifications have been found to be advantageous in many contexts, including phenotypic evolution and cancers , but few experiments have addressed whether *cis*-regulatory mutations could also lead to similar effects. In one study, random mutations were introduced into yeast and assayed for their effects on the expression of a fluorescent reporter . Strains with altered reporter expression were classified as either having *trans* or *cis* mutations, the *cis* mutations being either

non-coding point mutations in the reporter construct or amplifications. As expected, amplifications increased reporter expression, but there was no difference in effect between these and fluorescence-increasing *cis*-regulatory point mutations (though only two such mutations were isolated). No strains with a reporter copy number greater than two were isolated, leaving open the possibility that higher copy number gene amplifications could enhance expression beyond that achieved by point mutation.

In order to study these questions directly, we turned to another case in which gene amplifications are adaptive. When *S. cerevisiae* is subjected to long term growth under sulfate-limitation, a region of chromosome II containing the sulfate transporter gene *SUL1* is recurrently amplified after ~50 generations . The amplification of this gene leads to a fitness increase of 37% -51% , depending on amplicon size and copy number. Amplification appears to be the preferred means by which fitness can be increased, as the appearance of coding or non-coding mutations at the *SUL1* locus is very rare. However, this preference for amplification of *SUL1* could have other possible explanations besides superior fitness, such as differences in the mutation rate between point mutations and amplifications. We sought to determine whether *cis*-regulatory mutations were capable of comparable effects by directly creating such mutations via mutagenesis, an approach that allowed us to skirt the potential effect of mutation rate.

The specifics of *SUL1*'s transcriptional regulation are largely unknown. *SUL1* is one of the 45 genes comprising the core Met4 regulon of genes involved in yeast sulfur metabolism (*164*). Met4 is a transcriptional activator, but has no DNA-binding activity. It is targeted to promoter sequences by combinations of three transcription factors – Cbf1, Met31, and Met32 – that

themselves lack transcriptional activation activity. Cbf1, Met31, and Met32 can induce transcription individually and in combination (*164*, *165*). *SUL1* is unannotated in many functional genomic studies on regulation, including RNA sequencing (*166*), DNase I mapping (*167*), and ChIP-Exo (*168*), presumably due to its low expression in the sulfate-rich conditions used by these studies.

Mutagenesis has long been applied to study the function of non-coding sequence. Early work in yeast established the structure of eukaryotic promoters, delineating upstream activating sequences as the binding sites for transcriptional regulators. These experiments changed promoter sequences by truncation, internal deletion, replacement of activating sequences, rearrangement, and random mutagenesis (*169*). However, these experiments were low-throughput, assaying promoter variants one at a time. More recent studies that couple mutagenesis with high-throughput sequencing have increased the resolution and throughput of the analysis of eukaryotic *cis*-regulatory elements (*83*, *84*, *86*). These studies assay thousands of variants simultaneously and provide detailed information on the relative mutational constraints on each base of a promoter, identifying transcription factor binding sites and other regulatory elements.

Here, we use the methods of mutagenesis, selection, and sequencing to assay the fitness of nearly all single mutations in the *SUL1* promoter. We show that while point mutations can increase fitness in sulfate limitation by up to approximately 10%, neither single mutations nor combinations of these mutations can increase fitness to the extent that *SUL1* amplification can. We also use these data to define potential transcription factor binding sites that regulate *SUL1*

expression and to identify point mutations that create new regulatory sites. Additionally, our assay is sensitive to post-transcriptional effects of *cis*-regulatory mutations and identified detrimental mutations that create new upstream open reading frames in the *SUL1* 5' untranslated region.

2.2 Materials and Methods

Oligonucleotides, yeast strains, and plasmids used in this study. Oligonucleotides used in this study can be found in Table 7.2. The *S. cerevisiae* strain used in this study was FY3, a *MATa* uracil auxotroph (*ura3-52)* of the S288c background. The strain deleted for *SUL1* was obtained by transformation with a PCR fragment containing a *NatMX* cassette and two flanking regions with homology to the *SUL1* locus. The transformant was backcrossed three times to FY2 to select for a clone containing both the *sul1* deletion and the *ura3-52* allele. A list of strains and plasmids used in this study can be found in Table 7.3.

Fitness estimates of individual strains. Fitness measurements of individual clones were performed as previously described (*7*) in sulfate-limited chemostats using a prototrophic FY strain where the *HO* locus had been replaced with *eGFP* (*MATa*: YMD1214).

Promoter truncation and mutagenesis. Promoter truncations were created by amplifying the *SUL1* locus from genomic DNA using oligos 314-319 as forward primers and oligo 266 as the reverse primer and the following PCR conditions: 98ºC for 1 minute; then 25 cycles of 98ºC for 10 seconds, 65ºC for 15 seconds, and 72ºC for 15 seconds; then a final incubation at 72ºC for 5

minutes. These PCR products were cloned into an *EcoR*I- and *Sac*I-digested pRS416 vector by Gibson assembly and their sequences were confirmed by Sanger sequencing. Yeast was transformed by lithium acetate transformation with each plasmid individually. At least two independent transformants were used for fitness measurements. The plasmid containing the 493 base promoter and remainder of *SUL1* was named pMR002.

The template used for mutagenesis was first amplified from pMR002 using oligos 295 and 297. Promoter mutagenesis was performed using the GeneMorph II Random Mutagenesis Kit (Agilent) according to manufacturer's recommendations. One and 10 nanograms of template were amplified in mutagenesis reactions using oligos 295 and 297 and the following conditions: 95ºC for 2 minutes; then 25 cycles of 95ºC for 30 seconds, 55ºC for 30 seconds, and 72ºC for 30 seconds; then a final step of 72ºC for 10 minutes. These libraries were gel-extracted with a MinElute Gel Extraction kit (Qiagen), mixed one-to-one, and cloned into a *BamH*I and *Sac*I-digested pMR002 plasmid using Gibson assembly. One microliter of this reaction was transformed directly into ElectroMAX DH10B electrocompetent *E. coli* (Life Technologies).

Barcodes were created by annealing oligos 283 and 296, then performing a single cycle of extension using the Klenow fragment of DNA polymerase I (NEB). These fragments were cleaned and concentrated using a DNA Clean and Concentrator-5 kit (Zymo Research). pMR002 library plasmids were digested overnight with *Hind*III, dephosphorylated with calf intestinal phosphatase (NEB), and gel purified. Barcode fragments were mixed in a 10-fold molar excess with cut library plasmids, and cloned using Gibson assembly. ElectroMAX DH10B electrocompetent *E. coli* (Life Technologies) was transformed with 1 μl of this reaction. In total,

the library comprised 157,159 variants tagged by 634,639 barcodes. Yeast strains *FY3* and *FY3 sul1* were transformed with barcoded plasmids using a high-efficiency protocol (*170*), resulting in over 5 million transformed cells per library.

Continuous culture in chemostats. Nutrient-limited media (sulfate-limited and glucose-limited) were prepared as described (*171*). The 200 ml chemostat vessels were inoculated with 1 ml of each pool (~$2x10^7$ cells). The pools were grown in chemostats for 25 hours in batch and then switched to continuous culture at a dilution rate of 0.17±0.01 volumes/hour at 30°C. The cultures reached steady state after ~6 generations and were maintained for ~40 generations. A sample was taken immediately after the initiation of pumping and was designated Generation 0 (G0). Samples (25 to 50 ml of culture at a density of $2x10^3$ cells/µl) for cell counting and DNA extraction were passively collected once or twice daily, every 3 to 6 generations on average.

Mapping barcodes to promoter sequences. To map barcodes to promoter sequences, 25 ng of pMR002 was amplified for 9 cycles with KAPA HiFi Hotstart Readymix using oligos 327 and 328 and the following cycling conditions: 98˚C for 20 seconds; then 9 cycles of 98˚C for 30 seconds, 65˚C for 15 seconds, and 72˚C for 25 seconds. This reaction was cleaned with a DNA Clean and Concentrator kit (Zymo Research) and quantified using a Qubit florometer (Life Technologies). The products were sequenced with 300-base paired-end reads and a 12-base index read on a MiSeq (Illumina). Oligo 325 was used as the sequencing primer for both read 1 and read 2, and oligo 326 was used as the sequencing primer for read 3. 12-base barcode sequences were removed from read 1, and read 1 was trimmed to 270 bases using Prinseq (*172*). The resulting forward read and read 3 were merged and mapped to their consensus barcode (*67*,

152). To remove unbarcoded plasmids and truncations occurring during Gibson assembly, barcode sequences were aligned to pMR002, and all barcodes that mapped were removed from the dataset.

Barcode sequencing library preparation. Yeast samples (5 ml**)** were harvested and flash-frozen after growth in chemostats, and plasmids were extracted using the Zymoprep Yeast Miniprep II kit (Zymo Research). Sequencing libraries were created using two amplification steps. First, 4 μl of each plasmid miniprep was amplified in a 25 μl reaction by quantitative PCR with oligos 327 and 344 using KAPA HiFi HotStart ReadyMix (KAPA) on a BioRad MiniOpticon (BioRad). The following PCR conditions were used: 98ºC for 2 minutes; then cycles of 98ºC for 10 seconds, 65ºC for 15 seconds, and 72ºC for 25 seconds. Reactions were stopped after 10 to 13 cycles to avoid over-amplification. One μl of each reaction was then diluted into a 25 μl KAPA HiFi HotStart ReadyMix reaction with oligos P5 and NexV2ad2_N to add sequencing indices and amplify full-length products. As with the first reaction, quantitative PCR was used to avoid over-amplification, and each reaction was stopped after either 7 or 8 cycles. The same cycling conditions were used as in the first amplification reaction. These reactions were cleaned with Ampure XP beads (Agencourt) and sequenced with 25 base reads on a MiSeq, NextSeq, and HiSeq 2000 (Illumina), using oligo 325 as the read 1 sequencing primer.

Barcode sequencing analysis. Twelve-base barcode sequences were extracted from reads and analyzed using custom software implemented in Python. Barcode sequences that perfectly matched the consensus barcodes were counted, and counts were converted to frequencies within each round of selection. Barcode frequencies were converted to log-ratios between each round

and the input. The fitness of each barcode was calculated as the slope of the ordinary least squares regression of these ratios and the number of generations elapsed for each sample. Variant fitnesses were normalized to wild type fitness by subtracting the wildtype slope from the slope of each barcode. The fitness of a mutant is the average fitness of all barcodes that map to that mutant. To create a set of high-confidence variants, we compared the fitness and input read counts of the barcodes mapping to wild type promoters. We qualitatively set a minimum read count threshold at a point to maximize the number of wild type barcodes, yet minimize the variance of the wild type fitness scores and dependence of fitness scores on input count number. Using this heuristic, we set the read count threshold for sulfate limitation experiments at 50 input reads, and for glucose-limitation experiments at 15 input reads (Figure 2.1).

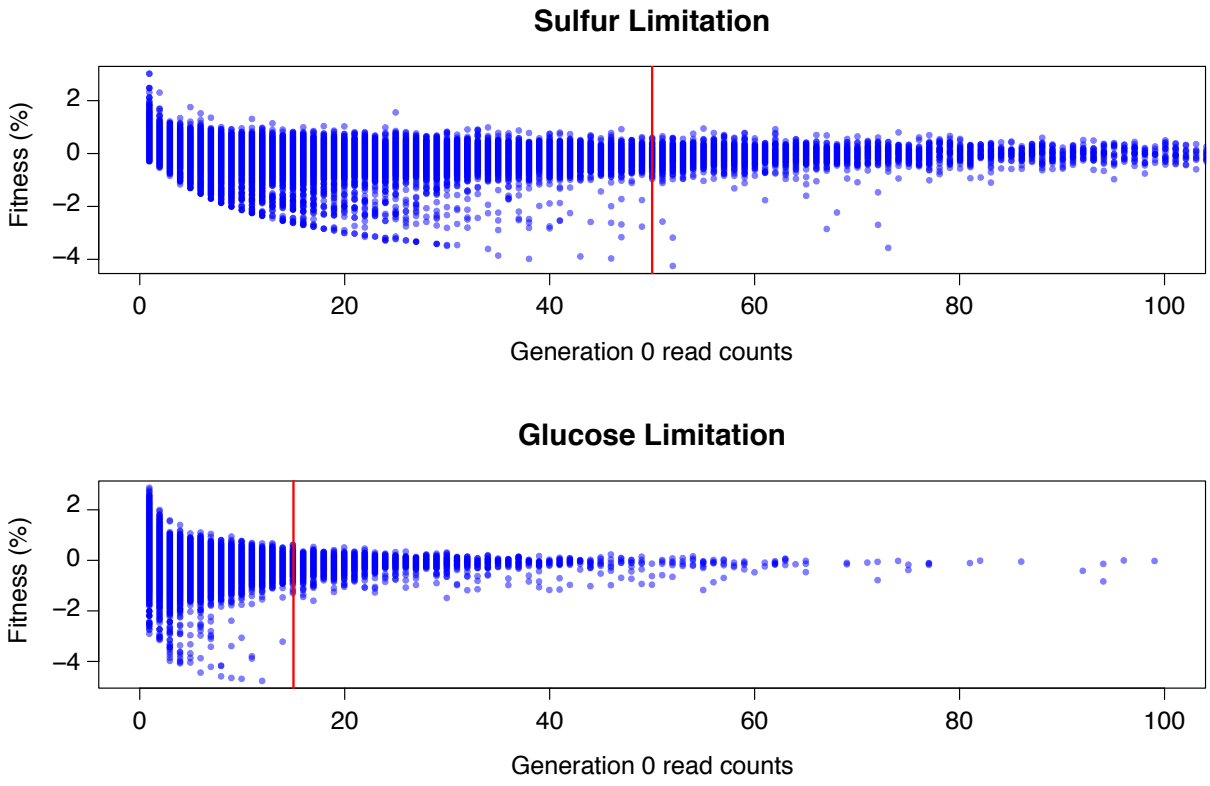

Figure 2.1. Determining minimum read count threshold for variant filtration. Barcodes linked to wildtype SUL1 promoters were used to set a minimum

generation 0 read count for variant filtration. The fitness of each wildtype barcode is plotted against that barcode's generation 0 read count. Filtration thresholds (red lines) were set at a point where the dependency between read count and barcode fitness was not apparent.

Creation and selection of a combinatorial library of high-fitness mutations. A library of promoters containing only combinations of 5 high-fitness mutations and neutral variation at one position was built using 5 sequential steps of PCR, each using the product of the previous reaction as a primer. Inadvertently, oligo 381 encoded either C or T, both neutral variants, at position -246. The primers and products used in each reaction were: (1) oligos 380 and 381; (2) the product from (1) and 237; (3) the product from (2) and 382; (4) oligos 237 and 382; (5) the product from (4) and M13F; (5) M13F and 237. All reactions were 25 µl KAPA HiFi HotStart ReadyMix reactions using the following PCR conditions: 98ºC for 2 minutes; then 25 cycles of 98ºC for 10 seconds, 65ºC for 15 seconds, and 72ºC for 15 seconds; then 72ºC for 5 minutes. Only fifteen cycles of PCR were used in reaction 5. The final product was gel extracted using a MinElute Gel Extraction kit (Qiagen), cloned into pMR002 using Gibson assembly, used to transform ElectroMAX DH10B electrocompetent *E. coli* (Life Technologies), and then used to transform YMD3017 using a high-efficiency yeast transformation (*170*). Selections and plasmid extractions were performed as before. Fragments for sequencing were amplified using 400 nM each of a custom forward index primer (oligos 395-402, one per sample) and oligo 328, using the following PCR conditions: 98ºC for 2 minutes; then cycles of 98ºC for 10 seconds, 65ºC for 15 seconds, and 72ºC for 25 seconds. Quantitative PCR was used to avoid over-amplification, and reactions were stopped after 8-16 cycles, depending on the sample.

Libraries were sequenced with overlapping 150-base reads on a Miseq (Illumina) using oligo 442 as a forward primer and oligo 387 as a reverse primer. These reads covered five of the six mutated sites in the library. We sequenced the final mutated site with an 8-base index read (using oligo 443), and used the second index read to demultiplex samples. Reads were genotyped at each mutated site, and analyzed similarly to our barcode sequencing data.

Generation and genomic integration of promoter variants. We first replaced the genomic *SUL1* promoter with *URA3*. *URA3* was amplified from FY3 genomic DNA with oligos 678-681, and this PCR product was used to transform YMD3107 via high-efficiency yeast transformation (*170*). We created single mutants of the *SUL1* promoter in two sequential PCR steps. We amplified the 3' end of the promoter with a primer containing each mutation (oligos 517-522, 603-605) and oligo 237, using pMR002 as a template. These products were gel-extracted with a FastGene PCR Extraction Kit (Nippon Genetics) and used as primers along with M13F to add the 5' end of the promoter to the fragment (using pMR002 as template). These reactions were gel extracted and cloned into pMR002 by Gibson assembly. The promoter-*SUL1* fragments of these plasmids were amplified with oligos 237 and 685 and used to transform YMR002, creating strains YMR008-YMR012 and YMR017-YMR020. These strains were made prototrophic (URA^+) by backcrossing to YMD3018, creating strains YMR022 and YMR024-YMR030.

Matching fitness data to known transcription factor motifs. Single mutant log-normalized frequencies at 0 and 40 generations were estimated using the slope and y-intercept of the linear fit used to calculate variant fitness, creating a log-likelihood of finding each variant in the dataset. These ratios were then ordered by position and identity (*i.e.*, A, C, T, or G). The

estimated wild type ratio at 40 generations was used for the wild type base at each position, and a value of 1.0 was used for missing data. Any extrapolated values less than 0.001 were set to 0.001. Specific position ranges were extracted from this matrix and compared to known transcription factor motifs (*173*, *174*) using Tomtom version 4.10.0 (*175*).

Finding newly created transcription factor binding sites. We enumerated all possible single mutations in the *SUL1* promoter in the context of a 25mer (each mutation was flanked by 12 upstream and downstream wild type bases), as well as their wild type alternatives. We searched this set of sequences for occurrences of motifs with FIMO (*176*), relaxing the significance threshold for reporting motif matches to P=0.01. This relaxed threshold allowed us to identify even weak matches to a motif that could be improved by point mutations. We compared the significance of each mutant match to the significance of that motif's match to the wild type 25-mer for all motif matches that overlapped the middle position in the sequence and calculated a log-normalized ratio of these two scores.

Scripts and raw data can be found at https://github.com/msr2009/Rich2016. Raw sequencing reads can be found in the NCBI Sequence Reads Archive, Bioproject ID PRJNA273419. A flowchart describing the experiments presented here can be found in Figure S7.

2.3 Results

2.3.1 *The extent of the SUL1 promoter.*

We first sought to define the *SUL1* regulatory region because little functional annotation is available apart from the identification of three putative TATA boxes at -199, -93, and -91 (*177*,

178). We created a centromeric plasmid-borne copy of *SUL1* that included the coding region, 687 bases upstream of the start codon and 270 bases downstream of the termination codon. This region (chrII:788548-792107) begins 118 bases downstream of the *VBA2* termination codon and stops 57 bases upstream of the first base of ARS228. Deletion of this genomic region causes a 19% decrease in sulfate-limited fitness (*179*). Expression of plasmid-borne *SUL1* in the deletion background increased fitness by 20.4%. To identify the functional extent of the *SUL1* promoter, we measured the fitness in sulfate-limited media of *sul1Δ* yeast strains transformed with plasmids containing one of a set of truncated promoters, each progressively deleting about 100 bases (Figure 2.2). Truncation to a 592 base promoter caused a small fitness decrease (-6.45%) compared to the full-length regulatory sequence, though truncation further to 493 bases restored fitness. All truncations that created a promoter shorter than 493 bases were unable to complement the deletion of the endogenous *SUL1*, leading to substantial fitness defects (-33% to -53%). We defined the 493 bases (chrII:788742-789235) as the minimal *SUL1* promoter and used this region for mutagenesis. Since the transcription start site of *SUL1* is also unannotated, mutations will be numbered relative to the *SUL1* start codon; for example, -493 is chrII:788742 and -1 is chrII:789235.

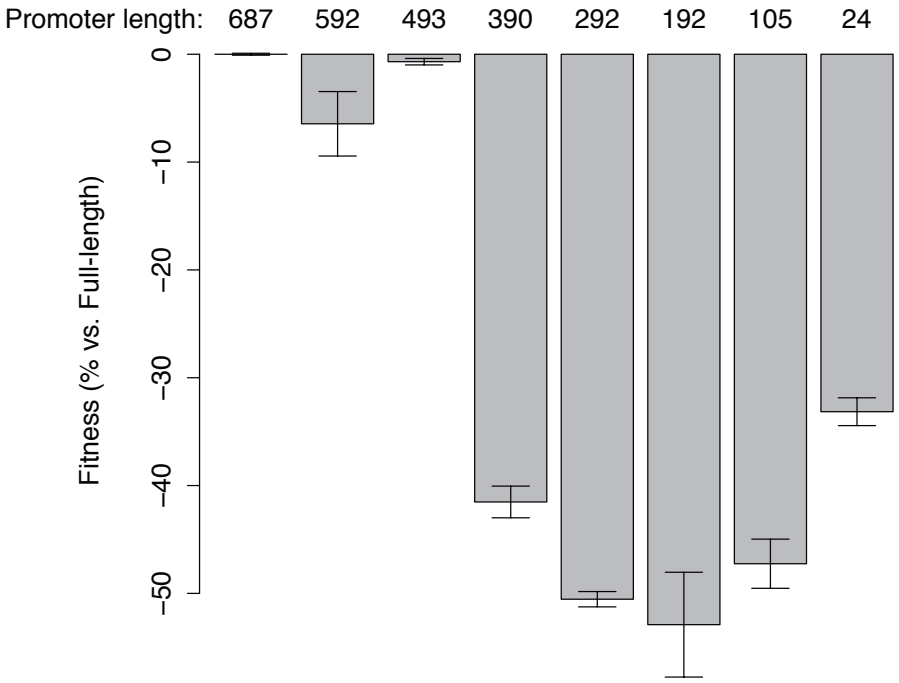

Figure 2.2. Determining the extent of the *SUL1* through truncation. The fitness of *sul1Δ* strains harboring plasmids with full-length and truncated *SUL1* promoters was assayed by competition in sulfate-limited chemostats against a wild type (SUL^+) strain. Fitness values are normalized to the full-length promoter (687 bases). Promoter lengths are measured from the *SUL1* start codon. Error bars represent standard deviation (n=2).

2.3.2 *SUL1 promoter mutagenesis and selection.*

We used error-prone PCR to create random mutations in the *SUL1* promoter and cloned these promoter variants scarlessly by Gibson assembly into the *SUL1* construct. After uniquely barcoding plasmids and linking the plasmid barcodes to the sequence of the promoter variants, we obtained sequences of 152,723 variants uniquely tagged by 630,517 barcodes. Each uniquely-barcoded variant had on average 2.2 mutations. The wild type promoter was tagged by 92,753 barcodes (15.1% of the total barcodes, Figure 2.3A). We transformed both SUL^+ and *sul1Δ* strains with this library. Transformants were competed as a pool during growth in sulfate-limited

chemostats, and the activity of each barcoded promoter, as approximated by the relative fitness of each strain in the pool, was calculated as described in the Methods. After stringent filtering for high-confidence variants (Methods), we calculated a wild type-normalized fitness value for 29,906 promoter variants. The pooled transformants as a whole had a median fitness decrease of -1.7%. Single mutant fitnesses were specific to sulfate-limitation, as fitness values after selection in glucose limitation, in which *SUL1* activity does not drive competitive fitness, were neutral and the distribution of all fitness scores in glucose limitation was not statistically different from the fitness score of the 94,912 wild type promoter sequences in sulfate-limited medium (P=0.38, t-test).

Both the distribution of wild type barcodes as well as the distribution of all variants were centered on zero and had a long negative tail (Figure 2.3B). Unlike the distribution for wild type barcodes, the distribution of variant fitness had a large shoulder corresponding to variants with fitness decreases between -15% and -5%, consistent with many positions in the promoter being sensitive to mutation.

We also examined the effect of mutations in a strain in which the endogenous copy of *SUL1* was not deleted. The fitness effects of variants were generally highly correlated between this SUL^+ strain and the *sul1Δ* strain (Spearman's rho=0.859, Figure 2.4). Fitness values were correlated between the two backgrounds for variants with wild type-normalized fitness greater than -15%. However, variants with wild type-normalized fitness less than -15% in *sul1Δ* were not as unfit in a SUL^+ background (Figure 2.4).

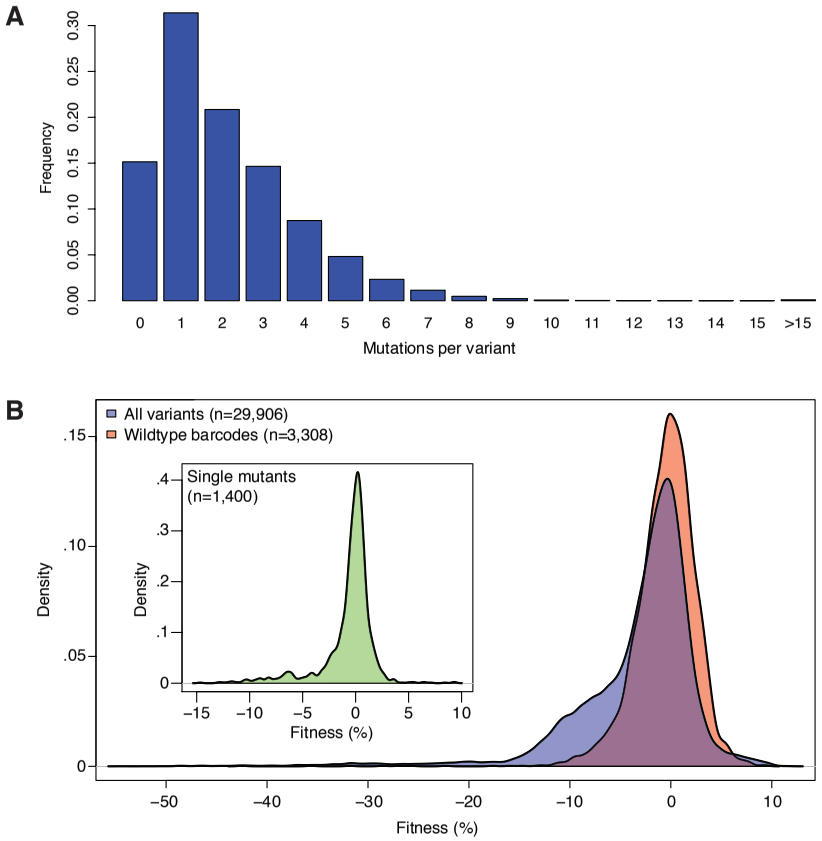

Figure 2.3. *SUL1* promoter variant distributions. (A) Histogram showing number of mutations per variant. (B) Density plot of variant fitness change distributions.

