GENETIC BACKGROUND EFFECTS ON ADAPTATION AND GENE FUNCTION EVOLUTION

MONICA ROSE SANCHEZ

A Dissertation Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

The University of Washington

2017

Reading Committee:

Maitreya Dunham, Chair

M. K. Raghuraman

Harmit Malik

Program Authorized to Offer Degree: Molecular and Cellular Biology

University of Washington

Abstract

Genetic background effects on adaptation and gene function evolution

Monica Rose Sanchez

Chair of the Supervisory Committee:

Maitreya Dunham

Associate Professor of Genome Sciences

Department of Genome Sciences

The ability to predict phenotype from genotype is the ultimate goal of genomic research. Achieving this goal requires understanding, on a systems level, how natural variation between genetic backgrounds can influence overall allelic effects. Understanding the relationship between sequence variation and phenotypic variation for complex traits will provide insights that are important for not only predicting adaptive evolution outcomes, but also for predicting disease risks in human populations and help improve personalized therapeutic treatments and selective breeding in agriculturally important plants and animals. In this thesis, I have investigated the impact of genetic background on adaptation and gene function evolution across diverged species of yeast spanning the *Saccharomyces* clade. In chapter one I give an overview of molecular mechanisms of adaptation and gene function characterization in the context of genome

dependencies. I discuss general concepts of comparative genomic studies and adaptation and provide background information on gene function characterization. Chapter two of this thesis describes my work addressing the dependence of cellular fitness and adaptation on genetic backgrounds of different yeast species. The comprehensive work in this chapter tests several hypotheses to explain differential amplification events between paralogs and illustrates how changes in regulatory sequence amid divergent genetic contexts can influence adaptive routes taken to achieve increased cellular fitness in sulfate-limited growth conditions. To expand comparisons of divergence to all orthologs between *S. cerevisiae* and *S. uvarum*, I describe a random mutagenesis method applied to *S. uvarum* to interrogate gene dispensability in chapter three of this thesis. Using this method, I created a pool of ~50,000 mutants in a diploid strain of *S. uvarum* and made comparisons against a haploid pool of ~40,000 mutants to: 1) prioritize candidate essential genes, 2) identify genes that differ in dispensability between species, and 3) investigate *cis* vs. *trans* effects to explain differential essentiality using cross-species complementation assays. Ongoing work described in chapter four explores centromere sequences to compare required functional elements between *S. cerevisiae* and *S. uvarum* as a means to explore centromeric evolution between these two species. I conclude with chapter five where I summarize the results of my thesis work, discuss ideas for future projects, and address the implications of these results in a broader context of functional comparative genomic research. Collectively, the work presented in this thesis furthers our understanding of genetic context and its impact on phenotypic outcomes associated with molecular evolution.

AKNOWLEDGEMENTS

The work present in this thesis is highly collaborative and requires gratitude to several individuals as well other institutions. I would like to thank the members of the Dunham lab for providing a supportive and pleasant environment and for their advice about experimental design and data representation. There are people in particular from the Dunham lab that deserve additional recognition. I would like to thank Celia Payen for her extremely helpful guidance and camaraderie throughout every aspect of my graduate career. I would like to Anna Sunshine for her patience and willingness to help me with sequencing analysis. I would like to thank Ivan Liachko for sharing his cloning wisdom with me and challenging me to think critically. I would also like to thank Caiti Heil for her editorial help with writing, daily advice, and for her meaningful friendship. I owe a huge thank you to Noah Hanson and Giang Ong who are talented technicians who have helped me with technical aspects of my project, Emily Mitchell for her expertise in performing chemostat experiments and Maitreya Dunham in particular for her exceptional mentoring, ideas, encouragement and her infectious enthusiasm for science!

On a more personal note, I would like to thank my friends, classmates and family for their encouragement and support. I want to particularly thank my parents, Stephen and Rose Mascareñas, my big brothers George and Michael, their wives Trish and Barbara as well as my nieces and nephews. I thank my husband Joe Sanchez for not only believing in me but also for sharing this experience together. Finally, I would like to thank Maggie Werner-Washburn for her undergraduate mentoring and encouraging me to pursue an advanced degree in science and her continual support throughout my career.

TABLE OF CONTENTS

LIST OF FIGURES

Supplemental Figure 3.9 Complementation assay confirming two examples of genes that differ in essentiality but complement the viability phenotype in both genetic backgrounds

LIST OF TABLES

DEDICATION

I dedicate this work in loving memory of my father who continues to inspire me to do my best every day.

CHAPTER 1: INTRODUCTION TO COMPARATIVE FUNCTIONAL ANALYSIS, ADAPTATION AND GENE FUNCTION EVOLUTION

1.1 Comparative analysis

Comparative genomics can be defined as a field of biological research in which genomic features such as DNA sequences, genes, synteny, regulatory sequences and other genomic features are compared across organisms, offering new insight into the impacts of natural selection on genome evolution. Sequence comparisons across species allow the identification of conserved sequences likely to be constrained due to evolutionary pressures, implying a biological function (Alföldi and Lindblad-Toh, 2013). Conserved sequences that are maintained across species provide information about selection and functional divergence based on the extent to which these sequences are conserved (Ellegren, 2008). Comparative genomic analysis is one major approach used to annotate newly sequenced genomes and has provided information about functionally conserved sequences across many species, including humans (Miller et al., 2004). Furthermore, researchers can apply comparative genomic approach to gain a more complete understanding about novel phenotypes, evolutionary processes and the effects of purifying selection on non-coding regions of the genome. The following sections highlight specific applications of comparative research, and explain how this approach was applied to further our understanding of genetic background effects on phenotypic outcomes across diverged organisms.

1.2 Comparative analysis and functional genomics: How are functional elements determined?

1.2.1 *Predicting gene function*

Defining the molecular function of each component in an organism's genome is the ultimate goal of functional genomics. Advancements in technology, genetic toolkits and plummeting costs of sequencing have pushed our progress forward; however, assigning function to genomic sequence remains a significant bottleneck for gene annotation. Even within well-studied systems, functional annotation has plateaued, suggesting that new methods are required to better interrogate phenotypic information. Comparative sequence-based studies can predict the function of conserved sequence elements in newly sequenced organisms through sequence homology. However, direct experimental testing of sequence features is required to validate the predicted function. Furthermore, the genetic context is important to account for divergent function of sequences features, which may be attributed to modifier loci that vary between and within species. Thus, *in silico* methods are often used to help predict function by comparing sequence, expression pattern or protein structure similarity (Eisen, 1998). Based off the guilt by association principle, if an uncharacterized sequence feature is very similar to a characterized sequence, they are likely to have similar function.

For example, co-expression analysis is a common method used to predict gene function from transcriptomic data such as RNA microarrays or RNA-seq to group genes based on similar expression patterns (Usadel et al., 2009). This method has proved to be useful for assigning biological processes to genes of unknown function. However, there are known caveats to this method, including the abundance of false positives. Applying

comparative analysis approaches have shown to yield more accurate gene function predictions, since biologically relevant associations are likely to be independently observed in different species, whereas false associations are less likely to be repeatedly observed (Stuart et al., 2003).

Although *in silico* predictions have proved to be useful in predicting function, there is strong need for more high throughput methods to systematically test the function of genes in uncharacterized species to validate *in silico* predictions. There are examples of differential functions between predicted paralogs that would have gone undiscovered without experimental validation of gene function. One such study investigated the function of four paralogs involved in cyanobacterial photoprotection in *Anabaena* and identified three paralogs to have sub-functionalized, while one paralog was determined to no longer be involved in photoprotection at all (López-Igual et al., 2016). A purely *in silico* approach would have predicted all four paralogs to have similar function. Functional characterization across a wide range of species may elicit phenotypic information that may be masked in model systems. Thus, extending functional tests outside of model systems will improve predictive models and help assign gene function to newly sequenced systems. I will review a subset of methods that have proved to be instrumental in determining gene function across a diverse range of organisms.

1.2.2 *Experimental testing of gene function: forward and reverse genetics*

Systematically testing gene function through the observation of mutant phenotypes is a powerful approach to validate predicted gene function and to dissect biological pathways. Forward and reverse genetics are two known strategies that are

utilized to study gene function on a global scale. Classical forward genetics defines gene function by generating random mutants, often through exposure to a mutagen (ionizing radiation or chemical agents) and screening for a particular phenotype of interest, followed by the identification of the mutated genes (Coelho et al., 2000; Forsburg, 2001). Since the nature of the mutated gene is unknown, downstream analysis to identify the gene and validate the function requires time-consuming methods that are difficult to scale. While forward genetic screens have led to many primary biological discoveries, there are major limitations of this approach that make it less amenable to high-throughput identification of gene function.

Reverse genetic screening is an alternative method that creates defined mutations in all genes first, followed by the characterization of the effects of the targeted mutations. Targeted gene replacement, targeted induction of double stranded breaks and double-stranded-RNA-mediated gene silencing fall into this category and rely on *a priori* knowledge about the genetic loci of the targeted genes (Adams and Sekelsky, 2002). Throughout the last decade, several projects have generated genome-scale gene disruption libraries and have rapidly advanced our knowledge of many model organisms (Hardy et al., 2010). The following sections will highlight current methods of reverse genetics and their applications in diverse species.

1.2.3 Mutant collections are valuable high throughput tools for determining global gene function

1.2.3.1 Genome-wide deletion collections

Since George W. Beadle and Edward L. Tatum confirmed Sir Archibald Garrod's hypothesis that describes the fundamental relationship between gene and gene product in 1941, numerous techniques have been developed to experimentally test the function of individual genes (Beadle and Tatum, 1941). Classical genetic approaches require a significant amount of effort to identify the genotype responsible for a particular phenotype from random mutagenesis. As a result, current methods have been established to more easily and comprehensively create changes or disruptions of genomic sequences to identify the resulting phenotype. Targeted libraries of loss of function mutants have been created in model organisms by introducing particular mutations, inducing targeted double stranded breaks or systematically deleting entire ORFs (Baba et al., 2006; Dowell et al., 2010; Kim et al., 2010a; Noble et al., 2010; Shalem et al., 2014; Varshney et al., 2013; Winzeler et al., 1999). These resources have proved to be highly useful in understanding gene function, complex network interactions and synthetic lethal interactions (Tong et al., 2001). However, considerable labor and resources have been implemented to generate the near-complete deletion sets, and their applicability toward a wide range of organisms is problematic and currently not a scalable approach (Hughes et al., 2000; van Opijnen et al., 2009; Winzeler et al., 1999). Furthermore, deletion collections cannot provide information about undiscovered genes, non-coding sequences and partial-loss or gain-of-function mutations (Guo et al., 2013). While these whole-genome deletion collections have been instrumental in furthering our understanding of gene function in model systems, the following section describes a method that can be applied to a variety of previously under-studied species.

1.2.3.2 Transposon mutagenesis as a genetic tool

Transposable elements are a class of selfish genetic elements that contain repetitive DNA sequences that sometimes have the ability to change positions throughout a genome. Since Barbara McClintock discovered the first transposable element in maize (*Zea mays*) (McClintock, 1950), transposons have been found to influence genome evolution, driving adaptation and causing disease states (Chénais et al., 2012; Reilly et al., 2013). Due to their mobile characteristics, transposable elements have been an especially useful tool in molecular biology as a means of mutagenesis. Transposon insertions are used to produce random mutations, providing common flanking sequences that can be used to identify the location of the mutated allele.

Unlike targeted mutagenesis, transposon-mediated insertional mutagenesis can create random mutations throughout the genome in an unbiased manner (Kumar et al., 2004). Historically, this method has been used in various species, ranging from flies to microorganisms and led to the identification of essential genes and thousands of virulence genes (Bachmann and Knust, 2008; Hensel et al., 1995; Lau et al., 2001; Mei et al., 1997; Salama et al., 2004). These studies required laborious PCR, sequencing or microarray detection for each clone, limiting throughput and dynamic range of the method. However, pairing transposon insertional mutagenesis with massively parallel sequencing alleviates these limitations, making transposon-directed insertion site sequencing (TraDIS) an appealing platform to use in a diverse set of organisms, and includes a whole family of Tn-seq approaches (Langridge et al., 2009). Of course, insertion site sequencing has its own set of pitfalls. The insertion density and location of the insert determines the degree of deficiency of the mutant gene, and subsequent validation is

required to determine the effect of the predicted insertion mutations. Additionally, the method relies on the absence of an insertion as a metric for determining essential genes (Coelho et al., 2000). However, this type of approach is useful for prioritizing candidate genes for further analysis, allowing functional experimentation to be performed across different strains and species, providing a framework for functional comparisons.

1.2.3.3 Phenotypic readouts

The previous section described a method to create mutants to screen for phenotypes. Observing phenotypes as a result of a mutation is the first step in determining the function of a gene. Although there are several measurements that can be used to determine different phenotypes (i.e. fitness profiles, protein localization, RNA expression, etc.), one of the most reproducible and straightforward phenotypes to assay is cell viability, which can be used as a preliminary indicator of similar gene function. Identifying genes that have differential requirements for growth can be used as a first step to identifying genes that have diverged in function. To test if two orthologs from different species differ in gene function, cell viability can be used as a robust phenotype in cross-species complementation assays to test the ability to rescue an inviable phenotype.

1.2.4 Identification of other functional elements

Although comparative sequence analysis has helped address in the challenge of discovering functional protein coding regions, an even greater challenge is identifying functional elements that do not code for protein. Such sequences include, but are not limited to, sequences of nonprotein-coding RNAs plus sequences that regulate gene

expression, govern chromosome replication and maintain structure and stability. Due to the diverse nature of these sequence elements (i.e., usually short, independent of orientation, residing at various distances away from target gene), comparison of sequences across species are difficult to identify such functional elements (Cliften et al., 2001). Functional elements that are defined require additional experimentation to validate the predicted function of the proposed sequence feature. Sequence features, such as replication origins (Nieduszynski et al., 2006) and telomere length (Liti et al., 2009), have been assayed in different backgrounds and are also important for getting a complete picture of cellular function of all sequence features across the genome (Parts, 2014).

For example, centromeres in particular pose a significant challenge in defining functional DNA elements using standard comparative methods. Centromeric repeats are comprised of the most rapidly evolving DNA sequences in eukaryotic genomes, and differ even between closely related species (Csink and Henikoff, 1998; Haaf and Willard, 1997; Murphy and Karpen, 1998). Although centromeres are involved in an essential process that ensures proper segregation of genetic material to daughter cells, their precise boundaries are difficult to delineate, especially in plants and animals where they are likely epigenetically maintained and are located in highly repetitive regions (Henikoff et al., 2001). Previous studies in budding yeast, where point centromeres are defined by a ~125 bp sequence, have tested different regions of the genome to identify functional sequences that maintain plasmids (Fitzgerald-Hayes et al., 1982; Fleig et al., 1995). Open questions still remain about centromere sequence evolution that may be explained through comparative functional testing between diverged yeast species. Extending functional tests across a genus of species may aid in characterizing centromeric satellites

and may lead to the identification of an optimal fixed centromeric motif and help further our understanding of how centromeric sequences are distinguished from other parts of the chromosome.

The experiments mentioned in this section describe methods that rely on disrupting gene function to probe for molecular function. These approaches are successful in identifying genes that encode components of biochemical and physiological processes. However, they provide little insight into improved function. The following section highlights ways that evolution experiments offer opportunities to study beneficial mutations and allows researchers to study how general adaptation is dependent on the genetic context.

1.3 Comparative analysis and adaptation: How do different genomes evolve?

1.3.1 *Comparative experimental evolution*

The comparative method can simply be described as a comparison of two phenotypes across a range of species or higher taxa (Felsenstein, 1985). Comparative physiological analysis relies on retrospective analysis taken over long periods of time, often on the order of millions of years. These historically based comparisons analyze physiological observations along with the phylogeny of the taxa to identify patterns of characteristics distributed across species (Garland et al., 1999). This approach relies on the reconstruction of historical events to make inferences about evolutionary mechanisms, but provides an advantage of examining diverse taxa that have evolved in complex natural environments that are difficult to reconstruct in the lab and are often selecting for multiple traits. However, the problem with most comparative studies is the

reliance on correlation rather than causation and their interpretation depends on the assumed parsimonious phylogenic relationships. Through the use of comparative experimental evolution, a wide range of genetic backgrounds can be used to test adaptive outcomes as a result of selection to one defined parameter.

Beyond just making observations about evolution and adaptation in action, more direct experimentation is required to test hypotheses about evolutionary processes. Experimental evolution experiments are important tools to study evolution and adaptation in real time, allowing precise tracking of genotypic and phenotypic changes throughout many generations, monitored over very fine timescales. Many of these early experiments originated in fly species but soon expanded out to include diverse taxa such as fungi, plants, vertebrates and especially microorganisms. Microorganisms specifically are an ideal system to study adaptation in a laboratory setting due to their short generation times, ease of culturing in the lab and large population sizes. Although there are numerous experimental approaches to studying evolution in the lab, I will primarily focus on continuous culture experiments of microorganisms, first introduced in the 1950's by Monod (1950) and Novick and Szilard (1950) using the chemostat.

Briefly, the chemostat is a culture vessel that keeps cell populations growing at a reduced rate governed by a limiting nutrient or other factors restricting growth rate over unlimited timescales (Novick and Szilard, 1950). Typically, an initial population of isogenic cells are grown in the culture vessel at a constant growth rate that is equal to the rate at which fresh medium drips into the culture vessel, while an equivalent amount is being expelled. The steady state inside the vessel maintains a constant selective pressure, acting on random mutations accumulating in the population, and allows reproducibility.

Samples are taken throughout the course of the experiment to determine the fitness and genetic changes of clones from populations at particular time points. Since relative fitness measurements depend not only on the genotype but also on the environment in which there were measured, competition assays are performed in chemostats under the same selective pressure to reliably measure the fitness of evolved clones when competed against an ancestral strain. Although I focused on continuous culture approaches to study laboratory evolution, this concept more broadly provides tractable methods for observing dynamic adaption events in real time and offers insights into evolutionary innovations in response to strictly defined environmental conditions (Barrick and Lenski, 2013; Kawecki et al., 2012).

Although many laboratory evolution experiments allow cell populations to grow over indefinite timescales, it falls short of timescales observed in most comparative studies, especially when trying to mimic macroevolution. Therefore, merging adaptive laboratory evolution and comparative genomics provides a comprehensive view of genomic changes that underlie evolution, which addresses some of the discussed limitations of strictly comparative or experimental evolution methods. Taken together, comparative experimental evolution provides a method to monitor the process of adaptation to a specific selective pressure of diverged species that span varying evolutionary distance. This approach provides the opportunity to observe different adaptation trajectories due to the natural variation across different genetic backgrounds, linking genetic variation to phenotypic differences.

1.3.2 Genetic background and adaptation

Recent experimental evolution studies have identified examples of genetic background effects influencing phenotypic outcomes and evolutionary trajectories. For instance, Vogwill *et al.* investigated how genetic background influences the evolution of rifampicin resistance in eight different strains of *Pseudomonas,* varying in genome size between 4.6 and 7.1 Mb. Adaptation across strains occurred by 47 mutations at conserved sites in *rpoB,* the target of rifampicin, which resulted in different effects on growth rate in different strains. They conclude that parallel evolution occurs more frequently within-strains than between different strains across the *Pseudomonas* genus, implying that genetic background has a detectable impact on adaptation in this system, resulting in differential phenotypic outcomes (Vogwill et al., 2014).

1.3.3 Ploidy and adaptation

Other than large differences between divergent strain backgrounds, how might ploidy affect adaptability of otherwise isogenic strains? Selmecki *et al.* recently addressed this question by taking an experimental evolution approach to measure the acquisition and spread of beneficial mutations across haploid, diploid and tetraploid populations of *S. cerevisiae*. Tetraploid strains grown on poor carbon sources underwent faster adaptation, driven by higher rates of beneficial mutations with stronger fitness effects, including whole chromosome aneuploidy. They conclude that polyploidy increases genetic diversity due to the mutagenic effects of aneuploidy that results from the loss of chromosomes, and thus provides genetic plasticity that aids facilitation of rapid adaptation (Selmecki et al., 2015). Following this same logic, high rates of

aneuploidy induced by whole-genome duplication may further increase the rate at which beneficial mutations are acquired, including regions where selective retention of duplicate sequences has occurred. To elaborate on this point further, the following section describes known mechanisms of adaptation through gene duplicates and amplifications.

1.3.4 The role of gene duplicates and amplifications in adaptation and gene function evolution

Understanding how neutral, redundant copies of genes can be maintained in the genome for extended periods of time has been a topic of extensive study in the field of evolutionary biology. Although there are several examples of adaptive gene duplications to various environmental conditions across species, it is still uncertain what role long-term persistence of gene duplicates plays in specific mechanisms of adaptation. It was Ohno who proposed the redundancy hypothesis about gene duplicates, reasoning that one copy of a gene duplicate pair performs the necessary function while the other is completely redundant and free from selection (Hughes, 1994). This idea led to an ongoing effort to understand how the redundant copies are maintained without being eliminated by mutation.

One explanation for retention of gene duplicates may simply be the division of ancestral function into two genes, complementing degenerative mutations (Qian and Zhang, 2014). Secondly, both gene copies may achieve fixation through positive selection as a result of increased protein dosage. Furthermore, there are several examples of gene duplications conferring an adaptive response to nutrient limitation, heat, and cold

stress that have been identified in yeast, bacteria, *Arabidopsis* and even Antarctic cod (Brown et al., 1998; Chen et al., 2008; Christ and Chin, 2008; DeBolt, 2010). Additionally, the amplification of the *Plasmodium falciparum* multidrug resistance gene (pfmdr1) has been shown to be a target of adaptive evolution due to the widespread use of anti-malaria drugs, duplicating in nature multiple times yielding increased resistance to different drugs throughout the world (Duraisingh and Cowman, 2005; Foote et al., 1989; Triglia et al., 1991). Finally, and perhaps at the heart of interest towards gene duplicates, is the principle that each duplicate can evolve independently, potentially leading to functional novelty (neofunctionalization) and increased fitness.

1.3.5 Yeast genetics: a tool for studying comparative functional genomics

Understanding human genetic variation and its relationship to human disease has been recognized as an important factor for uncovering how the same allele can present different phenotypes between individuals. Discovering the heritability of many complex diseases will rely on a more comprehensive analysis of genetic background effects, ranging from large structure variants to understanding the joint effects of many loci of small effect (Eichler et al., 2010). Connecting genotype to phenotype amongst individuals is complicated by multiple genetic loci interactions observed as epistasis, buffering, or robustness, making it difficult to parse out causality (Hartman et al., 2001). Due to the complex nature of genetic background effects on phenotypes across human populations, researchers have turned to model organisms to study genetic interactions and modifying loci on a systematic scale.

Yeast are an ideal group of organisms to pilot comparative functional genomic studies. Several species spanning large evolutionary distances have been sequenced to high coverage and deletion collections have been made in a subset of these species (Sherman et al. 2009; Scannell et al. 2011; Dujon 2010). Fungi are also known to pose public health problems and identifying commonalities across them may guide antifungal therapeutics (Karkowska-Kuleta et al. 2009). Additionally, numerous comparative studies have demonstrated that interesting evolutionary features are revealed when processes are compared within the *Saccharomyces* genus (Doniger 2005; Tsong et al. 2006; Kvitek et al. 2008). Finally, all members of this genus can inter-mate to create hybrids, allowing us to leverage the large genetic toolsets established in the eukaryotic model organism, *S. cerevisiae* (Sebastiani et al. 2002; Fischer et al. 2000).