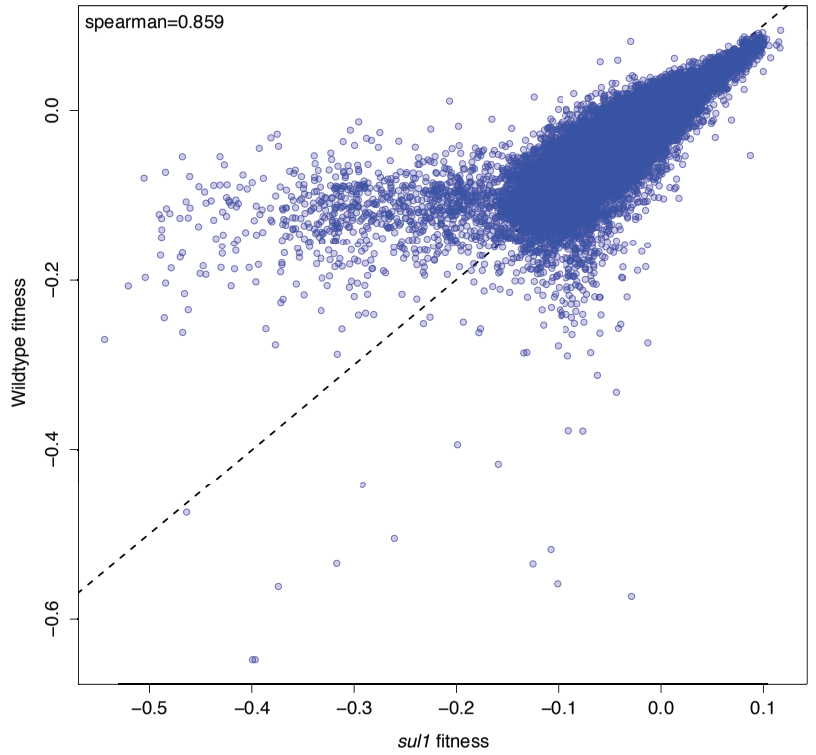

Figure 2.4. Correlation between fitness changes of variants in FY3 and FY3*sul1Δ*. Only variants passing filtration in both datasets (n=29,763) are shown. The dashed line shows y=x.

2.3.3 *The effect of point mutations in the SUL1 promoter.*

SUL1 amplification is present in all populations at generation 100 during laboratory experimental evolution in sulfate limitation (*171*, *180*, *181*). Accumulating more than one mutation in the 493 bases of the *SUL1* promoter during those 100 generations is unlikely, so we first limited the bulk of our analysis to the fitness effects from single mutations. We assayed 1400 of the 1479 (94.6%) possible single mutants of the *SUL1* promoter (Figure 2.5). Most mutations had little effect on fitness. Based on the fitness distribution of uniquely barcoded wild type promoters, we established an empirical false discovery rate of 5% to be fitness values less than a -7.1% decrease and greater than a 4.3% increase. Only eleven single mutations increased fitness more than 4.3%, whereas 50 single mutations decreased fitness by more than 7.1%. Overall, single mutations had a narrower distribution than the set of all variants, including a much shorter negative tail. The effect size of fitness-increasing mutations was similar between singly- and multiply-mutated variants; only 46 variants with multiple mutations had fitness increases greater than the maximum single-mutant fitness (9.4%) and these variants increased fitness up to 11.7%. All 46 of these variants contained at least one of -353T>G, -372T>C, -404C>T, and -458T>A, and 39 contained -372T>C, the second-most fit variant in the library. The effects of fitness-decreasing variants appeared to be additive, as 1054 multiply-mutated variants decreased fitness more than the most detrimental single mutation (-14.7%).

46 of the 50 single mutations that decreased fitness by more than 7.1% were found in four sites in the promoter (A-D in Figure 2.5). These sites appear as low-fitness “stripes” on the heatmap, with nearly every mutation that occurs in each site decreasing fitness. These sites are short, between 8 and 20 bases, and likely correspond to binding sites for the transcription factors that

regulate *SUL1* expression (see below). Truncation of the *SUL1* promoter to 393 bases, which removes two of these sites, resulted in the failure of the shortened promoter to complement a *SUL1* deletion, presumably because it provides insufficient expression.

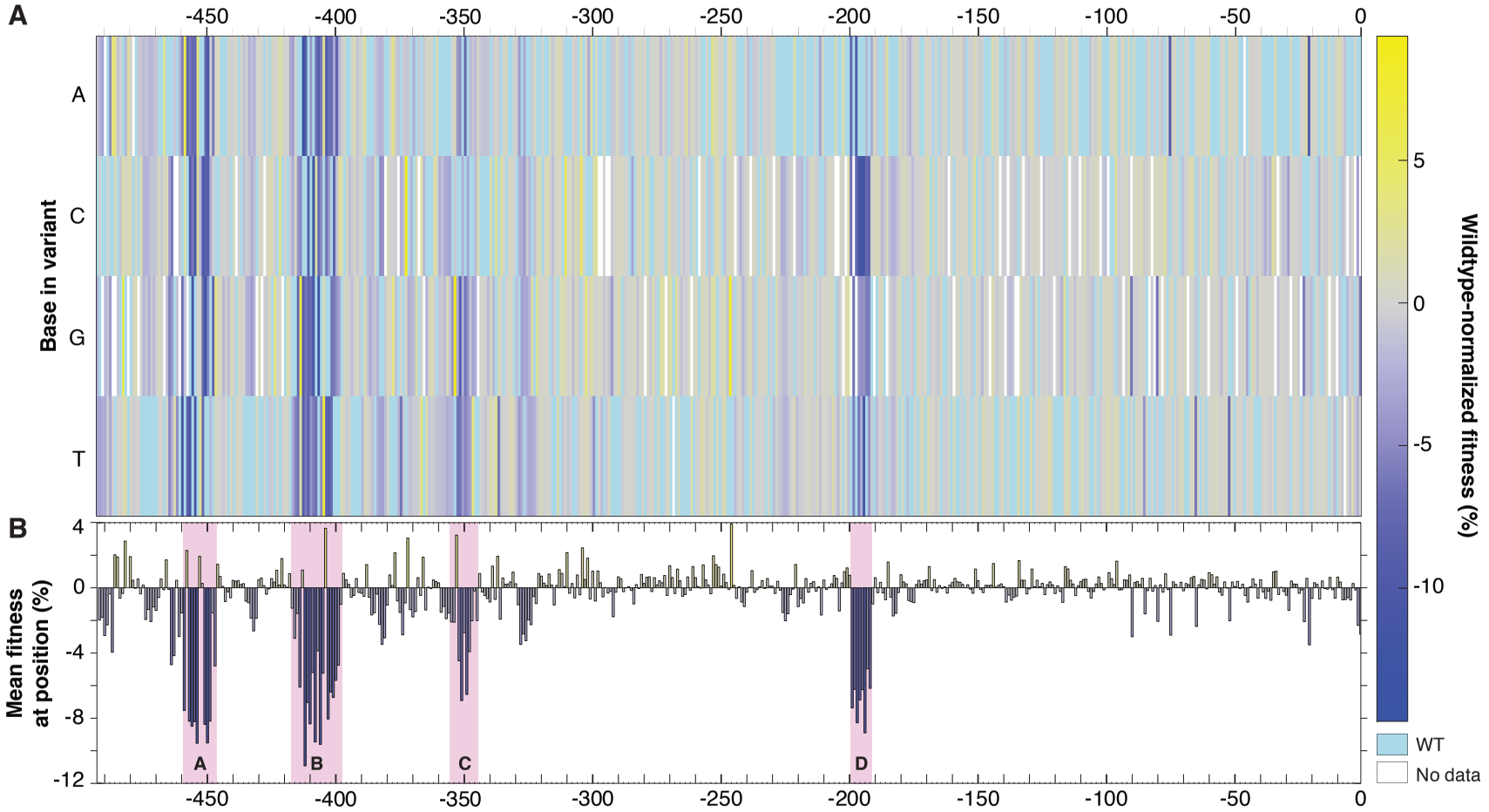

Figure 2.5. The effect of single mutations in the *SUL1* promoter on sulfate-limited fitness. (A) The fitness change of each single mutation in our dataset is plotted as a heatmap. At each position in the promoter along the x-axis, the base present in each single mutant is ordered on the y-axis, and every cell is shaded based on the fitness change of that mutation. Missing data are white cells in the heatmap and the wild type base at each position is teal. (B) Mean fitness change of all variants that are mutated at that position. Sites examined in Figure 3 are marked with pink background.

2.3.4 *Discovering SUL1 transcriptional regulatory sites.*

We used the single mutation data to identify the likely transcription factors that regulate *SUL1* expression. The wild type sequence of site D (-200 to -193) is TATAAATA, matching a canonical yeast TATA box (*177*). Using the slope and intercept data for each single mutant in the site, we created a matrix of the log-likelihoods of finding each single mutant in the site after 40 generations. We searched transcription factor motif datasets (*173*, *174*) to find significant matches to the log-likelihood matrix. This analysis identified Spt15, the yeast TATA-binding protein, as a match for this site (Figure 2.6, panel D), albeit one with a high false-discovery rate (q=0.27), likely driven by missing data in the first three positions of the site. Because we conservatively assigned missing values a log-likelihood of 1, *i.e.,* no change in frequency during the selection, we decreased the significance of our matches to this site.

We then applied this methodology to the other three sensitive sites (Figure 2.6, panels A-C). The log-likelihood matrix for site A, spanning positions -465 to -448, weakly matched the motif for Cbf1, a known regulator of sulfate metabolism that recognizes CACGTG (*182*), and Tye7, a glycolytic activator that is implicated in Ty1-mediated gene expression. A second site, from positions -452 to -450, flanking a hypothetical Cbf1 motif (CTCGTG), was also sensitive to mutation. These positions partially match the RYAAT motif, which is necessary for full induction of genes during growth in limited sulfate by enhancing binding of the Cbf1-Met28-Met4 regulatory complex (*183*).

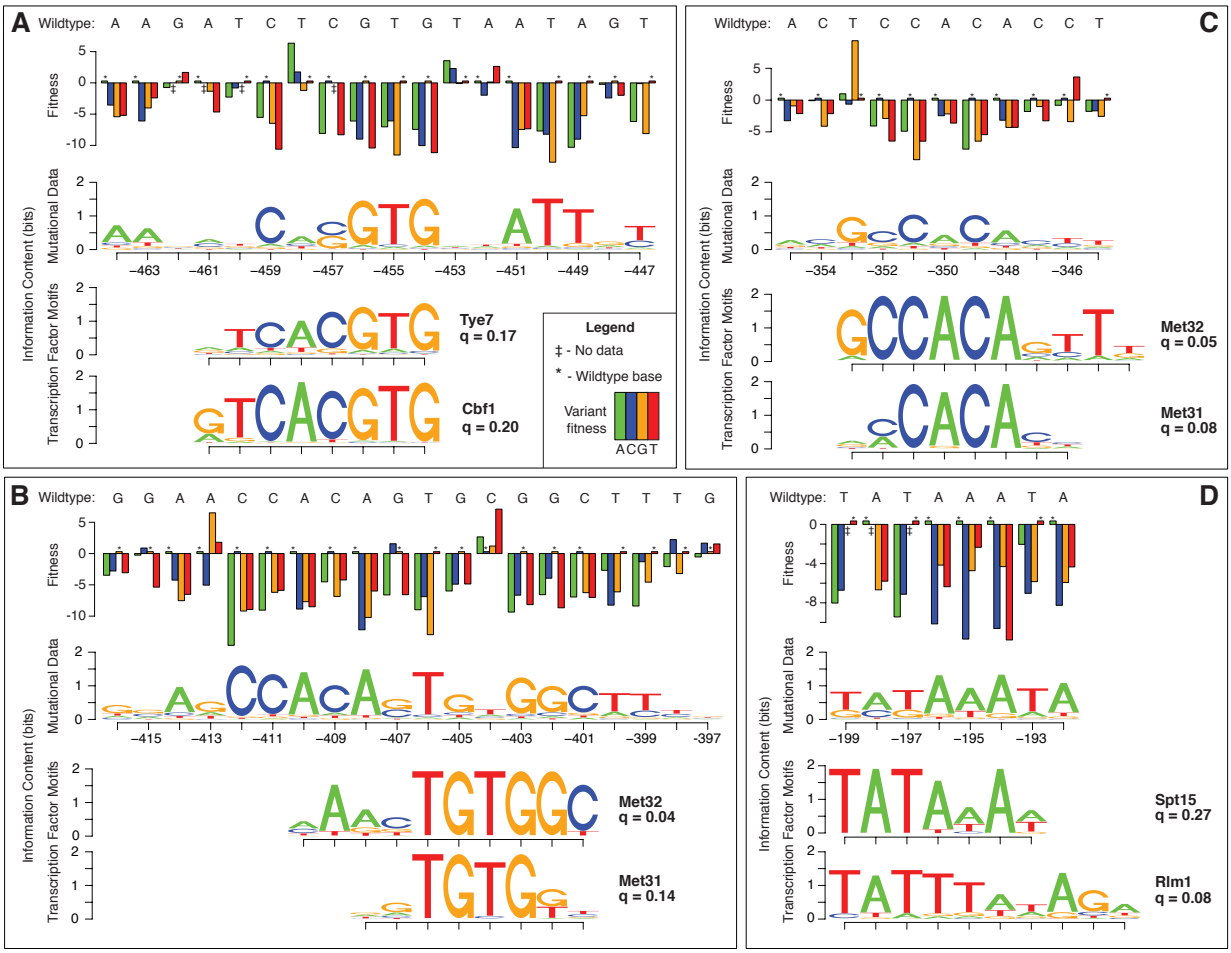

Figure 2.6. Determining *SUL1* transcriptional regulatory sites through mutagenesis. To identify the transcription factors binding at the four sites highlighted in Figure 2.5, we compared position weight matrices based on the fitness values for single mutations at each site with a database of known yeast transcription factor binding motifs. The wild type sequence for each region is shown at the top of each panel. Below this is a barplot showing the fitness value for the single mutations at each site. Below the barplot is a set of sequence logos, the first of which is based on the mutational data for the region (calculated as described in Methods). Underneath this logo are the significant ($q < 0.2$, except in D) matches between the fitness-based PWM and yeast transcription factor binding sites. Panel labels are consistent with the labels in Figure 2.5.

The log-likelihood matrices for both site B and site C contain the CCACA motif recognized by Met31 and Met32 (*184*). Site B matched Met32 (q=0.03) and Met31 (q=0.09), and appears to be

an incomplete palindrome (GCCACA[CG]TGTGGC) centered on position -407. Site C was the most significant match of the analysis and matched the Met32 motif (q=0.005).

Although sites A, B, and C were generally sensitive to most mutations, some mutations yielded large (>5%) increases in fitness. Comparison of the wild type sequence to the highest-fitness variant at these positions showed that each of these binding sites was one mutation away from the consensus binding site for the implicated transcription factor. For example, mutation -458T>A (*i.e.,* mutation at position -458 from T to A) creates the consensus binding site of Cbf1 (C<u>A</u>CGTG), and increased fitness by 6.3% in our competition and increased Cbf1 binding strength approximately 140-fold, as measured *in vitro* by the MITOMI assay (*185*). Mutations -413A>G and -404C>T had similar effects in site B, creating a consensus binding site for Met32 and increasing fitness 6.5% and 7.1%, respectively. Mutation -353T>G, which also creates a consensus Met32 binding site, caused the largest fitness increase in the dataset (9.4%). The importance of -413A>G and -356T>G distinguishes these sites as Met32 binding sites, rather than similar Met31 binding sites, as the importance of a 5' guanine is found only in the Met32 site.

Other short sites in the promoter (*e.g.,* -329 to -321) showed similar trends of 8-20 contiguous positions being sensitive to mutation, but with smaller fitness effects. When we performed a search for the log-likelihood of these sites against known transcription factor binding sites, we found no significant matches (data not shown).

2.3.5 *Creation of new regulator binding sites*

The methodology we used to find endogenous regulators depends on a signature of purifying selection throughout a binding site, and so is not generally applicable to finding mutations that create new transcription factor binding sites. Endogenous *SUL1* regulatory sites are highly sensitive to mutation across the entire site, except for rare mutations that optimize active sites (sites A, B, and C – Figure 2.5). This signature of purifying selection would not be found in inactive binding sites, and point mutations that activate these sites would be independent of the surrounding positions. As such, we took a different approach to find binding sites that occur or are lost upon single mutation.

To identify such sites, we first enumerated all 25-mers and all possible single mutations centered in 25-mers *in silico,* and then searched these 25-mers for matches to transcription factor binding sites using relaxed parameters to allow for weak matches. In total, 124,316 motifs matched at least one 25-mer at a threshold of $p<0.01$. An arbitrary p-value of 0.01 was used for motifs that matched either the mutant or wild type 25-mer, but not both. We calculated the log-ratio between p-values for the wild type and a mutant motif match, $\log_2(P\_wt/P\_mut)$, and used this ratio as a measure of motif strength.

Of the 588 motifs with a match significance of $p<0.0001$, 347 were either strengthened or weakened at least 10-fold by mutations (Figure 2.7). The majority (52/56) of mutations that decreased fitness by more than 5% and altered the significance of a motif match by at least 10-fold occurred in either the TATA box or sites A-C, or created a new upstream open reading frame in the 5' UTR. Seven mutations matching 11 different motifs increased fitness by more

than 5% and had at least a ten-fold increase in motif significance when compared to wild type. Three of these (-353T>G, -404C>T, and -458T>A) we identified above in sites C, B, and A, respectively, as mutations that optimize endogenous Met32 and Cbf1 binding sites. One of the remaining mutations, -372T>C, creates a Met32 binding site (GCCACA), increasing the significance of the motif match by 23-fold and fitness by 9.3%, the second-highest fitness increase from a single mutation in our dataset. Another, -246T>G, creates a one-off Cbf1 site (CATGTG), increasing the significance of the motif match by 32-fold and increasing fitness by 8.4% (Figure 2.7, panel B).

Other mutations that increased fitness more than 5% do not have clear-cut biological explanations. For example, -310A>T increased fitness by 5.5% and the significance of the match to Hap2, a glycolytic activator, by 20-fold. The -482A>G mutation might strengthen the binding of Aft1 or Aft2 (22.5- and 96-fold, respectively) or weaken the binding of Xbp1 by 43-fold. Aft1 and Aft2 regulate iron homeostasis, and Aft1 interacts physically with Cbf1 (*186*); the fitness increase caused by -482A>G may be dependent on Cbf1 binding to the neighboring binding site (site A). Alternatively, the fitness increase could also be caused by decreased Xbp1-mediated repression, as Xbp1 is known to repress other Met4-responsive genes (*164*, *187*).

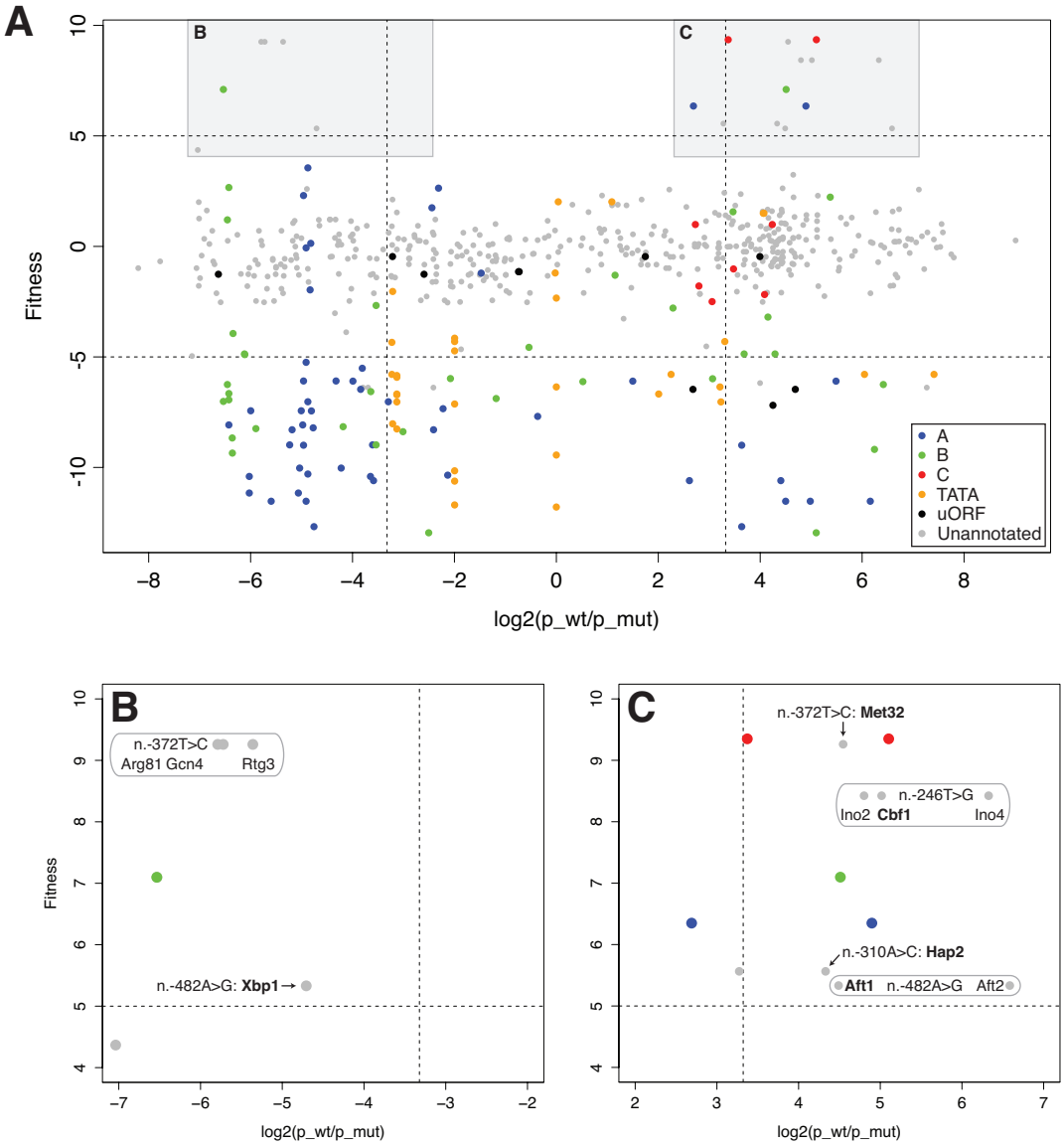

Figure 2.7. Identifying cryptic regulatory sites uncovered by mutations. (A) All motifs that strongly match either a wildtype or mutant 25-mer (P<0.001). Motifs are plotted by the change to match significance of each mutation (x-axis) and the fitness of each single mutant (y-axis). Points are colored based on their annotated function. (B) Inset showing motif matches that are weakened by mutations that confer fitness increases. Unannotated mutations that match multiple motifs are grouped the transcription factor motif matched is noted. Transcription factors that we hypothesize to be functionally important at that site are in bold.

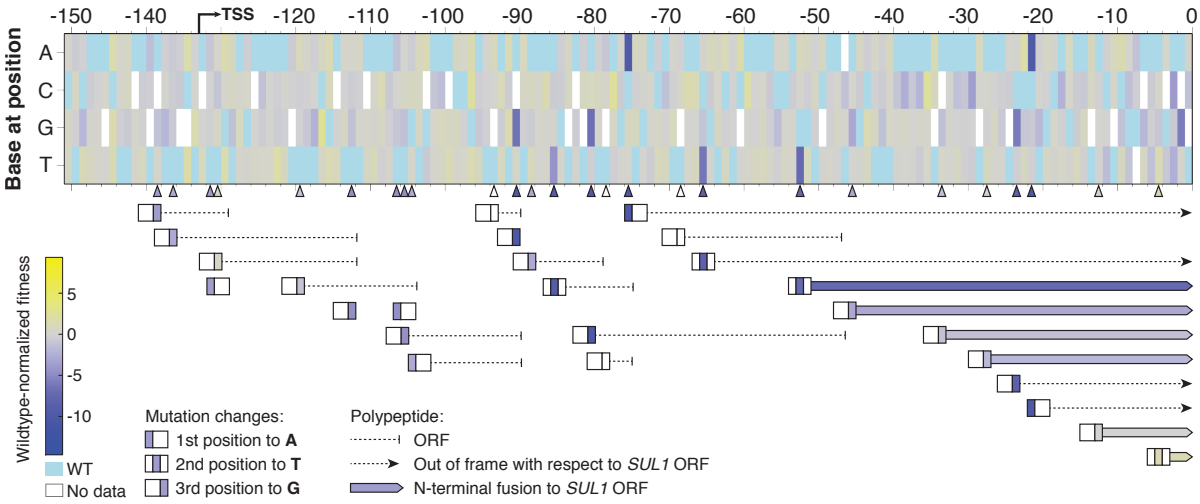

Figure 2.8. The effects of upstream open reading frames on Sul1 expression. We searched the 150 bases upstream of the *SUL1* start codon for sites that could mutate to form an upstream ATG. The heatmap of this region is shown, with triangles marking the positions of mutations to upstream start codons. Below each triangle is a cartoon representation of the upstream open reading frame created by each ATG. The position of the colored box in each cartoon represents which base was mutated to create an ATG, and each box is colored based on the fitness change of that mutation. Dotted lines extending to right of each box show the extent of the hypothetical polypeptide encoded by each uORF. In the case that the uORF is in-frame with the *SUL1* coding sequence, these polypeptides are drawn as colored boxes. The effect of upstream open reading frames on SUL1 fitness.

Short open reading frames starting upstream of a gene's coding sequence can post-transcriptionally regulate gene expression, as the ribosome creates unproductive polypeptides instead of the correct protein (*188*, *189*). No upstream open reading frames are present in the wild type *SUL1* 5' untranslated region. Because our assay selects on the amount of Sul1 protein in the cell, it should identify mutations that cause a fitness defect due to post-transcriptional regulation of *SUL1* expression. Therefore, we searched for mutations that create an upstream open reading frame (Figure 2.8), *i.e.,* the mutation of a 3-mer to an AUG start codon within the 5' untranslated region of the *SUL1* transcript (*190*). Our data contained 23 of the 26 possible

upstream ORFs, 8 of which decreased fitness by at least 5%; none of the upstream ORF mutations increased fitness. Four upstream ORFs created a polypeptide that was out of frame with the *SUL1* coding sequence and read past the *SUL1* start codon. These mutations were invariably deleterious (decreasing fitness by 6% or more). Six upstream ORFs created an in-frame fusion to Sul1, with lengths ranging from 2 to 18 amino acids. The longest fusion decreased fitness by 7.2%, and as the fusions shortened, their effect on fitness decreased, with fusions adding 5 or 2 amino acids being neutral (0.3% and 1.3% fitness increases, respectively).

2.3.6 *Combinatorial analysis of high-fitness mutations.*

Six mutations in our dataset yielded greater than a 6% increase in fitness: -353T>G, -372T>C, -246T>G, -404C>T, -413A>G, and -458T>A. Based on our analyses above, these mutations may increase the affinity of a transcriptional activator. Individually, these mutations conferred less than 10% fitness increases, much less than the increased fitness (>37%) of evolved strains with amplifications containing *SUL1* (*171*, *181*). No variants that passed filtration combined two or more of these mutations. To investigate the combinatorics of these mutations, we created another library in which five of these sites (excluding -246T>G) were present as either the reference base or the high-fitness mutation. Our library contained all 32 possible mutation combinations. As before, we transformed a *sul1Δ* strain with this library, competed the population in sulfate-limited chemostats, and calculated variant fitness.

Fitness effects from single mutations were additive when combined in pairs, though double mutants increased fitness only 10.5%, a fitness increase of 1.15% above the most-fit single mutant (). The addition of more mutations to each variant did not increase fitness significantly

above the fitness of the double mutants, with 11% the maximum increase in fitness reached by any mutation combination.

Either -353T>G or -372T>C were necessary to reach the fitness plateau of about a 10% increase. Four combinations without these mutations had a wild type-normalized fitness increase of 8.4%±0.2% on average. All combinations including -353T>G or -372T>C (N=26) had a wild type-normalized fitness increase of 10.4±0.3%.

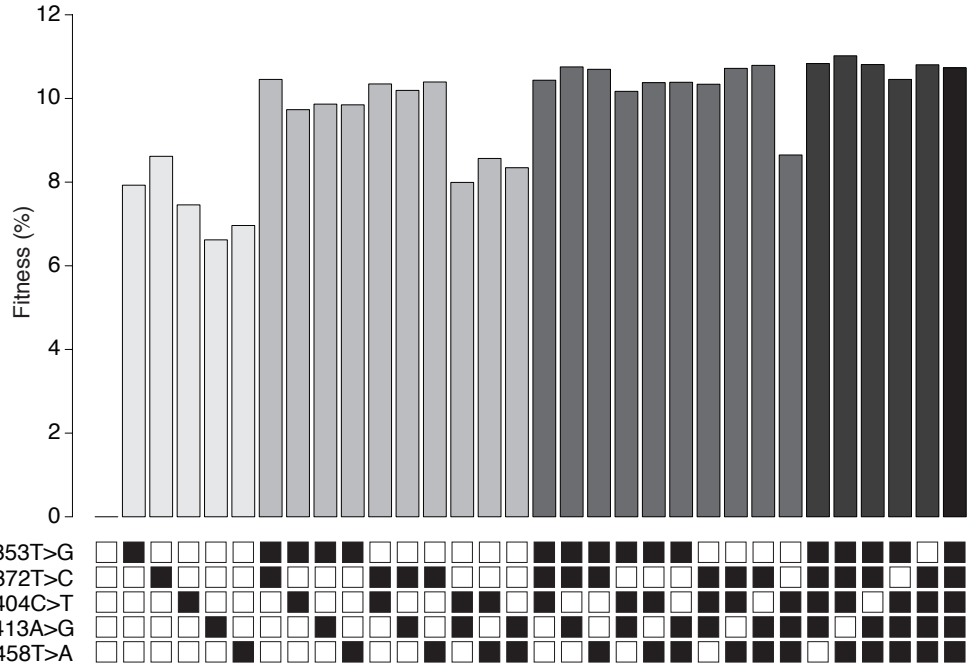

Figure 2.9. Combinatorial effects of high-fitness mutations. The wild type-normalized fitness for each of 32 combinations of five high-fitness mutations is shown as a bar. The genotype of each variant is identified on the X-axis: a black box in each row indicates the presence of each mutation in that variant.

We tested a set of eight single mutants and the combinatorial variant in their native genomic context. Integration of the 493 base promoter which deletes ~200 bases of the intergenic

sequence *SUL1* and *VBA2*, led to strain fitness approximately that of a *sul1Δ* strain (data not shown). We therefore integrated eight mutations and the combinatorial allele in the context of the entire 687 base promoter. The integrations resulted in a scarless replacement of the *SUL1* promoter. The fitnesses of integrated single mutants were highly correlated with pooled measurements of fitness (Pearson coefficient=0.872, p=0.002, Figure 2.10A). All variants except -246T>G were contained within a 95% confidence interval based on the pooled fitness values of the barcodes mapping to each variant (Figure 2.10B). We could not calculate a confidence interval for the 5-mutation variant because it was not found in the barcoded library and the combinatorial library did not employ barcodes.

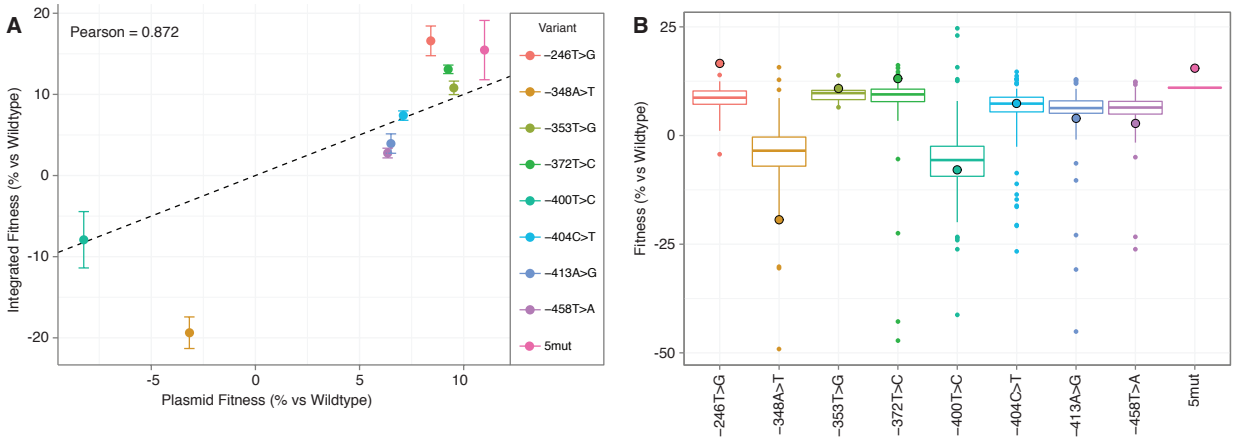

Figure 2.10. Correlation of fitness measurements between pooled plasmid assays and genomic integrations. (A) Scatterplot showing fitness measurements for strains based either on a pooled assay in which promoter variants are found on centromeric plasmids (x-axis) or individual measurements in which promoter variants were integrated scarlessly at the *SUL1* promoter (y-axis). Error bars represent standard deviations (n=2). The dashed line shows y=x. (B) Boxplots showing the distribution of fitness values amongst redundantly barcoded variants tested in (A). Whiskers for boxplots span 95% of the total distribution. Larger outlined points mark the fitness change of strains where each variant has been integrated at the *SUL1* promoter. In both panels, fitness measurements are normalized to the wildtype fitness. The variant "5mut" comprises five mutations: -353T>G, -372T>C, -404C>T, -413A>G, and -458T>A.

2.4 Discussion

Amplification has been proposed to be a "quick and dirty" solution for adaptation, allowing more stable mutations to be selected over a longer period of time (*191*, *192*); indeed, amplification of stress-specific genes may be detrimental during non-stress growth. The fitness increase due to *SUL1* amplification (up to 51% (*181*)) is much higher than that due to any of the single mutations, whose largest increase in fitness was 9.35%. Single mutations were not additive when

combined, and the fitness increase of any combination of high-fitness mutations plateaued at approximately 11%. This lack of additivity may imply that there is a limit to the maximal rate of transcription achievable from the *SUL1* promoter, limiting the effect of *cis*-regulatory mutations. Thus, for the *SUL1* promoter, amplification and *cis*-regulatory mutations are not equivalent, inconsistent with the "quick and dirty" hypothesis of amplification. Two coding mutations in *SUL1* have been isolated that increase sulfate-limited fitness by 23%, though it is not known how they affect protein function. These mutants were isolated in a strain in which gene amplification was prevented by deletion of the recombinase gene *RAD51* (*7*). It remains to be tested whether beneficial mutations in the 3' UTR can surpass gene amplification.

In this study, we did not investigate the potential for *trans*-acting mutations to increase *SUL1* expression, in part because the transcription factors governing *SUL1* expression were largely unknown. Our data show that there are three sites in the *SUL1* promoter that match binding sites for sulfate-regulatory transcription factors, two for Met32 and one for Cbf1. Both these transcription factors provide DNA-binding specificity for Met4, which recruits other transcriptional machinery. Met4 is a strong transcriptional activator (*193*), such that little Met4 occupancy in the promoter, perhaps only a single binding event, may be sufficient to near-maximally activate *SUL1*. Mutations that create new binding sites for Met32 or Cbf1 presumably add an additional site for Met4 occupancy and conferred fitness increases up to 9.3%, which is a much smaller effect than that of amplification. Amplification would increase the total copies of *SUL1* available to be transcribed, leading to more protein than from a single copy.

We determined that the Met32 and Cbf1 binding sites are the primary regulators of *SUL1* expression, and that these sites are not the consensus binding site for either factor. Consistent with our analysis, *met32* strains have decreased fitness under sulfate limitation, whereas *met31* strains do not (*7*). Our data also strengthen the observation that Met31 and Met32 are not fully redundant (*194*, *195*). Because each binding site in the *SUL1* promoter is at least one mismatch away from its consensus site, their annotation by sequence-based motif-finding methods is difficult. Studies showing combinatorial control by Cbf1 and Met31/32 of Met4-dependent genes failed to identify by standard motif-finding approaches the Cbf1 binding site we identified in the *SUL1* promoter (*164*, *196*), though combinatorial regulation of *SUL1* can be parsed by analyzing transcription after inducing or repressing sulfate-metabolism regulators (*165*, *197*). We searched the *SUL1* promoter sequence using two yeast transcription factor databases, YEASTRACT (*198-201*) and YeTFaSCo (*202*), to compare our results to the results from *in silico* analyses. Our search yielded 99 possible motifs, only one of which, a Met31 motif centered at -410 (overlapping a site we hypothesize to be a Met32 site), aligned with any of our predictions.

Why are none of the transcription factor binding sites in the *SUL1* promoter consensus sequences? A 5% increase in fitness during sulfate limitation would be a strong evolutionary force, and these mutations should therefore become fixed in a population. While the effects of these mutations were specific to sulfate-limitation, they do not appear to be detrimental in permissive, sulfate-rich conditions, and thus they would not be under purifying selection during non-stressful growth. However, sulfate limitation, at least to the extent of the experimental design here, may not be a selective pressure frequently experienced by yeast in the wild. Alternatively, balancing selection may be working against maximal activation of *SUL1;* as

another activity, Sul1 also transports toxic heavy metals (*203*). *SUL1* also appears to be dispensable in rich media conditions, as lager strains carry loss of function mutations in the gene (*204*).. Our assay measures fitness in steady state, so it is possible that the wild type promoter may have favorable kinetic qualities for fast adaptation to sulfate limited environments; any temporally-dependent aspects of *SUL1* regulation would go uninvestigated by our methodology.

The use of plasmids may confound some of our results. A 493 base *SUL1* promoter on a centromeric plasmid complemented the genomic deletion of *SUL1*, but not when it was genomically integrated. Centromeric plasmids have a low copy number, but not necessarily a copy number of 1 (*205*). An increase in plasmid and *SUL1* copy number may have masked some of the effects of truncation and mutagenesis of the *SUL1* promoter, with multiple copies of low-fitness promoters complementing a *SUL1* deletion. We saw this effect when measuring the fitness of variants in the context of a SUL^{+} strain, where the endogenous copy of *SUL1* masked variants that were detrimental in a *sul1Δ* background, and this could be the cause of the discrepancy in the plasmid and integrated fitness measurements for variants like -348A>T. Promoter strength has been shown to affect plasmid copy number (*205*). Our assay essentially counts the number of plasmids in a population, so biasing plasmid copy number dependent on promoter strength would also bias our results.

Alternatively, background mutations in the strain may alter fitness independent of the *SUL1* promoter genotype. Transformation is known to be mutagenic (*206*), and mutations can accumulate during strain construction (*207*). We advise that future studies carefully control for plasmid copy number. Copy number is not an issue with integrations, but minor fitness effects of

variants may be difficult to discern in integrants above the effects of extraneous mutations in the strain background. This problem may be circumvented by obtaining large numbers of barcoded integrants.

By using chemostats, we were able to perform sensitive analyses of the fitness effects of *cis*-regulatory mutations, an analysis not accessible to other methodologies that assay only transcriptional output. This approach should be applicable to studying transcriptional regulation more generally, as the *SUL1* promoter can be replaced with other *S. cerevisiae* promoters, allowing for the same mutational analysis as long as these other promoters are functional during growth under limited sulfate. Finally, because this approach is sensitive to post-transcriptional effects on protein expression, it could be adapted as a platform for studying post-transcriptional regulation in yeast, allowing, for example, assays of the effects of 5' untranslated regions on protein expression.

Chapter 3. A massively parallel fluorescence assay to detect the effects of synonymous mutations on *TP53* expression

Although synonymous mutations can affect gene expression, they have generally not been considered in the analysis of mutations that increase the risk of cancer. However, mounting evidence implicates some synonymous mutations as driver mutations in cancer. We used a massively parallel assay based on cell sorting of a reporter containing a segment of p53 fused to a fluorescent protein to measure the effects of nearly all synonymous mutations in exon 6 of *TP53*. Several mutations within the exon caused strong expression defects. We show that these effects can be largely attributed to errors in splicing, including exon skipping, intron inclusion, and exon truncation, resulting from mutations both at exon-intron junctions and within the body of the exon. These mutations are found at extremely low frequencies in healthy populations and are enriched a few-fold in cancer genomes, suggesting that some of them may be driver mutations in *TP53*. This assay provides a general framework to identify previously unknown detrimental synonymous mutations in cancer genes.

Contributions Geetha Bhagavatula and S.F. developed the idea. G.B. built mutagenesis libraries, performed cell culture experiments, FACS sorting, and library sequencing. G.B. and David Young performed initial analysis to tabulate variants. M.S.R. performed all other analysis. M.S.R. and S.F. wrote this manuscript. This work has been accepted for publication at *Molecular Cancer Research.*

3.1 Introduction

A central pursuit of cancer genomics is to identify somatic mutations that serve as "driver" mutations, which confer a growth advantage to cells. As high-throughput sequencing technologies now enable comprehensive identification of the mutational spectrum of tumors, many recurrently-mutated driver genes and mutations have been identified (*208*). While significant headway has been made in identifying such mutations, the focus of these studies typically has been missense, nonsense and frameshift alleles. However, synonymous mutations, which do not change the protein sequence of a gene product, were enriched in some oncogenes and the tumor suppressor *TP53* (*209*), based on an analysis of data from The Cancer Genome Atlas. This result implies that many synonymous mutations could, in fact, be drivers of oncogenesis.

Synonymous mutations can affect gene expression and protein function in multiple ways (*210*). Synonymous mutations have widely-studied effects on splicing (*211*). Those that fall in splice junctions can often abrogate splicing at that site or uncover cryptic splice sites. Similarly, synonymous mutations can fall in splicing enhancers or suppressors – exonic sequences that modulate the inclusion of an exon. The rate of protein translation can be modulated by many properties of the mRNA, including secondary structure and codon bias due to limitations in the tRNA pool. These factors alter ribosomal progression through the mRNA, which can affect the efficiency of protein folding and modification, leading to a decrease in protein expression. Synonymous mutations in miRNA binding sites can affect protein expression; for example in Crohn's disease (*212*) and melanoma (*213*), miRNA targeting was abrogated in the mutant copy,

leading to increased protein expression. Disease-associated synonymous mutations are also enriched in sites of RNA-protein interaction (*214*).

Considering the range of their possible effects, methods to readily annotate synonymous mutations would be useful. One avenue to identify possible drivers is via massively parallel assays to determine the effects of mutations (*215*). The resulting functional information yields greater predictive power than state-of-the-art *in silico* predictions of mutation function, like CADD (*216*) or PolyPhen (*217*). However, massively parallel assays also have largely ignored synonymous variants, often using these mutations as neutral variation to evaluate the significance of the changes of functional scores caused by non-synonymous mutations (*61*).

Here, we demonstrate a general method to detect the effect of synonymous mutations in a cancer-associated gene. We chose *TP53*, the gene most frequently mutated in cancer, as a test case, and focused on exon 6 of this gene because synonymous mutations at the terminal 3' splice site of this exon are enriched in tumor samples (*209*). Although no splicing enhancer or silencer has been functionally identified in this exon, a previous study found that synonymous mutations in *TP53* are enriched in predicted splicing enhancers (*218*). Using a massively parallel fluorescence-based deep mutational scanning assay, we measured the protein level of variants containing nearly all synonymous mutations in this exon. In addition to splice site-disrupting mutations, distinct internal mutations caused defects in splicing. These mutations are rare in healthy populations, and enriched a few-fold in cancer patients. Our methodology provides a general workflow for measuring the splicing effects of mutations, and could be used widely to uncover and catalog possible driver mutations in other cancer genes.

3.2 Materials and Methods

Primers and other oligonucleotides. A list of oligonucleotides used in this study is found in Table 7.4.

Creation of a mini-gene variant library of *TP53* exon 6. We first created a reporter construct that encodes the p53 DNA-binding domain (residues 97-292) fused to EGFP. A cassette containing *TP53* exons 5-7 (chr17:7675238-7674115) was amplified from NA12892 genomic DNA (Coriell Biorepository) with TP53_Exon5,6,7_F01 and TP53_Exon5,6,7_R01 primers. This product was cleaned and reamplified using TP53_expand_F1 and TP53_expand_R1 to add flanking sequence corresponding to the 31 N-terminal amino acids (residues 97 to 127) and 31 C-terminal amino acids (residues 262 to 292). EGFP was amplified from pAG414GAL1-EGFP (*219*) with primers GFP_Cloning_F01 and GFP_Cloning_R03. These products were cloned by Gibson assembly (*220*) into the *Afl*II and *Hind*III sites in pcDNA™5/FRT and transformed into chemically-competent DH5α *E. coli* (NEB). *BamH*I and *Nsi*I sites flanking exon 6 were added to this plasmid by inverse PCR with TP53pcDNA5_BamHI_F02 and TP53pcDNA4_NsiI_R01, followed by phosphorylation and ligation of the PCR product. This plasmid was named TP53E-GFP-pcDNA5.

We replaced *TP53* exon 6 in our reporter with a library of synonymously-mutated variants of the exon. The library was created by doped oligonucleotide synthesis (Trilink Biotechnologies, San Diego). Each position that could be mutated synonymously was doped at 3% to each possible

synonymous variant. For example, position 11 in the exon (the third base of a proline codon, CCT) was doped with 1% each of A, C, and G, whereas position 17 in the exon (the third base of a glutamic acid codon, CAG) was doped with 3% A. This oligonucleotide was made double-stranded by annealing the primer *TP53*_libraryamp_R01 and extending with Phusion polymerase (NEB), and the double-stranded DNA was amplified with *TP53*dopedoligo_RE_F03 and *TP53*dopedoligo_RE_R02. This product was cleaned with a Zymo Clean and Concentrator 5 kit (Zymo Research). Both *TP53*E-GFP-pcDNA5 and the doped library were digested with *Nsi*I and *BamH*I, and ligated together using T4 DNA ligase (NEB). This ligation reaction was cleaned and electroporated into ElectroMAX DH10B *E. coli* (Invitrogen), and cultured overnight in 50ml of LB + 100μg/ml ampicillin. This transformation resulted in a library containing approximately 242,000 colony forming units. DNA used for transfection was harvested using a PureLink HiPure Plasmid Midiprep Kit (Invitrogen).

Specific clonal variants, used for assay validation or splicing analysis, were created by site-directed mutagenesis using inverse PCR with a mutation-containing primer, followed by phosphorylation with T4 polynucleotide kinase and T4 DNA ligase (NEB). See Table 7.4 for primer sequences.

Flp-mediated integration of variant libraries. The reporter assay that we used requires that only a single variant is integrated into each cell. Thus, we took advantage of the Flp-In™-293 cell line, a derivative of the human embryonic kidney cell line (HEK-293T, Invitrogen), which has a single Flp-FRT site integrated in the genome. We integrated the library of *TP53* exon 6 variants into the genome of the cell line by co-transfection with 22 μg of pOG44 (Invitrogen) and

75 μg of the *TP53* library plasmid pool using Lipofectamine 3000 (Thermo Fisher). Media was changed after 24 hours, and selection for flipped-in cells was started after 48 hours by adding 100 μg/ml hygromycin. Cells were cultured under hygromycin selection for 14 days, at which point they were harvested. The 293 cell line library contained approximately 12,000 colony-forming units.

Fluorescence-activated cell sorting. The amount of reporter protein for each synonymous variant was measured by fluorescence-activated cell sorting. Library cells were washed, trypsinized, and resuspended in PBS + 1 μg/ml DAPI. Libraries were sorted on a FACSAria (BD Biosciences) into four bins, and cells were collected in PBS. Gates were set by comparing the distributions of fluorescence from a cell line expressing either a negative control (G113A) or the wild type exon. The lowest gate was set to bisect the negative distribution. The highest gate was set to bisect the wild type distribution. A final gate was set to bisect the range between the other two gates.

Sequencing library preparation and analysis. Variant frequencies were calculated by high-throughput sequencing of the integrated library in each sorted bin. Genomic DNA and total RNA were extracted from sorted cells using Trizol. Library variants were amplified from genomic DNA using three sequential reactions. Phusion polymerase was used for all reactions, and each reaction were performed on a BioRad MiniOpticon and monitored to avoid over-amplification.

The integrated library was amplified from 250 ng of genomic DNA (to avoid the endogenous *TP53* locus), using primers specific for library locus (pcDNA5/FRT_seq_F01 and

pcDNA5/FRT_seq_R01) and the following PCR conditions: 98°C for 30 seconds, then 25 cycles of 98°C for 10 seconds, 62°C for 30 seconds, 72°C for 65 seconds. This reaction was cleaned with a Zymo Clean and Concentrator 5 kit. 100 ng of this product was used as template for a second PCR to amplify a short fragment containing only exon 6, using oligos TP53_exon6amp_F01 and TP53_exon6amp_R01 and the following PCR conditions: 98°C for 30 seconds, then 8-10 cycles (depending on the sample) of 98°C for 10 seconds, 65°C for 30 seconds, 72°C for 5 seconds. This reaction was cleaned as before. Sequencing and demultiplexing adaptors were appended by a final PCR. 150 ng of the previous PCR product was amplified with *TP53*_seqadapter_F1 and *TP53*_seqadapter2_R1-4 and the following PCR conditions: 98°C for 30 seconds, then 8-10 cycles (depending on the sample) of 98°C for 10 seconds, 67°C for 30 seconds, 72°C for 5 seconds. Reactions were cleaned and sequenced on an Illumina MiSeq using paired-end 160 base reads, using *TP53*_libraryseq_F01 as a forward primer, *TP53*_libraryseq_R02 as a reverse primer, and *TP53*_indexseq_F01 as an index read primer.

The effect of each variant on fluorescence was calculated as a bin-weighted sum of inferred cell counts for each variant. Variant counts in each bin were tabulated using Enrich (*221*), and each variant's frequency in a bin was used to estimate the number of cells sorted in that bin for that variant. The frequency of total cell counts in each bin was multiplied by that bin's median GFP intensity. The weighted frequencies were summed to calculate a GFP score, and this score was divided by the GFP score of the wild type sequence to create a normalized GFP score. The normalized GFP scores for each of two replicates were averaged to compute the score reported here. Non-synonymous variants (likely due to synthesis or sequencing errors) were removed

after GFP score calculations. We filtered variants with fewer than 7,500 total cell counts in each replicate, as this threshold allows for the most variants with a high correlation between replicate GFP scores (data not shown).

Downstream analyses were performed in Python. Raw sequencing reads can be found in SRA Bioproject PRJNA384242.

Splicing analysis by PCR. RNA was extracted from clonal cell lines, reverse transcribed, and amplified. These products were separated by polyacrylamide gel electrophoresis. Fragments of interest were gel-extracted, re-amplified with KAPA Robust2G (KAPA Biosciences) using primers TP53_exon6amp_F01 and TP53_exon6amp_R01, cloned into pGEM-T (Promega), and Sanger sequenced to identify isoforms.

3.3 Results

3.3.1 *Developing an assay to measure the effects of synonymous mutations in* TP53

To test the effects of synonymous variants in exon 6 of *TP53*, we developed a fluorescent reporter assay based on FlowSeq (*92*), which uses FACS sorting to separate variants of differing expression levels followed by next generation DNA sequencing to calculate a score for each variant based on its enrichment or depletion in each FACS bin. The reporter consisted of a cassette containing the genomic sequence spanning exons 5-7 of *TP53* (encoding residues 128 to 261), as well as flanking cDNA encoding the rest of the p53 DNA-binding domain, such that in total the translated protein product of the cassette spanned residues 97-292 of p53 (Figure 3.1A).

The C-terminus of this cassette was fused to GFP, and the construct was expressed from a constitutive *CMV* promoter. The pre-mRNA from this construct can undergo only two correct splicing events, fusing exon 5 to 6, and exon 6 to 7 (Figure 1A). We created a library of exon 6 synonymous variants that fall in the third position of a codon by doped oligonucleotide synthesis (Figure 3.1B), and cloned these variants in the place of the wild type exon (*i.e.,* chr17:7578289–7578177) in the expression construct. The library was integrated into the genome of Flp-In™-293 cells to ensure that a single exon variant was expressed per cell.

Median fluorescence in this assay spanned an approximately 20-fold range, from a low due to a known splice site-disrupting mutation (G113A; Figure 3.1C, red) to a high due to wild type (Figure 3.1C, blue). The population of cells in the Flp-In library had a large peak overlapping the G113A peak, and a large shoulder of cells with intermediate fluorescence (Figure 3.1C, green). We sorted these cells into four bins that were established based on the fluorescence properties of the G113A mutant and the wild type, sequenced the exon 6 integrated in the cells of each bin, and calculated a score for each variant as a wild type-normalized weighted sum of GFP intensity across the bins (S_{GFP}, see *Materials and Methods*). In total, we calculated S_{GFP} for 16,916 synonymous variants of the exon (including multiply-mutated variants), which averaged 3.2 mutations per variant.