Targeted gene deletion collections have been created in two yeast species, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe,* spanning ~420 million years of evolution (Wood et al. 2002). The nearly complete libraries identified that 20% and 26% of genes in these two species respectively, are essential for growth in rich media. Comparisons of orthologous essential genes between these two species revealed that 17% of the essential genes are species-specific (Giaever et al. 2002; Kim et al. 2010). A recent study proposed a model describing that proteins are essential mostly due to their involvement in essential functional protein complexes (Ryan et al. 2013). This pattern was highlighted through analyzing differentially essential genes that switch between the two species, which can be grouped into modules that perform particular functions. This difference was notably seen in mitochondrial translation machinery that was found to be essential in *S. pombe* and non-essential in *S. cerevisiae.* Since *S. cerevisiae* can survive

without mtDNA as "petite" mutants in contrast to *S. pombe,* this result suggests that the flip of essentiality may be due to a difference in lifestyle (Heslot et al. 1970). Unfortunately, the ability to uncover causative genetic loci for differences in gene essentiality has been limited because of the vast evolutionary distances involved.

In addition to the identification of genes that differ in dispensability between species, Dowell *et al.* (2010) investigated the genotype-to-phenotype problem by studying the effects of strain background on gene essentiality in *Saccharomyces cerevisiae*. Although the divergence between the two strains was roughly equal to the divergence between two human genomes (Wang et al., 2008), they discovered a subset of conditionally essential genes that were dependent on the strain background, despite the low divergence. This study systematically investigated genetic mechanisms that can lead to different phenotypes for the same mutation in two different genetic backgrounds, providing a framework to test general principles of background effects on phenotypes due to natural variation amongst individuals (Dowell et al., 2010). Extending these types of analyses to investigate the effects of natural variation between two species that diverged 20 million years ago will enhance the effect of greater genetic variability but will constrain the effects of large differences in metabolic pathways due to their similarity in fermentative properties (Dujon 2006).

Genome evolution and adaptation have been investigated using a variety of methods that include comparative genomics, functional screens, *in silico* theoretical approaches as well as laboratory evolution experiments. Taken together, these examples identify the impact genetic variation has on variable phenotypic outcomes. Applying comparative experimental approaches has proven to be extremely valuable to generate

insight into evolutionary processes, such as genome evolution and mechanisms involved in retention of gene duplicates, and will continue to be fruitful. Moreover, comparative experimental approaches help identify the molecular basis of adaptation by linking genetic variation to phenotypic differences.

1.4 Dissertation objectives

The work presented in this thesis merges my interests in understanding the effects of genetic background on gene function evolution and in gene function characterization. Specifically, I investigated the impact of divergent genomic backgrounds on evolutionary outcomes by examining paralog divergence across species of yeast in the *Saccharomyces* clade. I also took a whole genome approach to study gene function evolution between *S. cerevisiae* and *S. uvarum* using transposon mutagenesis.

There were many advantages to using this particular clade of yeast to successfully carry out my research aims. First of all, I was able to leverage the many genetic tools established in *S. cerevisiae* to test hypotheses in *S. uvarum*. Second, these species have a high level of sequence divergence, while maintaining largely syntenic genomes and similar genetic manipulation approaches. Comparisons between species that have sufficiently diverged but are close enough to maintain elements that might have been lost through evolution are ideal for making functional predictions of non-protein coding functions (Cliften et al., 2001). Lastly, approximately four hundred *S. uvarum* deletions strains were already in existence in the Rine lab and were generously shared with us to use as positive controls **(Supplemental Table 1.1**).

Despite the wealth of information summarized above about adaptation and gene function determinants, there are many questions that remain about how gene functions evolve in the face of new challenges, and what the impacts are for different genetic backgrounds. How we can determine gene function in previously uncharacterized species, and how doing so can help further our knowledge of general mechanisms of molecular evolution are examples of questions that remain to be answered.

In addressing the questions I stated above, I have aimed to accomplish the following objectives in this dissertation:

1. To investigate the effect of divergent genomic backgrounds on paralog divergence by
    a. Characterizing the fitness effects and genomic changes in species across the *Saccharomyces* clade when grown in sulfate-limited chemostats
    b. Testing hypotheses to describe differential paralog amplification events between *S. cerevisiae* and *S. uvarum*
    c. Determining if genetic background affects the fitness benefit of amplified paralogs
    d. Identifying coding or non-coding regions of the paralogs that account for differences in the selective effects between species
2. To identify evidence of gene function evolution across orthologs by
    a. Creating a transposon mutagenesis library in *S. uvarum* to assay viability phenotypes genome-wide
    b. Identifying and confirming candidate genes that differ in essentiality between species

c. Performing cross-species complementation assays

3. To determine divergent centromeric sequence function across species by
    a. Creating a library consisting of diverse lengths and regions of centromeric sequences from *S. cerevisiae* and *S. uvarum*
    b. Performing quantitative measurements of correctly segregating centromere sequences in *S. uvarum* and *S. cerevisiae* genetic backgrounds
    c. Identifying and validating the most efficient centromeric sequences and length for each genetic background.

CHAPTER 2: *DIFFERENTIAL PARALOG DIVERGENCE MODULATES GENOME EVOLUTION ACROSS YEAST SPECIES*

This chapter is based on the following manuscript with the same title, published in *PLoS Genetics 13(2): e1006585. https://doi.org/10.1371/journal.pgen.1006585*

2.1 ABSTRACT

Evolutionary outcomes depend not only on the selective forces acting upon a species, but also on the genetic background. However, large timescales and uncertain historical selection pressures can make it difficult to discern such important background differences between species. Experimental evolution is one tool to compare evolutionary potential of known genotypes in a controlled environment. Here we utilized a highly reproducible evolutionary adaptation in *Saccharomyces cerevisiae* to investigate whether experimental evolution of other yeast species would select for similar adaptive mutations. We evolved populations of *S. cerevisiae, S. paradoxus, S. mikatae*, *S. uvarum,* and interspecific hybrids between *S. uvarum* and *S. cerevisiae* for ~200-500 generations in sulfate-limited continuous culture. Wild-type *S. cerevisiae* cultures invariably amplify the high affinity sulfate transporter gene, *SUL1*. However, while amplification of the *SUL1* locus was detected in *S. paradoxus* and *S. mikatae* populations, *S. uvarum* cultures instead selected for amplification of the paralog, *SUL2*. We measured the relative fitness of strains bearing deletions and amplifications of both *SUL* genes from different species, confirming that, converse to *S. cerevisiae*, *S. uvarum SUL2* contributes more to fitness in sulfate limitation than *S. uvarum SUL1*. By measuring the fitness and gene expression of chimeric promoter-ORF constructs, we were able to delineate the cause of this

differential fitness effect primarily to the promoter of *S. uvarum SUL1*. Our data show evidence of differential sub-functionalization among the sulfate transporters across *Saccharomyces* species through recent changes in noncoding sequence. Furthermore, these results show a clear example of how such background differences due to paralog divergence can drive changes in genome evolution.

2.2 INTRODUCTION

Understanding how organisms adapt to their environment is a fundamental goal of evolutionary biology. This goal has been complicated by the dependence on the reconstruction of historical events to make inferences about selective pressures and evolutionary mechanisms. Furthermore, it can be difficult to pinpoint genetic variation that causes new phenotypes of interest amid very divergent genomes. One approach to circumventing this limitation is to study evolution in the laboratory, where growth, environment, and population parameters can be controlled and dynamic adaptation events can be followed in real time (Adams and Rosenzweig, 2014; Barrick and Lenski, 2013; Colegrave and Collins, 2008; Dettman et al., 2012; Kawecki et al., 2012). However, experimental evolution has its own limitations, such as being too far removed from natural environmental factors and extending over only limited time scales. Merging laboratory evolution and comparative genomics could provide a more comprehensive view of processes that underlie evolution. In addition, comparative experimental evolution allows us to determine to what degree genetic background may result in differential functional innovation in the future (Dettman et al., 2012; Wood et al., 2005).

One source of genetic novelty that may vary across divergent species is gene duplication. Gene duplicates can have different fates, either through dosage effects of an extra copy, splitting ancestral functions or regulatory patterns over duplicates (sub-functionalization), or acquiring novel function (neo-functionalization) (VanderSluis et al., 2010; Voordeckers et al., 2012). Alternatively, they can provide genetic redundancy to endow organisms with mutational robustness (Conant and Wolfe, 2008; Gu et al., 2003; Ohno S., 1970). Duplications occur frequently during evolution and are commonly linked to genome innovations that result in an adaptive or phenotypic change to a particular environment (Lynch and Force, 2000; Wapinski et al., 2007). After a duplication event, adaptation may result through the accumulation of mutations in the non-coding or protein coding regions of the genome, which may alter gene function, protein-protein interactions, or expression profiles. Accumulation of mutations in the coding region of each paralog may potentially modify active sites, affecting biochemical functionality, or alter binding interfaces and thus their interaction specificity (Gagnon-Arsenault et al., 2013). Mutations in the non-coding region of each paralog may cause regulatory interactions in networks to be lost or re-wired, potentially leading to expression divergence between paralogs (Guan et al., 2007; Hittinger and Carroll, 2007; Teichmann and Babu, 2004).

The *Saccharomyces* clade of species provides a particularly appealing platform for comparative studies of gene function. The last common ancestor of this group existed approximately 20 million years ago, with approximately 80% identity in coding sequences between *S. cerevisiae* and *S. uvarum* (Dujon, 2010; Kellis et al., 2003; Scannell et al., 2011). The *Saccharomyces* species are experimentally tractable, have high

quality genome sequences (Cliften et al., 2003; Kellis et al., 2003; Scannell et al., 2011), contain largely syntenic chromosomes (Fischer et al., 2001), and can mate to form hybrids, including with the laboratory workhorse *S. cerevisiae*, providing access to a huge knowledge base and extensive toolkit of genetic and genomic resources. Additionally, the *Saccharomyces* genus is a result of a well-studied whole genome duplication event, which occurred just before the separation of *Vanderwaltozyma polyspora* from the *S. cerevisiae* lineage (Scannell et al., 2007) and was itself probably a result of a hybridization event (Marcet-Houben and Gabaldón, 2015).

In this study, we compared the evolutionary outcomes upon sulfate-limited growth in chemostat culture between *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. uvarum*, and *S. cerevisiae/S. uvarum* hybrid strains and used whole genome sequencing and species-specific microarrays to identify resultant genetic changes. We discovered differential amplification of sulfate transporter gene paralogs *SUL1* and *SUL2* in the different species. The species-specific amplification preference correlated with the selective effects of amplification and deletion of each sulfate transporter gene. Analysis of functional divergence of the two paralogs across these species provides evidence for differential sub-functionalization between the *SUL1* and *SUL2* paralogs of *S. cerevisiae* and *S. uvarum,* driven largely by lineage-specific acquired changes in the non-coding region of *SUL1* in *S. uvarum*. In this work, we discovered an example of recent paralog divergence between two gene duplicates with altered gene expression between *S. cerevisiae* and *S. uvarum,* and demonstrated that such differences can alter the genetic mechanisms by which these species adapt to future challenges.

2.3 RESULTS

2.3.1 Adaptation through differential gene amplification: experimentally evolved *S. cerevisiae* and *S. uvarum* populations amplify different sulfate transporter genes

As described previously (Brewer et al., 2015; Gresham et al., 2008; Miller et al., 2013; Payen et al., 2013), evolved clones of *S. cerevisiae* selected during long-term continuous culture under sulfate-limitation reproducibly carry amplification events near the right telomere of chromosome II containing the high affinity sulfur transporter gene *SUL1* (representative event shown in **Fig 2.1B**). This mutation confers one of the highest (20-40% increase) and most reproducible (25/25 populations) fitness advantages known in the experimental evolution literature (Brewer et al., 2015; Gresham et al., 2008; Miller et al., 2013; Payen et al., 2013). In order to determine whether other yeast species would follow this same evolutionary path, we performed two evolution experiments with a sister species, *S. uvarum,* in chemostats using the same condition in which the *SUL1* amplification has been observed for *S. cerevisiae*. Each experiment was initiated with a prototrophic diploid *S. uvarum* strain that had never before been exposed to long-term sulfate limitation in the laboratory (see materials and methods).

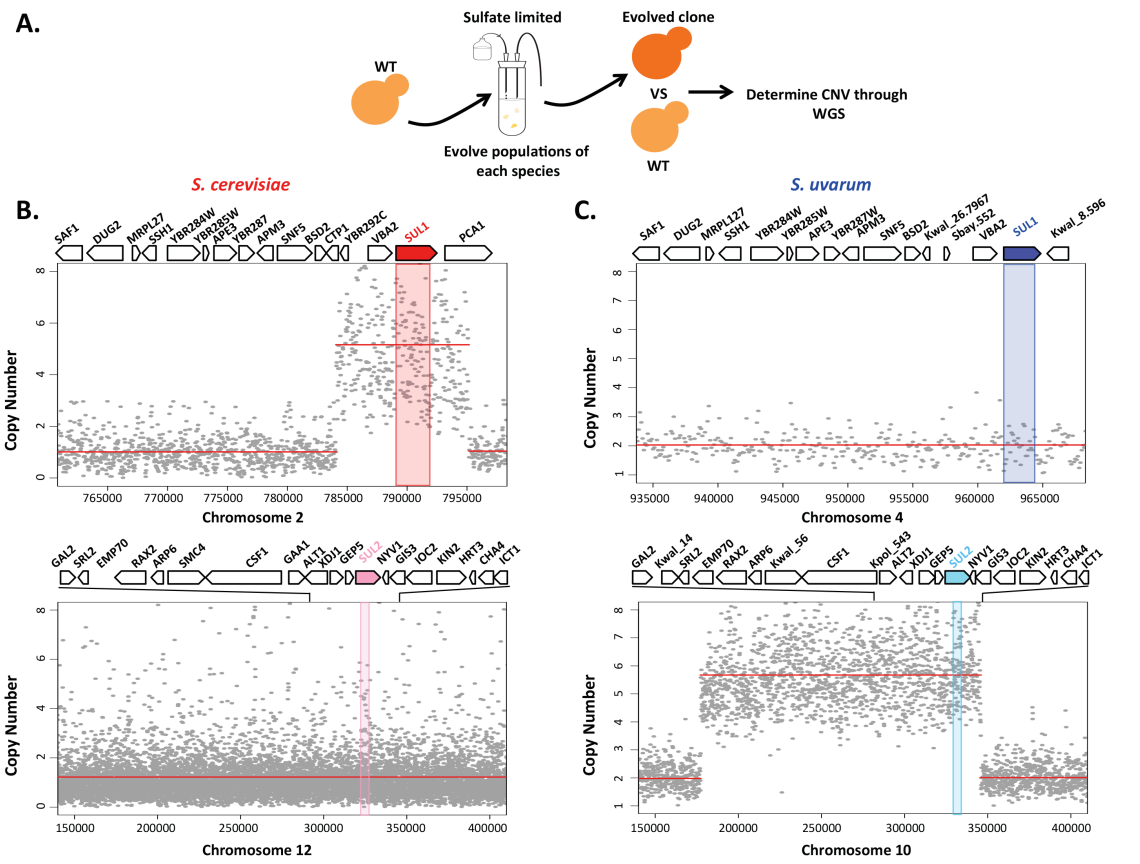


Figure 2.1. Adaptation through differential gene amplification between *S. cerevisiae* and *S. uvarum*. A) Schematic illustrating how evolved strains were derived and analyzed by sequencing to generate copy number plots shown below. B) Copy number of relevant genomic segments surrounding the *SUL1* (top) and *SUL2* (bottom) loci in a representative evolved strain of *S. cerevisiae*. Copy number plots were calculated by sequencing-depth ratios between evolved and parental genomes in *S. cerevisiae* at 188 generations. Gray dots represent the per nucleotide read-depth averaged across 25 bp windows. Segmentation-derived regions of equal copy number are indicated in red. Segmentation defines an ~11 kb region with a copy number of 5. The region of the sulfate transporter gene *SUL1* gene is shaded red. C) Sequencing-depth ratios between evolved and parental genomes in an *S. uvarum* clone isolated at 510 generations are plotted at relevant genomic segments surrounding the *SUL1* (top) and *SUL2* (bottom) loci. A large segmental amplification of an evolved clone defining a ~168 kb region with a copy number of ~5 includes the locus containing the sulfate transporter gene *SUL2*, shaded in blue. Genes aligned at the top represent the loci in the expanded panel.

In contrast to the amplification of *SUL1* in the *S. cerevisiae* clones, no amplification of this locus was observed in the two populations of *S. uvarum* evolved under sulfate limitation for 500 generations. However, the locus containing the gene *SUL2* was amplified in both populations as determined through microarray-based comparative genomic hybridization (aCGH) (**S2.1 Fig**). Two clones from one population were analyzed further by deep sequencing, revealing an internal segment of chromosome X containing the gene *SUL2* at an increased copy number of 5 in one of the two clones (**Figure 2.1C)**. The fitness benefit of this evolved clone was 20% when competed against the ancestral strain (n=4, **Table 1**, see Materials and Methods for further details of how fitness was measured in **S2.10 Fig**). Although the exact function of the protein Sul2 has never been experimentally tested in *S. uvarum*, Sul2 has been identified as a lower affinity transporter of sulfate in *S. cerevisiae* (Cherest et al., 1997).

Table 2.1. Fitness coefficient of evolved clones

Species	Generations	Clones	Fitness Coefficient, %	Gene Amplified	Copy Number
S. cerevisiae	210	Clone1	44.2 ± 8.2 (n = 2)	*SUL1*	5
S. uvarum	510	Clone1	21.8 ± 2.37 (n = 4)	*SUL2*	5

We next set out to explain the differential amplification of *SUL1* and *SUL2* in these closely related species. We hypothesized that the different evolutionary outcomes could result from divergence in gene function—*SUL2* may encode the higher affinity transporter gene in *S. uvarum* and so its amplification causes a higher fitness benefit—or from changes in chromosomal context that affect amplification rate or amplicon fitness. We test these hypotheses below.

2.3.2 The genomic context of *SUL1* and *SUL2* in *S. uvarum*

We hypothesized that the preference for the amplification of *SUL1* in *S. cerevisiae* could be due to changes in chromosomal context between the two species that might affect the propensity of the region to amplify. Although the *SUL1* orthologs are largely syntenic between the two genomes, some differences do exist. *SUL1* in *S. cerevisiae* is located on the right arm of chromosome II, near the telomere. The *S. uvarum* ortholog is located in a syntenic region, on chromosome IV where, as compared with the *S. cerevisiae* genome, the left portion of this chromosome contains a reciprocal translocation with a region syntenic to the right arm of *S. cerevisiae* chromosome IV (Fischer et al., 2001; Ryu et al.). The regions immediately adjacent to *SUL1* are largely syntenic, though the gene just distal to *SUL1* in *S. cerevisiae, PCA1,* is missing in *S. uvarum* (**S2.2 Fig**). Adjacent sequences to the telomeric repeats, including X and Y' elements as well as subtelomeric gene families, have been shown to be rapidly evolving across species of the *Saccharomyces* clade, possibly contributing to a difference in mutation rate (Kellis et al., 2003; Liti et al., 2005; Martin et al., 2009; Pedram et al., 2006). In *S. cerevisiae*, this region also contains a DNA replication origin (ARS228),

which we previously demonstrated to be involved in (though not necessarily required for) the generation of the amplification (Brewer et al., 2011; Rienzi et al., 2012). To test for replication origin function, we cloned the corresponding region from *S. uvarum* and tested it for the ability to support plasmid replication (i.e., an assay for Autonomously Replicating Sequences, or ARSs). Like *S. cerevisiae, S. uvarum* does contain an ARS in this region (**S2.3 Fig)**. However, there do appear to be differences in activity among a minority of ARSs between *S. cerevisiae* and *S. uvarum*, determined through whole genome replication assays (Müller and Nieduszynski, 2012).

The *SUL2* gene is located on chromosome XII in *S. cerevisiae* and X in *S. uvarum,* though the immediate surrounding region is mostly syntenic. From comparisons with the reconstructed ancestral genome, *SUL2* appears to be the ancestral copy of the sulfur transporter, with *SUL1* being a more recent gene duplicate after a small-scale duplication (SSD) event (Byrne and Wolfe, 2005). Amino acid conservation between *SUL1* and *SUL2* in *S. cerevisiae* is 62.5% and 61.3% shared identity in *S. uvarum*, whereas *SUL1* from *S. cerevisiae* and *SUL1* from *S. uvarum* share 84% identity and *SUL2* from *S. cerevisiae* and *SUL2* from *S. uvarum* share 87% identity, indicating that the sulfate transporter genes are correctly annotated.

2.3.3 *SUL1* in *S. uvarum* can be amplified in the absence of *SUL2*

Although the origin of replication is present, there may be other differences near *SUL1* in *S. uvarum* that might explain why this region has not been observed to amplify in the evolved strains. To test if *SUL1* is capable of amplification, we evolved four haploid *sul2*Δ strains of *S. uvarum* in sulfate-limited media and tested the evolved

populations for copy number variation using aCGH. At 260 generations, we identified an amplification of the *SUL1* locus in one of the four populations and no other amplifications in the other three populations (**Figure 2.2D)**. This result indicates that the *SUL1* locus in *S. uvarum* has the capacity for amplification, but does not attain high frequency in populations initiated with strains containing both *SUL1* and *SUL2* genes.

Alternatively, the *SUL2* locus cannot amplify in *S. cerevisiae.* To test if the *SUL2* locus can amplify in *S. cerevisiae,* we evolved four haploid strains of *S. cerevisiae* in which *SUL1* has been deleted (*sul1*Δ) in sulfate-limited media and tested the evolved populations for copy number variation using aCGH. We identified an amplification of the *SUL2* locus in all four populations (including a whole chromosome aneuploidy event that occurred in one population) indicating that *SUL2* can amplify in *S. cerevisiae,* but these amplifications do not attain high frequency in evolution experiments performed with strains in which the *SUL1* gene is present (**Figure 2.2C and S2.4 Fig)**. We note that these experiments leave open the possibility that differences in amplification rate might contribute to the observed differences in amplification propensity. We have so far been unable to measure the amplification rate of these loci and so have not tested this hypothesis. However, another possible explanation for these results is that *SUL2* amplification may have a greater selective effect in *S. uvarum.* To test this possibility, we performed additional experiments to determine the functional contribution of each gene from both species.

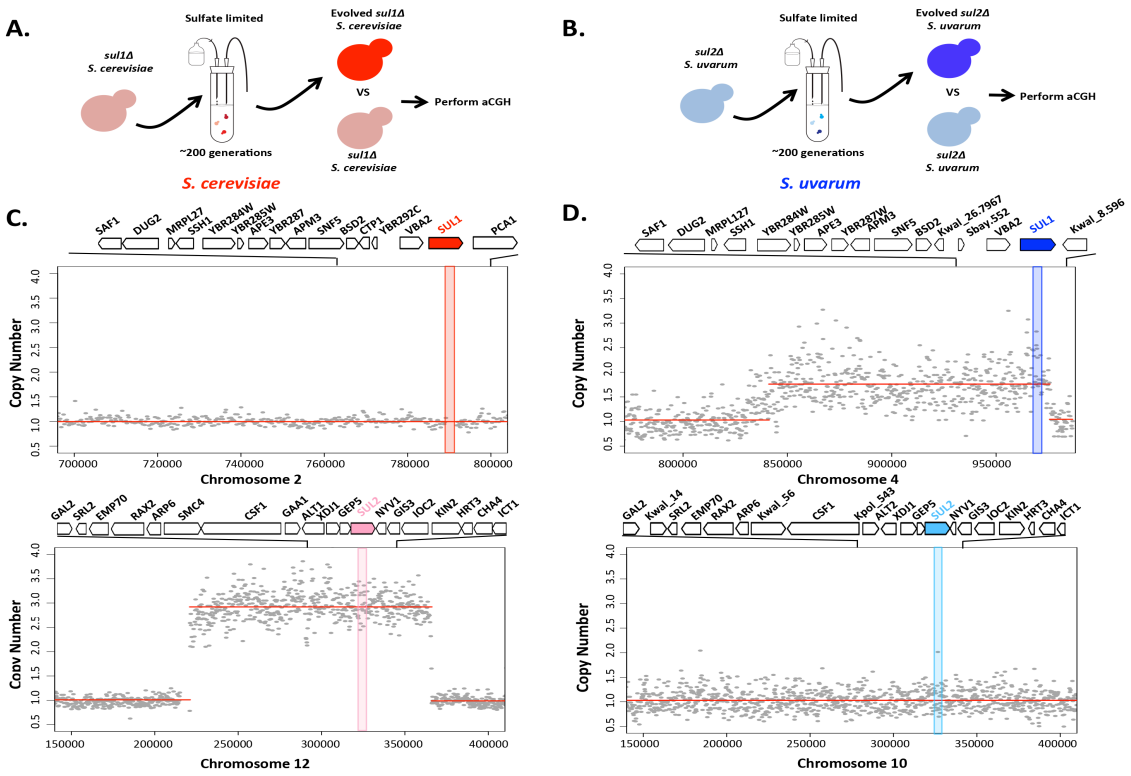


Figure 2.2. Alternate paralogs in *S. cerevisiae* and *S. uvarum* can amplify but such amplifications are not observed when preferred paralogs are expressed. A) Schematic illustrating how evolved strains were derived in *S. cerevisiae* and analyzed by aCGH to generate copy number plots shown below. B) Schematic illustrating how evolved strains were derived in *S. uvarum* and analyzed by aCGH to generate copy number plots shown below. C) Array CGH of evolved clone versus the parental genome of a *sul1Δ S. cerevisiae* strain at the relevant genomic segments surrounding the *SUL1* (top) and *SUL2* (bottom) loci. Array data (gray dots) indicate a copy number amplification. Segmentation-derived regions of average copy number are indicated in red. Segmentation defines a ~144 kb region of chromosome VII with a copy number estimation of 3. Genes along the top are represented in the locus in the expanded panel. D) Array CGH of evolved clone versus the parental genomes of a *sul2Δ S. uvarum* strain at the relevant genomic segments surrounding the *SUL1* (top) and *SUL2* (bottom) loci. Array data (gray dots) indicate a copy number amplification. Segmentation-derived regions of average copy number are indicated in red. Segmentation defines a ~134 kb region of chromosome IV with a copy number estimation of 2. Genes along the top are represented in the locus in the expanded panel.