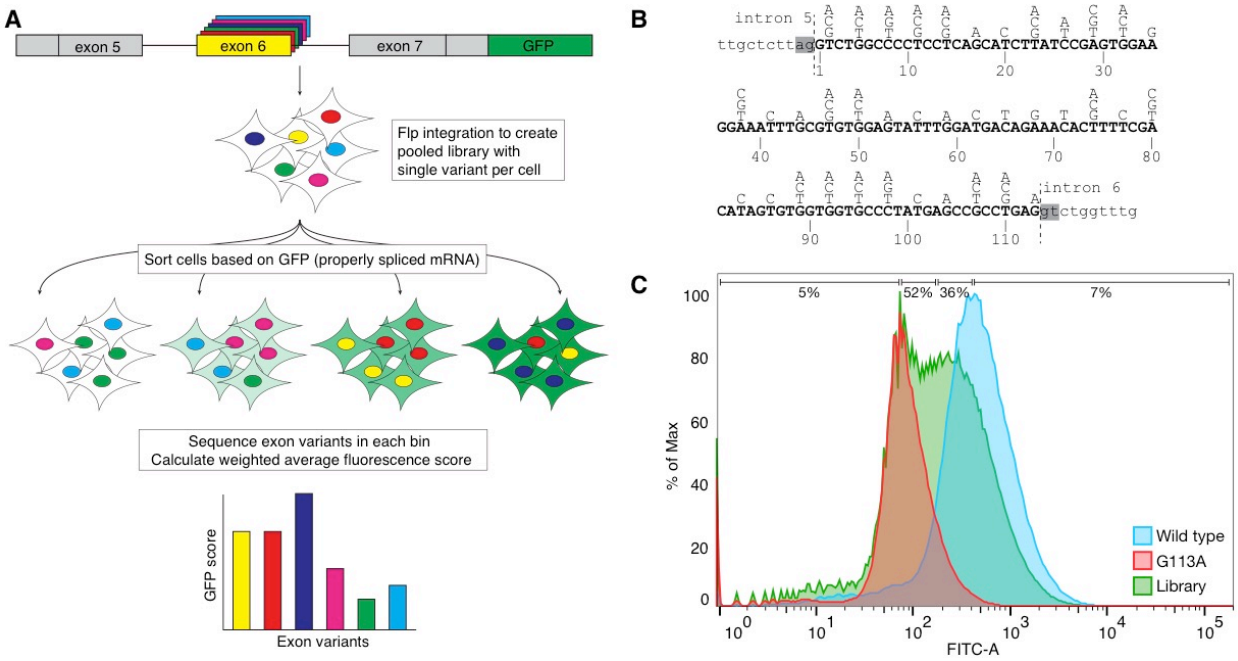

Figure 3.1. A massively parallel assay to measure the splicing effects of synonymous mutation in *TP53*. (A) Overview of the assay. A library of synonymous variants is created in a *TP53* reporter construct. This library is integrated into the genome of Flp-In^{TM}-293 cells, a derivative of the human embryonic kidney cell line (HEK-293T) using Flp-mediated recombination. Cells are sorted into bins based on GFP intensity, and the exon library from each bin is sequenced and genotyped. S_{GFP} is calculated as a weighted sum based on the frequency of each variant in each sorted bin. (B) Library design. Positions in the mutagenized exon (bold) are numbered, and degenerate positions are annotated with the possible mutant bases above the wild type sequence. The known splice donor and acceptor sites are highlighted with gray boxes. (C) Flow cytometry of *TP53* variant library. Cells were sorted based on GFP signal (x-axis); blue, cells expressing a wild type exon 6; red, negative control cells expressing the G113A variant of exon 6; green, *TP53* exon 6 synonymous variant library. Bins are represented by lines at top, annotated with the percentage of sorted cells in each bin.

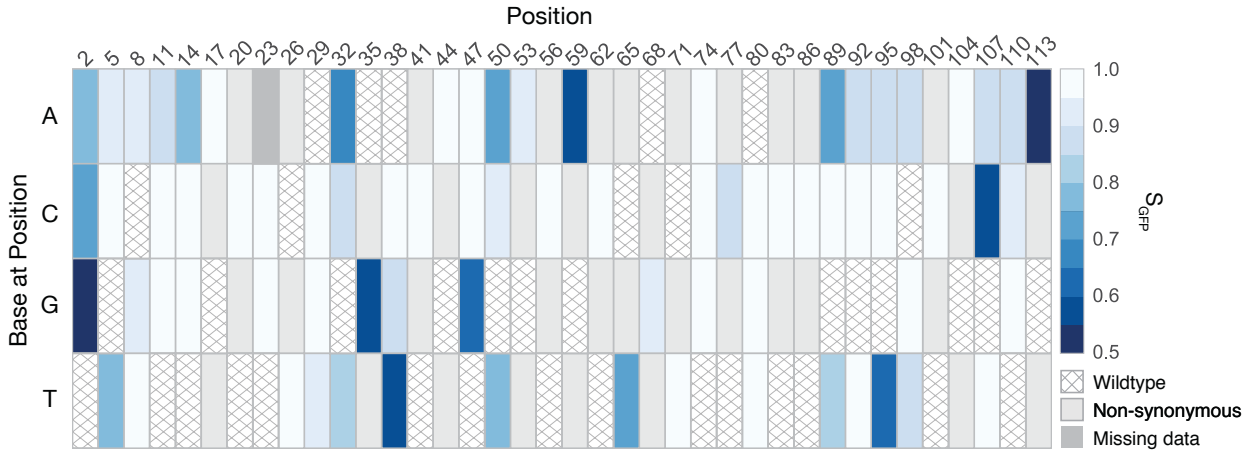

Figure 3.2. Synonymous mutations in exon 6 affect p53 expression. S_{GFP} scores for synonymous mutations in exon 6 of *TP53* plotted as a heatmap. At each codon's third position (x-axis), the base present in the exon variant (y-axis) is colored based on S_{GFP}. The wild type base at each position is hatched, non-synonymous mutations are light grey, and missing data is dark grey.

3.3.2 *Synonymous mutations in exon 6 of* TP53 *affect protein expression*

We focused our analysis on single synonymous mutants in the exon, of which we measured 76 of the 77 expected synonymous mutations. S_{GFP} scores ranged from 0.47 to 1.31 (Figure 3.2). Sixty-eight variants were neutral (considered here as having S_{GFP} 70% or above relative to wild type), but a subset of nine mutations caused a large decrease (S_{GFP} less than 70% of wild type). As expected, two of the mutations (T2G and G113A) disrupting either of the exon's canonical splice sites caused at least a 30% decrease in S_{GFP}. Additionally, seven other mutations in the interior of the exon (G32A, A35G, A38T, T47G, G59A, G95T, and G107C) had S_{GFP} scores less than 0.7. We tested the fluorescence of 20 variants clonally and found significant correlation with our pooled data (p=0.0002, Figure 3.3A). As an additional validation, we mined our data for doubly-mutated variants that contained these mutations, and found that, with the exception of T47G, the negative effects of the single mutants were epistatic to other synonymous mutations in the exon (Figure 3.3B).

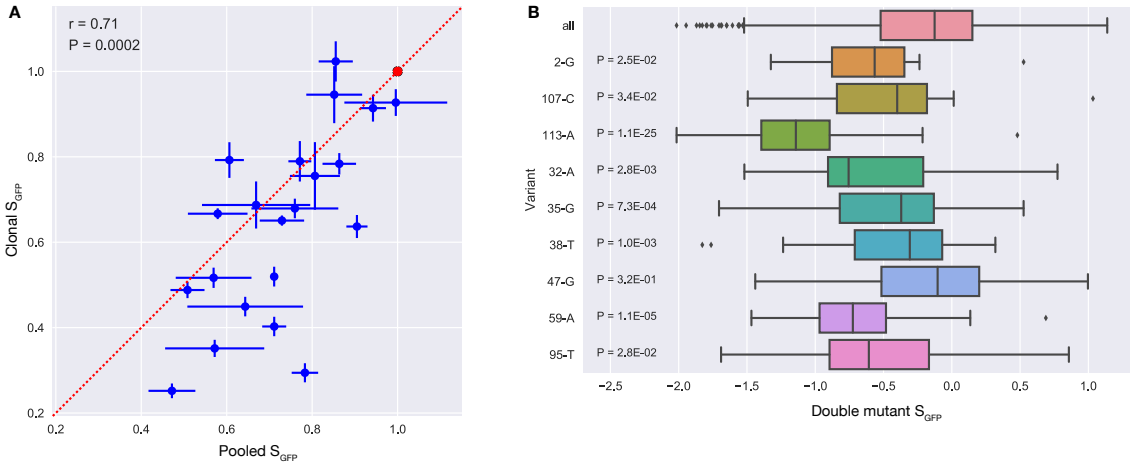

Figure 3.3. Validation of S_{GFP}. (A) Comparison between clonal and pooled S_{GFP}. Cell lines that expressed a single reporter variant were analyzed by FACS (y-axis) and were compared to the scores determined by pooled sorting and sequencing. A red line marks y=x. Wild type is shown as a red point. (B) S_{GFP} boxplots for double mutants (with total cell counts greater than 800 in replicate 1) containing a detrimental single mutant. P-values are based on a t-test comparing the distribution of each double mutant distribution to the distribution of all double mutants in the library (top row).

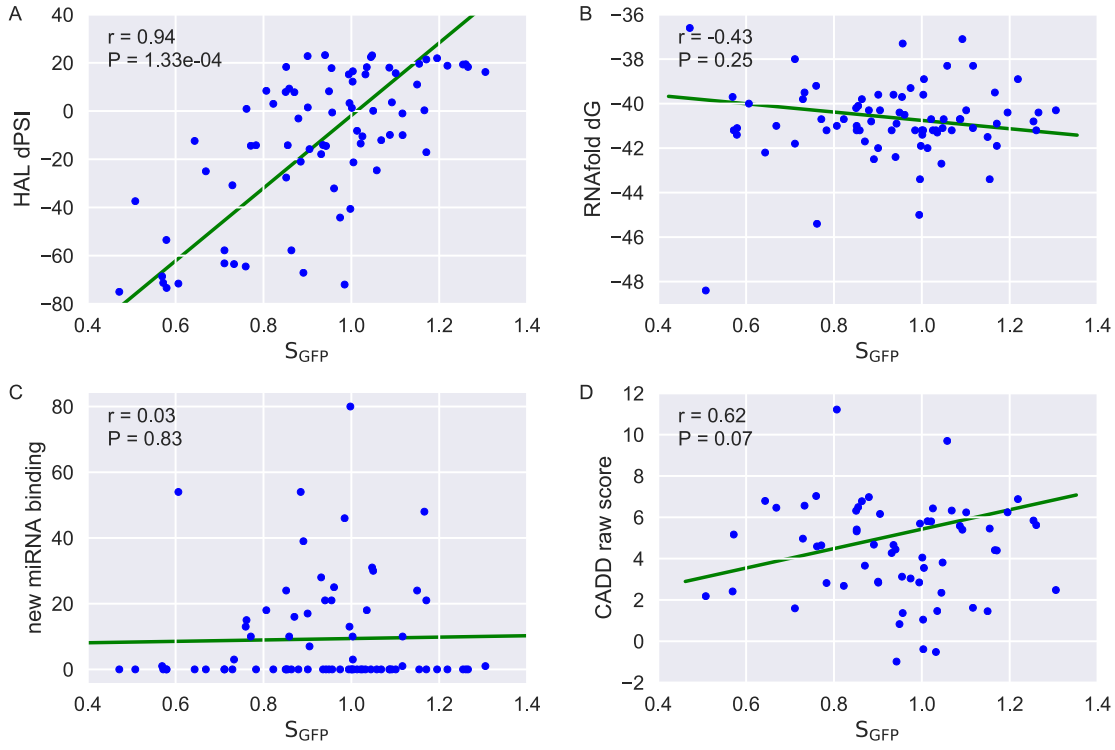

Figure 3.4. *In silico* predictions of synonymous variant effect. We used software prediction to measure the possible effects of synonymous mutations in exon 6 of *TP53* and compared those predictions (y-axes) to the S_{GFP} for each mutation (x-axis). Linear fits are shown as a green line, and the correlation coefficient and P-value for that correlation is shown in each plot. (A) Change in percent spliced-in compared to a wildtype exon based on HAL predictions (*128*). (B) Change in RNA structure, as measured by ΔG with RNAfold (*222*). (C) The predicted number of additional miRNAs targeting each synonymous variant using *rna22* (*223*) and the mature human miRNA sequences from miRBase (*224*). We removed the miRNAs that bound to the wild type exon sequence, and plotted the number of remaining miRNAs for each mutation (y-axis). (D) CADD (*216*) scores for each synonymous mutation.

Using *in silico* predictions, we asked whether mutations in the library would have predicted effects on splicing (*128*), RNA structure (*222*), or miRNA targeting (*223*), or were predicted to have strong functional effects (*216*). Of these, only splicing prediction yielded a significant

correlation with experimental results (p=1.3e-4, Figure 3.4). Increased inclusion of exon 6 results in increased GFP intensity, while skipping of exon 6 results in GFP being out of frame and therefore decreased GFP intensity. As such, splicing defects would be apparent in our assay. To examine splicing, we extracted RNA from 22 clonal samples and used RT-PCR to amplify the region between exon 5 and exon 7 (Figure 3.5). The wild type sample contained two major isoforms: a correctly-spliced 196 base product and an exon 6-skipped 83 base product. These two fragments were present in varying intensities across all 22 samples. In addition, mutant samples contained an unspliced, 845 base product that includes both flanking introns. As the 845 base product was not present in the sample from cells expressing the wild type exon, it most likely derived from unspliced mRNA, rather than trace genomic DNA contamination.

S_{GFP} quantifies the amount of p53 reporter protein present, independent of the amounts of reporter RNAs that do not give rise to detectable fluorescent protein. However, a lower proportion of correctly spliced RNA compared to other RNAs would be expected to lead to less reporter protein and thus lower S_{GFP}. The approximate intensity of the 845 base and 83 base isoforms relative to the correctly spliced 196 base isoform observed from each mutant was consistent with its S_{GFP}. For example, the known T2G splice junction mutant appeared enriched in the amount of 83 base exon-skipped product relative to the correctly-spliced 196 base isoform (Figure 3.5). Eighteen of 21 mutants also contained a 277 base fragment, which includes intron 5. The intensity of this band decreased in variants with mutations farther from the 5' splice site, consistent with this isoform being due to disruption of this splice site. The G113A splice junction mutation, which inactivates the 3' splice site of the exon, resulted in the use of a cryptic site that shifted the splice site 5 bases (*209*). The A38T mutation produced a unique truncated splice

isoform of 117 bases by creating a consensus splice donor. This isoform contains the first 36 bases of exon 6, causing a frameshift and premature stop codon 8 bases later. G47T created an approximately 180 base truncated isoform. We confirmed the sequence of these isoforms by cloning and Sanger sequencing the cDNAs (data not shown). Many mutants contained additional alternatively spliced fragments ranging in size between approximately 350 and 600 bases, which were not easily identifiable.

Based on the abundance of the cDNA products, some mutations (most notably A38T and G95T) appeared to result in significantly less mRNA relative to other variants. We did not measure RNA stabilities of these variants, but note that RNA degradation pathways, especially the nonsense-mediated decay pathway, may affect these mRNAs differentially.

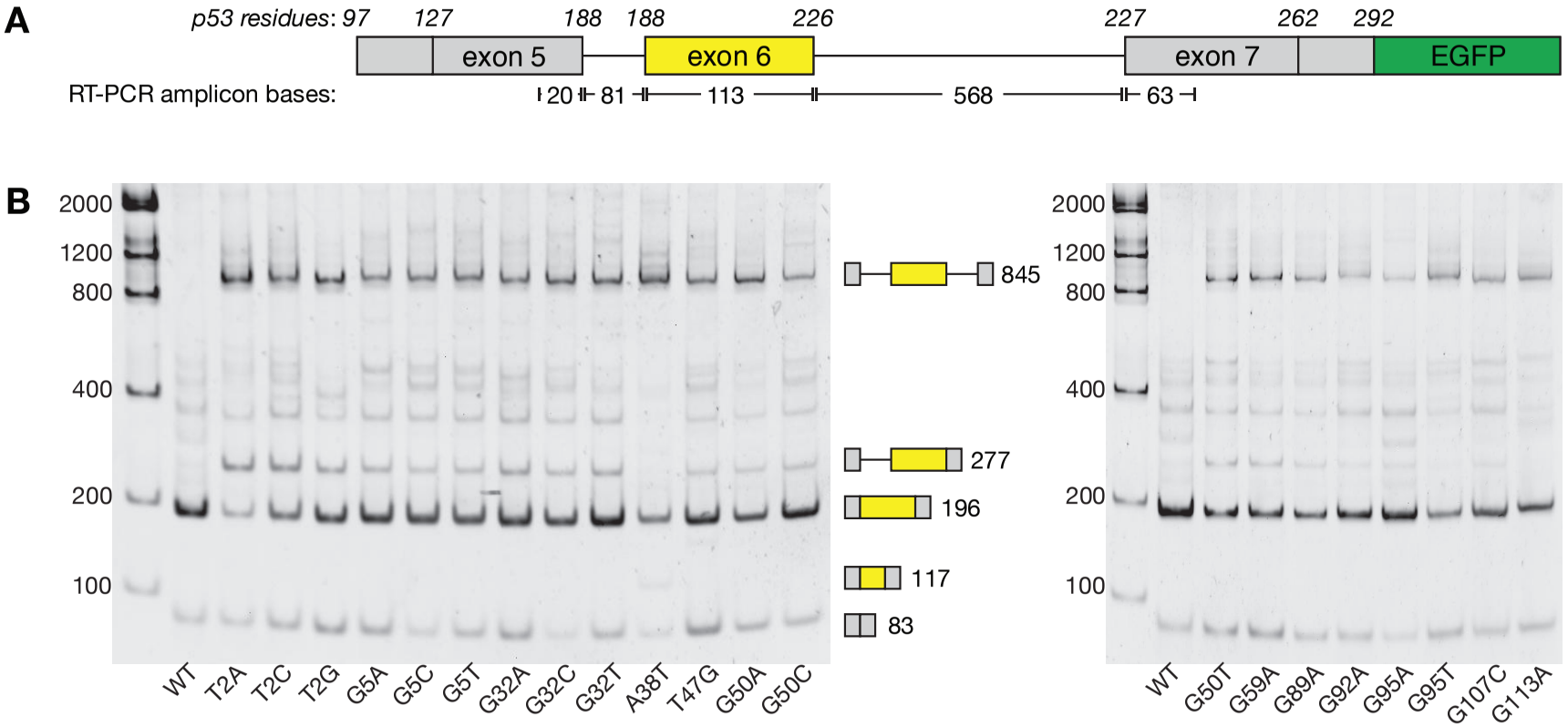

Figure 3.5. Synonymous mutations cause mis-splicing. (A) Model of the reporter construct. Amino acid residues in p53 are shown in italics above the model. Residue 188 spans the splice junction between exons 5 and 6. Lengths of relevant exonic and intronic fragments are below the model. (B) Polyacrylamide gel showing cDNA for clonal samples. *TP53* exon 6 mutations are shown below each well. Cartoon representations of major fragments and their lengths, matching the model in (B), are shown between the two panels.

3.3.3 TP53 *exon 6 synonymous variants are found in tumors*

We hypothesized that if the mutations identified in our assay were truly detrimental, they should be at low frequencies in healthy populations but present in tumor samples. We searched the genomes and exomes of healthy individuals using the gnomAD database ((*225*), Table 3.1), for variants in exon 6 of *TP53*. Eleven synonymous variants in exon 6 were found in the database, with 10 of the 11 variants present in fewer than 37 individuals (out of approximately 250,000 total haploid exomes and genomes). Only one of these was a detrimental variant, G59A (S_{GFP} =0.57), which was found in 9 heterozygous individuals. One neutral variant in our assay (A80G, S_{GFP} =1.02) was found at a high population frequency (1.2%), and was homozygous in 33 individuals.

We also searched for synonymous mutations in the COSMIC database (in which *TP53* was sequenced in 127,779 samples) to estimate the frequencies of mutations in tumor samples (*226*). Thirty-seven synonymous variants were found in COSMIC, found in between 1 and 9 samples. Four of these variants were detrimental in our assay – G32A (1 sample), G59A (1 sample), G107C (3 samples), and G113A (9 samples). Thirty-three other neutral (S_{GFP} >0.7) synonymous mutations were also present in the database, including T74C and A80G, both found in 8 samples each. The 3.4x enrichment of detrimental mutations in COSMIC compared to gnomAD correlates with the analysis of samples from The Cancer Genome Atlas showing that synonymous mutations were enriched in *TP53* approximately 4-fold compared to a matched, non-oncogenic gene set (*209*).

Table 3.1. Variants found in healthy individuals and in tumors

Variant	S_{GFP}	gnomAD occurrences	COSMIC occurrences
T2C	0.711192	0	1
C8T	1.047307	1	2
T11G	1.035109	0	1
T14A	0.761361	0	1
G17A	1.092307	0	2
T20C	1.049779	0	2
T23A	1.087732	0	2
T23C	1.086472	2	2
T23G	1.170654	0	1
A29C	1.001109	0	1
A29G	0.9847	0	3
G32A	0.668794	0	2
A38G	0.890892	0	2
T47G	1.149722	0	1
G50A	0.711093	0	2
G50T	0.759509	0	2
G53A	0.935543	5	2
T56C	1.003371	1	0
G59A	0.569321	9	1
T62C	1.260442	0	1
C71T	0.974411	0	1
T74A	1.116314	0	1
T74C	1.002911	0	8
A80G	1.025217	3407	8
T83C	1.195203	1	1
T86C	1.218696	0	2
G89C	1.044555	37	0
G92A	0.863519	0	1
G95A	0.855208	1	1
C98G	1.100993	0	1
C98T	0.884762	0	3
G104A	1.004803	0	2
G107A	0.851743	12	1
G107C	0.578939	0	3
G107T	1.012974	12	1
T110A	0.879456	0	1
T110C	0.90084	0	2
T110G	1.265524	0	1
G113A	0.471776	0	9

We searched gnomAD (healthy individuals) and COSMIC (tumor samples) for variants in our dataset. Variants with SGFP < 0.7 are shaded.

3.4 Discussion

Using a massively parallel approach, we identified synonymous, partial loss-of-function mutations in exon 6 of *TP53*. These mutations appeared to alter the splicing of the exon and surrounding introns, consistent with an accumulation of unspliced pre-mRNA or mis-spliced products that included introns. As a result, less full-length p53 protein would be expressed compared to the construct containing the wild type sequence of exon 6. These mutations are found at low frequencies in healthy human populations and are a few-fold more prevalent in tumor samples. Thus, synonymous mutations – including mutations that fall at canonical splice sites as well as at least 6 bases away from a splice site – may be drivers of oncogenesis.

A qualitative effect on splicing was evident for almost all of the severe mutations. As a general trend, variants with low fluorescence scores showed increased intensity for the unspliced 845 base or exon-skipped 83 base isoform relative to the intensity of the correctly-spliced 196 base isoform, although some variants did not follow this trend. Our analysis here does not take RNA stability into account, although some variants (*e.g.,* A38T and G95T) appeared to accumulate less RNA than most.

It is possible that our reporter construct, which has cDNA rather than genomic sequence flanking exons 5 and 7, differs in its splicing properties compared to the genomic copy of *TP53.* While the wild type version of this construct spliced correctly, we cannot rule out the possibility that the effects of mutations in our construct may be more or less severe if they were present in the genomic copy. The artificial nature of the construct may partly explain the accumulation of the 845 base isoform in the mutants, as this isoform may not enter the splicing pathway, rather than

being mis-spliced. However, it would be possible in future experiments to use endogenous gene structures or perform mutagenesis *in situ,* which could mitigate these effects.

A p53-negative tumor arises from two inactivating mutations of *TP53*, such as by mutation in one copy and complete deletion in the other (*227*). Could synonymous mutations be responsible for one of these "hits?" In most of the variants we assayed, the amount of exon-included splice isoform was qualitatively less abundant than in wild type cells but still present. As the tumor suppressor function of p53 is known to be haploinsufficient (*228*), a reduction in p53 protein from one allele, even if not complete, could render the cell susceptible to oncogenesis in the event of a mutation to the other allele. Alternatively, variants of *TP53* in which the protein is truncated in exon 6 or before exon 7 have been shown to promote cell proliferation and metastasis through a gain-of function activity (*229*, *230*). Synonymous mutations in exon 6 that cause p53 truncations would produce similarly structured proteins and may also have gain-of-function activities. These include mutations that lead to inclusion of intron 6, which creates a premature stop codon 39 bases into the intron. The A38T mutation mis-splices to delete the last 77 bases of exon 6 and leads to termination 12 bases after the aberrant splice junction with exon 7. Whether these isoforms are also gain-of-function variants remains to be tested.

Our assay measured the effects of mutations in a single *TP53* exon on expression, but the method should be general to studying other exons in *TP53* or other tumor suppressor or oncogenes. The major effects we found in our assay were due to splicing errors. Although other studies have directly measured splice isoforms of libraries of mutated exons using massively parallel RNA sequencing (*128*, *231*), which may be a more quantitative assay than our sorting methodology,

these methods do not account for other mechanisms by which synonymous mutations can affect protein expression. Our approach, on the other hand, is agnostic to the molecular mechanism of protein expression variation and should allow for a more complete understanding and characterization of synonymous mutations.

Chapter 4. Combining multiple barcodes in a single cell with *trans*-splicing ribozymes

Contributions M.S.R. developed the idea and performed all experiments described here.

4.1 Introduction

As reviewed in Chapter 1, high-throughput sequencing has enabled major advances in the massively parallel measurement of variant libraries. The variants in these libraries are generally expressed off plasmids in a cell (as in Chapter 2) or integrated into genomes at specific sites (as in Chapter 3) or randomly (as with lentiviral infection, for example). Library genotypes are assayed by high-throughput sequencing. However, a general limitation of these assays is the inability to obtain more than one genotype per cell, which prevents the measurement of combinations of variants. For example, we cannot perform all-by-all assays, as might be useful in the context of genetic or protein-protein interaction assays.

In order for multiplexed combinatorial library assays to be performed, such as all-by-all genetic interaction or protein-protein interaction screens, library members from distinct loci (*e.g.* plasmids or genomic loci) must be genotyped together in the same cell. In general, this grouping can be accomplished by physically linking the two loci, so that they can be amplified and sequenced in the same amplicon. In practice, the sequencing requires a short amplicon (*e.g.,* less than 1000 basepairs), requiring that disparate loci be brought into close proximity. In yeast-based assays, which are amenable to large combinatorial pools due to yeast's easy mating, Cre-based recombination has been used to link barcodes at distinct loci, both on plasmids (*109*) and in the

genome (*110*, *111*). These methods are limited in their scope for a number of reasons. First, they require significant initial genome engineering to provide, at minimum, both for Cre to be induced in the cell and for the necessary loxP sites for recombination. Second, Cre recombination can be inefficient due to the low success of recombining two low-copy loci. Third, as has been shown in the case of methods where the barcode loci are integrated into the yeast genome, Cre induction is non-specific, leading to aneuploidy in a significant fraction of cells, confounding fitness-based assays (*110*, *232*).

I demonstrate here a novel general barcoding methodology amenable to assaying pairwise and higher-order barcode combinations. In this method, we express barcodes as RNAs fused to fragments of the *Tetrahymena thermophila* ribosomal RNA intron ribozyme (*233*). This ribozyme is able to splice in *trans in vivo* (*234*), which physically links barcodes into a single RNA molecule that can be sequenced. I propose as a proof of principle using this *trans*-splicing ribozyme to pair barcodes in a small combination all-by-all yeast two-hybrid assay, with on the order of ~200 combinations assayed in parallel. In addition, I show that two ribozymes can be used to combine three barcodes *in trans*, allowing for higher-order combinations of barcodes to be assayed.

4.2 Materials and Methods

Oligonucleotides, yeast strains, and plasmids used in this study. Oligonucleotides used in this study can be found in Table 7.5. Yeast strains used in this study can be found in Table 7.6.

Cloning and barcoding ribozyme plasmids. Two-micron, small RNA expression plasmids were created by cloning a synthesized fragment containing the *SNR52* promoter and *SUP4* terminator between the *Sac*I and *Kpn*I sites in pRS425 and pRS426 (*235*). This fragment contained a multiple cloning site comprising *Spe*I, *Not*I, and *Xho*I sites. We synthesized *cis* (*233*) and *trans* (*234*) versions of the *Tetrahymena thermophila* ribozyme and cloned these into yeast expression plasmids. To create centromeric expression plasmids (in which the ribozymes were expressed from a *TEF2* promoter), we subcloned the *Spe*I-*Xho*I fragment containing each ribozyme into either p415TEF or p416TEF (*236*). pMR147 was created by replacing the *LEU2* sequence in pMR144 with the KanMX cassette. pMR144 was amplified with oligos 1029 and 1030, treated with *Dpn*I, and cleaned with a Zymo Clean and Concentrator-5 kit. KanMX was amplified from pFA6-KanMX6-pGAL1 (*237*) with oligos 1031 and 1032, *Dpn*I-treated and cleaned. These fragments were then joined by Gibson assembly.

Barcodes were created by annealing oligos (1105 and 1106 for the 5' exon, or 1082 and 1083 for the 3' exon) and performing a single cycle of extension with the Klenow fragment of DNA polymerase I (NEB). These fragments were cleaned using the Zymo Clean and Concentrator-5 kit (Zymo Research). *Trans* ribozyme plasmids were digested with *Eco*RI or *Nde*I, treated with Calf Intestinal Phosphatase (CIP, from NEB), and gel-extracted using a Qiagen MinElute Gel Extraction kit. Barcodes were mixed in a 10-fold molar excess with their respective plasmids and cloned by Gibson assembly. This reaction was cleaned with a Zymo Clean and Concentrator-5 kit, and one microliter of these reactions was transformed into DH5alpha *E. coli* (NEB).

Triple-barcode ribozymes were created by cloning a synthesized double-ribozyme into p414TEF. *Trans* versions of this construct were then cloned individually. The 5' fragment was created by annealing oligos 1099 and 1100 and performing a single cycle of extension with the Klenow fragment of DNA polymerase I (NEB). The internal fragment was created by amplifying pMR146 with oligos 1101 and 1102. These two fragments were digested with *Spe*I and *Xho*I and ligated into p416TEF (for the 5' fragment) and p415TEF (for the internal fragment) using T4 DNA ligase. Barcoding of these plasmids was performed as above, using the *Eco*RI site in pMR150, and the *Hind*III site in pMR152.

4.3 Results

4.3.1 *Combinatorial barcoding with* trans-*splicing ribozymes*

I followed the methodology of (*234*) in my initial designs of *trans*-splicing ribozymes, using the wild type sequence for the ribozyme (Figure 4.1). In this design, the intact *cis* ribozyme is split in the middle of the P1 hairpin. This split yields two fragments, one that contains the uracil of the required wobble base pair and the first half of the P1 helix (the "BC1 fragment"), and a second that contains the remainder of the ribozyme (the "BC2 fragment"). Each of these contains a unique restriction enzyme site used to clone barcodes into the plasmid. Additionally, I added an MS2 hairpin to the 5' end of the BC1 fragment and a PP7 hairpin to the 3' end of the BC2 fragment, in an effort to provide additional stability to the expressed RNA.

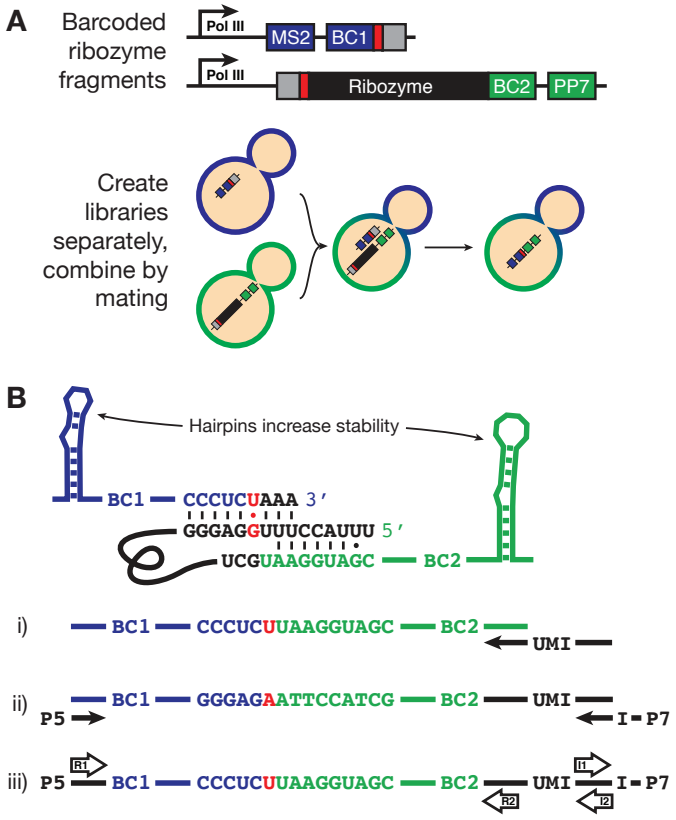

Figure 4.1. Combinatorial barcoding using *trans*-splicing ribozymes. (A) Schematic of barcoding constructs and yeast combinatorial library screening. Red boxes mark the wobble base pair necessary for splicing (also in red in B); grey boxes mark locations of homologous sequences. (B) Ribozyme active site and RNA sequencing methodology. *Trans*-splicing fragments use the endogenous *Tetrahymena thermophila* ribosomal RNA intron. Flanking the barcodes on each fragment are RNA hairpins, which should increase stability of the molecule. For RNA sequencing, spliced RNA are reverse transcribed with a barcode-specific primer containing a unique molecular identifier (UMI, i). After reverse transcription, cDNA molecules are amplified with primers containing Illumina sequencing adapters and demultiplexing indices (ii). The amplicons are sequenced with paired reads covering the two barcodes, as well as reads sequencing the UMI and index (iii).

These fragments were initially expressed using the same construction as for CRISPR guide RNAs in yeast (*21*). Fragments were expressed from high-copy 2-micron plasmids using the promoter from *SNR52*, a constitutive RNA polymerase III promoter. High copy plasmids were necessary to observe splicing (data not shown). Splicing efficiency can be increased by adding homologous to the ends of the *trans* fragments (*238*), in theory by creating additional base pairing to bring *trans* fragments into proximity. I designed homologous sequences to add to the 3' end of the BC1 fragment and the 5' end of the BC2 fragment to test whether the additional homology has an effect in this method. Using RNA structure prediction (*222*), I identified 20, 50, and 250 base unstructured regions of the lambda bacteriophage genome that were not present in the yeast genome and used these as homology. I also created a *cis* version of the barcoded ribozyme construct.

I co-transformed yeast with combinations of these plasmids and assayed *trans*-splicing by reverse transcription and quantitative PCR. All versions of the ribozymes spliced efficiently and specifically *in vivo*, yielding a robust band of the correct 81 basepairs (Figure 4.2). The sequence of these products was confirmed by Sanger sequencing (not shown). *Trans*-splicing combinations amplified about 3-4 cycles after the *cis* ribozyme. Adding homologous sequence to the *trans* fragments did not increase splicing efficiency, which agrees with the results from (*239*) that the effects observed in (*238*) are not a general rule.

Two-micron plasmids are not actively maintained at a specific copy number in cells, and so they are not generally used for fitness-based assays; a change in copy number of the plasmid could yield a large fitness advantage and mask the effects of the assayed constructs. As such, I wished

to determine whether *trans*-splicing RNAs expressed from centromeric plasmids would be detectable. I cloned the initial set of fragments into centromeric plasmids in which they were expressed from high expression *TEF2* promoters (*236*). When co-expressed in yeast, this set of fragments also yielded detectable spliced products (not shown), and I therefore used centromeric P_{TEF2} plasmids for the remainder of the experiments presented here.

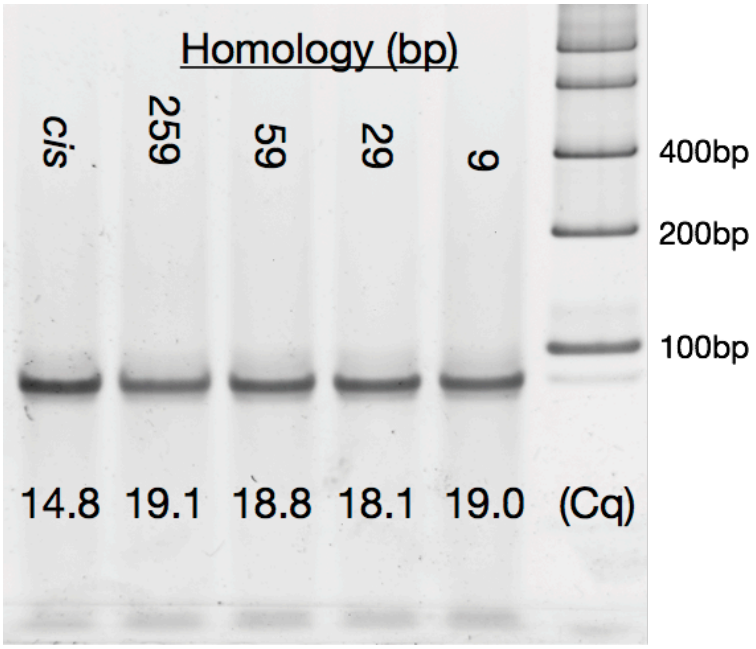

Figure 4.2. *Trans* ribozymes splice *in vivo*. Total RNA was harvested from yeast co-expressing ribozyme fragments. 2 µg of RNA was DNAse-treated, cleaned, and reverse-transcribed with a primer specific for the BC2 fragment. Homology lengths are written above each lane (9 bp is equivalent to the length of homology without additional homology sequence added). Quantitation values (Cq) for each sample are shown at the bottom of lanes. Products shown were amplified for 35 cycles of PCR.

4.3.2 *Splicing three barcodes by tiling trans-splicing ribozymes*

I hypothesized that using ribozymes to splice loci would be a general approach to link more than two barcodes in the same molecule. I adapted the two-barcode ribozyme to splice three barcodes by fusing a second ribozyme to the 5' barcode and adding a new 5' barcode fragment (Figure 4.3). In order to ensure that the ribozymes spliced in a specific order, we used a synthetic recognition sequence from a study of *in vivo trans*-splicing (*238*). This double ribozyme construct spliced specifically and efficiently in *cis,* as only a single, correctly-spliced band was present after RT-PCR. We created *trans* fragments of the construct, but have yet to test whether they *trans*-splice specifically.

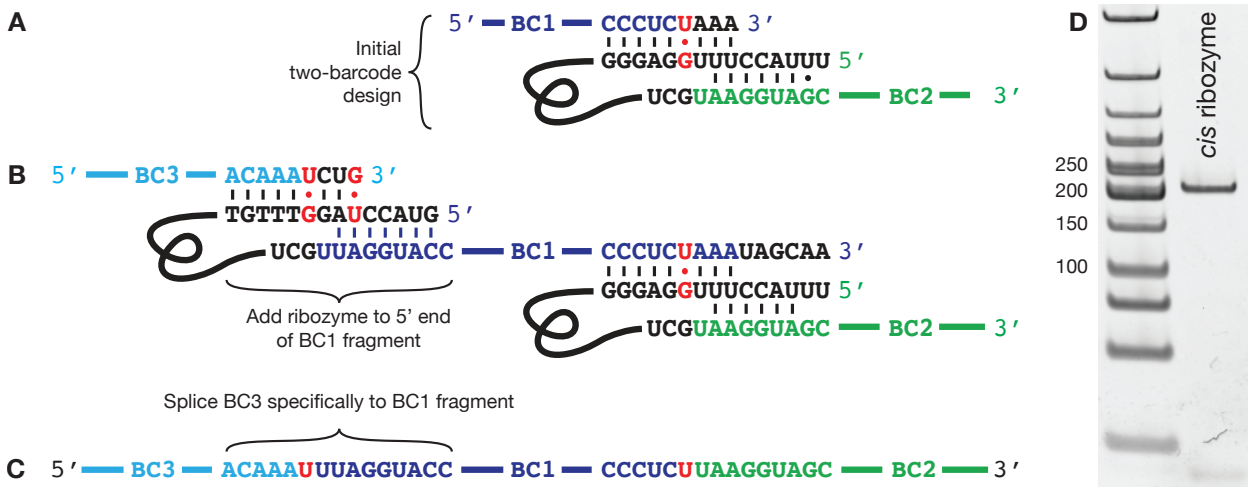

Figure 4.3. *Trans*-splicing three barcodes into a single molecule. To adapt the initial two-barcode ribozyme (A) to splice more than two barcodes, we added an additional ribozyme, using the synthetic recognition site used in (*238*), to the 5' end of the BC1 RNA (B). This structure should, in theory, allow for specific ordering of *trans*-splicing into a single RNA molecule (C). In (D), I show the RT-PCR product from a *cis* triple barcode ribozyme. The fragment is the expected size.

4.4 Discussion

I have made initial headway in developing a method for combinatorial cell barcoding using *trans*-splicing. By using the *trans*-splicing fragments of the *Tetrahymena thermophila* ribozyme as plasmid barcodes, rather than static DNA barcodes, I was able to physically link barcodes into a single RNA molecule. I plan to apply this methodology to a small all-by-all yeast two-hybrid selection as a pilot experiment. I also showed that ribozymes can be tiled to splice barcodes from three plasmids in the same cell. It is probable that this construction will be general to any number of barcodes, albeit with decreasing efficiency.

While this work was performed in yeast, I believe that this barcoding methodology should be applicable to most other model systems because the *trans*-splicing reaction requires only cofactors found commonly in a cell (magnesium ions and guanosine bases) and no proteins. Indeed, *trans*-splicing with the ribozyme has been applied to mammalian cells in attempts to repair mRNA coding for defective proteins (*240-242*). These methods suffered from low splicing efficiencies, but our needs – potentially accumulation of only a small number of spliced barcodes per cell – are much simpler than the clinical task of splicing a repair exon into significant proportion of a specific mRNA in a cell.

Single cell methods could also identify barcode combinations, but these assays sparsely measure the RNA in only thousands of cells. In order to quantitatively measure many thousands or millions of barcode combinations for large-scale assays, it will be necessary to assay millions of cells, which is currently accessible only by specifically sequencing linked barcodes.

Once a ribozyme-barcoded library is created, it can be used in conjunction with any other library that it can splice into. For example, CRISPR guide RNAs can be heavily modified and still retain function, and thus they could be simply modified with the 5' fragment from this method, allowing CRISPR screening to be combined with a separate barcode-based screen or selection. Similarly, by fusing a *trans*-splicing ribozyme to the guide RNA itself, which can tolerate relatively large 5' fusions (*42*), dual CRISPR screens could be performed without necessitating the current need for the design and cloning of paired guides.

Chapter 5. Modulating yeast flocculation by DNA shuffling of natural *FLO8* alleles

5.1 Introduction

Flocculation and invasive growth are complex yeast phenotypes that show a high level of diversity between strains, both in the flocculin proteins that physically cause adherence phenotypes and in the transcription factors that regulate them. Classical yeast genetics as well as quantitative trait loci mapping have shown that the transcription factor Flo8 is a predominant regulator of these phenotypes (*243*, *244*). *FLO8* alleles from different strains are capable of exclusively driving either flocculation or invasion when expressed in the same *flo8*-null laboratory strain. To study how Flo8 mediates these phenotypes, we used DNA shuffling (*139*, *140*), a molecular technique that has been applied primarily to directed evolution, to create many chimeric *FLO8* alleles that contain an assortment of genotypes from two strains. These strains have divergent phenotypes: K11, a sake strain that flocculates but does not invade agar, and DBVPG1106, a wine strain that invades agar but does not flocculate. We phenotyped a large cohort of these strains and found that flocculation segregates in a Mendelian fashion, whereas the intensity of invasion spans a continuous distribution. To isolate the specific polymorphisms driving each trait, we pooled the shuffled strains based on their flocculent or invasive phenotypes and used high-throughput shotgun sequencing to assay the parental allele frequencies among the phenotypically-similar strains in each pool. Using this method on ~1000 strains, we identified sites in the 3' end of the *FLO8* locus, as well as two non-synonymous polymorphisms, that are linked to the non-flocculent phenotype of DBVPG1106. To determine whether our results were valid, we recloned the *FLO8* variants present in either K11 or DBVPG1106, but could not detect

a phenotypic difference dependent on which of the two alleles was present. Therefore, it appears likely the original phenotypic differences were caused by a plasmid effect, possibly created during cloning. While this result was unfortunate, the methodology we developed here should be applicable to the fine mapping of genetic traits with phenotypes that can be obtained by selection or screening.

Contributions M.S.R., Elyse Hope, M.J.D. and S.F. developed the idea. E.H. performed initial phenotyping of *FLO8* alleles. M.S.R. created the DNA shuffling library, performed phenotypic screens, prepared sequencing libraries, and analyzed data.

5.2 Materials and methods

DNA shuffling of *FLO8* alleles.

The FLO8 locus was amplified from genomic DNA from a set of diverse natural isolates of *S. cerevisiae* (*245*) and cloned into pRS316 (*235*) by ligation. An S288C yeast strain (flo-) was transformed with these plasmids and phenotyped for flocculation and invasion (Elyse Hope, data not shown).

K11 and DBVGP1106 alleles were shuffled as in (*145*). Briefly, *FLO8* alleles were amplified from plasmids using M13F and M13R primers and KAPA HiFi Hotstart Readymix polymerase (KAPA). These PCR products were cleaned and digested with DNaseI (NEB). The extent of digestion was determined empirically. Digested fragments were gel-extracted, mixed 1:1 of each allele, and assembled using polymerase cycling assembly with KAPA HiFi Hotstart Readymix for 25 cycles. 1µl of this reaction was used as template for a second PCR, this time with M13F

and M13R primers, to enrich for full-length products. Clean, full-length amplicons were cloned into pRS316 by Gibson assembly (*220*) and electroporated into ElectroMAX DH10B *E. coli* (Invitrogen). This culture was grown overnight in LB+ampicillin and miniprepped. FY3 yeast were transformed with this library using a high-efficiency lithium acetate protocol (*170*). Single colonies were picked into 96 well plates to perform phenotyping.

Strain phenotyping. Strains were phenotyped for flocculation after growth in 96-well plates followed by shaking for 30 s on an orbital shaker. Invasion was phenotyped by plating cells on SC-uracil agar plates, growing them at 30°C for six days, and then washing colonies from a plate and observing how many cells remain (i.e., invaded into the agar). Plates were scanned before and after washing. Invasion intensity was calculated as the ratio of the pixel intensity before washing to the cells remaining after washing, using a custom workflow in ImageJ.

Sequencing analysis. After phenotyping, strains were pooled based on their flocculation phenotype or deciles of invasion intensity. Each pool's plasmids were extracted using a Zymoprep Yeast Plasmid Miniprep II kit (Zymo Research) and amplified by transformation into *E. coli*. Pool plasmids were miniprepped, fragmented with Nextera transposase (Illumina), and sequenced on an Illumina MiSeq. Reads were aligned to the K11 FLO8 sequence with Bowtie2 using relaxed parameters and read depths at polymorphic sites were called using Samtools.

5.3 Results

5.3.1 *DNA shuffling can assort FLO8 alleles*

Natural isolates of *S. cerevisiae* have a wide variety of biofilm-related phenotypes (*246*). Indeed, expressing the *FLO8* alleles from these strains in a flo- lab strain can induce a variety of combinations of flocculation phenotypes (Elyse Hope, data not shown). We decided to use DNA shuffling (*139*, *140*) and techniques from bulk segregant analysis (*247*) to perform fine genetic mapping on these alleles in an attempt to identify the specific residues in Flo8 that were responsible for such diverse phenotypes. Briefly, by randomly assorting the mutations in *FLO8* alleles and phenotyping strains, we can apply bulk segregant analysis methodologies to identify sites responsible for a phenotype (Figure 5.1).

We focused on the *FLO8* alleles from two strains, K11, a sake brewing strain isolated from a brewery in Japan, and DBVGP1106, a wine strain isolated from a grape in Australia, which induce an intriguing combination of phenotypes. Expression of $FLO8_{K11}$ causes S288C to flocculate but not invade, and $FLO8_{1106}$ causes S288C to invade but not flocculate (Figure 5.2A). Since these strains are otherwise isogenic, we hypothesized that mutations in *FLO8* must cause these differential phenotypes. There are 43 mutations between the $FLO8_{K11}$ and $FLO8_{1106}$, 12 of which are nonsynonymous (Figure 5.2B).

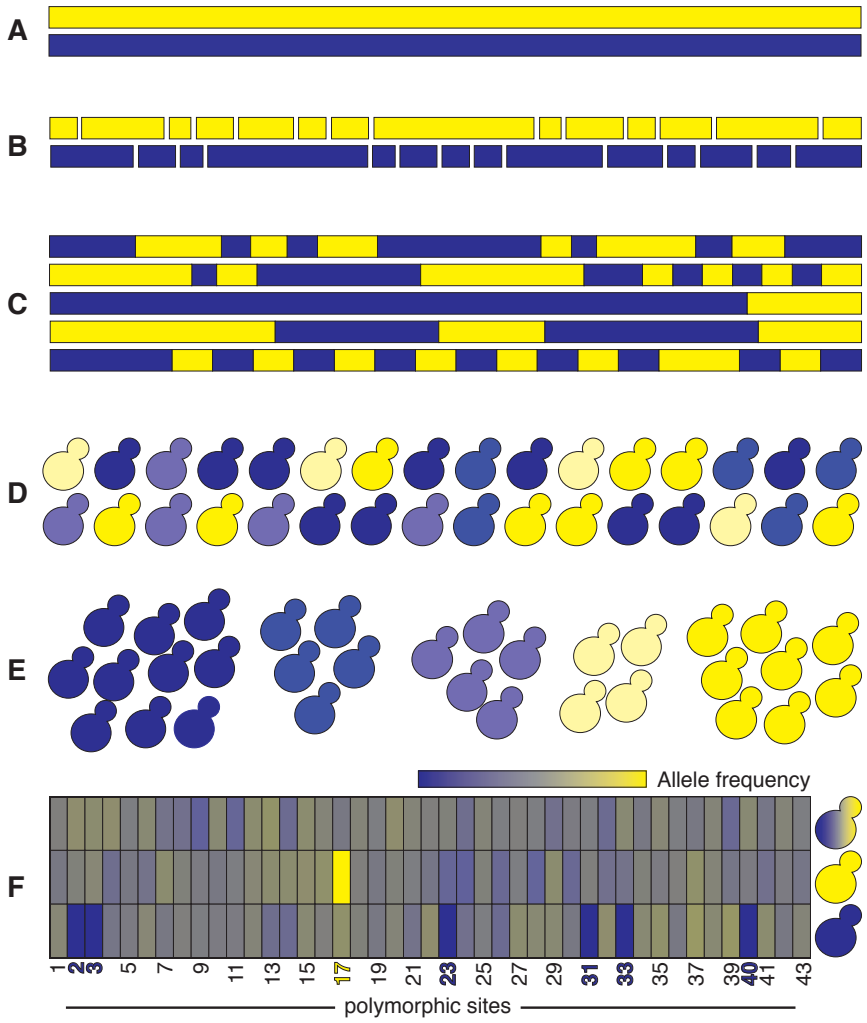

Figure 5.1. DNA shuffling and sequencing methodology. (A) Alleles are mixed in an equimolar ratio and digested with DNAse I (B) to 50-500 bp fragments. (C) Chimeric alleles are created by polymerase cycling assembly and PCR, cloned into an expression vector and transformed into yeast (D). Yeast are phenotyped and (E) divided into pools whose alleles are sequenced. Allele frequencies are calculated, with the expectation that required polymorphisms will be at high frequencies and neutral mutations at ~50%. The idealized data (F) show the entire library of alleles (top), a simple trait with a single polymorphism at position 17 driving the phenotype (middle), and a complex trait in which many polymorphisms are at high frequencies (bottom).