2.3.4 Extra copies of sulfur transporter genes from *S. cerevisiae* and *S. uvarum* confer differential fitness effects

To test whether the functions of these genes may have diverged between these species, we measured the fitness effects of having additional copies of each gene. Previous studies have shown that the addition of *SUL1* on a low copy plasmid in *S. cerevisiae* increases the fitness of the strains by ~40% (Payen et al., 2013). To determine the effect of additional copies of *SUL1* and *SUL2* from *S. cerevisiae* and *S. uvarum,* we transformed *S. cerevisiae* with ARS/CEN plasmids individually containing each *SUL* gene along with 500 bp upstream of the coding region. We performed chemostat competition experiments between GFP+ and dark strains harboring additional copies of each gene in *S. cerevisiae* (**Figure 2.3A).** The fitness cost of expressing GFP, determined by competing isogenic wt strains with and without a GFP construct, is negligible (-0.02). The pairwise competitions provided fitness data that allowed us to more precisely determine the rank order of the fitness benefit of each gene amplification. The strain with an extra copy of *SUL1* from *S. cerevisiae (ScSUL1)* outcompeted all other strains, followed by *SUL2* from *S. cerevisiae (ScSUL2),* which had a comparable fitness effect to *SuSUL2*. The strain with the *SuSUL1* gene had the lowest fitness effect of all genes tested (**Figure 2.3B**). This result suggests that *SUL2* may have maintained a similar function between the two species, but *SUL1* function may have diverged. In support of our original hypothesis, the *SUL2* gene from *S. uvarum* (*SuSUL2*) conferred a greater fitness effect than the *S. uvarum SUL1 (SuSUL1)*. This result is also consistent with our predictions based on the evolution experiments, suggesting that *SuSUL2* amplification may have a greater selective benefit than amplification of *SuSUL1*.

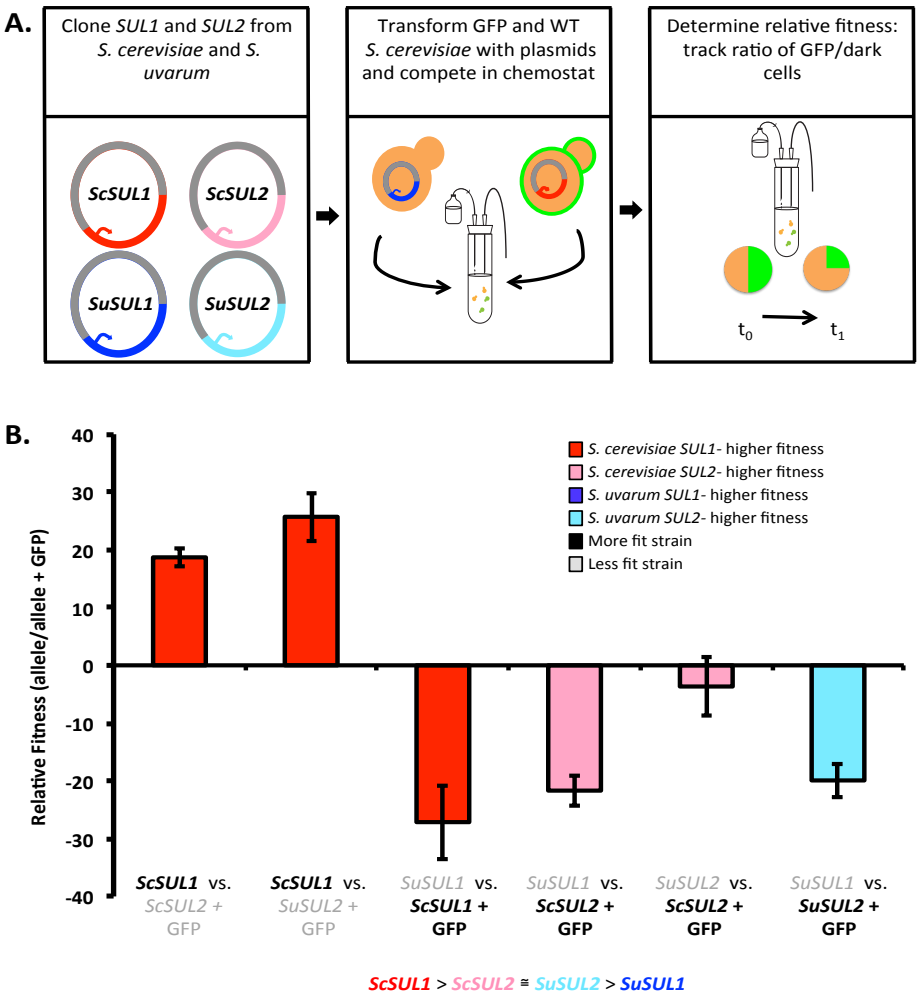


Figure 2.3. *SUL1* and *SUL2* have differential fitness effects between *S. cerevisiae* and *S. uvarum*. A) Schematic describing the strains that were competed against one another in the chemostat. B) The relative fitness of four strains of *S. cerevisiae* containing CEN plasmids with a *SUL1* or *SUL2* allele from *S. cerevisiae* or *S. uvarum* was determined in a pairwise manner. The fitness was measured in the chemostat under sulfate-limited conditions against a GFP-marked lab strain also containing CEN plasmids with either *SUL1* or *SUL2* from *S. cerevisiae* or *S. uvarum*. Each bar corresponds to the mean of 4 or more replicates ±SD. The direction of the bar illustrates the strain that is more fit and depends on the strain that is GFP+. The labels on the x-axis describe the dark strain over the GFP+ strain competed against each other, with the most fit strain bolded in black. The fitness ranking of all the alleles is indicated at the bottom.

To determine if these results were consistent across genetic backgrounds, we performed chemostat competition experiments between GFP+ and dark strains harboring additional copies of each gene integrated at the *URA3* locus in *S. uvarum* (**Figure 2.4A)**. The strain with an extra copy of *SUL1* from *S. cerevisiae (ScSUL1)* outcompeted all other strains, followed by *SUL2* from *S. uvarum (ScSUL2),* which had a greater fitness effect than *SuSUL1* (**Figure 2.4B)**. The strain with the additional copy of *ScSUL2* gene had the lowest fitness effect of all genes tested, which differs from the *S. cerevisiae* background results. These results suggest that other epistatic interactions may also contribute to the differences in the fitness effects of each allele between genetic backgrounds (**Figure 2.4B**).

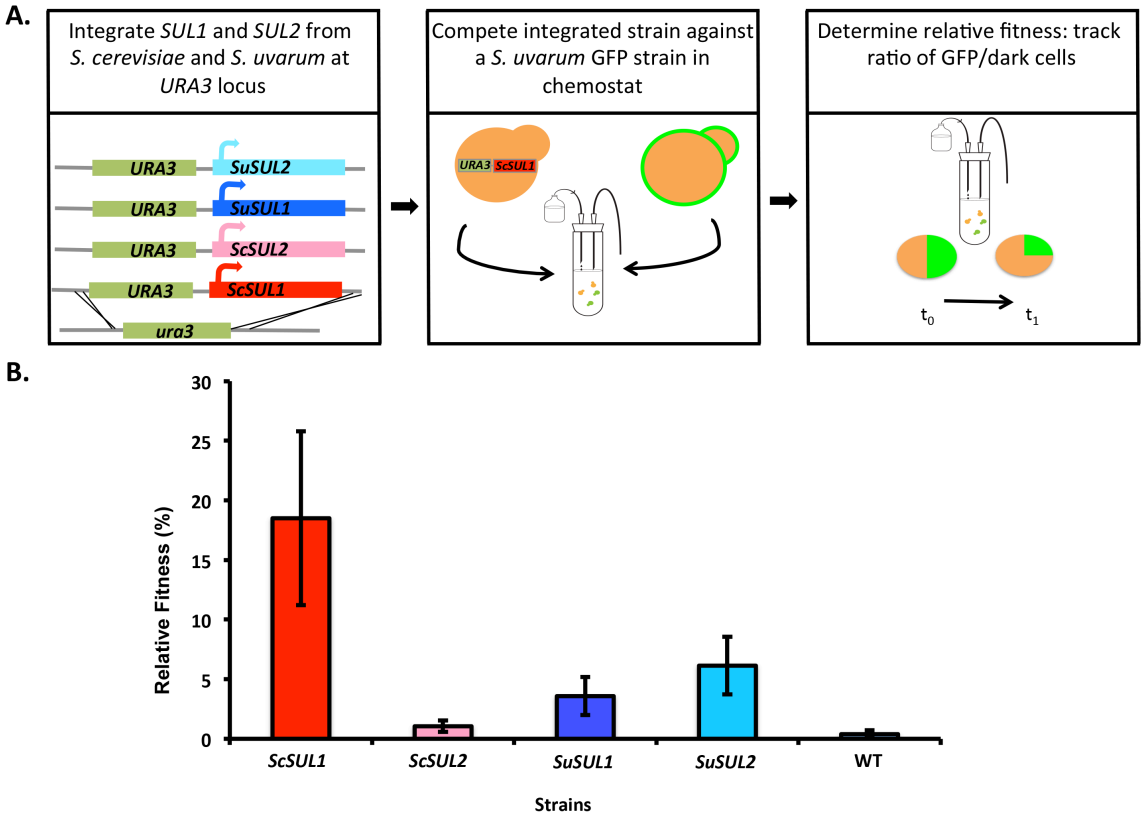


Figure 2.4. Fitness effects of *S. uvarum* strains harboring additional copies of *SUL* alleles. A) Schematic describing the strains that were competed against one another in the chemostat. B) The relative fitness of four strains of *S. uvarum* containing an additional integrated copy of either *SUL1* or *SUL2* from *S. cerevisiae* or *S. uvarum* was determined. The fitness was measured in the chemostat under sulfate-limited conditions against an isogenic GFP-marked strain. Each bar corresponds to the mean of 6 replicates ±SD.

2.3.5 *S. cerevisiae* x *S. uvarum* hybrid strains amplify the *ScSUL1* allele

In addition to testing the fitness effects of each *SUL1* and *SUL2* gene independently, we also investigated the amplification preference in the context of having all alleles present in one genome. Given the results from the single gene plasmid experiments above, we predicted that *ScSUL1* would be the preferred allele for amplification. We had previously created *de novo S. cerevisiae/S. uvarum* hybrid strains and subjected them to hundreds of generations of growth in sulfate-limited continuous culture. Evolved strains were then analyzed by aCGH to determine differences in genome content from their ancestral strains (see [38] for additional analysis).

Amplification of segments containing the *SuSUL1* or *SuSUL2* gene was never observed in 16 clones from 8 independent populations, and *SuSUL1* was even found deleted in one evolved clone, displaying loss of heterozygosity at this locus (**S2.5 Fig)**. In contrast, the *S. cerevisiae* copy of *SUL1* was found amplified in 14/16 evolved clones (**Figure 2.5B**). Copy numbers estimated from the array CGH data ranged from 3 to as many as 20 copies of *SUL1*. Centromere-proximal breakpoints varied from population to population, but amplicons extended to the most distal telomeric probe in all cases. Additional rearrangements were rarely observed in these strains (**S2.5 and S2.6 Figs**). When all four alleles are present in the same genome, *ScSUL1* amplifications are preferentially recovered, suggesting that *ScSUL1* amplification yields the greatest fitness advantage in this particular environment and genomic context.

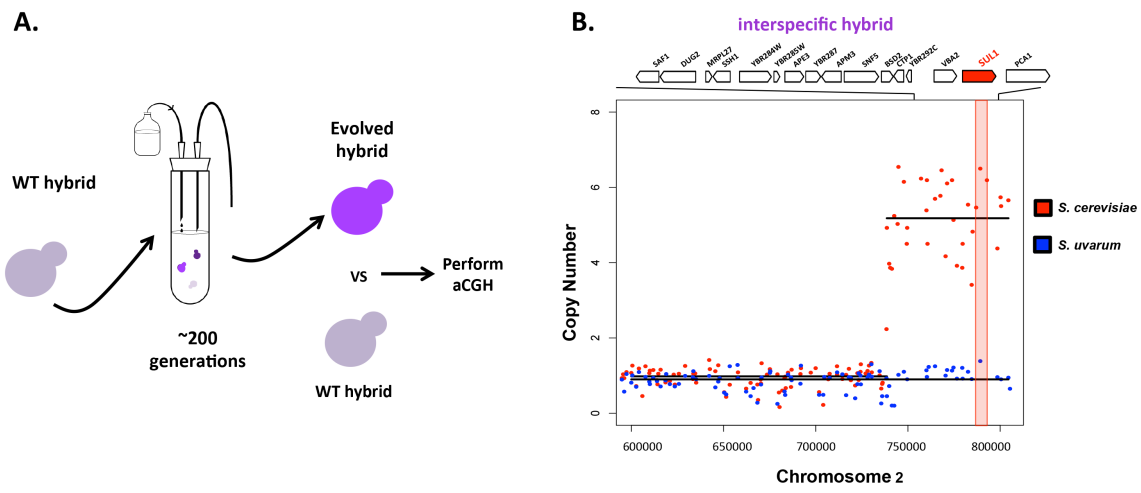


Figure 2.5. *S. cerevisiae* x *S. uvarum* hybrid strains amplify *S. cerevisiae SUL1*. A) Schematic illustrating how evolved strains were derived and analyzed by aCGH to generate neighboring copy number plot. B) Array CGH of evolved hybrid clone versus the parental *S. cerevisiae*/*S. uvarum* hybrid genome. Array data from *S. cerevisiae* (red dots) and *S. uvarum (*blue dots) are plotted according to the hybrid genome coordinates. Data support a copy number amplification of the *SUL1* locus of the *S. cerevisiae* allele. Black lines indicate the segmentation-derived regions of average copy number. Segmentation defines a ~65 kb region of chromosome II with a copy number estimation of 5. The region of the *S. cerevisiae SUL1* gene is shaded red.

2.3.6 Deletions of sulfate transporter genes display differential fitness effects between *S. cerevisiae* and *S. uvarum* genetic backgrounds

We have shown that the addition of extra copies of each gene results in an increased fitness in *S. uvarum* and *S. cerevisiae*, with *ScSUL1* yielding the greatest fitness increase, a result that corresponds to the amplification preferences in evolved strains derived from an interspecific hybrid. In addition, we deleted *SUL1* and *SUL2* in both *S. cerevisiae* and *S. uvarum* backgrounds to determine the relative fitness contributions of these loci in each background. We created *sul1*Δ and *sul2*Δ haploid strains and measured the competitive fitness of each null mutant in sulfate-limited conditions. We competed the *sul1*Δ and *sul2*Δ strains within each species against each other to calculate the fitness effect of each mutant. In *S. cerevisiae,* the *sul2*Δ strain outcompeted the *sul1*Δ strain, suggesting that *SUL1* in *S. cerevisiae* is the gene that is more important for growth in sulfate-limited conditions. Conversely, in *S. uvarum,* the *sul1*Δ strain outcompeted the *sul2*Δ strain, suggesting that *SuSUL2,* rather, is the gene that is more important for growth in sulfate-limited conditions (**Figure 2.6B**). Taken together with the fitness data from increasing the copy number of each gene, these data suggest differential *SUL1* and *SUL2* fitness contributions across these two species despite the genes' similarity in amino acid composition and genomic context.

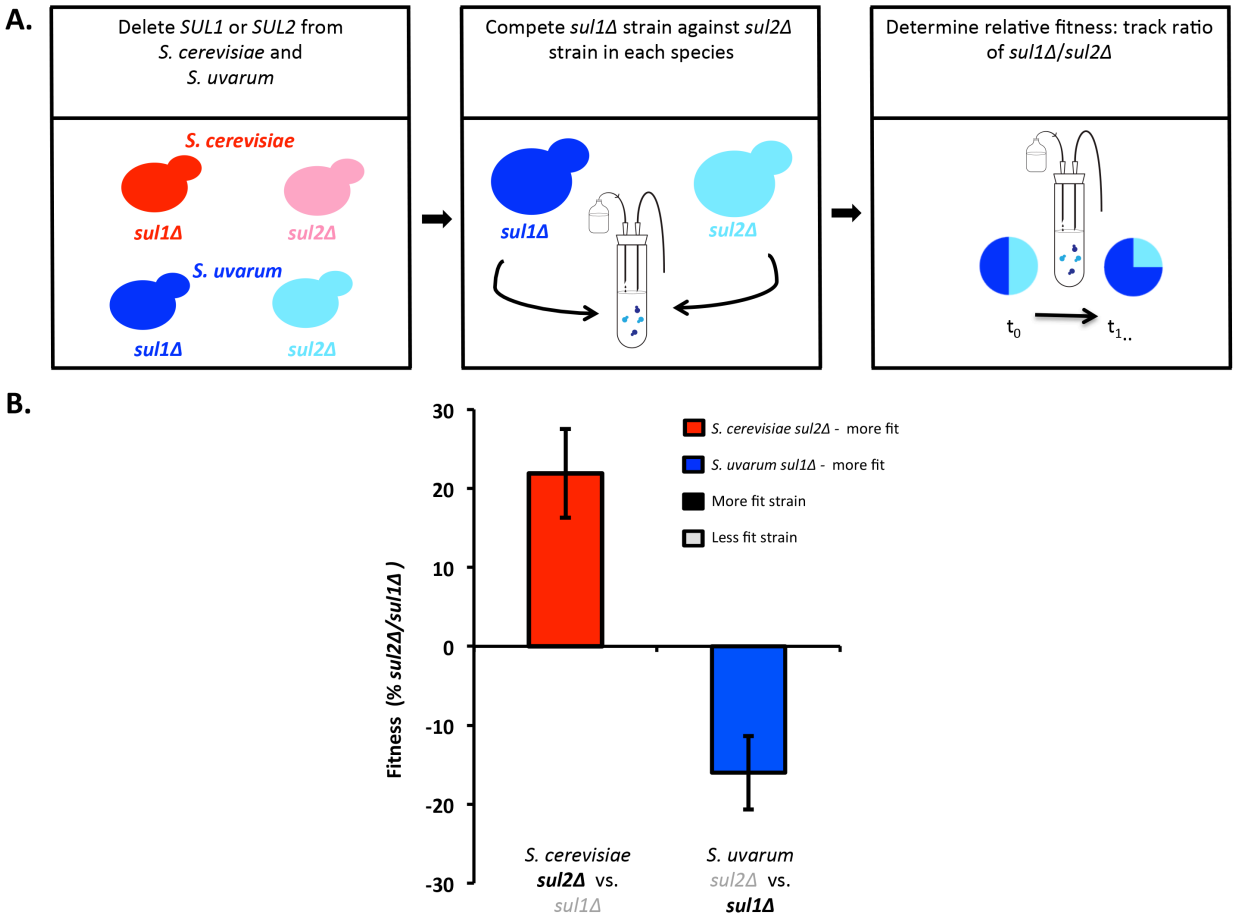


Figure 2.6. *sul1Δ* and *sul2Δ* have differential fitness effects between *S. cerevisiae* and *S. uvarum*. A) Schematic of strains used in the competition assay to measure the relative fitness of the null alleles in the *S. cerevisiae* and *S. uvarum* backgrounds. B) The relative fitness of two strains containing either a *sul1Δ* or *sul2Δ* allele in *S. cerevisiae* (red) and *S. uvarum* (blue) respectively was determined in a pairwise manner within each species. The fitness was measured in the chemostat under sulfate-limited conditions against a strain containing *sul2Δ* or *sul1Δ* allele in *S. cerevisiae* (red) and *S. uvarum* (blue) respectively. The x-axis indicates which strains were competed and the bolded strain represents the more fit strain. The proportion of each strain was determined by monitoring canavanine resistance, which differentially marked the competing strains and has previously been shown to be neutral (Gresham et al., 2008; Paquin and Adams, 1983). Each bar corresponds to the mean of 6 or more replicates ±SD. In *S. cerevisiae, sul2Δ* outcompetes *sul1Δ* and in *S. uvarum sul1Δ* outcompetes *sul2Δ.*

2.3.7 *SUL1* amplification in other species of the *sensu stricto* clade

In order to determine where the divergence in relative fitness effects between *SUL1* and *SUL2* in *S. cerevisiae* and *S. uvarum* occurred in evolutionary history, we tested the fitness of *SUL1* and *SUL2* from *S. paradoxus* and *S. mikatae*—two other species of the *sensu stricto* clade—and *SUL2* from *Naumovozyma castellii,* a more distant species that has not undergone gene duplication of this locus. We cloned the genes along with 500 bp upstream of the coding region from each species into an ARS/CEN plasmid and determined the relative fitness effect of the addition of the *SUL* genes in *S. cerevisiae* when competed against a plasmid-free strain. This experiment allowed us to calculate the relative fitness coefficient of each strain. All strains showed significantly higher fitness than wild type *S. cerevisiae*, with the relative fitness coefficients ranging from 18.1% to 43.8%, after correcting for the cost of carrying a plasmid (-5.4% ± 0.59). The *S. cerevisiae SUL1 (ScSUL1)* plasmid conferred a fitness benefit of 42.6% (**Figure 2.7B**). The strains containing *SUL1* from *S. paradoxus* and *S. mikatae* conferred a greater fitness advantage than *SUL2* from the respective species. In *N. castellii,* the singleton *SUL2* conferred the greatest fitness advantage of 43.8% (**Figure 2.7B**). One possible scenario to explain these results is that the new *SUL1* duplicate in the last common ancestor of *S. cerevisiae, S. paradoxus, S. mikatae* and *S. uvarum* may have maintained the high affinity function of the ancestor, while *SUL2* subfunctionalized or lost specificity. Alternatively, the *S. uvarum SUL1* paralog may have acquired mutations that decreased its fitness only in that lineage.