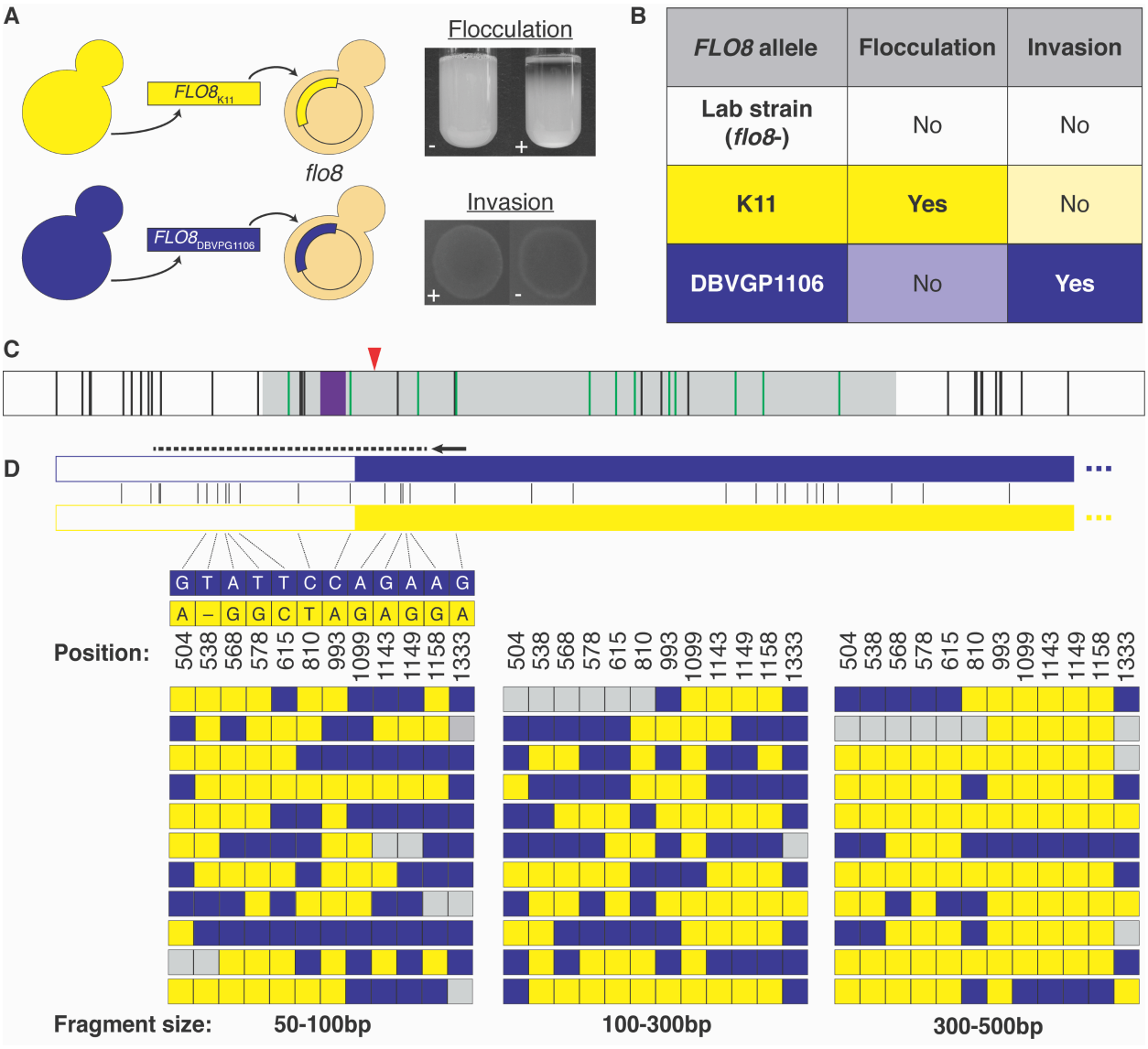

Figure 5.2. DNA shuffling of phenotypically-diverse natural *FLO8* alleles. (A) *FLO8* alleles were cloned from the genomic DNA of two natural yeast strains and transformed into a *flo8* lab strain (FY3). (B) Flocculation and invasion phenotypes of *FLO8* alleles. (C) Of 43 polymorphisms in the *FLO8* locus (vertical lines), 11 are in the 1kb upstream, 12 are in the 1kb downstream, and 20 are in the *FLO8* coding sequence (shown in grey). 12 nonsynonymous variants are shown as green lines. The locations of a nonsense codon found in the lab strain (red triangle) and the only recognizable protein domain, a LISH dimerization domain (purple box), are marked. (D) Alleles of the two strains were shuffled and fragments of the indicated ranges were used as templates for assembly. Polymorphic positions

were genotyped by Sanger sequencing of one region (dotted line). Boxes correspond to polymorphic positions, rows denote a single chimeric variant: K11 (yellow) and DBVPG1106 (blue); not determined (gray).

As a first step, we determined whether we could shuffle *FLO8* alleles. $FLO8_{K11}$ and $FLO8_{1106}$ alleles were amplified and digested with DNaseI, and we extracted fragments in three size windows: 50-100 basepairs, 100-300 basepairs, and 300-500 basepairs. We performed polymerase cycling assembly with these fragment pools separately, cloned the resulting full-length chimeric *FLO8* alleles, and used Sanger sequencing on 11 clones of each pool to test whether mutations were shuffled. As expected, larger fragments created qualitatively longer haplotypes, though single mutations still disrupted long haplotypes even in the 300-500bp library. There appeared to be little qualitative difference between the 50-100bp and 100-300bp libraries, potentially due to the lower efficiency with which shorter fragments should assemble. We chose to pool the 50-100bp and 100-300bp libraries and use these for phenotyping experiments.

We transformed FY3 (an S288C derivative that is flo- due to a frameshift in *FLO8*, see Figure 5.2B) with the library of chimeric alleles and phenotyped approximately 940 strains individually. While agar invasion can be phenotyped in a pool (*62*), pooled flocculation assays are confounded by the ability of flocculent cells to act *in trans*, *i.e.*, form heterogeneous flocs with non-flocculent cells. As such, we developed high-throughput phenotyping assays performed in 96-well plates to allow for efficient screening.

5.3.2 *Phenotyping shuffled FLO8 alleles*

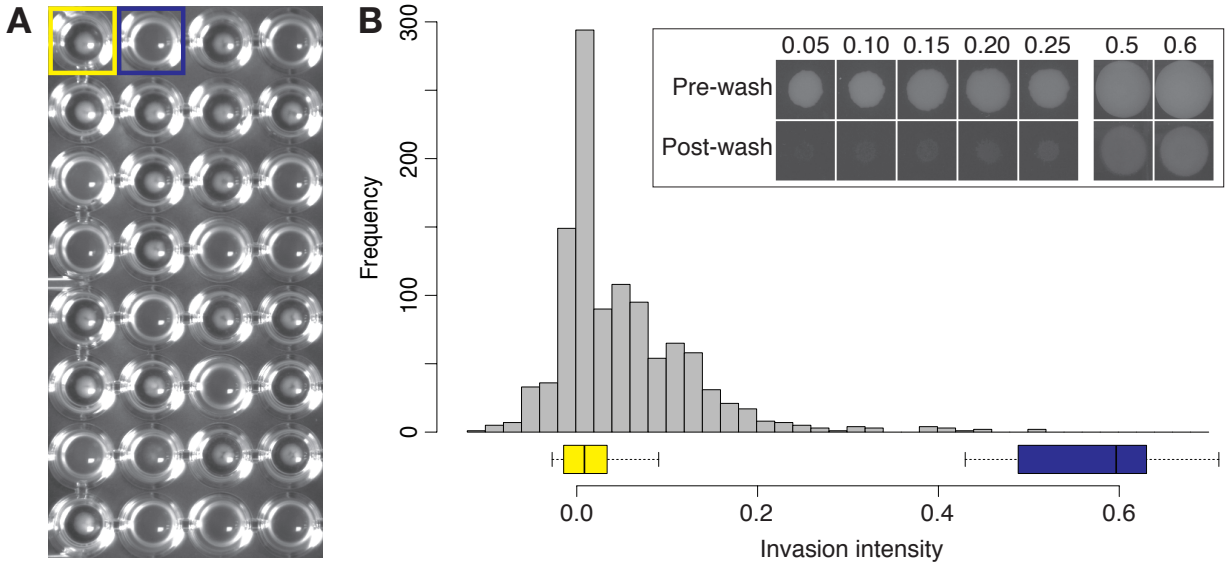

Figure 5.3. Phenotyping of shuffled *FLO8* alleles. (A) A representative example of the flocculation assay. Strains either formed a clump of cells (flocculent, yellow) or remained suspended (non-flocculent, blue). Each well contains a unique yeast transformant. (B) A histogram of the invasion intensity. Box plots show the distribution of parental scores (yellow: K11; blue: DBVGP1106). Inset, representative pre- and post-wash colonies are shown for a range of scores.

Flocculation was assayed by growing strains to saturation in 96-well plates, shaking the plate on a rotary shaker for 30 seconds, then measuring the amount of clumped cells at the bottom of well. There were no strains with intermediate flocculation phenotypes; strains were either flocculent or not. Overall 55.1% of strains were flocculent, consistent with this trait due to a single mutation (or multiple closely-linked mutations) in *FLO8*.

Invasion was assayed by growing strains on agar plates for 6 days, then physically washing cells off the surface of the plate. Images taken before and after washing were used to calculate an intensity metric for the number of cells that have invaded into the agar (Figure 5.3B). Strains

ranged from non-invasive (invasion intensity close to 0) to strongly invasive (invasion intensity close to 0.5), though most strains were either non- or weakly invasive.

5.3.3 *Flocculation allele frequencies are skewed in the FLO8 3' UTR*

We pooled strains based on their phenotypes and used high-throughput shotgun sequencing to genotype each mutation. Allele frequencies of each site for some of these pools can be found in Figure 5.4. Mutations in $FLO8_{1106}$ were highly enriched for non-coding mutations at the 5' and 3' ends of the cloned *FLO8* locus.

5.3.4 *Phenotypic differences were potentially caused by plasmid biases.*

To confirm these results, we recloned and retested the *FLO8* alleles from each strain. Instead of the divergent phenotypes we saw originally, after retesting, all strains were mildly invasive and flocculent. I performed additional quality control tests (not shown here) that were consistent with our early results being a plasmid artifact from the initial cloning of $FLO8_{1106}$.

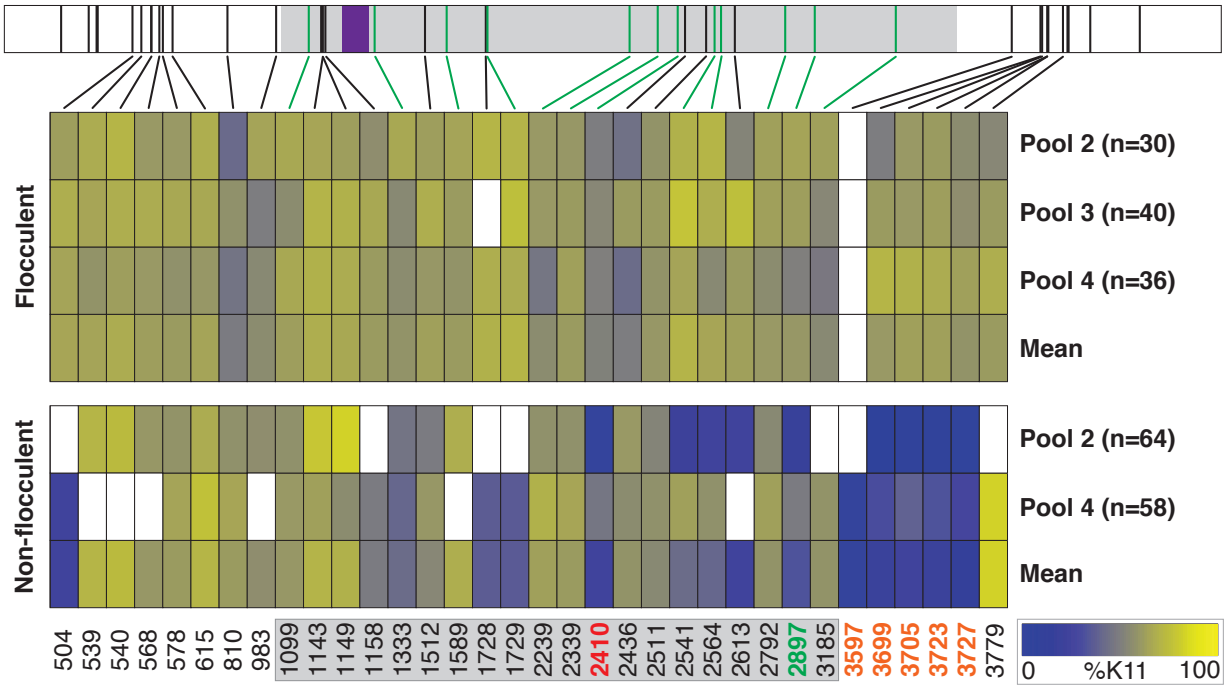

Figure 5.4. Allele frequency analysis of pooled strains based on flocculation. Above, cartoon representation of *FLO8,* with polymorphic sites marked as in Figure 5.1. Below, heatmaps showing allele frequencies in flocculent and non-flocculent pools. Sites are colored based on their allele frequency (100% K11: yellow, 100% DBVPG1106: blue). White cells represent sites with low coverage (less than 10% of the mean depth of mapped reads) or sites were not identified as polymorphic during analysis.

5.4 Discussion

We attempted to finely map mutations in natural variants of *FLO8* that were causative for divergent biofilm-related phenotypes. Our results here were confounded by problems with the initial phenotyping, but we developed a robust methodology capable of phenotyping many strains in high-throughput. This methodology could be used to map the genetic basis of other traits, as the DNA shuffling methodology is general to any similar sequences with lengths that

are able to be amplified by PCR. Indeed a similar methodology was used to map the causative mutations of different Argonaute proteins (*144*) in a mammalian cell culture assay.

The assays we performed here were not massively parallel, though I developed them to be high-throughput and easy for a single researcher to perform. It is possible to assay invasive growth as a pool (*62*), and I believe that there are ways to assay flocculation in a pooled manner, as well. For instance, flocculation is sometimes measured as the rate at which a culture settles (*246*). Flocculent cells should therefore settle faster than non-flocculent cells, and while clumps of cells should be heterogeneous, samples taken from the top of a culture should retain mostly non-flocculent cells. Sampling the top fraction of a culture over time, or with varying concentrations of flocculation inhibitors like $CaCl_2$, and bulk sequencing these samples may be a viable bulk flocculation assay.

DNA shuffling methodologies should also be expanded to use array-synthesized oligonucleotides as templates, rather than pre-amplified homologous genes. Synthetic oligonucleotides have been used as a shuffling substrate (*248*), but only in the context of protein engineering, and only using column-synthesized oligos rather than a large pool. Array synthesis enables a new diversity of substrates to be used for shuffling libraries, allowing for both natural and non-natural variants to be combined without up-front library construction. Additionally, programmed oligonucleotides can easily "unlink" variants that are in close proximity in a sequence by synthesizing oligonucleotides with local mutations in all possible combinations, which can be difficult for standard fragmentation methods.

Chapter 6. The future of massively parallel assays

High-throughput biology is a rapidly moving field. Massively parallel assays have been developed to measure many molecular phenotypes, including protein function and stability, regulatory sequence function, and pre-mRNA splicing. In this dissertation, I described my work to measure the effects of mutations on a yeast promoter as well as the effects of synonymous mutations in a cancer gene. I also described the basis for a novel method to assay combinations of barcodes within a single cell, and methods development for fine mapping traits in yeast using large libraries of chimeric sequences. Through this work, I have identified some areas that I believe will be important for the field to consider over the coming years, as we continue to develop new massively parallel methods for studying cellular phenotypes as they relate to evolution, basic biology, and human disease.

6.1 Is a lookup-table for the effects of all mutations feasible?

Our ability to catalog human genetic variation is tremendous. One promise of massively parallel assays is that we may be able assign functional significance to each of these mutations (*249*), but is that a reasonable pursuit? As I discussed in Chapter 1, massively parallel assays (like those described here) require a selectable phenotype that can be assayed using high-throughput sequencing. Generally, this process requires spending a significant amount of effort to develop assays for each protein or regulatory sequence of interest, making a genome-wide catalog of functional effects an insurmountable task. This pursuit is further confounded by proteins that have multiple functions. Multiple deep mutational scanning experiments have examined *BRCA1* (*32*, *64*), which is known to function in the DNA damage response. These studies have assayed

various molecular phenotypes of *BRCA1*, including mRNA stability, binding to partners, ubiquitin ligase activity, and homology-directed repair, all of which are necessary to understand to predict the effects of *BRCA1* variants in oncogenesis.

Model organisms can help with this task.[3] Organisms like yeast are easily manipulated and have many genetic tools that are not yet well developed for mammalian cell culture systems. Additionally, basic biological mechanisms (which are often those disrupted in human disease) in these organisms are often highly conserved and well understood. Recently, studies have sought to "humanize" yeast by determining which human genes can complement the deletion of their yeast orthologs (*250-252*). These studies have focused primarily on the complementation of essential yeast genes, but it would be possible to apply this technique more broadly using synthetic lethal genetic interactions.

Similarly, assaying regulatory variant effects can be simple, as in the case of massively parallel reporter assays studying a single enhancer-gene interaction. In reality, enhancers can have long-range effects on many genes, and quantitatively assaying these (especially given that in most cases, many of the transcripts regulated by a single enhancer are not known) is a difficult problem. It is possible that single-cell genomics will provide a method by which we can assay the multi-dimensional space necessary to link such regulatory variation to molecular phenotypes, but to quantify the genome-wide effects of variants in a single cell, many single cells will need to be assayed. Assays capable of this pooled throughput have been developed (*253*, *254*), and given the rate at which the field of single-cell genomics is progressing, it is probable that in the near

[3] Despite what many manuscript reviewers may be convinced of.

future single cell techniques may eclipse much of the methodology I have described here, allowing for high dimensional analysis of the effects of human genetic variation.

6.2 How can we best assay multiple molecular processes simultaneously?

The work I presented in Chapter 2 and Chapter 3 demonstrates how to assay the effects of mutations on single phenotypes (fitness, as a proxy for expression, in Chapter 2, and protein expression in Chapter 3). The biology of a cell, though, is not a single phenotype. For example, multiple steps of the Central Dogma can be affected by variation in mRNA. We measured only protein levels in Chapter 3, and had to clone single variants and test their RNA levels and splicing patterns individually to be able to make conclusions about the mechanism of decreased protein expression. Similarly, in Chapter 2 we measured strain fitness, which we assumed was correlated with *SUL1* expression, but at no point did we measure the mRNA expression of *SUL1* variants. In these experiments, the plasmid barcode was not transcribed, but a transcribed barcode would have allowed us to measure fitness and transcription simultaneously. I have since attempted to move the location of the plasmid barcode into the 3' UTR of the *SUL1* transcript, but managed to cause only an extreme fitness defect with barcodes at that location (not shown). It is possible that a more exhaustive search of the 3' UTR would yield a fitness-neutral barcoding location.

It is important, therefore, to design massively parallel experiments that are able to assay multiple types of data. Some methods, like SortSeq (*92*, *93*), were developed with this in mind – cells are sorted on a fluorescent marker, then both their RNA and DNA are sequenced *a la* MPRA to

calculate both mRNA expression and protein expression from a single experiment. I think we could have assayed fluorescence, mRNA expression, and splicing in a single experiment for our work in Chapter 3 by barcoding variants and employing a method similar to that of Rosenberg and colleagues (*128*). I think that it should be possible, with some ingenuity, to design experiments that can assay many of the molecular mechanisms involved in a phenotype of interest, and these experiments will provide richer and more complete datasets with which to probe biology.

6.3 Can we perform massively parallel assays in higher eukaryotes?

To this point, massively parallel assays have been performed almost exclusively *in vitro, in vivo* in single-celled organisms like bacteria and yeast, or in cell culture. This limited set of conditions has primarily been due to two restrictions: assayable population sizes and the ease of introducing mutations. In order to quantify the effects of mutations on a sequence, we must assay a large number of cells bearing a variant, making it necessary to assay millions of cells per experiment. Similarly, variants can be easily introduced into single cells, where there is not the concern of cell-type specific expression differences that might be problematic in higher eukaryotes like nematodes, fruit flies, or mice.

Higher organisms can produce complex, interesting phenotypes that single-celled organisms are unable to provide. For instance, *C. elegans* and *D. melanogaster* exhibit many behavioral phenotypes that could be candidates for deep mutational scanning experiments. While making variant libraries *in vitro* is relatively straightforward (see Section 1.2), introducing these libraries

into these organisms is an extreme bottleneck. In *C. elegans,* for example, DNA must either be physically injected or bombarded into a parental worm, yielding very few transformants. As such, it is likely that *in situ* mutagenesis methods, like those I described in Section 1.2.7, may be necessary to create variant libraries of specific sequences of sufficient size. In these methods mutations accumulate *in vivo,* so multiple generations of growth should create libraries of variants that are roughly equivalent in complexity to the population. There are currently very few possible mutations can be created by CRISPR-based *in situ* mutagenesis methods. For example, in the case of cytosine deaminases fused to Cas9, the majority of mutations are C>A and C>G transversions at a specific location relative to the PAM. Many other proteins, like uracil deglycosylases, which leave abasic sites in the genome, or single oxygen generators (*255*) may be able to induce mutations when fused to Cas9.

Assaying variants in higher organisms may also require a significant amount of technology development. Selections based on fitness should still be straightforward, as variant frequencies will change as beneficial mutants increase in the population. However, assaying higher-order phenotypes, like behaviors, cannot be simply selected for; in these cases, a variety of possible microscopy-based barcoding schemes to assay variants might be employed. Expressing multiple fluorescent markers in distinct organelles has been used in principle to barcode yeast (*256*), and I believe this methodology could be adapted to organisms like *C. elegans,* in which distinct cell or tissue types are combinatorially and fluorescently labeled as a barcode. Additionally, barcodes could be assayed by *in situ* RNA sequencing (*257*) or combinatorial barcoding by fluorescent *in situ* hybridization (*258*).

Overall, significant advances in our understanding of biology can be accessed using massively parallel assays and high-throughput mutagenesis techniques. I believe it is the way forward for biological inquiry, and there are many tools still to develop.

REFERENCES

1. J. Shendure, E. L. Aiden, The expanding scope of DNA sequencing. *Nat Biotech*. **30**, 1084–1094 (2012).

2. E. A. Winzeler *et al.*, Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. **285**, 901–906 (1999).

3. M. E. Hillenmeyer *et al.*, The chemical genomic portrait of yeast: uncovering a phenotype for all genes. **320**, 362–365 (2008).

4. A. M. Smith *et al.*, Competitive genomic screens of barcoded yeast libraries. (2011), doi:10.3791/2864.

5. C.-H. Ho *et al.*, A molecular barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds. *Nat Biotechnol*. **27**, 369–377 (2009).

6. L. Magtanong *et al.*, Dosage suppression genetic interaction networks enhance functional wiring diagrams of the cell. *Nat Biotechnol*. **29**, 505–511 (2011).

7. C. Payen *et al.*, High-Throughput Identification of Adaptive Mutations in Experimentally Evolved Yeast Populations. *PLoS Genet*. **12**, e1006339 (2016).

8. M. A. Jacobs *et al.*, Comprehensive transposon mutant library of Pseudomonas aeruginosa. *Proc Natl Acad Sci USA*. **100**, 14339–14344 (2003).

9. A. Kumar *et al.*, Large-scale mutagenesis of the yeast genome using a Tn7-derived multipurpose transposon. *Genome Res*. **14**, 1975–1986 (2004).

10. A. A. Caudy *et al.*, A new system for comparative functional genomics of Saccharomyces yeasts. *Genetics*. **195**, 275–287 (2013).

11. M. Baym, L. Shaket, I. A. Anzai, O. Adesina, B. Barstow, Rapid construction of a whole-genome transposon insertion collection for Shewanella oneidensis by Knockout Sudoku. *Nat Commun*. **7**, 13270 (2016).

12. Y. Guo *et al.*, Integration profiling of gene function with dense maps of transposon integration. *Genetics*. **195**, 599–609 (2013).

13. A. H. Michel *et al.*, Functional mapping of yeast genomes by saturated transposition. *Elife*. **6**, e23570 (2017).

14. H. Wang, D. Mayhew, X. Chen, M. Johnston, R. D. Mitra, Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins. *Genome Res*. **21**, 748–755 (2011).

15. H. Wang, D. Mayhew, X. Chen, M. Johnston, R. D. Mitra, “Calling cards” for DNA-binding proteins in mammalian cells. *Genetics*. **190**, 941–949 (2012).

16. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. **337**, 816–821 (2012).

17. P. Mali *et al.*, RNA-guided human genome engineering via Cas9. **339**, 823–826 (2013).

18. L. Cong *et al.*, Multiplex genome engineering using CRISPR/Cas systems. **339**, 819–823 (2013).

19. M. Jinek *et al.*, RNA-programmed genome editing in human cells. *Elife*. **2**, e00471 (2013).

20. W. Jiang, D. Bikard, D. Cox, F. Zhang, L. A. Marraffini, RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat Biotechnol*. **31**, 233–239 (2013).

21. J. E. Dicarlo *et al.*, Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res* (2013), doi:10.1093/nar/gkt135.

22. J. Z. Jacobs, K. M. Ciccaglione, V. Tournier, M. Zaratiegui, Implementation of the CRISPR-Cas9 system in fission yeast. *Nat Commun*. **5**, 5344 (2014).

23. D. J. Dickinson, J. D. Ward, D. J. Reiner, B. Goldstein, Engineering the Caenorhabditis elegans genome using Cas9-triggered homologous recombination. *Nat Methods*. **10**, 1028–1034 (2013).

24. S. J. Gratz *et al.*, Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in Drosophila. *Genetics*. **196**, 961–971 (2014).

25. H. Yang, H. Wang, R. Jaenisch, Generating genetically modified mice using CRISPR/Cas-mediated genome engineering. *Nat Protoc*. **9**, 1956–1968 (2014).

26. S. Khatodia, K. Bhatotia, N. Passricha, S. M. P. Khurana, N. Tuteja, The CRISPR/Cas Genome-Editing Tool: Application in Improvement of Crops. *Front. Plant Sci.* **7**, 569 (2016).

27. O. Shalem *et al.*, Genome-scale CRISPR-Cas9 knockout screening in human cells. **343**, 84–87 (2014).

28. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, Genetic screens in human cells using the CRISPR-Cas9 system. **343**, 80–84 (2014).

29. G. Korkmaz *et al.*, Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol*, 1–10 (2016).

30. N. Rajagopal *et al.*, High-throughput mapping of regulatory DNA. *Nat Biotechnol*. **34**, 167–174 (2016).

31. N. E. Sanjana *et al.*, High-resolution interrogation of functional elements in the noncoding genome. **353**, 1545–1549 (2016).

32. G. M. Findlay, E. A. Boyle, R. J. Hause, J. C. Klein, J. Shendure, Saturation editing of genomic regions by multiplex homology-directed repair. *Nature*. **513**, 120–123 (2014).

33. P. Mali *et al.*, CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol*. **31**, 833–838 (2013).

34. L. A. Gilbert *et al.*, CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* (2013), doi:10.1016/j.cell.2013.06.044.

35. S. Konermann *et al.*, Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*. **517**, 583–588 (2015).

36. S. Kiani *et al.*, Cas9 gRNA engineering for genome editing, activation and repression. *Nat Methods* (2015), doi:10.1038/nmeth.3580.

37. J. G. Zalatan *et al.*, Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell*. **160**, 339–350 (2015).

38. P. I. Thakore *et al.*, Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat Methods*. **12**, 1143–1149 (2015).

39. A. J. Keung, C. J. Bashor, S. Kiriakov, J. J. Collins, A. S. Khalil, Using targeted chromatin regulators to engineer combinatorial and spatial transcriptional regulation. *Cell*. **158**, 110–120 (2014).

40. I. B. Hilton *et al.*, Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat Biotechnol*. **33**, 510–517 (2015).

41. N. A. Kearns *et al.*, Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat Methods*. **12**, 401–403 (2015).

42. D. M. Shechner, E. Hacisuleyman, S. T. Younger, J. L. Rinn, Multiplexable, locus-specific targeting of long RNAs with CRISPR-Display. *Nat Methods* (2015), doi:10.1038/nmeth.3433.

43. B. C. Cunningham, J. A. Wells, High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. **244**, 1081–1085 (1989).

44. V. E. Gray, R. J. Hause, D. M. Fowler, Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions. *Genetics*, genetics.300064.2017 (2017).