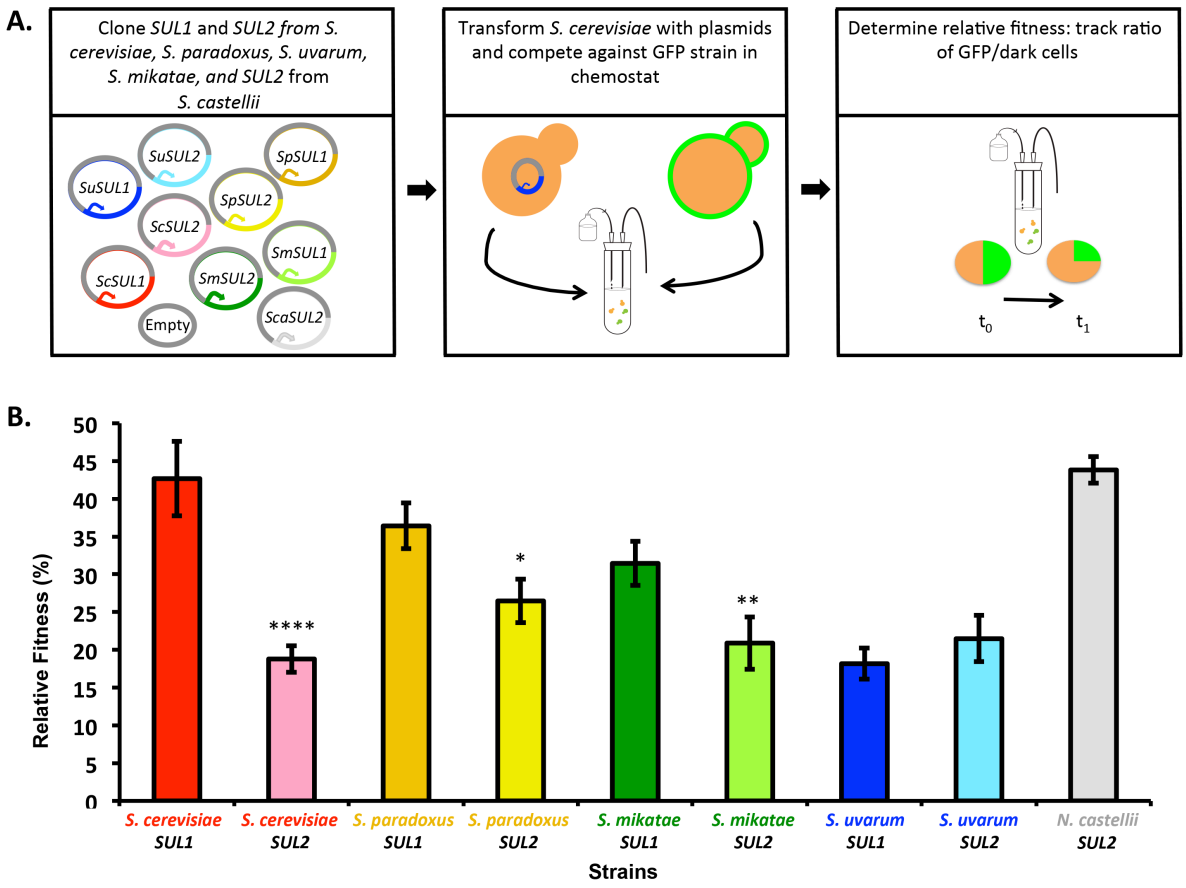


Figure 2.7. Plasmid-borne copies of *SUL1* cause a higher fitness benefit than *SUL2* from all species except *S. uvarum*. A) Schematic describing the strains that were competed against each another in the chemostat. B) The relative fitness of 9 *S. cerevisiae* strains containing a plasmid with *SUL1* and *SUL2* alleles from *S. cerevisiae, S. paradoxus, S. mikatae, S. uvarum,* and *N. castellii SUL2*. The fitness was measured in the chemostat under sulfate-limited conditions against a GFP-marked lab strain. Each bar corresponds to the mean of 4 or more replicates ±SD with asterisks indicating a p-value <0.05.

From these data, we can make predictions about the types of genomic events that would occur if we evolved *S. paradoxus* and *S. mikatae* under sulfate limited conditions. Since *SUL1* from both species resulted in the highest fitness benefit, we would expect to select for amplifications of the *SUL1* locus. To test this prediction, we grew four populations of *S. paradoxus*, *S. mikatae,* and *S. uvarum* for 200 generations in sulfate limited chemostats and determined the copy number variation between evolved populations and each ancestral strain using deep sequencing. We did not detect amplification events at the *SUL1* nor the *SUL2* locus in any of the four populations of *S. uvarum*. One explanation for this result could be due to the 200-generation timescale. The detection of the original *SUL2* amplification event occurred after 500 generations. However, consistent with expectations, we did identify two populations with an amplification containing the *SUL1* locus in *S. paradoxus* and one population in *S. mikatae* (**Figure 2.8**). Other aneuploidy and segmental amplifications occurred in addition to the *SUL1* locus amplification in the evolved populations (**S7 and S8 Figs**); however, none of these copy number variants included the *SUL2* locus. Overall, these data are consistent with the previous gene function measurements of each allele in *S. cerevisiae*, indicating that *SUL1* is more adaptive when amplified in *S. paradoxus* and *S. mikatae*.

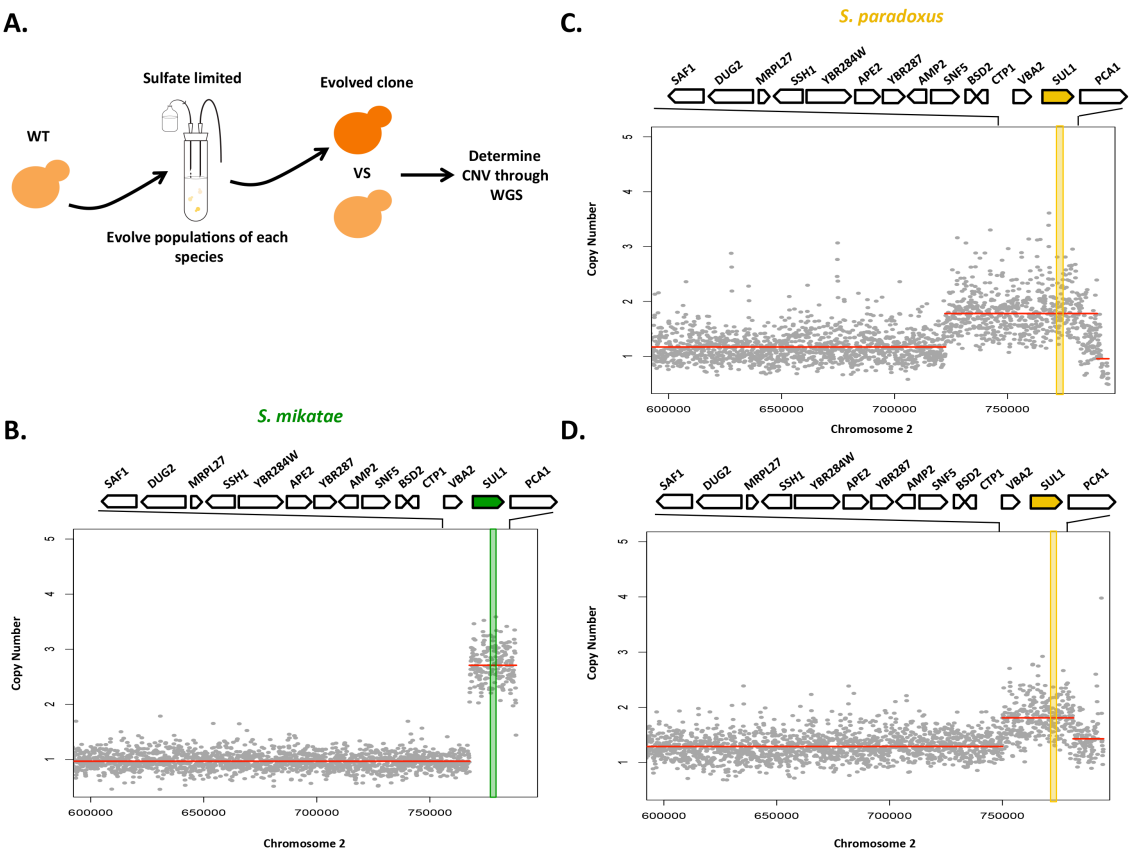


Figure 2.8. *SUL1* amplification in *S. paradoxus* and *S. mikatae* evolved populations. Copy number plots calculated with sequencing-depth ratios between evolved and parental genomes in 2 populations of *S. paradoxus* and one population of *S. mikatae* at ~200 generations. Gray dots represent the per nucleotide read-depth averaged across 100 bp window and normalized against the average coverage of the ancestral strain. Segmentation-derived regions of average copy number are indicated as red lines. Genes aligned along the top are represented in the locus in the expanded panel. The *SUL1* gene is shaded yellow in *S. paradoxus* and green in *S. mikatae*. A) Schematic illustrating how evolved strains of *S. paradoxus* and *S. mikatae* were derived and analyzed by WGS to generate neighboring copy number plots. B) Segmental amplification defines a ~64 kb region with a copy number of 2 in *S. mikatae*. C) Segmental amplification defines a ~20 kb region with a copy number of 2 in *S. paradoxus*. D) Segmental amplification defines a ~20 kb region with a copy number of 3 in *S. paradoxus*.

2.3.8 The species-specific relative fitness contributions among *SUL* genes are largely driven by promoter sequences.

Based on the similar results across *S. cerevisiae, S. mikatae,* and *S. paradoxus*, we decided to focus on understanding what is different about the paralogs in *S. cerevisiae* vs. *S. uvarum*. To identify the genetic region responsible for the differences in fitness effects of *SUL1* and *SUL2* between the two sister species, we created chimeric constructs composed of different combinations of the promoter and open reading frame (ORF) of each gene. Rich *et al* recently used a deep mutational scanning approach to identify the functional elements of the *ScSUL1* promoter that are crucial for growth in sulfate limitation (Rich et al., 2016). Based on their results, we cloned 500 bp upstream of each ORF (the region encompassing all elements that positively influence *SUL1*'s fitness contribution) and cloned the ORF until the stop codon. We then cloned all 12 chimeric combinations of promoter and ORF into a low copy ARS/CEN plasmid. Wild-type *S. cerevisiae* strains were transformed with the individual plasmids carrying chimeric *SUL* constructs and competed against a plasmid-free strain to calculate the relative fitness coefficient of each strain in sulfate-limited media. Additionally, the non-chimeric alleles were also tested against a plasmid-free strain, with a total of 16 alleles tested.

As seen in **Figure 2.9,** the fitness coefficient values ranged from 0.2 to 38% after correcting for the cost of carrying a plasmid (-5.4% ± 0.59), which was calculated by competing a strain with an empty plasmid against a WT strain. When placed under the same promoter, the *SuSUL1* ORF had a greater fitness advantage than the *SuSUL2* ORF, opposite to the result obtained when each ORF was driven by its native promoter. All chimeras containing the promoter region of *SuSUL1* showed substantial decreases in

fitness. This result suggests that expression differences between the two species may largely explain the differential fitness effects of the two *SUL1* genes. Interestingly, the chimeric allele containing the *SuSUL2* promoter with the *SuSUL1* ORF *(P_{SuSUL2}-SuSUL1*) recapitulates the fitness effect of *ScSUL1*. Additionally, strains containing the promoter of *ScSUL1* or *ScSUL2* resulted in similar fitness patterns when paired with the three other ORFs, with the *ScSUL1* coding region yielding the highest relative fitness. However, when promoters of *SuSUL1* or *SuSUL2* were paired with the other three ORFs, we identified a different ranking of fitness patterns, with the *SuSUL1* coding region yielding the highest fitness. We did not attempt to further dissect these apparent epistatic interactions between the promoters and coding regions; however, such complex genetic interactions have been observed in other contexts (Breen et al., 2012; Costanzo et al., 2010; Engle and Fay, 2012; Khan et al., 2011).

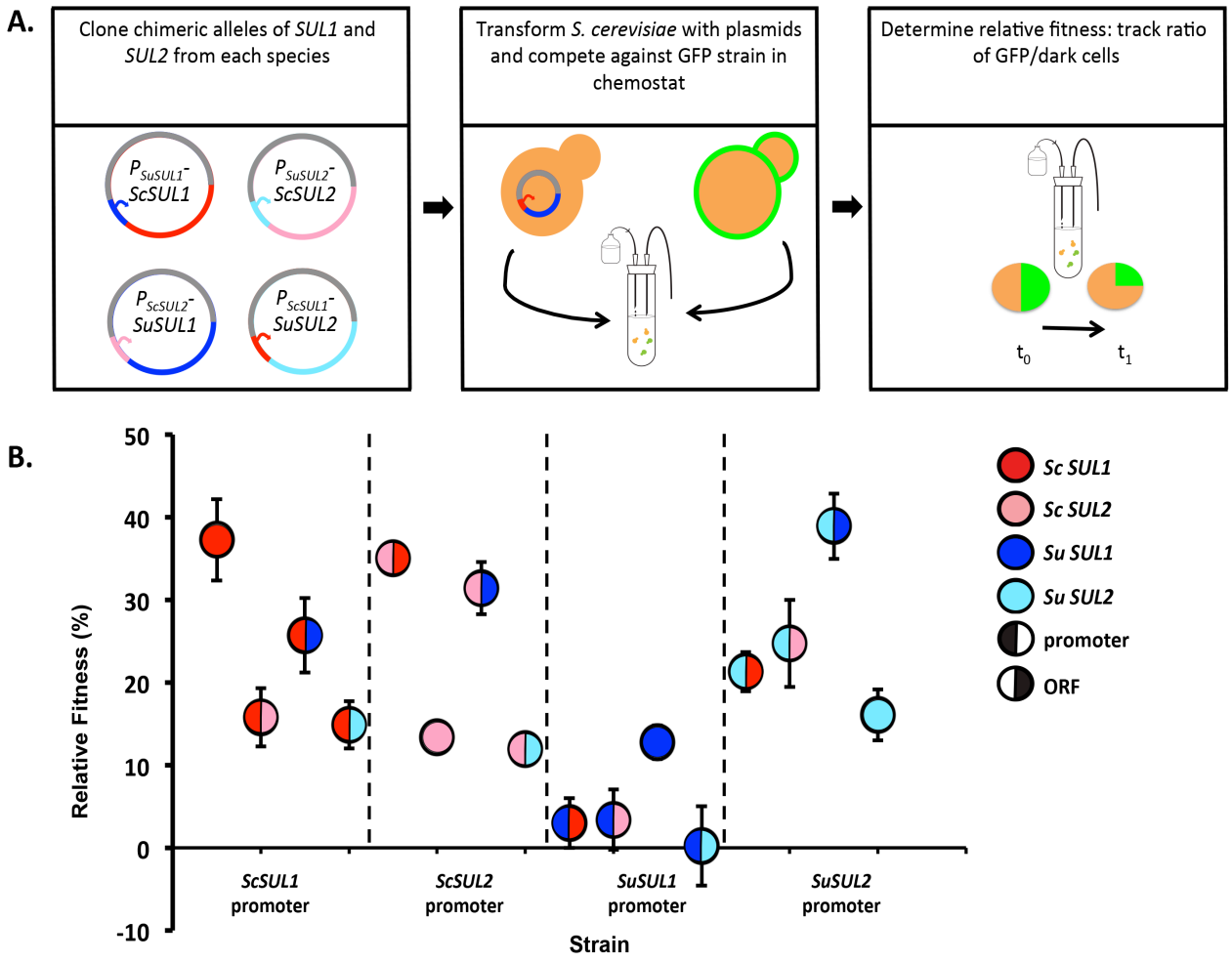


Figure 2.9. The promoter from *S. uvarum SUL1* reduces the fitness effect of *SUL1* and *SUL2* amplification. A) Schematic illustrating the strains that were used in the competition. B) The relative fitness of 16 *S. cerevisiae* strains containing a plasmid including chimeric constructs of *SUL1* and *SUL2* alleles from *S. cerevisiae* or *S. uvarum*. The figure is split into quadrants by promoter; *ScSUL1, ScSUL2, SuSUL1, SuSUL2* (left to right) where each circle represents one chimeric construct (Engle and Fay, 2012). The left portion of the circle represents the promoter allele whereas the right portion represents the coding allele. The fitness was measured in the chemostat under sulfate-limited conditions against a GFP-marked lab strain. Each bar corresponds to the mean of 4 or more replicates ±SD.

Since the results from the chimeric constructs suggested that the promoter region is largely responsible for the differences in fitness, we sought to measure gene expression levels driven by each promoter. We used reverse transcriptase real time PCR (RT-PCR) to determine the expression level of *ScSUL1* under the control of all four promoters in *S. cerevisiae* strains grown at steady state in sulfate-limitation. We found that the expression level of the *ScSUL1* chimera with the promoter from *SUL1* from *S. uvarum* (*P_{SuSUL1}-ScSUL1*) was significantly reduced in comparison to the other promoters (**Figure 2.10A**). We also found a modest correlation between expression level and the fitness value of each construct (R^2=0.55) (**Figure 2.10B**). This result demonstrates that the differences between the fitness contributions of the two transporter genes may be due to gene expression differences.

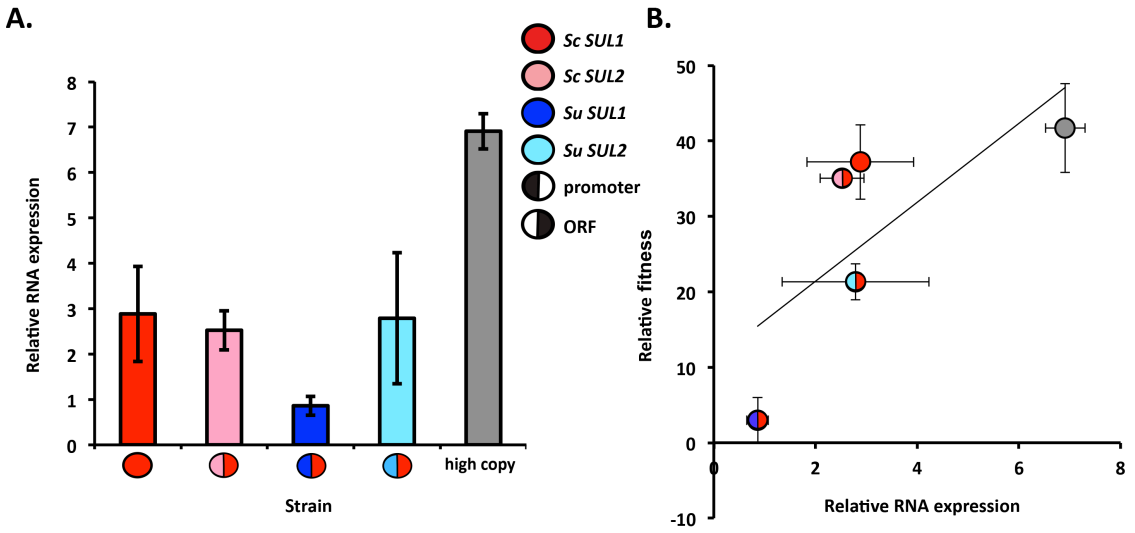


Figure 2.10. Reduced expression of chimeric construct with the *S. uvarum SUL1* promoter. A) The relative expression level (compared to a single copy gene) of four *S. cerevisiae* strains containing a plasmid including the coding region of *S. cerevisiae SUL1* and the non-coding region of *SUL1* or *SUL2* from *S. cerevisiae* or *S. uvarum.* Each split circle represents the chimeric construct. The left portion of the circle represents the promoter allele whereas the right portion represents the coding allele. The grey fitness value is a control strain of *S. cerevisiae* with five copies of *SUL1*. Each bar corresponds to the mean of 4 or more replicates ±SD. B) The relative expression (compared to a single copy gene) of the construct weakly correlates with the relative fitness (compared to wild-type) of the strain (R^2=0.55). The X and Y error bars indicate ±SD from fitness and expression data.

2.4 DISCUSSION

In this work, we used comparative experimental evolution to investigate how genetic background influences the genetic mechanisms of adaptation to sulfate limitation across different species of yeast in the *Saccharomyces* clade. We identified differential amplification of gene duplicates that encode sulfate transporters in *S. cerevisiae* and *S. uvarum*. Collectively, our results display an example of adaptation via amplification of different genomic loci, likely driven by regulatory divergence of paralogs.

Specifically, we have shown *SUL1* amplification during long-term growth in sulfate-limited conditions occurs in all species tested in the *Saccharomyces* clade except *S. uvarum*. While the number of *S. paradoxus, S. mikatae,* and *S. uvarum* populations that were used for the laboratory evolution experiments was small (n=4-6), we have repeatedly identified *SUL1* locus amplifications in all reported evolution experiments of wild type *S. cerevisiae* (n=25/25). Therefore, it is surprising that even within two evolved populations of *S. uvarum,* we did not identify *SUL1* amplification, but instead identified *SUL2* locus amplifications in both populations after 500 generations. Additionally, two of the evolved populations in *S. paradoxus* and one population of *S. mikatae* amplified *SUL1*. The other populations that did not amplify *SUL1* or *SUL2* may contain other events that may be equally or more beneficial than either amplification, or additional time may be required for the amplification event to occur and rise to high frequency (>200 generations) (**S2.7 and S2.8 Figs**). This point is further supported by additional evolution experiments we performed in *S. uvarum* for 200 generations where neither *SUL1* nor *SUL2* amplifications were detected, suggesting that amplification events are dynamic and may depend on longer time scales to occur and/or achieve high

frequency (**S2.9 Fig)**. These findings also demonstrate that other means of adaptation to sulfate limitation may exist, since populations from both *S. paradoxus* and *S. mikatae* amplify other regions of the genome in addition to *SUL1* or do not amplify either of the *SUL* genes at all **(S2.7 and S2.8)**. Further work will be required to understand the genetic differences that mediate these other evolutionary trajectories and connect them definitively to fitness changes.

Our results contribute to ongoing efforts to understand the mutations that drive adaptation, a long-standing question in evolutionary biology. There are examples of parallel molecular evolution that occur across genetic backgrounds for many traits (Barrick et al., 2009; Dettman et al., 2012; Lieberman et al., 2011; Ostrowski et al., 2008; Rokas and Carroll, 2008; Woods et al., 2006), suggesting that genetic background plays a relatively unimportant role in determining the outcome of adaptation at the molecular level. A more recent study, however, tested how genetic differences between strains of bacteria influence their adaptation to a common selection pressure and found that parallel evolution was more common within-strains than between-strains, implying that genetic background has a detectable impact on adaptation (Vogwill et al., 2014). Taken together, it is unclear to what degree genetic background impacts the mechanism and rate of adaptation to a novel selection pressure. Our study has identified differential locus parallelism between sulfate transporter loci in *S. cerevisiae* and *S. uvarum,* demonstrating one example where genomic background influences the route taken to adapt to sulfate limitation during experimental evolution.

To further investigate the effect of genetic context and whether this was due to coding or non-coding variation, we generated chimeric alleles of promoter and coding

regions between *S. cerevisiae* and *S. uvarum SUL1* and *SUL2* genes. We identified poor fitness outcomes associated with the non-coding region of the *SUL1* gene in *S. uvarum,* along with other complex interactions with the coding regions. These results suggest that the accumulation of mutations in the non-coding region of *S. uvarum SUL1* may have resulted in reduced expression, thus driving selection for *SUL2* amplification during adaptation of *S. uvarum* to sulfate limited conditions. Rich *et al* recently used a deep mutational scanning approach to identify the functional elements of the *ScSUL1* promoter that are crucial for growth in sulfate limitation (Rich et al., 2016). This same approach could be applied to the promoter region of *SUL1* in *S. uvarum* to determine which sequences are responsible for these differences in activity.

Many studies have aimed to determine whether adaptation and phenotypic change typically occur from mutations in non-coding or coding regions in the genome (Fraser et al., 2010; Hoffmann and Palmgren, 2016; Kvitek et al., 2008; Stapley et al., 2010; Wray, 2007). In the case of gene duplicates, it has been proposed that their retention provides genetic redundancy, buffering the mutational space to either acquire new function, or to partition the ancestral function between duplicates. Gradual stochastic changes in expression level may lead to an eventual imbalance in the selective pressure between the two duplicates (Gout and Lynch, 2015). These gradual changes in gene expression may play a significant role in shaping the adaptive landscape over time, resulting in different adaptation outcomes across diverse genetic backgrounds. Our results provide an example of divergent adaptation through changes in expression of one duplicate in the *S. uvarum* lineage in the *Saccharomyces* clade. In the case of nutrient limitation, a simple modification in expression may be more likely to suffice, since the metabolic pathway for

uptake and utilization already exists, and increasing uptake is a straightforward solution (Dettman et al., 2012). Alternatively, differential tradeoffs between toxic metal resistance and ion transport may exist between species and result in altered sulfur biosynthesis requirements to synthesize glutathione, a key factor in the cell's defense against oxidative stress and metal toxicity, and/or other sulfur-containing compounds (Chang and Leu, 2011; Kieliszek et al., 2015; Thorsen et al., 2007).

In addition to metal exposure, nutrient limitation is also a likely scenario experienced by wild and industrial yeast strains. Growing evidence suggests that domesticated *Saccharomyces* species have been exposed to sulfate related selective pressures through the selection for favorable characteristics associated with brewing. In lager brewing yeast, increased sulfite production is important for its antioxidant properties and for preserving favorable flavor profiles (Donalies and Stahl, 2002). *Saccharomyces pastorianus* is a lager brewing species found only in the brewing environment and appears to be an allotetraploid hybrid between *S. cerevisiae* and *S. eubayanus* (Dunn and Sherlock, 2008). Interestingly, *S. pastorianus* carries inactive copies of *SUL1* from *S. cerevisiae* and *S. eubayanus,* while retaining functional copies of *SUL2* which have been shown previously to improve sulfite production when overexpressed (Libkind et al., 2011; Nakao et al., 2009b, 2009a). Identifying the genetic basis of traits under selection in a particular environment may not only help highlight the emergence of new traits but also inform ways to engineer further improvement.

2.5 MATERIALS AND METHODS

2.5.1 Yeast strains, plasmids, and culture conditions

The strains used in this study are listed in **Supplemental Table 2.1**. The *S. cerevisiae* strains used in this study were from the FY series FY4 in the S288c background, with the exception of the interspecific hybrids, which utilized GRF167. The *S. uvarum* strains used were derived from the CBS 7001 background. The *S. mikatae* strain was IFO 1815 and the *S. paradoxus* strain was CBS 432. The *N. castellii* strain was CBS 4309. The *SUL1* and *SUL2* deletion strains were created in *S. cerevisiae* and *S. uvarum* by targeting 50 bp upstream of the ATG and 100 bp upstream of its stop codon. The deletions were confirmed with primers targeting approximately 175 bp upstream of the ATG (**Supplemental Table 2.3)**.

To test the fitness due to the amplification of *SUL1* or *SUL2* from each species, we transformed DBY7283, a *ura3 S. cerevisiae MAT*α strain, with a low-copy plasmid (Ho et al., 2009). Phusion PCR was used to amplify 500 bp upstream and 5 bp downstream of the stop codon of *SUL1* and *SUL2* from *S. cerevisiae, S. uvarum, S. paradoxus,* and *S. mikatae,* and *SUL2* from *S. castelli.* Each *SUL1* and *SUL2* gene was blunt cloned into pIL37 using primers listed in **Supplemental Table 2.3.** All plasmids used in this study are listed in **Supplemental Table 2.2.** The haploid *S. cerevisiae* strain used in the competition experiments was a haploid FY MATα where the *HO* locus had been replaced with *eGFP* as previously described (Payen et al., 2013). The diploid *S. cerevisiae* GFP+ strain was made by crossing the haploid FY MATα strain, where the *HO* locus had been replaced with *eGFP,* to a MATa FY strain. The *S. uvarum* GFP+ haploid strain was created by replacing the *HO* locus with eGFP by amplifying the NatMX-GFP construct from the plasmid YMD1139. The strain was verified using primers that target 600pb upstream of the *HO* locus. The fitness of the haploid *S. uvarum*

GFP+ strain, YMD2869, was 0.388% +/- 0.33 (n=2). The fitness of the diploid *S. uvarum* GFP+ strain, YMD2869, was 2.33% +/- 0.19 (n=2). To directly compete two strains each containing an additional copy of either *SUL1* or *SUL2* from *S. cerevisiae* or *S. uvarum,* a GFP+ *ura3 S. cerevisiae* strain was transformed with plasmids containing either *SUL1* or *SUL2* from *S. cerevisiae* or *S. uvarum.* These GFP+ strains were used in a competitive assay (see below) against strains also containing additional copies of each gene.

To test the fitness due to the amplification of *SUL1* or *SUL2* from each species in the *S. uvarum* background, we integrated each *SUL* allele into YMD2823, a *ura3*Δ *S. uvarum MAT*α strain, at the *URA3* locus due to the high loss rate of *S. cerevisiae* CEN plasmids (Scannell et al., 2011). We used primers listed in **Supplemental Table 2.3** to amplify 700 bp upstream and 214 bp down stream of each *SUL* allele ORF and cloned the construct into a CEN plasmid with the *URA3* marker. To create homology to the *URA* locus, we amplified each allele and the *URA3* marker using primers indicated in **Supplemental Table 2.3.** Strains were verified using primers that target 200 bp upstream and downstream of the *URA3* locus and Sanger sequenced.

The chimeric plasmids were created by amplifying 500 bp upstream of the start codons of the *SUL1* and *SUL2* ORFs from *S. cerevisiae* and *S. uvarum* and cloning each upstream region into YMD2307 using primers with added SnaBI sites at the 3' end **()**. Each plasmid was digested with SnaBI and *SUL1* or *SUL2* from *S. cerevisiae* or *S. uvarum* was ligated immediately adjacent to the previously cloned upstream region, creating a total of twelve different chimeric strains.

2.5.2 Creation of hybrids

de novo hybrids between *S. uvarum* and *S. cerevisiae* were created by mating. Pulsed field gel analysis of the resulting strains confirmed the presence of both sets of chromosomes with no apparent size polymorphisms. Microarray analysis (see protocol below) of the hybrid DNA versus purebred DNA from each species also confirmed that these strains contained a complete haploid genome from each parent. Microarray data are deposited in the Gene expression Omnibus (GEO) repository under accession number GSE87401 and in the Princeton Microarray Database.

2.5.3 Microarray design

The *S. cerevisiae* and *S. uvarum* genomes were downloaded from the *Saccharomyces* Genome Database and concatenated to create a hybrid genome. The program Array Oligo Selector was used to design 70mers to each open reading frame in both genomes. Under the default stringency settings, 711 genes were too similar to another sequence in the combined genomes for a sufficiently unique oligo to be designed. For these cases, the program was rerun in the context of each single genome in order to provide more complete coverage of the purebred genomes. 485 genes were still too similar to other sequences in the single genomes to pass this test and were left off the array. The resulting 4840 *S. uvarum* and 6423 *S. cerevisiae* 70mers were purchased from Illumina.

2.5.4 Microarray printing and preparation

70mer DNA was resuspended at 40 μM in 3X SSC and printed using a pin-style

arraying robot onto aminosilane slides in a controlled-humidity environment. Slides were UV crosslinked at 70 mJ. On the day of hybridization, the slides were blocked by agitating for 35 minutes at 65C with 1% Roche blocking agent in 5X SSC and 0.1% SDS. Slides were then rinsed with water for 5 minutes and spun dry.

2.5.5 Comparative genomic hybridization

Hybridization conditions were optimized to maximize specificity. DNA from *S. cerevisiae* was labeled with one fluor and DNA from *S. uvarum* labeled with another and competitively hybridized to the arrays under a variety of DNA quantity, hybridization volume and temperature, and wash stringency conditions. As expected because of the 2-tier design strategy, less than 5% (563/11263) showed evidence of cross-hybridization with signal significantly over background levels in both channels. These probes were filtered out of all hybrid datasets.

All microarray manipulations were performed in an ozone-free environment. 4 μg DNA was sonicated to a size range near 1 kb then purified by Zymo DNA clean and concentrator columns. Labeling of 2 μg sonicated DNA was done by random-primed klenow incorporation of Cy-nucleotides either with the Invitrogen Bioprime kit according to the manufacturer's instructions, or with individually purchased reagents as previously reported (Dunham et al., 2002). The labeled reactions were purified by Zymo columns and measured for labeling yield and efficiency using a nanodrop spectrophotometer. 1 μg of each labeled DNA were mixed with Agilent blocking reagent and 2X hybridization buffer in a total volume of 400 μl, heated at 95°C for 5 minutes, and hybridized to a prepared microarray using an Agilent gasket slide. Hybridizations were performed

overnight at 65°C in a rotating hybridization oven. Gaskets slides were removed in 1X SSC and 0.1% SDS solution. Arrays were agitated for 10 minutes in a 65°C bath of the same wash buffer, then washed on an orbital shaker for 10 minutes in a new rack in 1X SSC, ending with 5 minutes in 0.1X SSC. Arrays were then spun dry and scanned in an Agilent scanner. The resulting images were analyzed using Axon Genepix software version 5. Complete microarray data are available for download from the Princeton Microarray Database and GEO under accession GSE87401.

Data were linearly normalized and filtered for spots with intensity of at least 2 times over background in at least one channel. Manually flagged spots were also excluded. These filters were adequate to routinely filter out >95% of empty spots and retain >95% of hybridizing spots.

2.5.6 Continuous culture evolution experiments

A single colony of *S. mikatae* and *S. paradoxus* and *S. uvarum* was inoculated into sulfate-limited chemostat medium with ura supplemented, grown overnight at 30°C, and 100 μL of the culture was inoculated into a ministat chamber (Miller et al., 2013) containing 20 mL of the same medium at 30°C. After 30 hr, the flow of medium was turned on at a dilution rate of 0.17 ± 0.01 hr^{-1}. Four chemostats were inoculated from four individual colonies for each species and cell samples (glycerol stock and dry pellet) were passively collected every day from fresh effluent for ~200 generations. DNA was isolated by a modified Smash-and-Grab protocol from each endpoint population (Hoffman and Winston, 1987). Whole genome sequencing of the evolved and ancestral populations was performed as described below.

Longer term *S. uvarum* and hybrid evolution experiments were performed in ATR Sixfors fermentors modified to run as chemostats, as described (Gresham et al., 2008), with the exception that *S. uvarum* populations were held at 25°C. Prior experiments comparing this system with the ministat system demonstrated that they are nearly equivalent (Miller et al., 2013).

To determine if *SUL1* would amplify in *S. uvarum,* four individual colonies of a *sul2*Δ *S. uvarum* strain were inoculated into four sulfate-limited ministat chambers as previously described. Array CGH was performed on the four populations after 260 generations using the ancestral *sul2*Δ deletion strain as the reference.

Yeast samples for real-time PCR analysis were collected directly from the culture vessels, when the cultures reached steady state (approximately 3 days at ~25 generations). The cells were filtered on Nylon membrane (0.45 μm pore size) and immediately frozen in liquid nitrogen and stored at -80ºC until RNA extraction.

2.5.7 Competition experiments

The pairwise competition experiments were performed in ministats. Each competitor strain was cultured individually. Upon achieving steady state, the competitors were mixed in 50:50 ratio. Each competition was conducted in two biological replicates for 15 generations after mixing. Samples were collected and analyzed three times daily. The proportion of GFP+ cells in the population was detected using a BD Accuri C6 flow cytometer (BD Biosciences). The data were plotted as ln[(dark cells/GFP+ cells)] vs. generations. The relative fitness coefficient was determined from the slope of

the linear region by the use of linear regression analysis (see schematic in **S2.10 Fig**)(Dykhuizen and Hartl, 1983).

The gene deletion competition assays were performed using two different drug resistant markers. For testing the fitness of either the *sul1*Δ or *sul2*Δ deletion strain in *S. cerevisiae* or *S. uvarum,* a spontaneous canavanine-resistant mutant (Can^R) was selected. Two 20 mL chemostats were inoculated with either deletion strain marked with either Can^R or the canavanine sensitive (Can^S) strain containing the alternate deleted allele. Cultures were brought to steady-state conditions over a period of 15 generations. 10 mL from the chemostat containing the canavanine sensitive (Can^S) strain (containing the alternate deleted allele) was removed and replaced with 10 mL from the chemostat containing the Can^R marked clone. We sampled the chemostat an average of every 5 generations for approximately 30 generations. Cells were sonicated, diluted, plated on rich nonselective media, and grown for 2 days at 30°C. We counted >200 colony forming units using sterile methods. Cells were then replica-plated to synthetic complete minus arginine media containing 60 mg/L canavanine and allowed to grow at 30°C or 25°C for 3 days. Can^R cells were identified as fully formed colonies (Gresham et al., 2008).

2.5.8 Total RNA extraction and Quantitative RT-PCR

RNA was extracted from the filtered sample by acid phenol extraction and quantified using a nanodrop spectrophotometer. 90 µg of RNA was cleaned-up using the Qiagen RNA easy kit according to the manufacturer's instructions (Qiagen).

Contaminating DNA was removed by using Rapid DNase out removal kit on 2 µg of RNA in a 100 µL reaction (Thermo).

Oligonucleotides for real-time PCR are listed in **Supplemental Table 2.3**. One microgram of total RNA was reverse-transcribed into cDNA in a 20 µL reaction mixture using the SuperScript VILO cDNA synthesis kit (Life). The cDNA concentrations were then analyzed using the nanodrop. For the RT-PCR, each sample was tested in triplicate in a 96-well plate using SYBR. The reaction mix (19 µL final volume) consisted of 10 µL of LightCycler 480 SYBR Green I Master (Roche), 2 µL of each primer (5 mM final concentration), 5 µL of H_2O, and 1 µL of a 1/100 dilution of the cDNA preparation. A blank was also incorporated in each assay. The thermocycling program consisted of one hold at 95°C for 4 min, followed by 50 cycles of 10 sec at 95°C and 45 sec at 56°C. The quantification of the expression level of *SUL1* was normalized with *ACT1* and the standard deviation was taken between four replicates.

2.5.9 Nextera libraries and whole-genome sequencing

Genomic DNA libraries were prepared for Illumina sequencing using the Nextera sample preparation kit (Illumina). Barcoded libraries were quantified on an Invitrogen Qubit Fluorometer and submitted for 150 bp paired end sequencing on an Illumina HiSeq 2000. Read data have been deposited at the NCBI under the Bioproject accession number PRJNA297229. The reads were aligned against the reference strain of *S. uvarum* (CBS 7001), *S. mikatae* (IFO 1815), and *S. paradoxus* (CBS 432) using Burrows-Wheeler Aligner (Li and Durbin, 2009) . The sequence coverage of the nuclear genome

ranged from 70 to 300x. Copy-number variations (CNVs) were detected by averaging the per-nucleotide read depth data across 100 bp windows. For each window, the $\log_2$ratio in read depth between the evolved and parental strain was calculated. The copy number was calculated from the $\log_2$ratios and plotted using the R package DNAcopy (Araya et al., 2010).

2.6 ACKNOWLEDGEMENTS

We thank members of the Dunham lab, especially Caiti Heil for helpful discussions. Thanks to Yixian Zheng and Doug Koshland for contributing to the initial experimental design, creating yeast strains, and purchasing the oligonucleotides used for the microarrays. Thank you to Celia Payen for providing the fitness data for a control strain used in **Figure 2.9B.** I thank Aaron Miller, Ivan Liachko, Anna B. Sunshine, Bryony Lynch, Mei Huang, Erica Alcantara, Christopher G. DeSevo, Dave A. Pai, Cheryl M. Tucker, Margaret L. Hoang for their involvement in this project.

CHAPTER 3: TRANSPOSON INSERTIONAL MUTAGENESIS IN *SACCHAROMYCES UVARUM*: DISSECTING THE GENETIC BASIS OF DIFFERENTIAL GENE DISPENSIBILITY BETWEEN TWO YEAST SPECIES

This chapter is based on the following manuscript with the same title, in preparation for submission to *Genome Research*

3.1 ABSTRACT

To understand how complex genetic networks perform and regulate diverse cellular processes, the function of each individual component must be defined. Comprehensive phenotypic studies of mutant alleles have been wildly successful in model organisms in determining what processes depend on the normal function of a gene. These results are often translated to the increasing number of newly sequenced genomes by using sequence homology. However, sequence similarity does not always mean identical function or phenotype, suggesting that new methods are required to functionally annotate newly sequenced species. We have implemented comparative functional analysis by high-throughput experimental testing of gene dispensability in *Saccharomyces uvarum*, a sister species of *S. cerevisiae.* We created haploid and heterozygous diploid Tn7 insertional mutagenesis libraries in *S. uvarum* to identify species-specific essential genes with the goal of detecting genes with divergent function. Using deep sequencing, we identified 46,326 and 42,904 independent insertion sites found in 79% and 72% of orthologous coding sequences in the diploid and haploid libraries respectively. We predict 717 genes to be essential in *S. uvarum*, with 412 of those genes also known to be essential in *S. cerevisiae*. Comprehensive gene dispensability comparisons with *S. cerevisiae* revealed that approximately 12% of conserved orthologs had diverged dispensability, and these genes were enriched for gene

ontology categories that include DNA replication, protein binding and structural constituent of the ribosome. Despite their differences in essentiality, these genes are capable of cross-species complementation, demonstrating that differences in genetic background must contribute to differential gene essentiality. This data set provides direct experimental evidence of gene function across species, which can inform comparative genomic analyses, improve gene annotation and be applied across a diverse set of microorganisms to further our understanding of gene function evolution.

3.2 INTRODUCTION

The ability to accurately predict gene function based on DNA sequence similarity is a valuable tool, especially in the current stage of genomic research where numerous genomes are increasingly becoming sequenced. It has become crucially important to predict gene function based on sequence similarity due to the lack of experimentally determined functional information associated with each newly sequenced genome. Most functional predictive methods rely on similarities of DNA sequence homology, co-expression patterns as well as protein structure help assign function to uncharacterized genes, using genes where known functions have been previously characterized (Eisen, 1998; Usadel et al., 2009). However, these methods come with their own set of limitations and often produce a substantial number of predictive errors, highlighting the importance of implementing experimental methods to directly test gene function of previously uncharacterized genomes to improve current methods of gene function annotation.

The gold standard of gene function characterization relies on targeted deletions of predictive coding sequences to probe the contributions of each gene to specific biological processes. To get a global view of gene function within an organism, several genome-wide deletion collections have been created in model species, particularly in bacteria and yeast, (Baba et al., 2006; Berardinis et al., 2008; Porwollik et al., 2014; Winzeler et al., 1999) including highly diverged species (Dowell et al., 2010; Kim et al., 2010b; Schwarzmüller et al., 2014) as well as different strains of yeast (Dowell et al., 2010). These systematic deletion collections are powerful tools for investigating molecular mechanisms of gene function, biological pathways, and genetic interactions, especially in the genetic workhorse *S. cerevisiae,* were gene function characterization and gene dispensability comparisons have been extensively performed amongst various deletion collections of yeast (Costanzo, 2016; Dowell et al., 2010; Kim et al., 2010b; Tong et al., 2001).

However, considerable effort and resources are required to create these targeted, systematic libraries and are not a practical approach for validating gene function across a wide range of non-standard genetic backgrounds in a high-throughput manner. Alternative approaches to targeted gene deletion libraries are transposon based insertional mutagenesis methods used to create random insertional mutant collections, eliminating requirements for *a priori* knowledge about defined coding regions and providing information about partial loss of function or gain of function mutations. Random insertional profiling has been widely applied across various species and has been instrumental in understanding virulence genes, stress tolerance mechanisms and even understanding tumor suppressor genes in mice (DeNicola et al., 2015; van Opijnen and

Camilli, 2013; de la Rosa et al., 2017; Weerdenburg et al., 2015; Yung et al., 2015). Several transposon libraries have also been implemented across diverged yeast species, providing useful information about gene function, growth inhibiting compounds and even essential functional protein domains (Gangadharan et al., 2010; Guo et al., 2013; Michel et al., 2017a; Oh et al., 2010; Ross-Macdonald et al., 1999) .

Here we utilize a random insertional method that has allowed us to assay gene dispensability of approximately 50,000 mutants in *Saccharomyces uvarum,* a species that diverged from *S. cerevisiae* approximately 20 million years ago and contains approximately 80% identity in coding sequences to *S. cerevisiae*(Dujon, 2010; Kellis et al., 2003; Scannell et al., 2007). These species can inter-mate to create hybrids, allowing us to leverage the large genetic toolsets established in *S. cerevisiae* to more fully explore the genetic basis for possible differential gene dispensability among these species. Genes with different dispensability patterns between these two species can be used as a preliminary indicator of divergent gene function, providing a model for investigating gene function evolution between two diverged species of yeast. While this Tn7 insertional density is modest in comparison to other insertional mutant libraries (50,000 compared to > 300,000 insertions) we successfully validated a subset of predicted differentially essential genes, proving this approach to be useful for prioritizing genes for testing viability. Furthermore, this Tn7 transposon mutagenesis library provides a valuable resource for studying *S. uvarum* gene function and serves as a framework for comparative functional genomics studies across newly sequenced, previously uncharacterized species.

3.3 RESULTS

3.3.1 Generating Tn7 insertional libraries in *S. uvarum*

One initial step to identifying genes with divergent gene function is to identify mutant phenotypes that are different between two diverged species. One of the most straightforward phenotypes to characterize is cell viability or gene dispensability. Therefore, we first sought to characterize gene essentiality in *S. uvarum*, with the final aim of identifying genes that are differentially essential between *S. cerevisiae* and *S. uvarum*. Instead of creating a library of individual knock out strains, we applied a high-throughput approach of creating random insertional mutants and leveraged the power of sequencing to identify the insertion sites in a pooled collection. The design of the transposon mutagenesis library is described below.

The Tn7 mutagenesis library approach described by Kumar *et al.,* (2004) was used to create a collection of *S. uvarum* mutant strains. The details of this library have been previously described by Caudy *et al.,* (2013). Briefly, *in vitro* transposition of the Tn7 transposon was performed in a plasmid library containing random *S. uvarum* genomic fragments. The Tn7 transposon was designed to carry a Clonnat resistance and marker carries stop codons in all reading frames near both termini. The interrupted genomic fragments were excised out of the plasmid and integrated through homologous recombination at their corresponding genomic positions, likely producing truncations when inserted within coding regions (**Figure 3.1A)**. The plasmid library contains ~50,000 unique genomic insertion sites that were integrated into a diploid and a haploid *MATa* strain. We isolated mutants with the ClonNat resistance marker on plates and created a final diploid pool of ~500,000 (10X coverage of plasmid library) transformants

(**Figure 3.1B)**. The haploid pool was obtained from the Caudy lab containing ~300,000 transformants. We added an additional 200,000 transformants to the haploid pool to increase the total coverage. While we could have created a haploid library directly from sporulating the diploid pool, eliminating unsporulated diploids would have been a challenge. Any diploid contamination in the haploid pool would make it difficult to determine which insertions were falling out of the population due to an integration into an essential gene. Pools of each Tn7 library were grown in rich media in batch cultures to identify mutants that drop out of the pool of mutants, as described in materials and methods. Insertion sites were determined using sequencing methods described below.