45. L. M. Gregoret, R. T. Sauer, Additivity of mutant effects assessed by binomial mutagenesis. *Proc Natl Acad Sci USA*. **90**, 4246–4250 (1993).

46. G. A. Weiss, C. K. Watanabe, A. Zhong, A. Goddard, S. S. Sidhu, Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc Natl Acad Sci USA*. **97**, 8950–8954 (2000).

47. D. M. Fowler *et al.*, High-resolution mapping of protein sequence-function relationships. *Nat Methods*. **7**, 741–746 (2010).

48. A. Melnikov, P. Rogov, L. Wang, A. Gnirke, T. S. Mikkelsen, Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res*. **42**, e112–e112 (2014).

49. M. A. Stiffler, D. R. Hekstra, R. Ranganathan, Evolvability as a Function of Purifying Selection in TEM-1 ?-Lactamase. *Cell*. **160**, 882–892 (2015).

50. J. R. Klesmith, J.-P. Bacik, R. Michalczyk, T. A. Whitehead, Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli. *ACS Synth Biol*. **4**, 1235–1243 (2015).

51. J. O. Kitzman, L. M. Starita, R. S. Lo, S. Fields, J. Shendure, Massively parallel single-amino-acid mutagenesis. *Nat Methods*. **12**, 203–6– 4 p following 206 (2015).

52. A. I. Podgornaia, M. T. Laub, Pervasive degeneracy and epistasis in a protein-protein interface. **347**, 673–677 (2015).

53. M. S. Rich *et al.*, Comprehensive Analysis of the SUL1 Promoter of Saccharomyces cerevisiae. *Genetics*. **203**, 191–202 (2016).

54. L. Brenan *et al.*, Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants. *Cell Reports*. **17**, 1171–1183 (2016).

55. M. B. Doud, J. D. Bloom, Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. *Viruses*. **8**, 155 (2016).

56. B. Thyagarajan, J. D. Bloom, The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *Elife*. **3**, e03300 (2014).

57. M. B. Doud, S. E. Hensley, J. D. Bloom, Complete mapping of viral escape from neutralizing antibodies. *PLoS Pathog.* **13**, e1006271 (2017).

58. O. Ashenberg, J. Padmakumar, M. B. Doud, J. D. Bloom, Deep mutational scanning identifies sites in influenza nucleoprotein that affect viral inhibition by MxA. *PLoS Pathog.* **13**, e1006288 (2017).

59. H. K. Haddox, A. S. Dingens, J. D. Bloom, Experimental Estimation of the Effects of All Amino-Acid Mutations to HIV's Envelope Protein on Viral Replication in Cell Culture. *PLoS Pathog.* **12**, e1006114 (2016).

60. R. T. Hietpas, J. D. Jensen, D. N. A. Bolon, Experimental illumination of a fitness

landscape. *Proceedings of the National Academy of Sciences*. **108**, 7896–7901 (2011).

61. D. Melamed, D. L. Young, C. E. Gamble, C. R. Miller, S. Fields, Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA*. **19**, 1537–1551 (2013).

62. M. W. Dorrity, J. T. Cuperus, J. A. Carlisle, S. Fields, C. Queitsch, Trait specificity mediated by alternative DNA-binding preferences of a single transcription factor in yeast. *bioRxiv*, 117911 (2017).

63. S. Fields, O. Song, A novel genetic system to detect protein-protein interactions. *Nature*. **340**, 245–246 (1989).

64. L. M. Starita *et al.*, Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics*. **200**, 413–422 (2015).

65. J. B. Kinney, A. Murugan, C. G. Callan, E. C. Cox, Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences*. **107**, 9158–9163 (2010).

66. K. S. Sarkisyan *et al.*, Local fitness landscape of the green fluorescent protein. *Nature*. **533**, 397–401 (2016).

67. L. M. Starita *et al.*, Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences*. **110**, E1263–72 (2013).

68. C. E. Tinberg *et al.*, Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*. **501**, 212–216 (2013).

69. J. R. Klesmith, J.-P. Bacik, E. E. Wrenbeck, R. Michalczyk, T. A. Whitehead, Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proceedings of the National Academy of Sciences*. **114**, 2265–2270 (2017).

70. P. Koenig *et al.*, Deep Sequencing-guided Design of a High Affinity Dual Specificity Antibody to Target Two Angiogenic Factors in Neovascular Age-related Macular Degeneration. *J Biol Chem*. **290**, 21773–21786 (2015).

71. P. Koenig *et al.*, Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. *Proceedings of the National Academy of Sciences*. **114**, E486–E495 (2017).

72. R. M. Adams, T. Mora, A. M. Walczak, J. B. Kinney, Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *Elife*. **5**, e23156 (2016).

73. P. A. Romero, T. M. Tran, A. R. Abate, Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences* (2015), doi:10.1073/pnas.1422285112.

74. M. P. Guy *et al.*, Identification of the determinants of tRNA function and susceptibility to rapid tRNA decay by high-throughput in vivo analysis. *Genes Dev*. **28**, 1721–1732 (2014).

75. C. Li, W. Qian, C. J. Maclean, J. Zhang, The fitness landscape of a tRNA gene. **352**, 837–840 (2016).

76. J. D. Buenrostro *et al.*, Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat Biotechnol*. **32**, 562–568 (2014).

77. I. Liachko, R. A. Youngblood, U. Keich, M. J. Dunham, High-resolution mapping, characterization, and optimization of autonomously replicating sequences in yeast. *Genome Res*. **23**, 698–704 (2013).

78. X. F. Kong, X. H. Zhu, Y. L. Pei, D. M. Jackson, M. F. Holick, Molecular cloning, characterization, and promoter analysis of the human 25-hydroxyvitamin D3-1alpha-hydroxylase gene. *Proc Natl Acad Sci USA*. **96**, 6988–6993 (1999).

79. K. Struhl, Regulatory sites for his3 gene expression in yeast. *Nature*. **300**, 285–286 (1982).

80. L. Guarente, R. R. Yocum, P. Gifford, A GAL10-CYC1 hybrid yeast promoter identifies the GAL4 regulatory region as an upstream site. *Proc Natl Acad Sci USA*. **79**, 7410–7414 (1982).

81. W. Chen, K. Struhl, Saturation mutagenesis of a yeast his3 “TATA element”: genetic evidence for a specific TATA-binding protein. *Proc Natl Acad Sci USA*. **85**, 2691–2695 (1988).

82. R. P. Patwardhan *et al.*, High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol*. **27**, 1173–1175 (2009).

83. J. C. Kwasnieski, I. Mogno, C. A. Myers, J. C. Corbo, B. A. Cohen, Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proceedings of the National Academy of Sciences*. **109**, 19498–19503 (2012).

84. A. Melnikov *et al.*, Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol*. **30**, 271–277 (2012).

85. J. Ernst *et al.*, Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat Biotechnol*. **34**, 1180–1190 (2016).

86. R. P. Patwardhan *et al.*, Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol*. **30**, 265–270 (2012).

87. C. D. Arnold *et al.*, Genome-wide quantitative enhancer activity maps identified by

STARR-seq. **339**, 1074–1077 (2013).

88. C. D. Arnold *et al.*, Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat Genet* (2014), doi:10.1038/ng.3009.

89. C. M. Vockley *et al.*, Massively parallel quantification of the regulatory effects of non-coding genetic variation in a human cohort. *Genome Res* (2015), doi:10.1101/gr.190090.115.

90. J. van Arensbergen *et al.*, Genome-wide mapping of autonomous promoter activity in human cells. *Nat Biotechnol*. **35**, 145–153 (2017).

91. F. Inoue *et al.*, A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res*. **27**, 38–52 (2017).

92. S. Kosuri *et al.*, Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proceedings of the National Academy of Sciences*. **110**, 14024–14029 (2013).

93. D. B. Goodman, G. M. Church, S. Kosuri, Causes and effects of N-terminal codon bias in bacterial genes. **342**, 475–479 (2013).

94. E. Sharon *et al.*, Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol*. **30**, 521–530 (2012).

95. S. Lubliner *et al.*, Core promoter sequence in yeast is a major determinant of expression level. *Genome Res* (2015), doi:10.1101/gr.188193.114.

96. S. Dvir *et al.*, Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proceedings of the National Academy of Sciences*. **110**, E2792–801 (2013).

97. O. Shalem *et al.*, Systematic dissection of the sequence determinants of gene 3' end mediated expression control. *PLoS Genet*. **11**, e1005147 (2015).

98. S. F. Levy *et al.*, Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* (2015), doi:10.1038/nature14279.

99. H.-E. C. Bhang *et al.*, Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nat Med* (2015), doi:10.1038/nm.3841.

100. S. H. Naik *et al.*, Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature*. **496**, 229–232 (2013).

101. L. Perié, K. R. Duffy, L. Kok, R. J. de Boer, T. N. Schumacher, The Branching Point in Erythro-Myeloid Differentiation. *Cell*. **163**, 1655–1662 (2015).

102. C. Wu *et al.*, Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. *Cell Stem Cell*. **14**, 486–499 (2014).

103. A. McKenna *et al.*, Whole-organism lineage tracing by combinatorial and cumulative genome editing. **353**, aaf7907–aaf7907 (2016).

104. S. D. Perli, C. H. Cui, T. K. Lu, Continuous genetic recording with self-targeting CRISPR-Cas in human cells. **353**, aag0511–aag0511 (2016).

105. R. Kalhor, P. Mali, G. M. Church, Rapidly evolving homing CRISPR barcodes. *Nat Methods*. **14**, 195–200 (2017).

106. K. L. Frieda *et al.*, Synthetic recording and in situ readout of lineage information in single cells. *Nature*. **541**, 107–111 (2017).

107. J. M. Kebschull *et al.*, High-Throughput Mapping of Single-Neuron Projections by Sequencing of Barcoded RNA. *Neuron*. **91**, 975–987 (2016).

108. I. D. Peikon *et al.*, Using high-throughput barcode sequencing to efficiently map connectomes. *Nucleic Acids Res* (2017), doi:10.1093/nar/gkx292.

109. N. Yachie *et al.*, Pooled-matrix protein interaction screens using Barcode Fusion Genetics. *Mol Syst Biol*. **12**, 863–863 (2016).

110. M. Jaffe, G. Sherlock, S. F. Levy, iSeq: A New Double-Barcode Method for Detecting Dynamic Genetic Interactions in Yeast. *G3 (Bethesda)* (2016), doi:10.1534/g3.116.034207.

111. U. Schlecht, Z. Liu, J. R. Blundell, R. P. St Onge, S. F. Levy, A scalable double-barcode sequencing platform for characterization of dynamic protein-protein interactions. *Nat Commun*. **8**, 15586 (2017).

112. C. A. Hutchison *et al.*, Mutagenesis at a specific position in a DNA sequence. *J Biol Chem*. **253**, 6551–6560 (1978).

113. H. Liu, J. H. Naismith, An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnol.* **8**, 91–91 (2008).

114. P. C. Jain, R. Varadarajan, A rapid, efficient, and economical inverse polymerase chain reaction-based method for generating a site saturation mutant library. *Anal. Biochem.* **449**, 90–98 (2014).

115. E. Firnberg, M. Ostermeier, PFunkel: efficient, expansive, user-defined mutagenesis. *PLoS ONE*. **7**, e52031 (2012).

116. T. A. Kunkel, Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc Natl Acad Sci USA*. **82**, 488–492 (1985).

117. E. E. Wrenbeck *et al.*, Plasmid-based one-pot saturation mutagenesis. *Nat Methods*. **13**, 928–930 (2016).

118. L. Tang *et al.*, Construction of "small-intelligent" focused mutagenesis libraries using well-designed combinatorial degenerate primers. *BioTechniques*. **52**, 149–158 (2012).

119. S. Kille *et al.*, Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS Synth Biol*. **2**, 83–92 (2013).

120. G. Pines *et al.*, Codon compression algorithms for saturation mutagenesis. *ACS Synth Biol*. **4**, 604–614 (2015).

121. R. C. Cadwell, G. F. Joyce, Randomization of genes by PCR mutagenesis. *PCR Methods Appl.* **2**, 28–33 (1992).

122. J. H. Spee, W. M. de Vos, O. P. Kuipers, Efficient random mutagenesis method with adjustable mutation frequency by use of PCR and dITP. *Nucleic Acids Res*. **21**, 777–778 (1993).

123. M. Zaccolo, D. M. Williams, D. M. Brown, E. Gherardi, An Approach to Random Mutagenesis of DNA Using Mixtures of Triphosphate Derivatives of Nucleoside Analogues. *J Mol Biol*. **255**, 589–603 (1996).

124. B. D. Biles, Low-fidelity Pyrococcus furiosus DNA polymerase mutants useful in error-prone PCR. *Nucleic Acids Res*. **32**, e176–e176 (2004).

125. A. R. Oliphant, K. Struhl, An efficient method for generating proteins with altered enzymatic properties: application to beta-lactamase. *Proc Natl Acad Sci USA*. **86**, 9094–9098 (1989).

126. A. R. Oliphant, K. Struhl, Defining the consensus sequences of E.coli promoter elements by random selection. *Nucleic Acids Res*. **16**, 7673–7683 (1988).

127. M. S. Horwitz, L. A. Loeb, Promoters selected from random DNA sequences. *Proc Natl Acad Sci USA*. **83**, 7405–7409 (1986).

128. A. B. Rosenberg, R. P. Patwardhan, J. Shendure, G. Seelig, Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. *Cell*. **163**, 698–711 (2015).

129. J. T. Cuperus *et al.*, Deep Learning Of The Regulatory Grammar Of Yeast 5′ Untranslated Regions From 500,000 Random Sequences. *bioRxiv* (2017), doi:10.1101/137547.

130. G. Haller *et al.*, Massively parallel single-nucleotide mutagenesis using reversibly terminated inosine. *Nat Methods*. **13**, 923–924 (2016).

131. F. Sanger, S. Nicklen, A. R. Coulson, DNA sequencing with chain-terminating

inhibitors. *Proc Natl Acad Sci USA*. **74**, 5463–5467 (1977).

132. J. Tian *et al.*, Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*. **432**, 1050–1054 (2004).

133. E. M. LeProust *et al.*, Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res*. **38**, 2522–2540 (2010).

134. S. Kosuri *et al.*, Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat Biotechnol*. **28**, 1295–1299 (2010).

135. J. C. Klein *et al.*, Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res*. **44**, e43–e43 (2016).

136. N. D. Taylor *et al.*, Engineering an allosteric transcription factor to respond to new ligands. *Nat Methods* (2015), doi:10.1038/nmeth.3696.

137. H. H. Wang *et al.*, Programming cells by multiplex genome engineering and accelerated evolution. *Nature*. **460**, 894–898 (2009).

138. S. Higgins, S. Ouonkap Yimga, D. Savage, Rapid and Programmable Protein Mutagenesis Using Plasmid Recombineering. *bioRxiv*, 124271 (2017).

139. W. P. Stemmer, DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc Natl Acad Sci USA*. **91**, 10747–10751 (1994).

140. W. P. Stemmer, Rapid evolution of a protein in vitro by DNA shuffling. *Nature*. **370**, 389–391 (1994).

141. A. Crameri, S. A. Raillard, E. Bermudez, W. P. Stemmer, DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature*. **391**, 288–291 (1998).

142. A. Crameri, E. A. Whitehorn, E. Tate, W. P. Stemmer, Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nat Biotechnol*. **14**, 315–319 (1996).

143. H. Zhao, F. H. Arnold, Optimization of DNA shuffling for high fidelity recombination. *Nucleic Acids Res*. **25**, 1307–1308 (1997).

144. N. Schürmann, L. G. Trabuco, C. Bender, R. B. Russell, D. Grimm, Molecular dissection of human Argonaute proteins by DNA shuffling. *Nat. Struct. Mol. Biol.* **20**, 818–826 (2013).

145. H. Zhao, F. H. Arnold, Functional and nonfunctional mutations distinguished by random recombination of homologous genes. *Proc Natl Acad Sci USA*. **94**, 7997–8000 (1997).

146. K. D. Belsare *et al.*, A Simple Combinatorial Codon Mutagenesis Method for Targeted Protein Engineering. *ACS Synth Biol*. **6**, 416–420 (2017).

147. J. A. Vidigal, A. Ventura, Rapid and efficient one-step generation of paired gRNA CRISPR-Cas9 libraries. *Nat Commun*. **6**, 8083 (2015).

148. M. Gasperini *et al.*, Paired CRISPR/Cas9 guide-RNAs enable high-throughput deletion scanning (ScanDel) of a Mendelian disease locus for functionally critical non-coding elements. *bioRxiv*, 092445 (2016).

149. J. P. Shen *et al.*, Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. *Nat Methods*. **14**, 573–576 (2017).

150. J. Eid *et al.*, Real-time DNA sequencing from single polymerase molecules. **323**, 133–138 (2009).

151. C. L. Ip *et al.*, MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research*. **4**, 1075 (2015).

152. J. B. Hiatt, R. P. Patwardhan, E. H. Turner, C. Lee, J. Shendure, Parallel, tag-directed assembly of locally derived short sequence reads. *Nat Methods*. **7**, 119–122 (2010).

153. C. A. Kowalsky *et al.*, High-resolution sequence-function mapping of full-length proteins. *PLoS ONE*. **10**, e0118193 (2015).

154. P. M. Magwene, J. H. Willis, J. K. Kelly, The statistics of bulk segregant analysis using next generation sequencing. *PLoS Comput Biol*. **7**, e1002255 (2011).

155. A. D. Ellington, J. W. Szostak, In vitro selection of RNA molecules that bind specific ligands. *Nature*. **346**, 818–822 (1990).

156. A. M. Levin, G. A. Weiss, Optimizing the affinity and specificity of proteins with molecular display. *Mol Biosyst*. **2**, 49–57 (2006).

157. K. A. Matreyek, J. J. Stephany, D. M. Fowler, A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res*. **45**, e102–e102 (2017).

158. K. Berns *et al.*, A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature*. **428**, 431–437 (2004).

159. K. Nishida *et al.*, Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science*. **353**, aaf8729–aaf8729 (2016).

160. A. C. Komor, Y. B. Kim, M. S. Packer, J. A. Zuris, D. R. Liu, Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*. **533**, 420–424 (2016).

161. H. Nishimasu *et al.*, Crystal structure of Cas9 in complex with guide RNA and target

DNA. *Cell*. **156**, 935–949 (2014).

162. H. A. Rees *et al.*, Improving the DNA specificity and applicability of base editing through protein engineering and protein delivery. *Nat Commun*. **8**, 15790 (2017).

163. Y. B. Kim *et al.*, Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nat Biotechnol*. **35**, 371–376 (2017).

164. T. A. Lee *et al.*, Dissection of combinatorial control by the Met4 transcriptional complex. *Mol Biol Cell*. **21**, 456–469 (2010).

165. A. A. Petti, R. S. McIsaac, O. Ho-Shing, H. J. Bussemaker, D. Botstein, Combinatorial control of diverse metabolic and physiological functions by transcriptional regulators of the yeast sulfur assimilation pathway. *Mol Biol Cell*. **23**, 3008–3024 (2012).

166. U. Nagalakshmi *et al.*, The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. **320**, 1344–1349 (2008).

167. J. R. Hesselberth *et al.*, Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods*. **6**, 283–289 (2009).

168. H. S. Rhee, B. F. Pugh, Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*. **483**, 295–301 (2012).

169. K. Struhl, Molecular mechanisms of transcriptional regulation in yeast. *Annu. Rev. Biochem.* **58**, 1051–1077 (1989).

170. R. D. Gietz, R. H. Schiestl, High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat Protoc*. **2**, 31–34 (2007).

171. D. Gresham *et al.*, The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet*. **4**, e1000303 (2008).

172. R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. **27**, 863–864 (2011).

173. J. Zhu, M. Q. Zhang, SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*. **15**, 607–611 (1999).

174. K. D. MacIsaac *et al.*, An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*. **7**, 113 (2006).

175. S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, W. S. Noble, Quantifying similarity between motifs. *Genome Biol*. **8**, R24 (2007).

176. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics*. **27**, 1017–1018 (2011).

177. A. D. Basehoar, S. J. Zanton, B. F. Pugh, Identification and distinct regulation of yeast

TATA box-containing genes. *Cell*. **116**, 699–709 (2004).

178. M. Kellis, N. Patterson, M. Endrizzi, B. Birren, E. S. Lander, Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*. **423**, 241–254 (2003).

179. C. Payen *et al.*, Empirical determinants of adaptive mutations in yeast experimental evolution. *bioRxiv* (2015), p. 014068.

180. A. W. Miller, C. Befort, E. O. Kerr, M. J. Dunham, Design and use of multiplexed chemostat arrays., e50262 (2013).

181. C. Payen *et al.*, The Dynamics of Diverse Segmental Amplifications in Populations of Saccharomyces cerevisiae Adapting to Strong Selection. *G3 (Bethesda)* (2014), doi:10.1534/g3.113.009365.

182. S. J. Dowell, J. S. Tsang, J. Mellor, The centromere and promoter factor 1 of yeast contains a dimerisation domain located carboxy-terminal to the bHLH domain. *Nucleic Acids Res*. **20**, 4229–4236 (1992).

183. T. Siggers, M. H. Duyzend, J. Reddy, S. Khan, M. L. Bulyk, Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol Syst Biol*. **7**, 555 (2011).

184. P. L. Blaiseau, A. D. Isnard, Y. Surdin-Kerjan, D. Thomas, Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Mol Cell Biol*. **17**, 3640–3648 (1997).

185. S. J. Maerkl, S. R. Quake, A systems approach to measuring the binding energy landscapes of transcription factors. **315**, 233–237 (2007).

186. V. Measday *et al.*, Systematic yeast synthetic lethal and synthetic dosage lethal screens identify genes required for chromosome segregation. *Proc Natl Acad Sci USA*. **102**, 13956–13961 (2005).

187. B. Mai, L. Breeden, Xbp1, a stress-induced transcriptional repressor of the Saccharomyces cerevisiae Swi4/Mbp1 family. *Mol Cell Biol*. **17**, 6491–6501 (1997).

188. D. R. Morris, A. P. Geballe, Upstream open reading frames as regulators of mRNA translation. *Mol Cell Biol*. **20**, 8635–8642 (2000).

189. Y. Yun, T. M. A. Adesanya, R. D. Mitra, A systematic study of gene expression variation at single-nucleotide resolution reveals widespread regulatory roles for uAUGs. *Genome Res*. **22**, 1089–1097 (2012).

190. F. W. Smith, M. J. Hawkesford, I. M. Prosser, D. T. Clarkson, Isolation of a cDNA from Saccharomyces cerevisiae that encodes a high affinity sulphate transporter at the plasma membrane. *Mol Gen Genet*. **247**, 709–715 (1995).

191. A. H. Yona *et al.*, Chromosomal duplication is a transient evolutionary solution to stress. *Proceedings of the National Academy of Sciences*. **109**, 21010–21015 (2012).

192. H. Hendrickson, E. S. Slechta, U. Bergthorsson, D. I. Andersson, J. R. Roth, Amplification-mutagenesis: evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification. *Proc Natl Acad Sci USA*. **99**, 2164–2169 (2002).

193. B. Titz *et al.*, Transcriptional activators in yeast. *Nucleic Acids Res*. **34**, 955–967 (2006).

194. N.-Y. Su, I. Ouni, C. V. Papagiannis, P. Kaiser, A dominant suppressor mutation of the met30 cell cycle defect suggests regulation of the Saccharomyces cerevisiae Met4-Cbf1 transcription complex by Met32. *J Biol Chem*. **283**, 11615–11624 (2008).

195. L. Cormier, R. Barbey, L. Kuras, Transcriptional plasticity through differential assembly of a multiprotein activation complex. *Nucleic Acids Res*. **38**, 4998–5014 (2010).

196. E. Carrillo *et al.*, Characterizing the roles of Met31 and Met32 in coordinating Met4-activated transcription in the absence of Met30. *Mol Biol Cell*. **23**, 1928–1942 (2012).

197. R. S. McIsaac, A. A. Petti, H. J. Bussemaker, D. Botstein, Perturbation-based analysis and modeling of combinatorial regulation in the yeast sulfur assimilation pathway. *Mol Biol Cell*. **23**, 2993–3007 (2012).

198. P. T. Monteiro *et al.*, YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res*. **36**, D132–6 (2008).

199. D. Abdulrehman *et al.*, YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic Acids Res*. **39**, D136–40 (2011).

200. M. C. Teixeira *et al.*, The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res*. **34**, D446–51 (2006).

201. M. C. Teixeira *et al.*, The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. *Nucleic Acids Res*. **42**, D161–6 (2014).

202. C. G. de Boer, T. R. Hughes, YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res*. **40**, D169–79 (2012).

203. H. Cherest *et al.*, Molecular characterization of two high affinity sulfate transporters in Saccharomyces cerevisiae. *Genetics*. **145**, 627–635 (1997).

204. D. Libkind *et al.*, Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. *Proceedings of the National Academy of Sciences*. **108**, 14539–

14544 (2011).

205. A. S. Karim, K. A. Curran, H. S. Alper, Characterization of plasmid burden and copy number in Saccharomyces cerevisiae for optimization of metabolic engineering applications. *FEMS Yeast Res*. **13**, 107–116 (2013).

206. D. Shortle, P. Novick, D. Botstein, Construction and genetic characterization of temperature-sensitive mutant alleles of the yeast actin gene. *Proc Natl Acad Sci USA*. **81**, 4889–4893 (1984).

207. S. Wilkening *et al.*, An evaluation of high-throughput approaches to QTL mapping in Saccharomyces cerevisiae. *Genetics*. **196**, 853–865 (2014).

208. B. Vogelstein *et al.*, Cancer genome landscapes. **339**, 1546–1558 (2013).

209. F. Supek, B. Miñana, J. Valcárcel, T. Gabaldón, B. Lehner, Synonymous mutations frequently act as driver mutations in human cancers. *Cell*. **156**, 1324–1335 (2014).

210. Z. E. Sauna, C. Kimchi-Sarfaty, Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet*. **12**, 683–691 (2011).

211. L. Cartegni, S. L. Chew, A. R. Krainer, Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Publishing Group*. **3**, 285–298 (2002).

212. P. Brest *et al.*, A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat Genet*. **43**, 242–245 (2011).

213. J. J. Gartner *et al.*, Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proceedings of the National Academy of Sciences*. **110**, 13481–13486 (2013).

214. I. M. Silverman *et al.*, RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. *Genome Biol*. **15**, R3 (2014).

215. D. M. Fowler, S. Fields, Deep mutational scanning: a new style of protein science. *Nat Methods*. **11**, 801–807 (2014).

216. M. Kircher *et al.*, A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. **46**, 310–315 (2014).

217. I. A. Adzhubei *et al.*, A method and server for predicting damaging missense mutations. *Nat Methods*. **7**, 248–249 (2010).

218. G. Lamolle, M. Marin, F. Alvarez-Valin, Silent mutations in the gene encoding the p53 protein are preferentially located in conserved amino acid positions and splicing enhancers. *Mutat Res*. **600**, 102–112 (2006).