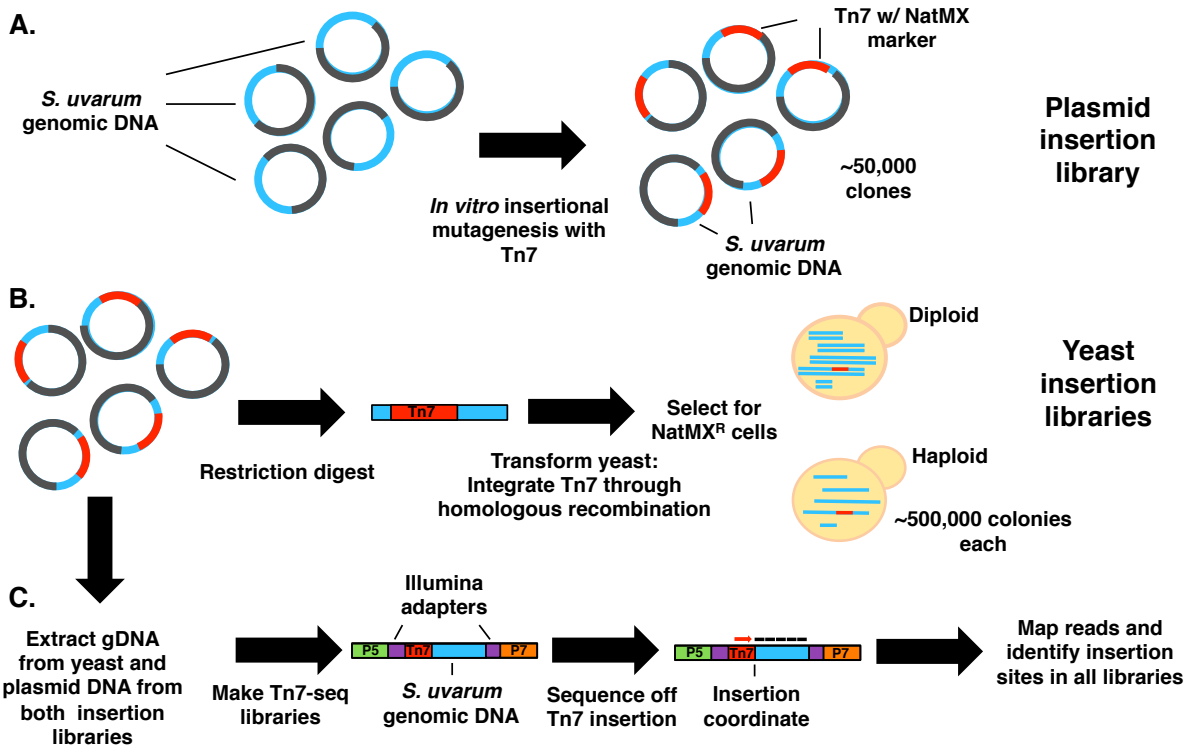


Figure 3.1 Schematic of Tn7 transposon mutagenesis library and insertion identification in *S. uvarum*. A) Simplified representation of *in vitro* transposition of the Tn7 transposon into a plasmid library containing random *S. uvarum* genomic DNA fragments. Approximately 50,000 plasmids containing the Tn7 transposon were pooled together to form the final library. B) Illustration of the Tn7 containing excised portion of the plasmid integrated into haploid and diploid yeast through homologous recombination. Approximately 500,000 Nat^R clones of each ploidy were pooled into two separate pools (Haploid pool and Diploid pool). C) Design of Tn7-seq libraries used to identify insertion sites through sequencing. Reads containing Tn7 sequence are enriched (PCR off common flanking region of the Tn7 and Illumina adapter sequences) and mapped to the genome to identify insertion sites.

3.3.2 Distribution of insertion sites across the *S. uvarum* genome

We first sequenced the plasmid pool of *S. uvarum* mutants to determine the genome coverage of genes containing insertion sites in the original plasmid library. We sequenced the library by extracting plasmid DNA and enriching for fragments of DNA containing the Tn7 sequence. Primers were designed to target the Tn7 sequence and the Illumina linker sequences that were ligated to randomly sheared plasmid DNA (**Figure 3.2C**). The PCR amplicons of the Tn7 library were sequenced, trimmed, mapped and processed through an in-house Ruby script to determine the position of the insertion site (Material and Methods). Insertion sites with fewer than 10 reads were filtered out. Detailed information about overall sequencing coverage is listed in **Supplementary Table 3.4**. We used this same method to determine the insertion sites in the haploid and diploid library by extracting genomic DNA from the pooled libraries. The distribution of haploid, diploid and overlapping insertion sites are evenly distributed throughout the *S. uvarum* genome, as illustrated in **Figure 3.2.**

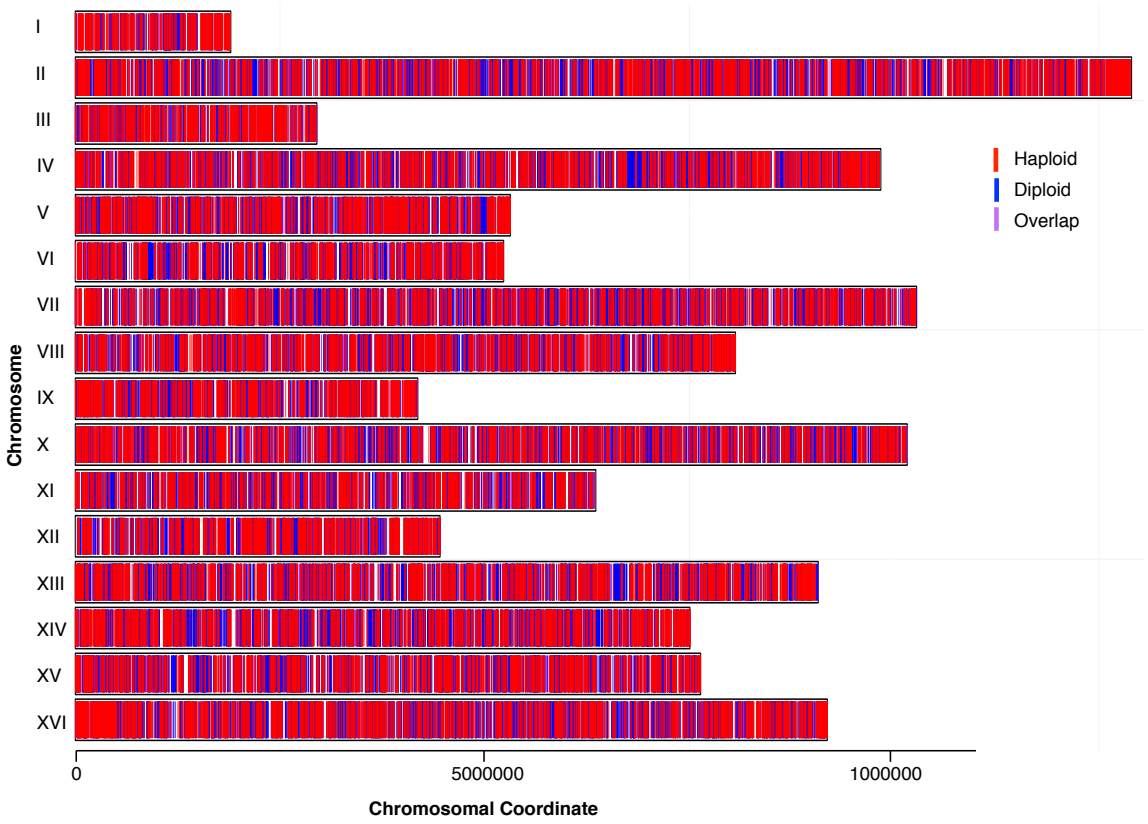


Figure 3.2 Distribution of haploid and diploid Tn7 insertions across the *S. uvarum* genome. Chromosomal map of sequenced transposon insertions wherein each chromosome is represented by a rectangle outlined in black and numbered one through sixteen from top to bottom. Haploid-only insertion sites are colored red, diploid-only inserts are blue. Insertion sites that were identified in both libraries are indicated in purple.

Once the insertion sites were determined in all three libraries, we counted the number of insertion sites in each annotated open reading frame using an in-house Python script (Materials and Methods). **Table 3.1** summarizes the number of insertion sites and the number of genes that contain insertion sites within the plasmid, haploid, and diploid libraries. All annotated *S. uvarum* genes containing the number of insertion sites from each library are fully listed in **Supplementary Table 3.5.** Of the 5,908 annotated genes, a total of 5,315 (90%) genes harbor insertion sites that were identified in at least one library. Comparisons between shared genes and unique genes with insertion sites are illustrated in **Figure 3.3.** The number of genes with insertion sites shared amongst all three libraries was 3,657 (69%) of the 5,315 genes summed across the libraries. There are subsets of genes that are library specific or shared between two libraries, likely due to differences in overall transformation coverage per library. There is a subset of 742 genes, however, that is shared only between the diploid and plasmid libraries (14%). Overall, only 3,933 genes harboring insertion sites were determined in the haploid pool, suggesting that at least some of the genes falling out of the haploid pool are likely to be essential based on their dispensability restrictions. We went on to test this assumption using the known essential gene set in *S. cerevisiae.*

Table 3.1 Summary of library coverage

Library Type	Number of inserts > 10 reads	Number of inserts in ORFs	Number of genes with insert	% Genome covered	Number of orthologs with inserts	% Orthologs with inserts
Plasmid	54,351	33,394 (61.4%)	4,944	83%	4,630	85%
Diploid	46,326	27,121 (58.5%)	4,557	77%	4,283	79%
Haploid	42,904	22,988 (54.5%)	4,190	71%	3,933	72%

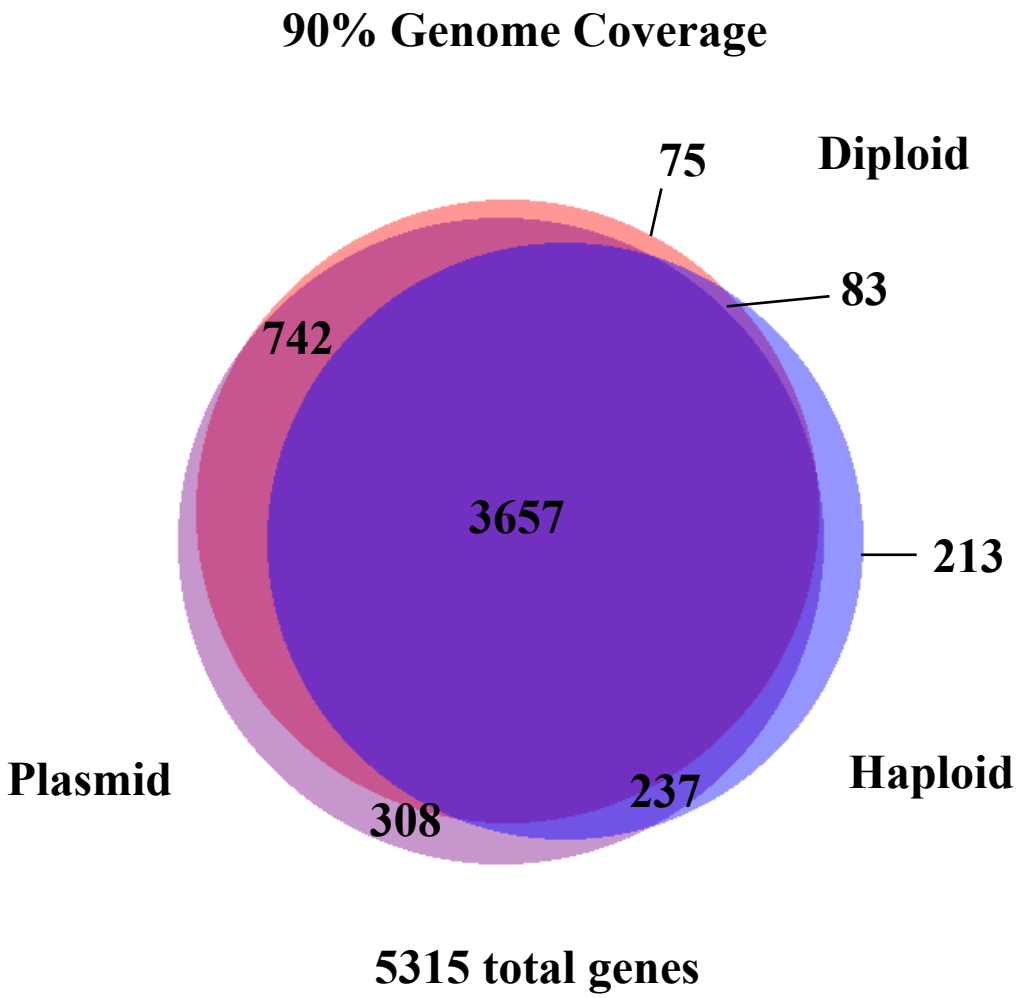


Figure 3.3 Proportional Venn diagram summarizing the number of insert-containing genes identified in each library. Summary of genes containing at least one insert sampled by 10 or more sequencing reads. The plasmid library is represented in purple, the diploid pool in orange and the haploid pool in blue. Non-overlapping regions represent genes that are library specific.

3.3.3 Known *S. cerevisiae* essential genes contain fewer inserts than known non-essential genes

Although the large number of dropouts suggest that we have identified essential genes, it is possible that these dropouts just reflect an incomplete library. To test if genes in the haploid library without (or with minimal) insertion sites were likely to be essential, we compared the haploid and diploid libraries to determine if differences in the number of insertions between known essential and non-essential genes exist. Due to the nature of the library, insertional events at different positions across a gene may result in a partial loss of function, meaning that even essential genes may still tolerate some insertions. Since the essentiality of most genes are expected to be conserved between *S. cerevisiae* and *S. uvarum,* we used the known essential set in *S. cerevisiae* to test if essential genes in the haploid library contain fewer insertion sites. We normalized the number of inserts within each gene by the length of the gene (inserts/kb) and plotted the distribution of normalized inserts within known essential genes and non-essential genes in both diploid and haploid libraries (**Figure 3.4 A&B)**.

The distribution between known *S. cerevisiae* essential and non-essential genes is similar in the diploid pool, with no significant differences between gene types (essential average inserts/kb=3.8 vs non-essential average inserts/kb = 4.2) (**Figure 3.4C**, Wilcoxon test p-value = 0.2001). However, the distributions in known *S. cerevisiae* essential and non-essential genes in the haploid library are significantly different, with known essential genes averaging fewer normalized inserts per kb (known essential average inserts/kb=0.88 vs non-essential average inserts/kb = 4) (**Figure 3.4D**, Wilcoxon test p-value $<2.2\ e^{-16}$). This result suggests that the known conserved essential genes can be predicted from the number of inserts in the haploid library. We note that known *S.*

cerevisiae essential genes in the haploid library harboring several insertion sites are detected as well. We predict that these genes are candidate *S. cerevisiae*-specific essential genes, and may not be essential in *S. uvarum.* We explore these genes more fully in the following sections.

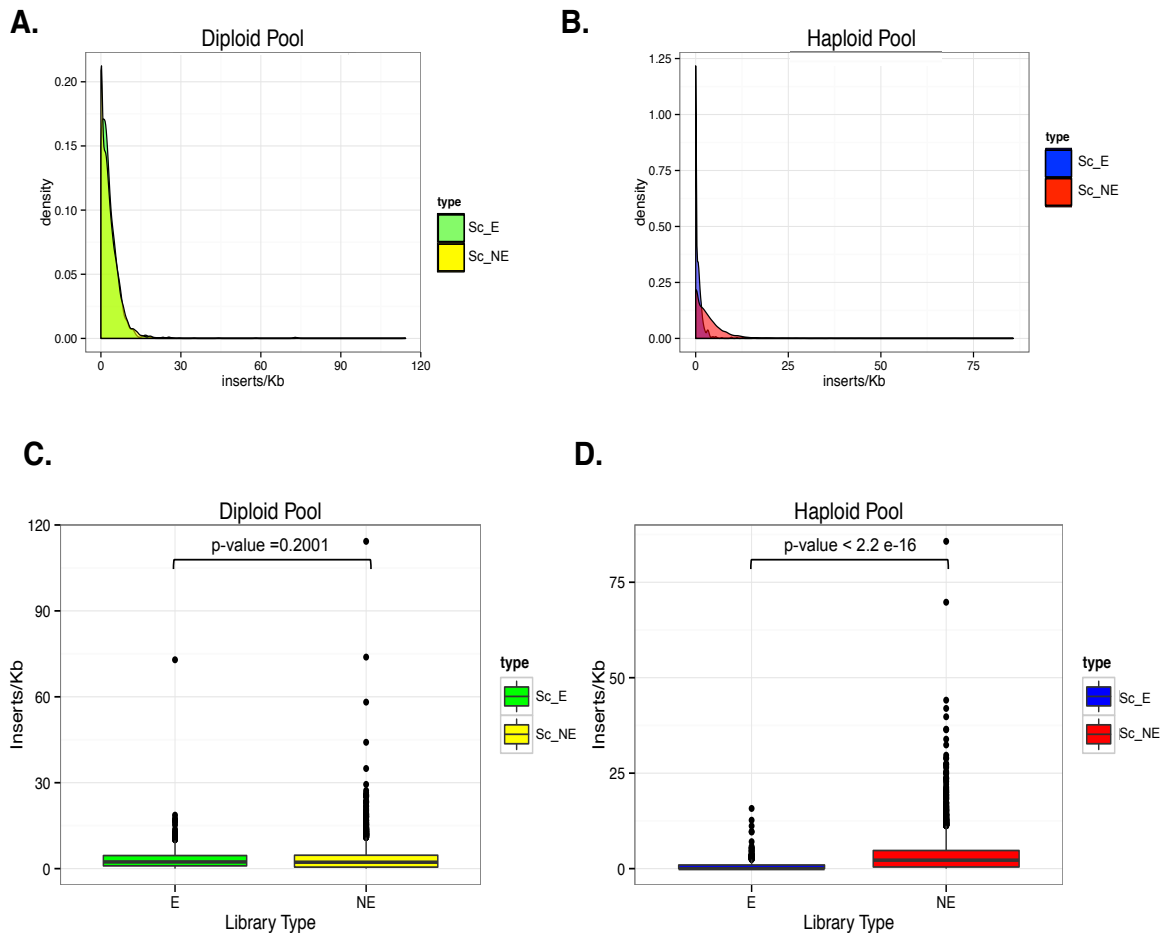


Figure 3.4. Comparison of insertion distributions between haploid and diploid libraries amongst known *S. cerevisiae* essential and non-essential genes. A) Density plot displaying the distribution of the normalized number of insertion sites in known *S. cerevisiae* essential and non-essential genes within the diploid and B) haploid library. C) Box plot of normalized insertions per kb in known *S. cerevisiae* essential and non-essential genes in the diploid library. No significant difference was detected between known essential and non-essential genes (Wilcoxon test p = 0.2001). Essential average inserts/kb= 3.8 (SD= 3.98) vs. non-essential average inserts/kb = 4.2 (SD=4.2). D) Box plot of normalized insertions per kb in known *S. cerevisiae* essential and non-essential genes in the haploid library displaying a significant difference between known essential and non-essential genes (Wilcoxon test $p < 2.2\ e^{-16}$). Essential average inserts/kb=0.88 (SD=1.28) vs. non-essential average inserts/kb = 4 (SD = 4.38).

3.3.4 Predicting *S. uvarum* essential and non-essential genes using an insertion ratio metric

Once we determined that essential genes contain significantly fewer insertion sites than non-essential genes in the haploid library, we created a metric for determining a cut off value to categorize predicted *S. uvarum* essential and non-essential genes. Due to the nature of the library, insertional events at different positions across a gene may result in a partial loss of function, meaning that even essential genes may still tolerate some insertions. Therefore, we relied on comparisons between the diploid and haploid libraries to make inferences about gene essentiality. Specifically, we calculated an insertion ratio using the number of inserts per gene in the haploid library divided by the number of inserts in the diploid library, which inherently normalizes for the length of the gene (Materials and Methods). Using the insertion ratio as a metric, we tested if significant differences exist between *S. uvarum* genes whose orthologs are known to be essential and non-essential in *S. cerevisiae,* as well as *S. uvarum* intergenic regions. Intergenic regions between convergent orientated genes are expected to not be essential, thus, the distribution of intergenic regions is expected to be similar to that of non-essential genes and represents our null distribution.

Figure 3.5A illustrates the distribution of each feature type, with *S. cerevisiae* known non-essentials having a similar distribution to *S. uvarum* intergenic regions. However, the left most shoulder of the non-essential distribution displays a distribution more similar to essential genes and is likely to reflect *S. uvarum*-specific essential genes. Furthermore, the opposite is true for a subset of known essential genes with larger insertion ratios. The differences between known *S. cerevisiae* essential genes and non-

essential gene insertion ratios were significant, as well the differences between the essential genes and intergenic regions (**Figure 3.5B**, Wilcoxon $p < 2.2e^{-16}$). We note the significant difference that also exists between non-essential genes and intergenic regions and attribute this difference to the possible genes that are differentially essential between species in this category in comparison to intergenic regions in *S. uvarum.* We note these differences are of a lesser magnitude than those that exist between known essential genes.

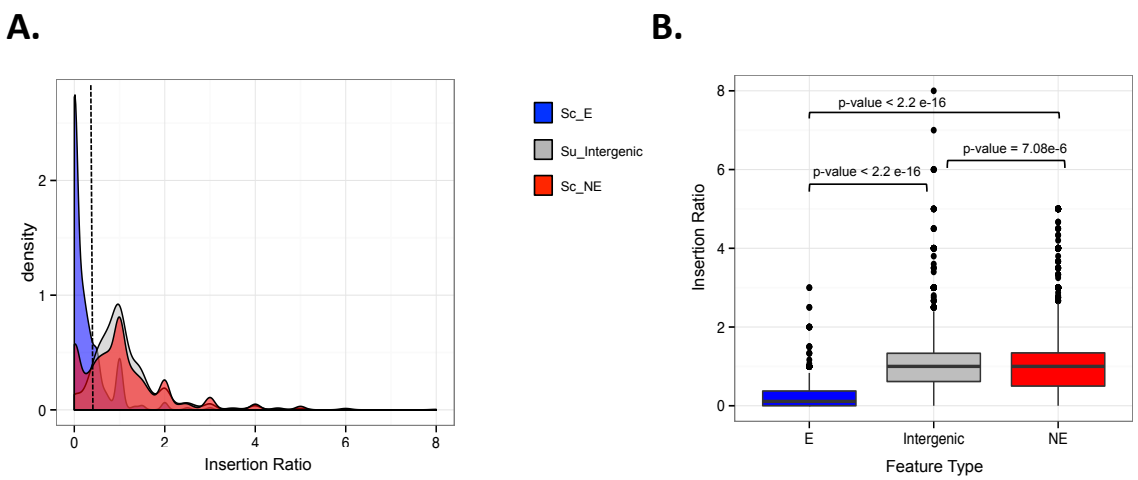


Figure 3.5. Insertion ratio distributions of *S. uvarum* intergenic regions and known *S. cerevisiae* essential and non-essential genes. A) Density plots displaying the distribution of insertion ratios across three feature types: *S. uvarum* intergenic regions between Watson and Crick oriented genes ranging from 7 kb-500 bp (grey) and *S. uvarum* genes whose orthologs are known *S. cerevisiae* essential (Sc_E in blue) and nonessential genes (Sc_NE in red). The dashed line represents an insertion ratio of 0.25 and defines the cut-off value to classify essential and non-essential genes. B) Box plots of insertion ratios by feature type described in plot A. Significant insertion ratio differences exist between known *S. cerevisiae* essential and non-essential genes and between *S. uvarum* intergenic regions. (Wilcoxon tests Sc_E:Su_Intergenic $p < 2.2e^{-16}$, Sc_E:Sc_NE $p < 2.2e^{-16}$, Sc_NE:Su_Intergenic $p = 7.08e^{-6}$).

Once we established the intergenic region as our null distribution, we ranked the insertion ratio value for each gene against the intergenic distribution and determined the proportion of intergenic regions whose insertion ratio was greater than the insertion ratio of that gene. Using this raking metric, we set a cut-off value of 0.25 to categorize all annotated *S. uvarum* genes into essential and non-essential categories. Using this cut-off value, 1170 genes were categorized as essential genes. We applied an additional cut-off metric (more details in Material and Methods) to remove a class of low coverage genes, resulting in a total number of 718 (13%) predicted essential genes and 3,838 (65%) genes that are predicted non-essential, with 1299 (22%) undetermined. A list of all genes with their predicted classification can be found in **Supplemental Table 3.5.** We proceeded to characterize each gene set and validate the dispensability of each of the predicted gene categories.

3.3.5 Analysis of predicted gene dispensability

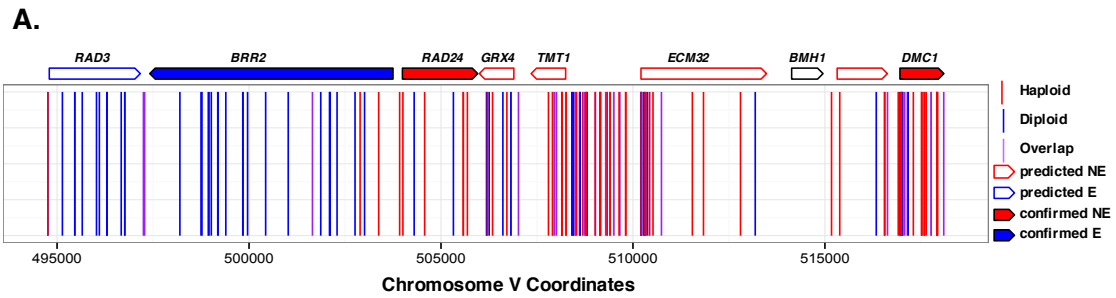
The predicted gene list of orthologous *S. uvarum* essential genes was compared to known essential genes lists from both *S. cerevisiae* and *S. pombe* to determine the amount of conservation that exists across diverged species. Of the predicted 718 *S. uvarum* essential genes, 297 genes (42%) are shared amongst all three sets, with a total of 487 genes (68%) shared with at least one other set (**Supplemental figure 3.1)**.

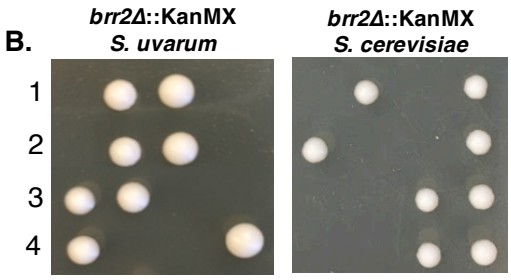
Similar to what has been previously shown in *S. cerevisiae,* essential genes in *S. uvarum* were more likely to be unique, with 91% of essential genes (656/718) being present in single copy compared to 76% of non-essential genes (2736/3604). Additionally, comparisons between Gene Ontology (GO) molecular function terms of

essential gene sets from both species show significant enrichment (p-value < 0.01) for fundamental biological functions. Processes such as DNA replication/binding, RNA and protein biosynthesis, as well as structural constituents of the ribosome and cytoskeleton were enriched in both sets of essential genes (**Supplemental Table 3.6)**. In contrast, non-essential genes were significantly (p-value < 0.01) enriched for regulatory functions (transcription factor activity) and conditional responsive processes, such as transmembrane transporter activity and cell signaling (kinase activity) (**Supplemental Table 3.7)**.

Once we determined that many of the features of the predicted essential genes were similar to confirmed essential genes in other species, we proceeded to create heterozygous deletions to validate 13 conserved essential genes. Sporulating each heterozygous deletion strain and performing tetrad analysis for cell viability confirmed essentiality for 12 (92%) of the 13 strains (**Table 3.2)**. One example of a confirmed essential gene can be found in **Figure 3.6A,** which illustrates the genomic positions of all insertion sites across a genomic locus of chromosome five that contains essential and non-essential genes. The color of the gene outline matches the predicted dispensability, which is determined by their insertion ratio. For example, the gene *BRR2* has an insertion ratio of 0.130 and is predicted to be a conserved essential gene **(Figure 3.6B)**. The tetrad analysis of a *BRR2* heterozygous deletion strain displays a 2 viable:2 inviable segregation pattern in both species, validating this gene as a conserved essential gene (**Figure 3.6B)**. Although we identified three insertion sites at the 5-prime end of *BRR2* in the haploid strain, there is an enrichment of diploid inserts that results in a small insertion ratio value. Images of all other confirmed essential genes are located in **Supplemental**

figure 3.2. We also tested three conserved non-essential genes and all three were confirmed as non-essential (100%) (**Supplemental figure 3.3) (Table 3.2)**. Additionally, we obtained an independent set of haploid deletion strains from the Rine lab, which was used as a validated non-essential gene set. Out of the total 356 gene deletions that were included in our library, 346 of those genes were predicted to be non-essential (97%) while the remaining 3% were predicted to be essential.





C

Gene	Haploid	Diploid	Insert ratio (hap/dip)	Prediction
RAD3	0	12	0.0	E
BRR2	3	23	0.130	E
RAD24	4	2	2.0	NE
GRX4	3	3	1.0	NE
TMT1	10	3	3.33	NE
ECM32	15	12	1.25	NE
BMH1	0	0	N/A	N/A
DMC1	26	14	1.85	NE

Figure 3.6. Validation of conserved essential and non-essential genes. A) Mapped chromosomal insertion positions are plotted across chromosome five. Haploid inserts are indicated in red, diploid inserts are blue and overlapping inserts are indicated in purple. Genes are indicated across the top are outlined according to predicted dispensability and filled in if confirmed. B) Tetrad analysis of a confirmed conserved essential gene *brr2*Δ in *S. cerevisiae* and *S. uvarum*. Segregants containing *brr2*Δ alleles are inviable in both species. C) Table indicating the number of insertional events per gene within each library. The final column lists the insertion ratio (number of haploid inserts by the number of diploid inserts) per gene.

3.3.6 Gene dispensability comparisons of orthologous pairs between *S. cerevisiae* and *S. uvarum*

Our main goal of this project is to identify genes with differential essentiality to test for evidence of divergent function. While the previous section categorized all annotated *S. uvarum* genes, we narrowed our analysis to 4,543 orthologous genes for which we had data in the *S. uvarum* dataset to make direct comparisons of dispensability between *S. cerevisiae* and *S. uvarum* (**Supplementary Table 3.5)**. Overall, 88% (4016/4543) of these genes display conserved dispensability between *S. cerevisiae* and *S. uvarum*. The remaining 12% of orthologs differ in essentiality between the two species, with 306 (7%) of these genes only essential in *S. uvarum* and 221 (5%) genes only essential in *S. cerevisiae* **(Figure 3.7A)**. Note the larger number of predicted essential genes in *S. uvarum* (305 in *S. uvarum* compared to 222 in *S. cerevisiae*). This difference may be attributed to the reliance on the absence of data (lack of insertion sites in a haploid gene) in a greater proportion of genes not previously characterized as an essential gene in *S. cerevisiae* (3687 non-essential genes vs. 765 essential genes), whereas, the latter category utilizes the presence of insertional data in a smaller proportion of genes that are known to be essential in *S. cerevisiae*. All predicted genes that differ in dispensability are listed in **Supplemental Table 3.8.** We compared a previously described metric used to quantify differences in gene expression between orthologous genes to determine if the difference in dispensability could be explained by gene expression. We compared this metric in known genes that differ in essentiality and did not find evidence of genes enriched for expression differences, suggesting that gene

expression alone cannot account for the differences in species dependent essential genes **(Supplemental Figure 3.7).**

To analyze the two categories of genes that differ in essentiality further, we compiled a list of 222 genes from the *S. cerevisiae*-specific category and a more restrictive list (Materials and Methods) of 220 *S. uvarum*-specific genes; using this list, we determined the proportion of *S. cerevisiae*-specific and *S. uvarum*-specific genes were annotated for each function by performed Gene Ontology (GO) term finder using the molecular function ontology. A more restrictive list was used to normalize the number of genes from each species represented in each functional category. The proportion of essential genes that differ between species for each functional category are represented in **Figure 3.7B,** illustrating a subset of all significant functional categories. Interestingly, the most striking difference is in the functional category of the structural constituent of the ribosome. This category is enriched for genes that are essential in *S. uvarum* (46/52). Additionally, differences exist between essential genes in the category of RNA polymerase activity, where 9/10 genes were identified from *S. cerevisiae* essential genes. Full lists of significant (p-value < 0.01) GO enrichment molecular function terms for each species individually are listed in **Supplemental Table 3.9.**

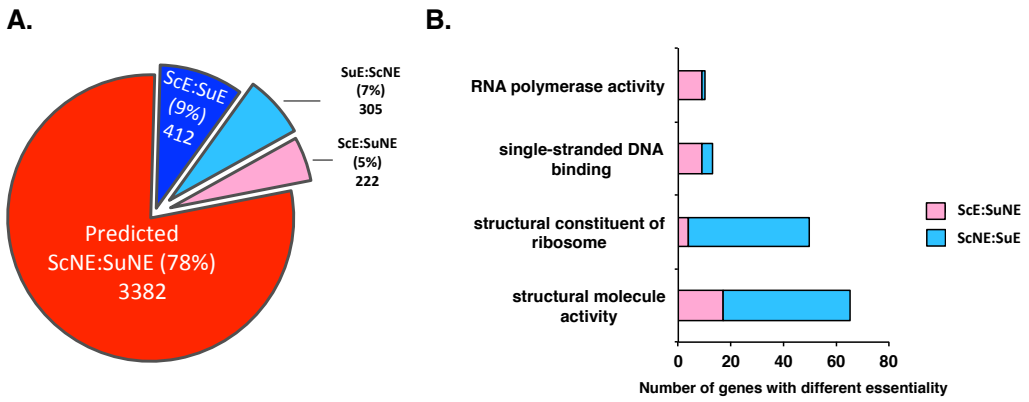


Figure 3.7 Orthologous gene comparisons between species A) Essentiality comparisons between 4,321 ortholog pairs amongst *S. cerevisiae* and *S. uvarum*. A total of 85% orthologs display conserved essentiality, with 12% showing differences in gene dispensability (NE= non-essential, E=essential). B) Functional enrichment of orthologs with differential essentiality. Gene Ontology (GO) enrichment was performed on genes that differ in essentiality and a subset of biological functions are represented (**Supplemental Table 3.10** for complete list). Each color indicates the proportion of total annotated genes categorized for each function, split by the number of genes represented by essential genes in each species. The number of *S. cerevisiae* essential genes in each functional category is indicated in pink and *S. uvarum* essential genes in light blue.

After each category of genes that differed in essentiality was determined, we proceeded to validate a subset of each category within both genetic backgrounds. Similar to the method previously described to confirm conserved essential genes, we sporulated heterozygous deletion strains and analyzed the viability pattern of the segregants in both species. We tested 28 genes in the *S. uvarum*-specific category and validated a total of 9 genes (32% correct) (**Table 3.2)**. The gene *SSQ1* is an example of a confirmed *S. uvarum*-specific essential gene, illustrated in **Figure 3.8**. Images of all 9 confirmed *S. uvarum*-specific essential genes compared to non-essential phenotypes in *S. cerevisiae* can be found in **Supplemental Figure 3.4.** Additionally, we tested the viability of 27 candidate *S. cerevisiae* essential genes and confirmed a total of 15 genes as differentially essential (56%) (**Table 3.2)**. An example of one confirmed *S. cerevisiae*-specific gene is *VTC4,* illustrated in **Figure 3.9.** Images of confirmed strains in *S. cerevisiae* and *S. uvarum* can be found in **Supplemental Figure 3.5.** Although our likelihood of accurately predicting genes in the differentially essential category was lower than the categories of conserved dispensability, we were able to validate a greater number than what would be expect by random chance, given that the expected probability of selecting an essential gene is 20%. All combined tetrad analysis results from the confirmation tests, also including false positives, are represented in **Supplemental Figure 3.6**.

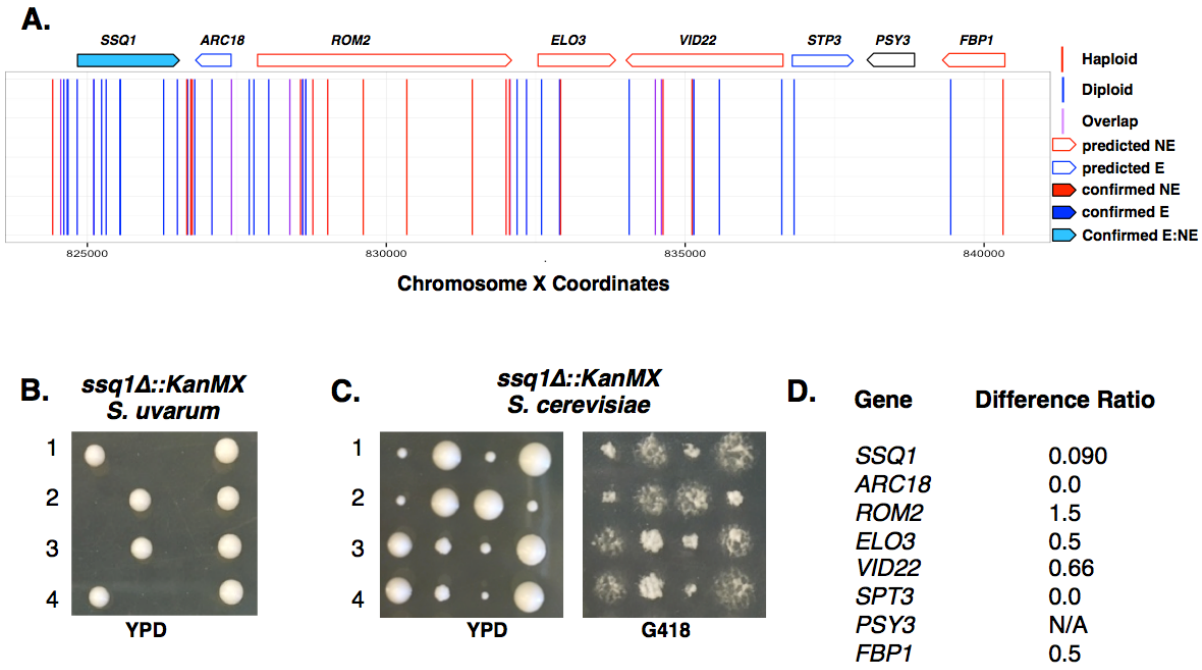


Figure 3.8 Validation of *S. uvarum*-specific essential gene *SSQ1*. A) Mapped chromosomal insertion positions are plotted across chromosome ten. Haploid inserts are indicated in red, diploid inserts are blue and overlapping inserts are indicated in purple. Genes are indicated across the top are outlined according to predicted dispensability and filled in if confirmed. Light blue filling indicates a gene that is essential in *S. uvarum* and non-essential in *S. cerevisiae* (confirmed E_NE). B) Tetrad analysis of a heterozygous *ssq1Δ::KanMX* strain displaying inviable segregants containing the *ssq1Δ* allele in *S. uvarum*. C) Tetrad analysis of a heterozygous *ssq1Δ::KanMX* strain in *S. cerevisiae* containing viable segregants plated on YPD and G418. D) Table indicating the number of insertional events per gene within each library. The final column lists the insertion ratio (number of haploid inserts by the number of diploid inserts) per gene.

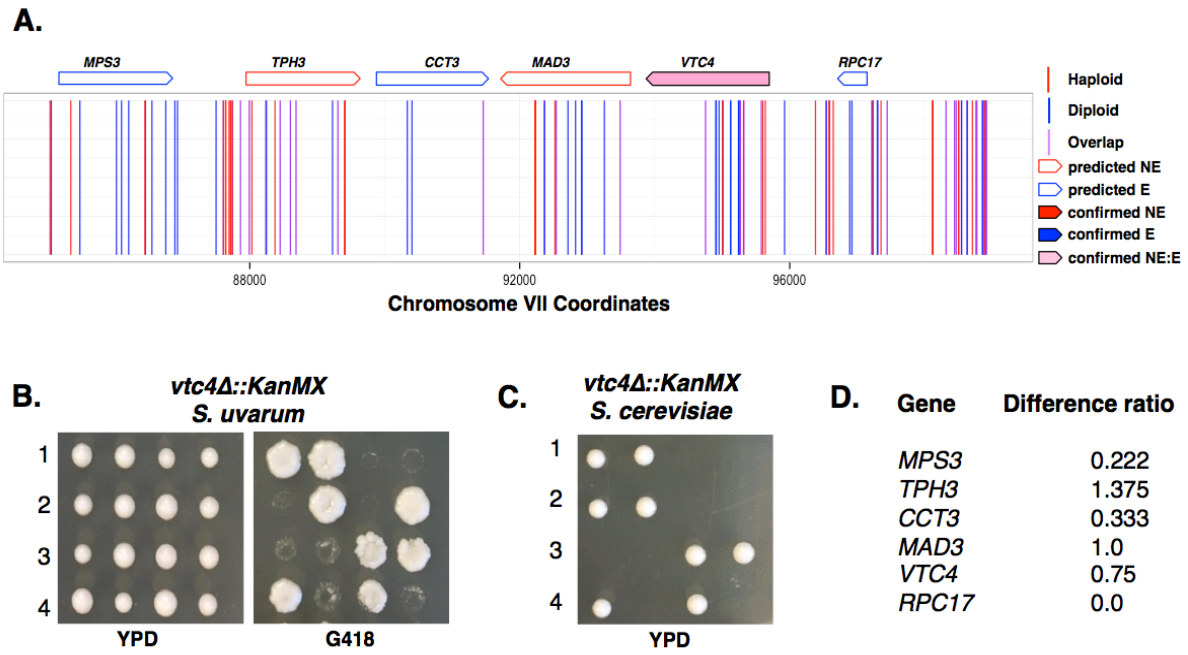


Figure 3.9 Validation of *S. cerevisiae*-specific essential gene *VTC4*. A) Mapped chromosomal insertion positions are plotted across chromosome twelve. Haploid inserts are indicated in red, diploid inserts are blue and overlapping inserts are indicated in purple. Genes are indicated across the top are outlined according to predicted dispensability and filled in if confirmed. Light pink filling indicates a gene that is essential in *S. cerevisiae* and non-essential in *S. uvarum* (confirmed NE_E). B) Tetrad analysis of a heterozygous *ssq1*Δ strain displaying inviable segregants with the *ssq1*Δ allele in *S. uvarum.* C) Tetrad analysis of a heterozygous *ssq1*Δ allele in *S. cerevisiae* resulting in viable segregants. D) Table indicating the number of insertional events per gene within each library. The final column lists the insertion ratio (number of haploid inserts by the number of diploid inserts) per gene.

Table 3.2

Predicted gene type	Number of genes tested	Number of genes confirmed	Number of false positives	% Correct
ScE_SuE	13	12	1	93%
ScNE_SuNE	3	3	0	100%
ScNE_SuE	28	9	19	32%
ScE_SuNE	27	15	12	56%

3.3.7 Paralog divergence and duplicate gene loss explain some background effects on differential gene dispensability

Once we confirmed a list of genes that differed in dispensability between species, we set out to determine what genetic background effects could be contributing to the differences in dispensability between *S. cerevisiae* and *S. uvarum*. One explanation could be genetic redundancy due to gene duplications, such that a gene is essential in one species due to the loss of the other duplicate gene, whereas the other species retained both genes. To investigate this possibility, we began investigating genes that differed in dispensability that contained paralogs. Of the 222 *S. cerevisiae*-specific essential genes, 11 were known to have paralogs. For example, *CDC25* is a *S. cerevisiae*-specific essential gene (nonessential in *S. uvarum)* and is a paralog to *SDC25*, which contains a premature stop codon in *S. cerevisiae.* We performed complementation assays by cloning *S. uvarum* alleles of both paralogs into a CEN/ARS plasmid and testing whether the *S. uvarum* alleles could rescue the inviable phenotype of segregants from a heterozygous *cdc25*Δ deletion in *S. cerevisiae*. We found that, *SDC25* from *S. uvarum* is functional and that both *SDC25* and *CDC25* alleles from *S. uvarum* can complement a *cdc25*Δ heterozygous deletion in *S. cerevisiae* (**Supplemental Figure 3.8)**. The results from the complementation assays suggest that perhaps *CDC25* is required for growth in *S. cerevisiae* due to the lack of redundancy as a consequence of the non-functional copy of *SDC25*.

Following this same logic, we attribute *CDC25* non-essentiality in *S. uvarum* to the redundancy provided by the functional copy of *SDC25* in this species. To test this idea, we created a double heterozygous *cdc25*Δ *sdc25*Δ mutant in *S. uvarum* and

performed segregation analysis on the dissected tetrads (**Supplemental figure 3.8).** Unexpectedly, the segregation pattern of a double mutant displays a lethal phenotype for not only the double mutant but also the single *sdc25*Δ mutant. We confirmed this result by constructing a *sdc25*Δ heterozygous mutant in *S. uvarum* and found a 2:2 segregation pattern suggesting that *SDC25* is an essential gene in *S. uvarum.* Although there is no clear explanation for the requirement of *SDC25* in *S. uvarum,* this comparison displays one clear example of paralog divergence in essentiality between these two species.

In addition to redundancy differences that are attributed to paralog pairs, gene gains and losses may also contribute to genetic background effects that result in differential dispensability between *S. cerevisiae* and *S. uvarum.* Previous studies have investigated gene gains and losses across the *Saccharomyces* clade and identified genes that lost their duplicate in some species but not others (Kellis et al., 2003; Scannell et al., 2011). For example, *ALR1* is found as a singleton in *S. cerevisiae* but has retained the duplicate copy in *S. uvarum.* The *ALR1* gene is a confirmed *S. cerevisiae*-specific essential gene, which may be explained by the loss of the other copy of the duplicate pair. We created a heterozygous deletion of both *ALR1* copies and tested the viability phenotypes of each mutant and of the double mutant separately. Both copies of *ALR1* in *S. uvarum* are non-essential. We also tested a strain with both copies of *ALR1* deleted to test if the duplicate copy buffered the requirement of *ALR1*. Surprisingly, a strain with both copies deleted is viable. While this does not follow our predictions, future experiments will determine if *ALR2* which is the paralog to *ALR1* is essential in *S. uvarum*.

3.3.8 Divergent gene dispensability is largely due to trans effects

While genetic redundancy or gene loss explains a fraction of differentially essential genes, the remaining much larger portion of genes remained unexplained. Because our main goal for this study was to find evidence of gene function divergence between these two species, we proceeded to further investigate the remaining genes for functional differences. To test a subset of genes that differ in essentiality that could not be explained by paralogs or gene loss, we performed complementation assays in both species to test for divergent function. We cloned five *S. cerevisiae* alleles from the list of *S. uvarum*-specific genes (*SAC3, TUP1, CCM1, SSQ1,* and *AFT1)* and seven *S. uvarum* alleles from the list of *S. cerevisiae*-specific genes (*ALR1*, *SHR3*, *CDC25*, *INN1*, *LCD1*, *SEC24*, *VTC24*) into a CEN/ARS plasmid to perform complementation tests in *S. uvarum* and *S. cerevisiae* **(Supplemental Figure 3.9)**. The results from these complementation tests revealed that all genes are able to complement the inviable phenotype, suggesting that the differences in essentiality are more likely to be due to *trans*-acting changes rather than functional differences of protein coding regions.

3.4 DISCUSSION

In this study, we applied a comparative functional approach to investigate how genetic background influences gene dispensability between two diverged species of yeast. Using insertional integration comparisons between haploid and diploid pools of mutants, we prioritized genes to validate predicted essential, non-essential, and differentially essential gene categories in *S. uvarum.* We identified approximately 12% of orthologs to differ in dispensability between *S. uvarum* and *S. uvarum* and validated 25

genes in this category. Surprisingly, however, most genes that differ in dispensability have retained their function between these two species, suggesting that differences in gene dispensability are likely due to *trans*-acting changes rather than the direct result of divergent coding sequence.

Specifically, our comparison of orthologous genes between *S. cerevisiae* and *S. uvarum* revealed that a majority of genes maintain conserved dispensability requirements (88%) while 12% of orthologs are predicted to be essential in one species but not the other. We confirmed 93% (15/16) of predicted conserved categories of essentiality and 44% (24/55) of genes predicted to be differentially essential. Although our rate of confirmed genes in this category was lower than the conserved category, we correctly identified a subset of genes that are differentially dispensable, despite the moderately dense insertional profile of the library and a less restrictive cut-off value applied to include more genes to be classified as this type. Further analysis of predicted species-specific genes revealed enriched GO ontology terms of molecular functions involved in structural constituent of the ribosome and DNA binding. Finally, we utilized genetic tools in *S. uvarum* to test hypotheses about genetic background effects that contribute to differences in essentiality. We find that differences can be explained by paralog divergence, gene loss and *trans*-acting changes.