219. S. Alberti, A. D. Gitler, S. Lindquist, A suite of Gateway cloning vectors for high-throughput genetic analysis in Saccharomyces cerevisiae. *Yeast*. **24**, 913–919 (2007).

220. D. G. Gibson *et al.*, Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*. **6**, 343–345 (2009).

221. D. M. Fowler, C. L. Araya, W. Gerard, S. Fields, Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics*. **27**, 3430–3431 (2011).

222. R. Lorenz *et al.*, ViennaRNA Package 2.0. *Algorithms Mol Biol*. **6**, 26 (2011).

223. K. C. Miranda *et al.*, A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*. **126**, 1203–1217 (2006).

224. A. Kozomara, S. Griffiths-Jones, miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*. **42**, D68–73 (2014).

225. M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. **536**, 285–291 (2016).

226. S. A. Forbes *et al.*, COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*. **43**, D805–11 (2015).

227. E. R. Fearon, B. Vogelstein, A genetic model for colorectal tumorigenesis. *Cell*. **61**, 759–767 (1990).

228. A. H. Berger, P. P. Pandolfi, Haplo-insufficiency: a driving force in cancer. *J. Pathol.* **223**, 137–146 (2011).

229. S. Senturk *et al.*, p53Ψ is a transcriptionally inactive p53 isoform able to reprogram cells toward a metastatic-like state. *Proceedings of the National Academy of Sciences*. **111**, E3287–96 (2014).

230. N. H. Shirole *et al.*, TP53 exon-6 truncating mutations produce separation of function isoforms with pro-tumorigenic functions. *Elife*. **5**, e17929 (2016).

231. W. F. Mueller, L. S. Z. Larsen, A. Garibaldi, G. W. Hatfield, K. J. Hertel, The Silent Sway of Splicing by Synonymous Substitutions. *J Biol Chem*. **290**, 27700–27711 (2015).

232. S. Venkataram *et al.*, Development of a Comprehensive Genotype-to-Fitness Map of Adaptation-Driving Mutations in Yeast. *Cell*. **166**, 1585–1596.e22 (2016).

233. T. R. Cech, Self-splicing of group I introns. *Annu. Rev. Biochem.* **59**, 543–568 (1990).

234. B. A. Sullenger, T. R. Cech, Ribozyme-mediated repair of defective mRNA by targeted, trans-splicing. *Nature*. **371**, 619–622 (1994).

235. R. S. Sikorski, P. Hieter, A system of shuttle vectors and yeast host strains designed for

efficient manipulation of DNA in Saccharomyces cerevisiae. *Genetics*. **122**, 19–27 (1989).

236. D. Mumberg, R. Müller, M. Funk, Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene*. **156**, 119–122 (1995).

237. M. S. Longtine *et al.*, Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae. *Yeast*. **14**, 953–961 (1998).

238. B. G. Ayre, U. Köhler, H. M. Goodman, J. Haseloff, Design of highly specific cytotoxins by using trans-splicing ribozymes. *Proc Natl Acad Sci USA*. **96**, 3507–3512 (1999).

239. B. G. Ayre, U. Köhler, R. Turgeon, J. Haseloff, Optimization of trans-splicing ribozyme efficiency and specificity by in vivo genetic selection. *Nucleic Acids Res*. **30**, e141 (2002).

240. J. T. Jones, S. W. Lee, B. A. Sullenger, Tagging ribozyme reaction sites to follow trans-splicing in mammalian cells. *Nat Med*. **2**, 643–648 (1996).

241. J. T. Jones, B. A. Sullenger, Evaluating and enhancing ribozyme reaction efficiency in mammalian cells. *Nat Biotech*. **15**, 902–905 (1997).

242. C. S. Rogers, C. G. Vanoye, B. A. Sullenger, A. L. George, Functional repair of a mutant chloride channel using a trans-splicing ribozyme. *J. Clin. Invest.* **110**, 1783–1789 (2002).

243. A. W. Teunissen, H. Y. Steensma, Review: the dominant flocculation genes of Saccharomyces cerevisiae constitute a new subtelomeric gene family. *Yeast*. **11**, 1001–1013 (1995).

244. K. J. Verstrepen, F. M. Klis, Flocculation, adhesion and biofilm formation in yeasts. *Mol. Microbiol.* **60**, 5–15 (2006).

245. G. Liti *et al.*, Population genomics of domestic and wild yeasts. *Nature*. **458**, 337–341 (2009).

246. E. A. Hope, M. J. Dunham, Ploidy-regulated variation in biofilm-related phenotypes in natural isolates of Saccharomyces cerevisiae. *G3 (Bethesda)*. **4**, 1773–1786 (2014).

247. J. W. Wenger, K. Schwartz, G. Sherlock, Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from Saccharomyces cerevisiae. *PLoS Genet*. **6**, e1000942 (2010).

248. J. E. Ness *et al.*, Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently. *Nat Biotechnol*. **20**, 1251–1255 (2002).

249. J. Shendure, S. Fields, Massively Parallel Genetics. *Genetics*. **203**, 617–619 (2016).

250. A. H. Kachroo *et al.*, Systematic humanization of yeast genes reveals conserved functions and genetic modularity. **348**, 921–925 (2015).

251. A. Hamza *et al.*, Complementation of Yeast Genes with Human Genes as an Experimental Platform for Functional Testing of Human Genetic Variants. *Genetics*. **201**, 1263–1274 (2015).

252. S. Sun *et al.*, An extended set of yeast-based functional assays accurately identifies human disease mutations. *Genome Res*. **26**, 670–680 (2016).

253. E. Z. Macosko *et al.*, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. **161**, 1202–1214 (2015).

254. J. Cao *et al.*, Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. *bioRxiv*, 104844 (2017).

255. X. Shu *et al.*, A genetically encoded tag for correlated light and electron microscopy of intact cells, tissues, and organisms. *PLoS Biol*. **9**, e1001041 (2011).

256. R. Chen *et al.*, A Barcoding Strategy Enabling Higher-Throughput Library Screening by Microscopy. *ACS Synth Biol*. **4**, 1205–1216 (2015).

257. J.-H. Lee *et al.*, Highly multiplexed subcellular RNA sequencing in situ. **343**, 1360–1363 (2014).

258. E. Lubeck, A. F. Coskun, T. Zhiyentayev, M. Ahmad, L. Cai, Single-cell in situ RNA profiling by sequential hybridization. *Nat Methods*. **11**, 360–361 (2014).

259. F. Winston, C. Dollard, S. L. Ricupero-Hovasse, Construction of a set of convenient Saccharomyces cerevisiae strains that are isogenic to S288C. *Yeast*. **11**, 53–55 (1995).

260. C. B. Brachmann *et al.*, Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast*. **14**, 115–132 (1998).

261. A. H. Tong *et al.*, Systematic genetic analysis with ordered arrays of yeast deletion mutants. **294**, 2364–2368 (2001).

262. P. James, J. Halladay, E. A. Craig, Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. *Genetics*. **144**, 1425–1436 (1996).

263. P. Uetz *et al.*, A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*. **403**, 623–627 (2000).

Chapter 7. Appendices

APPENDIX A

Table 7.2. Oligonucleotides used in Chapter 2

Oligo#	Oligo Name	Oligo Sequence
266	pRS416-SacI_chr2:792107_R	TCACTAAAGGGAACAAAAGCTGGATTCGGCTCAAGCTCAAGTTAACTTCTAGACC
283	pRS-Barcode@HindIII	GGCCCCCCCTCGAGGTCGACGGTATCGATANNNNNNNNNNNNAGCTTGATATCGGGATCC
294	pRS416-RI+HI_chr2:788548_F	ACGGTATCGATAAGCTTGATATCGGGATCCGACCGACAGTGCTGAACTGTGAG
295	pRS416-RI+HI_chr2:788742_F	ACGGTATCGATAAGCTTGATATCGGGATCCCGCCACCTCGAGTGCAC
296	fill-in_283_pMR002	CTCACAGTTCAGCACTGTCGGTCGGATCCCGATATCAAGCT
297	SUL1p-mut_R	AGCATCCTCCTGATTATGCACATATTCAGTCGAGCTCTTACGTGACATATTCTTTCCGAG
314	pRS416-RI+HI_chr2:788643_F	GACGGTATCGATAAGCTTGATATCGGGATCCGTGAATAAGATGTGGCTGTGATAACC
315	pRS416-RI+HI_chr2:788845_F	GACGGTATCGATAAGCTTGATATCGGGATCCGCAAATCGGAATTTGAGTCACAGATC
316	pRS416-RI+HI_chr2:788943_F	GACGGTATCGATAAGCTTGATATCGGGATCCTTTGAACACTTCTACCTGTTCATGTC
317	pRS416-RI+HI_chr2:789043_F	GACGGTATCGATAAGCTTGATATCGGGATCCGTGAAGTAAAATGTGTTGTAATGCACATGG
318	pRS416-RI+HI_chr2:789130_F	GACGGTATCGATAAGCTTGATATCGGGATCCGTGAAACCGATTATATAAAAGTATATTAGCTGAC
319	pRS416-RI+HI_chr2:789211_F	GACGGTATCGATAAGCTTGATATCGGGATCCCCCTGCAGAATACTCGGAAAGAAT
325	pRS416-HindIIIBC-seqF	GGGCCCCCCCTCGAGGTCGACGGTATCGATA

Oligo#	Oligo Name	Oligo Sequence
326	pMR002-seqR	TTCCCTATATGTGCGGTATTCTGATTCAAATACTTCGATATCAGCATCCTCCTGATTATGC ACATATTCAGTCGAGCTCTTACGTGACAT
327	P5-M13F	AATGATACGGCGACCACCGAGATCTACACTGTAAAACGACGGCCAGT
328	P7_SUL1-109_R	CAAGCAGAAGACGGCATACGAGATCCTCATCACCATTGTGAAGTCCGTCT
344	nexV2_SUL1-109_R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTCATCACCATTGTGAAGTCCGTCT
380	SUL1p_6mut_F1	TCTCGCTGGARCCACAGTGYGGCTTTGCAATTTTGCAAATCGGAATTTGAGYCACAGATC CCAGAAAAACKCCACACCTTCCCCACGCAG
381	SUL1p_6mut_R1	CACAGTGCATAATTTCAKATGACAGTGTTCGAGAAAAGACATGAACAGG
382	SUL1p_6mut_F2	GTGCACTTTTTTTAATAAAGATCWCGTGTAATTGTCCAAATCTGACTTTT
387	SUL1p_6mut_seqR	GGGTCTAATTTGACAACTTGTTCGTTATCGCTTGCTGCGTGGGGAAGGTGTGG
395	P5_X_M13_F[1]	AATGATACGGCGACCACCGAGATCTACAC[INDEX]TGTAAAACGACGGCCAGT
442	SUL1p_6mut_seqF2	CGATAAGCTTGATATCGGGATCCCGCCACCTCGAGTGCACTTTTTTTAATAAAGATC
443	SUL1p_6mut_seqI2	ATTAGACCCATAATAATTTTGAACACTTCTACCTGTTCATGTCTTTTCTCGAACACT
517	n.-353T>G	AGATCCCAGAAAAACGCCACACCTTCCCCAC
518	n.-372T>C	TCGGAATTTGAGCCACAGATCCCAG
519	n.-246T>G	GTCTTTTCTCGAACACTGTCATGTGAAATTATGCA
520	n.-404C>T	GGAACCACAGTGTGGCTTTGCAATT
521	n.-413A>G	AGTCTCGCTGGAGCCACAGTGCGGC
522	n.-458T>A	GTGCACTTTTTTTAATAAAGATCACGTGTAATTGTC
603	-33A>G	ggtcaatagctttaaaaataGaaataaatccctgcag
604	-400T>C	cagtgcggcCttgcaattttgc
605	-348A>T	gaaaaactccacTccttccccac
678	pSUL1::URA3_F	TCCCCCCACTTTTTTGTGTTACCGCCACCTCGAGTGCACgattcggtaatctccgaacag
679	pSUL1::URA3_R	ACATATTCAGTCGAGCTCTTACGTGACATgggtaataactgatataattaaattgaagct
680	pSUL1::URA3_F2	CACAAACCAATGAAAAAAGGCTTCAGAAAACTTGCTGTAAAGTCCCCCCACTTTTTTGTG
681	pSUL1::URA3_R2	attcaaatacttcgatatcagcatcctcctgattatgcacatattcagtcgagctcttac
685	pSUL1_687bp_F	gaccgacagtgctgaactgtgag
PGO40	SUL1::NatMX4_F	TTGAAACTAGATATTCTACTTGACACTAAACTTTTTTTGTATTTCGTGGACAGATTGTACT GAGAGTGCACCATA

Oligo#	Oligo Name	Oligo Sequence
PGO41	SUL1::NatMX4_R	TGAAAACGGCCCTTTTCGTGTACTAATGCTTTCAGGAAATTATAAAATTTCCTTACGCATC TGTGCGGTATTTCA
NexV2ad2_N	NexV2ad2_N	CAAGCAGAAGACGGCATACGAGAT[INDEX]GTCTCGTGGGCTCGGAGATGTGTATAAGA GACAG
M13F	M13F	TGTAAAACGACGGCCAGT
P5	P5	AATGATACGGCGACCACCGAGATCTACAC

Table 7.3. Yeast strains used in Chapter 2

Name	Mating type	Genotype	Reference
FY2	alpha	ura3-52	(*259*)
FY3	a	ura3-52	(*259*)
FY4	a	prototroph	(*259*)
YMD3017	a	SUL1::NatMX4, ura3-52 Gal+	This study
YMR002	a	FY3 P_{SUL1}::URA3 ura3-52	This study
YMD3018	alpha	sul1::NatMX	This study
YMR008	a	P_{SUL1}-246T>G ura3-52	This study
YMR009	a	P_{SUL1}-404C>T ura3-52	This study
YMR010	a	P_{SUL1}-458T>A ura3-52	This study
YMR011	a	P_{SUL1}-400T>C ura3-52	This study
YMR012	a	P_{SUL1}-348A>T ura3-52	This study
YMR017	a	P_{SUL1}-353T>G ura3-52	This study
YMR018	a	P_{SUL1}-372T>C ura3-52	This study
YMR019	a	P_{SUL1}-413A>G ura3-52	This study
YMR020	a	P_{SUL1}-5mut ura3-52	This study
YMR022	a	P_{SUL1}-246T>G	This study
YMR024	a	P_{SUL1}-458T>A	This study
YMR025	a	P_{SUL1}-400T>C	This study
YMR026	a	P_{SUL1}-348A>T	This study
YMR027	a	P_{SUL1}-353T>G	This study
YMR028	a	P_{SUL1}-372T>C	This study
YMR029	a	P_{SUL1}-413A>G	This study
YMR030	a	P_{SUL1}-5mut	This study
YMD1214	a	ho::kanMX-eGFP	(*7*)
YMD1139	a	pFA6a-TEF2Pr-eGFP-ADH1-Primer-NATMX4	D. Breslow

APPENDIX B

Table 7.4. Oligonucleotides used in Chapter 3

Oligonucleotide	Sequence
TP53_Exon5,6,7_F01	TCTGTCTCCTTCCTCTTCCTACA
TP53_Exon5,6,7_R01	GGGTCAGAGGCAAGCAGA
TP53_Exon5expand1_F01	TTGCATTCTGGGACAGCCAAGTCTGTGACTTGCACGTACTCCCCTGCCCTCAACA
TP53_Exon7expand1_R01	CTCAAAGCTGTTCCGTCCCAGTAGATTACCACTGGAGTCTTCCAGTGTGATGA
TP53_Exon5expand2_F01	CAGGGCAGCTACGGTTTCCGTCTGGGCTTCTTGCATTCTGGGACAGCC
TP53_Exon7expand2_R01	CGGTCTCTCCCAGGACAGGCACAAACACGCACCTCAAAGCTGTTCCGTCCC
TP53_Exon5expand3_F01	GCCACCATGGGCGTCCCTTCCCAGAAAACCTACCAGGGCAGCTACGGTTTC
TP53_Exon7expand3_R01	TTTCTTGCGGAGATTCTCTTCCTCTGTGCGCCGGTCTCTCCCAGGACAGG
TP53_Exon5E_cloning_F01	CAGCCTCCGGACTCTAGCGTTTAAACTTAAGCCACCAUGGGCGTCCCTTC
TP53_Exon7E_cloning_R01	TCCACCTGAGCCTCCAGAGCCACCTTTCTTGCGGAGATTCTCTTC
GFP_cloning_F01	GGTGGCTCTGGAGGCTCAGGTGGATCTAAAGGTGAAGAATTATTCACTGG
GFP_cloning_R01	TCTAGCTCGAGGCGGCCGCGTCGACGGTACTTTGTACAATTCATCCATACCATG
TP53dopedoligo_RE_R02	AACCCCTCCTCCCAGGGATCCCAGTTGCAAACCAGAC
TP53dopedoligo_RE_F03	TTGCCCAGGGTCCCCATGCATCTGATTCCTCACTGATTGCT
TP53pcDNA5_BamHI_F02	AGCAAGTGTTTTCTGAATTAGGATCCCTGGGAGGAGGGGTTAAGG
TP53pcDNA5_NsiI_R01	ATGCATGGGGACCCTGGGCAAC
TP53_libraryamp_R01	CCTCCCAGAGACCCCAGTT
TP53dopedoligo_RE_R02	AACCCCTCCTCCCAGGGATCCCAGTTGCAAACCAGAC
TP53dopedoligo_RE_F03	TTGCCCAGGGTCCCCATGCATCTGATTCCTCACTGATTGCT
doped oligo	GGCCTCTGATTCCTCACTGATTGCTCTTAGGtCTgGCcCCtCCtCAZCAFCTtATPCGa GTgGAXGGaAAFTTZCGtGTgGAZTAFTTZGAFGAPAGXAAPACtTTFCGaCAFAGFG TgGTgGTgCCcTAFGAZCCgCCtGAZGTCTGGTTTGCAACTGGGGTCTCTGGGAGG (see legend for details)
pcDNA5/FRT_seq_F01	CCTCCGGACTCTAGCGTTTA
pcDNA5/FRT_seq_R01	GTTTAAACGGGCCCTCTAGC
TP53_exon6amp_F01	TGATTCCTCACTGATTGCTCTTAG
TP53_exon6amp_R01	CCACCCTTAACCCCTCCTCC
TP53_seqadapter_F1	AATGATACGGCGACCACCGAGATCTACACCTCCGCCTAACCCGAGTCCACCCG TCCCNNNNTGATTCCTCACTGATTGCTCTTAG
TP53_seqadapter2_R1	CAAGCAGAAGACGGCATACGAGATACCACACCACCCTTAACCCCTCCTCC
TP53_seqadapter2_R2	CAAGCAGAAGACGGCATACGAGATAAATTTCCACCCTTAACCCCTCCTCC

Oligonucleotide	Sequence
TP53_seqadapter2_R3	CAAGCAGAAGACGGCATACGAGATAAGGAACCACCCTTAACCCCTCCTCC
TP53_seqadapter2_R4	CAAGCAGAAGACGGCATACGAGATTTCCTTCCACCCTTAACCCCTCCTCC
TP53_libraryseq_F01	GCCTAACCCGAGTCCACCCGTCCC
TP53_libraryseq_R02	CCACCCTTAACCCCTCCTCCCAGGGATCC
TP53_indexseq_R02	GGGTCCCTGGGAGGAGGGGTTAAGGGTGG
TP53E_T2A_F1	TGCTCTTAGGACTGGCCCCTC
TP53E_T2A_R1	ATCAGTGAGGAATCAGAGGCC
TP53E_T2C_F1	TGCTCTTAGGcCTGGCCCCTC
TP53E_T2C_R1	ATCAGTGAGGAATCAGAGGCCTG
TP53E_T2G_F1	TGCTCTTAGGgCTGGCCCCTC
TP53E_G5A_F1	TCTTAGGTCTaGCCCCTCCTC
TP53E_G5A_R1	GCAATCAGTGAGGAATCAG
TP53E_G5C_F1	TCTTAGGTCTcGCCCCTCCTC
TP53E_G5T_F1	TCTTAGGTCTtGCCCCTCCTC
TP53E_G32A_F1	TTATCCGAGTaGAAGGAAATTTGCGTGT
TP53E_G32A_R1	GATGCTGAGGAGGGGCCA
TP53E_G32C_F1	TTATCCGAGTcGAAGGAAATTTGCGT
TP53E_G32T_F1	TTATCCGAGTtGAAGGAAATTTGCGT
TP53E_G50A_F1	ATTTGCGTGTaGAGTATTTGGATG
TP53E_G50A_R1	TTCCTTCCACTCGGATAAGATG
TP53E_G50C_F1	ATTTGCGTGTcGAGTATTTGGATG
TP53E_G50T_F1	ATTTGCGTGTtGAGTATTTGGATG
TP53E_T47G_F1	GAAATTTGCGgGTGGAGTATTTG
TP53E_T47G_R1	CTTCCACTCGGATAAGATGCTG
TP53E_G89A_F1	GACATAGTGTaGTGGTGCCCTATG
TP53E_G89A_R1	GAAAAGTGTTTCTGTCATCCAAATAC
TP53E_G92A_F1	ATAGTGTGGTaGTGCCCTATGAGCC
TP53E_G92A_R1	GTCGAAAAGTGTTTCTGTCATCC
TP53E_G95A_F1	GTGTGGTGGTaCCCTATGAGC
TP53E_G95A_R1	TATGTCGAAAAGTGTTTCTGTCA
TP53E_G95T_F1	GTGTGGTGGTtCCCTATGAGC
TP53E_G107C_F1	CCTATGAGCCcCCTGAGGTCT
TP53E_G107C_R1	GCACCACCACACTATGTCGA
TP53E_G59A_F1	TGGAGTATTTaGATGACAGAAACAC
ß	CACGCAAATTTCCTTCCAC

Oligonucleotide	Sequence
TP53E_A38T_F1	GAGTGGAAGGtAATTTGCGTG
TP53E_A38T_R1	GGATAAGATGCTGAGGAG

For the doped oligonucleotide, lowercase a indicates 97% A, 1% G, 1% C and 1% T; Lowercase g indicates 97% G, 1% A, 1% C and 1% T; Lowercase c indicates 97% C, 1% A, 1% G and 1% T; Lowercase t indicates 97% T, 1% A, 1% G and 1% C; X indicates 97% A and 3% G; Z indicates 97% G and 3% A; P indicates 97% C and 3% T; F indicates 97% T and 3% C.

APPENDIX C

Table 7.5. Oligonucleotides used in Chapter 4

Oligo#	Oligo Name	Oligo Sequence
1029	LEU2::KanMX-switch_R	acgaaactatatacgcaatctacatacatttatcaagaag
1030	LEU2::KanMX-switch_F	aactgcggtcaagatatttcttgaatcag
1031	KanMX_leu2_R	cctgattcaagaaatatcttgaccgcagtttaagataatggggctctttacagatctGTT
1032	KanMX_leu2_F	tcttgataaatgtatgtagattgcgtatatagtttcgtTGGATGGCGGCGTTAGTATCGA
1082	BC2+nested	CTCGtaaggtagccaaatgGCANNNNNNNNNNNNNNTATGAGAGATGCACGACTGCGCG
1083	BC2+nested_fillin	CACGTAACGCTACCTGGATCGCTAAACGCGCAGTCGTGCATCTCT
1085	BC2_seqR.NdeI	GCTAAACGCGCAGTCGTGCATCTCTCATA
1086	P7_UMI_BC2_seqR.NdeI	CAAGCAGAAGACGGCATACGAGATGCNNNNNNNNGTATGCTCGCTAAACGCGCAGTCGTGC ATCTCT
1099	3xRz_frag1_F	tctagaACTAGTGCACATGAGGATCACCCATGTGCATCTTGTTTGCGTGTGCGCCAGATT
1100	3xRz_frag1_R	tacatgactcgagCAGATTTGTCCATTGTTATATCTCCTTCGGATCAAGCTTAATCTGGCGCACACGC AAACAAGAT
1101	3xRz_frag2_F	tctagaactagtATTCGAGTACCTAGGTTTGTAAAAGTTATCAGG
1102	3xRz_frag2_R	tacatgactcgagATTGCTATTTAGAGGGGAATTCTAAAGTGTCG
1103	BC1_seqF	CTGACGCACGTTCTGACTACAACGC
1104	P5_BC1_seqF	aatgatacggcgaccaccgagatctacacCTGACGCACGTTCTGACTACAACGC
1105	seqF-2_BC1_F	CGCGGCAGCGAATCGACACTTTAGCTGACGCACGTTCTGACTACAACGC
1106	BC1-RI_fillR	TTGCTATTTAGAGGGGAATTNNNNNNNNNNNNNNNGCGTTGTAGTCAGAACGTGCGTCAG
1107	P7_UMI_BC2_seqR.NdeI-2	CAAGCAGAAGACGGCATACGAGATGCNNNNNNNNccatacgcGCTAAACGCGCAGTCGTGCAT CTCT
1108	P7_UMI_BC2_seqR.NdeI-3	CAAGCAGAAGACGGCATACGAGATGCNNNNNNNNgatagttgGCTAAACGCGCAGTCGTGCATC TCT
1109	P7_UMI_BC2_seqR.NdeI-4	CAAGCAGAAGACGGCATACGAGATGCNNNNNNNNattagcatGCTAAACGCGCAGTCGTGCATC TCT
1110	P7_UMI_BC2_seqR.NdeI-5	CAAGCAGAAGACGGCATACGAGATGCNNNNNNNNtggatggaGCTAAACGCGCAGTCGTGCAT CTCT
1111	P7_UMI_BC2_seqR.NdeI-6	CAAGCAGAAGACGGCATACGAGATGCNNNNNNNNcgttgattGCTAAACGCGCAGTCGTGCATC TCT
P7	P7	CAAGCAGAAGACGGCATACGAGAT

VITA

Matthew Rich earned a Bachelor of Arts (AB) degree from Princeton University in 2009, majoring in Molecular Biology with certificates in Quantitative and Computational Biology and Music Performance. While at Princeton, he isolated and characterized yeast mutants resistant to antipsychotic drugs to try to identify the mechanism of the drugs' off-target effects, as well as studied the heat shock response in *Saccharomyces bayanus*. Before enrolling at the University of Washington, he spent a year performing proteomics research in Neil Kelleher's lab at the University of Illinois in Urbana-Champaign. Matthew spent 2015 in Tokyo, Japan, where he worked in Nozomu Yachie's lab at the University of Tokyo, tried to develop methods for massively parallel analysis of protein-protein interaction dynamics, ate a lot of sushi, and sang a lot of karaoke.

In Seattle, Matthew and his wife, Cindi (who earned a Ph.D. in the department of East Asian Languages and Literatures, studying Modern Japanese and Korean literature), enjoyed attending the Seattle Symphony, eating and drinking the city's offerings, and cheering for the Mariners (who went 519-565 from 2011-2017).