Applying a random insertional approach has proved to be useful in functionally profiling *S. uvarum* and will be useful for studying other understudied species, with the goal of adding information to gene annotation methods. While this study was performed in standard laboratory conditions, it is easily amenable for testing stressful conditions, other nutrient sources as well as naturally relevant conditions. The identification of

synthetic lethal interactions can also be determined by performing insertional profiling in the background of a particular mutation of interest relatively quickly and economically. Additionally, pooled competition experiments *en masse* can be used to determine the frequency of particular insertional mutants, providing quantitative measurements of cellular fitness across conditions. Furthermore, computational approaches are available to prioritize experimental conditions that are most likely to probe the most valuable phenotypic information for further functional characterization (Guan et al., 2010).

Gene regulation also plays a large role in evolution and is crucial for responding to environmental change (Carroll, 2005). Previous studies have aimed to functionally characterize differences in gene expression patterns between *S. cerevisiae* and *S. uvarum* and determining species-specific responses to osmotic stress, peroxisome biogenesis and autophagy, suggesting that each species may been exposed to different selective pressures within their respective evolution histories (Caudy et al., 2013). Future studies that more precisely functionally characterize all uniquely dispensable genes will lay the groundwork for connecting functional and regulatory differences to the sequence variants that have accumulated over diverged evolutionary timescales. Lastly, understanding how trans effects are coordinated to regulate genetic interactions that are dependent on natural variation amongst individuals may serve as a framework to understand regulatory principles of genetic interactions.

3.5 MATERIALS AND METHODS

3.5.1 Strains, plasmids and primers

The strains, plasmids and primers used in this study are listed in **Supplementary Tables 3.1**, **3.2** and **3.3** respectively. Unless specified below, yeast strains were grown at 25°C for *S. uvarum* strains and 30°C for *S. cerevisiae* strains and standard media recipes were used.

3.5.2 Construction of the Tn7 mutagenesis library

The construction of the Tn7 plasmid library has been previously described in detail and was obtained from the Caudy lab (Caudy et al., 2013). Briefly, this mutagenesis approach uses a plasmid library of *S. uvarum* genomic DNA, containing random Tn7 transposon insertions. The construct has a selectable marker for transformation into yeast, allowing the selection of disruption alleles.

To make the plasmid library, genomic DNA was isolated and fragmented by sonication to an average length of 3kb from a rho0 *S. uvarum* strain. The ends of the DNA were blunted and cloned into the pZero Blunt vector (Invitrogen). Approximately 50,000 colonies were recovered from the transformation into *E. coli* DH5α strain. The transformants were scraped from Kanamycin plates and pooled for plasmid purification. A version of the Tn7 transposon was constructed by amplifying the promoter from the Tet-on pCM224 (Bellí et al., 1998). The cassette of the Tet-on promoter and the ClonNAT resistance gene was amplified using PCR primers containing lox and BamHI sites and cloned into the BamHI site of the NEB vector pGPS3. This transposon construct was inserted into the *S. uvarum* genomic DNA library *in vitro* using the transposon kit from NEB. Initial selection (50,000 colonies) was on ClonNAT/Zeo. HindIII and XbaI were used to digest the pZero backbone to release the linearized genomic DNA for

efficient recombination. The library was then transformed into a haploid *S. uvarum* strain (ACY12) and a diploid strain (YMD1228) using a modified transformation protocol optimized for *S. uvarum*. Transformant colonies were plated to YPD-ClonNAT plates and allowed to grow for 5 days at 25°C. A total of ~ 500,000 colonies were scraped for each pool. Each final pool was well mixed at a 1:1 ratio with 5 % glycerol and 2 ml aliquots were stored at -80°C.

3.5.3 Pooled growth of Tn7 *S. uvarum* libraries

To determine the initial complexity of the integrated pools, genomic DNA was extracted directly from the glycerol stocks of both haploid and diploid pools using the Hoffman and Winston method (Hoffman and Winston, 1987). Additionally, we inoculated 500 µl of both libraries in separate YPD flasks for 24 hours to recover mutants after 24 hours of growth. Furthermore, to collect samples over time, we competed both pools under sulfate-limiting conditions in chemostats for approximately 30 generations at 25°C. A large-volume, ~300ml, sulfate-limited chemostat was inoculated with a single 2ml glycerol stock sample of each pool. After allowing the chemostat to grow at 25°C without dilution for ~24 hrs, fresh media was added to the chemostat at a rate of 0.17 h-1. This pooled growth assay was repeated twice, each including 5 time points with O.D. and dilution rate measurements as well as collected cell pellets for DNA extractions using the modified Hoffman-Winston prep referenced above.

3.5.4 Tn7 sequencing library preparation

Sequencing libraries were prepared by first extracting genomic DNA from pools of each library grown in YPD and sulfate limited conditions. Genomic DNA libraries were prepared for Illumina sequencing using a Tn7-seq protocol described previously (Wetmore et al., 2015). Briefly, the Covaris was used to randomly fragment DNA to approximately 200-800 bp in length. The fragments were blunt ended and A-tails were added to the fragments to ligate the Illumina adapter sequences. Custom index primers (listed in **Supplementary Table 3.3**) targeting Tn7-specific sequence and Illumina adapter sequence were used to enrich for genomic DNA with Tn7 insertion sites. The barcoded libraries were quantified on an Invitrogen Qubit Fluorometer and submitted for 150 bp-paired end sequencing on an Illumina HiSeq 2000 by JGI. This method was also applied to make the plasmid library, from linearized plasmid DNA.

3.5.5 Sequencing analysis

Sequencing reads from the fastq files were trimmed to remove Tn7 specific sequences and adapter sequences, restricting the minimal length of reads to 36 bp using Trimmomatic (Bolger et al., 2014) and FASTX-Toolkit. Trimmed fastq files were aligned against the reference strain of *S. uvarum* (CBS 7001) using Burrows-Wheeler Aligner with standard filters applied (Li and Durbin, 2009). Specifically, non-uniquely mapping reads, reads in which the pair did not map, reads with a mapping quality less than 30 and PCR/optical duplicate reads were filtered out; the samtools C-50 filter was applied as recommended for reads mapped with BWA. To limit the insertional analysis to actively growing cells, sam files were merged from the later time points in the growth assays of each pool using samtools (Li et al., 2009). The sequence coverage of the

nuclear genome ranged from 70 to 300x (**Supplementary Table 3.4**). Insertion sites were determined from sam files using an in-house ruby script. Insertion sites that had 10 reads or more were processed through an in-house python script that counted the number of insertion events in each coding region across the genome. This pipeline was applied to both libraries and further comparisons were made between the pools to determine essential genes. Read data have been deposited at the NCBI under the Bioproject accession number (XXX).

3.5.6 Predicting gene dispensability between species

In order to determine a list of predicted essential genes, comparisons were made between the haploid and diploid libraries. We calculated an insertion ratio by dividing the number of insertions in the haploid pool by the number in the diploid pool. This direct comparison inherently accounts for the length of the gene, since the length is constant in both libraries. Therefore, a decrease in insertion sites in the haploid library indicates a reduction in the presence of mutants containing insertional sites that impact cellular viability. Ratios closer to zero represent insertional mutants that reduce the frequency of haploids harboring insertional sites in a coding region that is required for cellular growth.

To make an insertion ratio cut-off value to categorize essential and non-essential genes, we analyzed the distribution of insertion ratios within intergenic regions between 7 kb and 500 bp in length and positioned between Watson and Crick coding regions (so chosen because these are less likely to contain promoter sequences). The distribution of the insertion ratio calculated for these regions was similar to that of known non-essential

genes in *S. cerevisiae.* Therefore, we used this distribution to rank the insertion ratios of all coding regions and set a cut-off value to 0.25 where 20% of the insertion ratio of coding regions fell below the intergenic distribution, which was similar to the kernel density estimates of known *S. cerevisiae* essential genes. The kernel density estimates were computed in R using ggplot2. To remove a class of low coverage genes in the essential gene category, we applied an additional cut-off value. Since the difference between 0 and 1 with a gene that is longer has a lower weighted difference than a shorter gene, we calculated the difference between the diploid pool and haploid pool and normalized this value to the length of the gene (normalized difference). Genes with less than a normalized difference of 2 were removed from the essential category.

3.5.7 Validating predicted essential and non-essential genes

We validated predicted essential genes by creating *S. uvarum* heterozygous diploid deletion mutants using primers listed in the **Supplemental Table 3.3.** Primers containing 50 bp of homology upstream and downstream of each candidate open reading frame were used to amplify the KanMX cassette from the pRS400 plasmid. The PCR product was used to integrate into the *S. uvarum* genome using a *S. uvarum* specific transformation protocol. The proper integration of the construct was validated through clone purifying positive clones for single colonies and extracting genomic DNA to perform PCR using diagnostic primers listed in the **Supplemental Table 3.3.** The diagnostic primers were designed to target ~150 bp upstream and ~150 bp downstream of the open reading frame to identify wild type and drug marker alleles. Positive clones were sporulated for 3-5 days at 25°C and tetrad analysis of 8 tetrads were screened for

2:2 viable segregation. Images were taken after 4 days of growth on YPD plates. Mutants conferring non-essential phenotypes were replicated on G418 plates and images were taken after 4 days of growth at 25°C (**Supplemental Figure 3.3&3.6**). This method was also applied to making double mutants. A collection of 440 *MATα S. uvarum* strains was used as confirmed non-essential genes that we generously obtained from the Rine lab.

3.5.7 Cross-species complementation assays

To determine if genes are diverging in gene function or in other trans-acting factors, we performed cross-species complementation assays with species-specific essential genes. Essential genes that were *S. cerevisiae* specific were tested in a heterozygous diploid deletion strain from the magic marker collection. Alleles of each *S. cerevisiae* essential gene were amplified from *S. cerevisiae* and *S. uvarum* genomes and cloned into a CEN ARS plasmid. Phusion PCR was used to amplify 500 bp upstream and 5 bp downstream of the stop codon of each gene from *S. cerevisiae* and *S. uvarum.* Each gene was cloned into pIL37 by Gibson assembly using primers listed in **Supplemental Table 3.3** using standard methods (Thomas et al., 2015)**.** All plasmids used in this study are listed in **Supplemental Table 3.2.** The *S. cerevisiae* heterozygous diploid deletion strains were transformed with a plasmid containing a corresponding allele from each species and selected on C-URA plates. Similarly, *S. uvarum* specific essential genes were also tested by making each heterozygous diploid deletion strain *ura3*Δ*/ura3*Δ*,* and transformed with a plasmid containing a corresponding *S. cerevisiae* allele from the MoBY-ORF collection (Ho et al., 2009).

Transformed strains were sporulated for 5 days at 30°C and 25°C for *S. cerevisiae and S. uvarum* species respectively and tetrad analysis was performed on YPD plates. After 3 days of growth, plates were replica plated on C-URA and YPD+G418 plates and imaged after 2 days of growth (**Figure 3.8)**.

3.6 ACKNOWLEDGEMENTS

Thank you to Noah Hanson for technical assistance with tetrad dissections. Thank you to Daniel Chee for help with optimization of Python code for the insertional analysis pipeline. Thank you to Celia Payen, Frances Cheong, Blake Hovde, Sarah Bissonnette, Jeffery Skerker, Rachel Brem, and Amy Caudy for their involvement in this project.

CHAPTER 4: CHARACTERIZATION OF FUNCTIONAL CENTROMERIC SEQUENCES IN *SACCHAROMYCES* YEAST SPECIES

This section is part of an ongoing project

4.1 ABSTRACT

Maintaining proper segregation of genetic material from one cell to the next is the primary function of the centromere. Despite its role is this very fundamental process, centromeric sequences are diverse and are known to evolve quickly. This study aims to functionally characterize centromeric sequences across the *Saccharomyces* clade of yeast to identify common features of centromeric sequences that are required for proper function. Using methods that quantify the amount of plasmid loss, we tested the ability of different strains and species to maintain a standard CEN/ARS plasmid. We identified an increase in plasmid loss in *S. uvarum, S. paradoxus* and one wild strain of *S. cerevisiae.* In *S. uvarum,* adding additional flanking regions around the centromeric region was sufficient to rescue the plasmid loss phenotype, thus, prompting the investigation into understanding the role length and sequence composition contribute to proper segregation of plasmids. Collectively, this work will not only highlight general principles of centromeric function across species, but will also be useful for designing universal plasmids that can be used to improve strains for industrial purposes across a wide span of diverse yeast species.

4.2 INTRODUCTION

Centromeres are required for the proper segregation fidelity of eukaryotic chromosomes during mitosis and meiosis. Specific chromatin structure and organization are characterized by centromere-associated proteins and histone variants forming a unique site for kinetochore attachment (Verdaasdonk and Bloom, 2011). Despite their role in this conserved process, the size range and complexity of centromeric regions are quite diverse, ranging from point centromeres that are approximately 125 bp in length in budding yeast, to megabases of repetitive satellites in human chromosomes (Malik and Henikoff, 2009). Furthermore, repetitive satellite DNA and proteins specific to centromeric chromatin are rapidly evolving, and can vary amongst closely related species (Bensasson et al., 2008). Even within well-defined "point" centromere sequences, variation exists between species within the *Saccharomyces* clade. This ongoing project aims to functionally characterize centromeric sequences in species across the *Saccharomyces* clade by testing the impacts of sequence level divergence on functional elements involved in a highly conserved process.

4.3 RESULTS

4.3.1 Increased plasmid loss in *S. uvarum*

To investigate if centromeric sequences from *S. cerevisiae* are sufficient to properly segregate chromosomes in *S. uvarum,* we tested the maintenance of plasmids as a proxy for proper segregation function. Plasmid propagation is stabilized during cell division by cloning the centromere along with an autonomous replication sequence (ARS) into the backbone of a plasmid (Hsiao and Carbon, 1981). We began by testing if

a standard pPR416 *S. cerevisiae* CEN6/ARS209 containing plasmid (125 bp from CEN6) can properly propagate in *S. uvarum* by calculating the percentage of plasmid loss per generation using a modified mini-chromosome maintenance assay. Plasmid loss is determined by comparing the proportion of plasmid-bearing cells in a given culture by plating some of the culture on YPD plate and some of the same culture on a dropout plate. After 24 hours of growth in non-selective growth, the same measurement is taken again to determine how many cells retained the plasmid. Although there is a background rate of plasmid loss in *S. cerevisiae* (~5%), the plasmid loss per generation is substantially greater (~22%) in *S. uvarum* **(Figure 4.1)**. Given this high plasmid loss in *S. uvarum,* we set out to determine which element of the plasmid was responsible for the high loss rate. We had two hypotheses to explain this result. The first focused on the ARS sequence as potentially not correctly replicating the plasmid. The second could be explained by the centromere and the failure of proper segregation due to differences in centromere sequence specificity or required elements outside the 125 bp sequence used in this plasmid. We proceeded to test the ability of the centromeric sequence to properly segregate by using of a GFP containing plasmid to monitor plasmid retention.

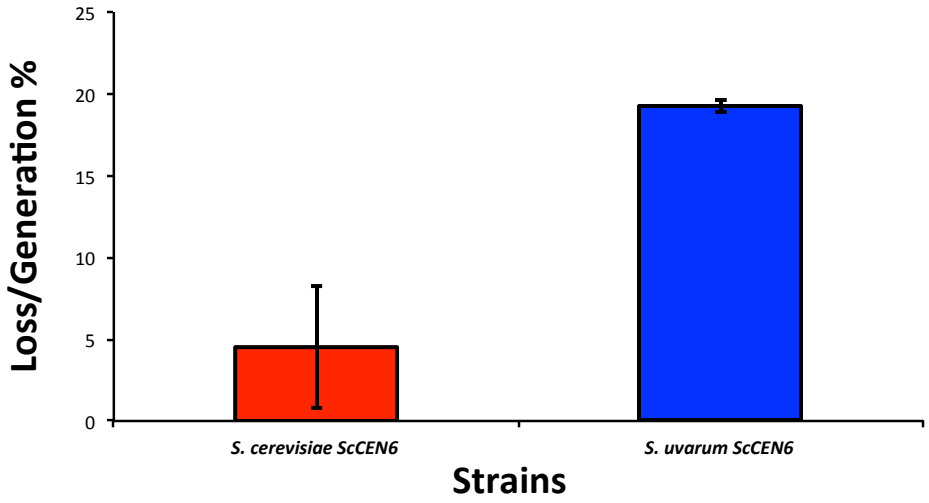


Figure 4.1 Increased percent plasmid loss in *S. uvarum*. Bar chart illustrating percent plasmid loss per generation in *S. cerevisiae* (red bar) and *S. uvarum* (blue bars) backgrounds. Both plasmids contain ARS206 and 125 bp of *S. cerevisiae* CEN6 sequence.

4.3.2 *S. cerevisiae* CEN6 is not sufficient to properly segregate plasmids in *S. uvarum*

We tested proper segregation of a CEN6/ARS206 plasmid in both *S. cerevisiae* and *S. uvarum* using a GFP expressing plasmid harboring an auxotrophic marker for *URA3*. Using flow cytometry, we measured what proportion of the population maintained a plasmid by measuring the florescence intensity of cells grown in restrictive (-ura) and non-restrictive (YPD) growth conditions. In *S. cerevisiae,* we identified a distribution of cells carrying approximately 1-2 copies of the plasmid, with a small proportion of cells undergoing plasmid loss in both restrictive and non-restrictive conditions (**Figure 4.2A)**. However, in *S. uvarum* there is a bi-modal distribution of cells that are either dark or carry several copies of GFP, measured by the population of cells that are highly fluorescent (blue distribution) (**Figure 4.2B)**. The population of cells harboring several copies of the plasmid is diminished under the non-restrictive growth condition, where a majority of the cells have lost the plasmid (red distribution). Taken together, these data suggest that the plasmid seems able to replicate but segregation and thus CEN activity is impaired.

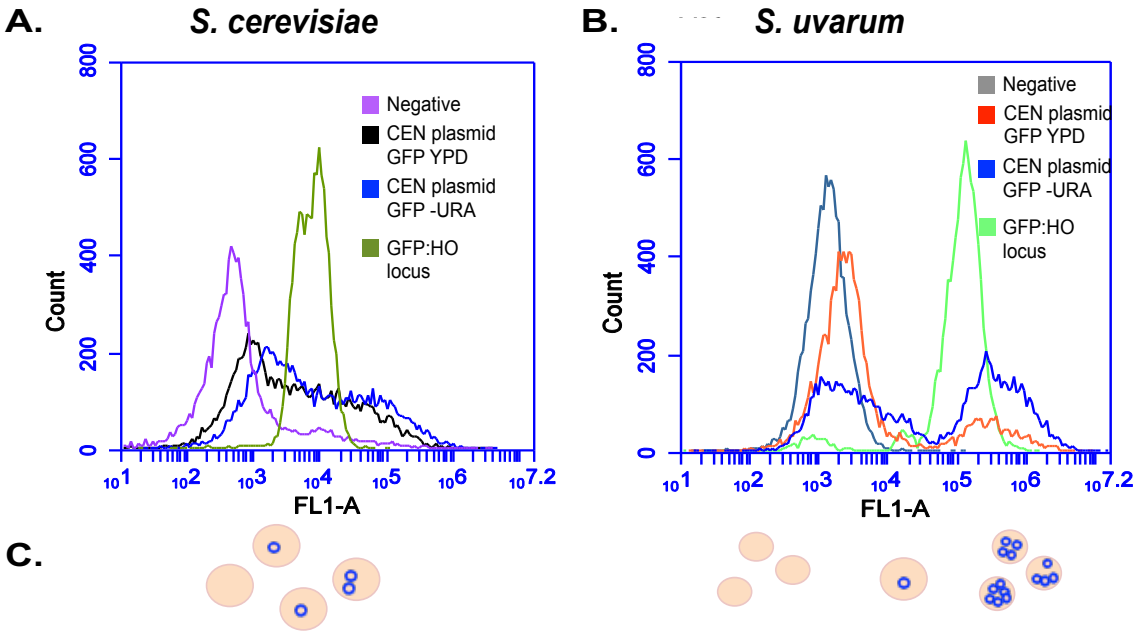


Figure 4.2 *CEN6* from *S. cerevisiae* is not sufficient in *S. uvarum* to properly segregate plasmid. Flow cytometric analysis of yeast strains propagating a GFP reporter CEN/ARS plasmid in *S. cerevisiae* and *S. uvarum*. A) Distribution of cells carrying the GFP plasmid in *S. cerevisiae,* x-axis is fluorescence intensity. Negative (dark) control is indicated in purple, positive control integrated GFP strain in green and strains transformed with the CEN plasmid with and without selection are indicated in blue and black respectively. B) Distribution of cells carrying the GFP plasmid in *S. uvarum,* x-axis is fluorescence intensity. Negative control is indicated in grey, positive control integrated GFP strain in green and strains transformed with the CEN plasmid with and without selection are indicated in blue and red respectively. Notice the bi-modal distribution. C) Schematic illustrating likely plasmid segregation patterns.

4.3.3 Diverse plasmid loss across strains and species in the *Saccharomyces* clade

The unexpected result of differential loss rate between species raised the question whether similar differences in CEN function would be seen between strains of the same species and between a wider set of species. Once we determined that the plasmid loss was due to improper segregation of the plasmid, we were interested in testing the loss rate across different natural isolate strains of *S. cerevisiae,* isolated from diverse ecological niches, and in different species across the *Saccharomyces* genus. Interestingly, one strain in particular, Y55, displayed a higher percentage of plasmid loss than the other three strains tested (~15%) (**Figure 4.3)**. To identify what might be causing the increase in plasmid loss in this strain, future experiments will use quantitative trait loci (QTL) analysis to investigate this result further, using plasmid loss as a quantitative trait in the progeny of a cross between Y55 and a low plasmid loss strain. Furthermore, we also detected a high percentage of plasmid loss (~20%) in *S. paradoxus*, a species more closely related to *S. cerevisiae* than *S. uvarum,* spanning ~2 million years of divergence.

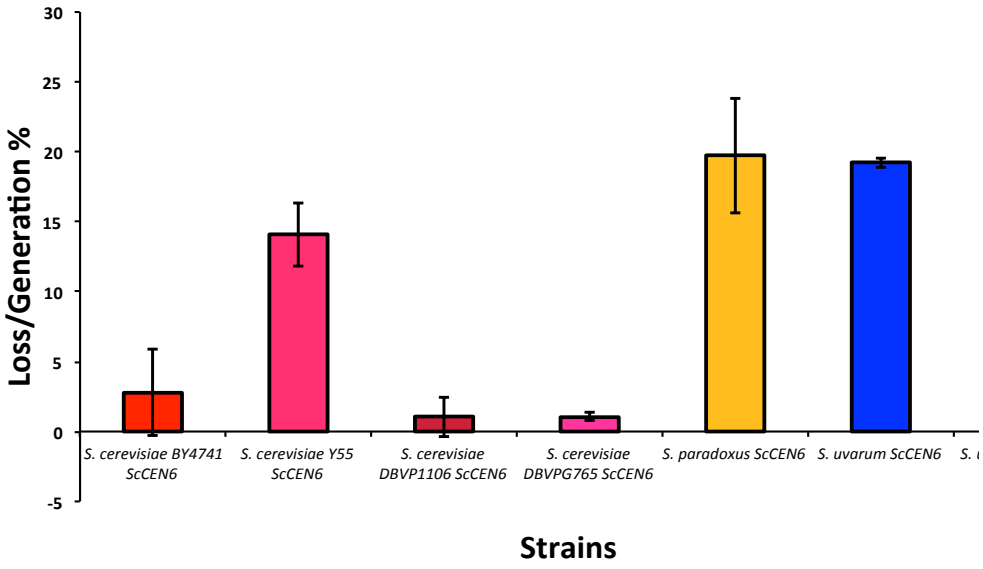


Figure 4.3 Diverse plasmid loss across strains of *S. cerevisiae* and *S. paradoxus*

Bar chart illustrating percent plasmid loss per generation in *S. cerevisiae* (shades of red), *S. paradoxus* (yellow) and *S. uvarum* (blue) backgrounds. All plasmids tested contained an ARS and indicated *CEN6* sequence (125 bp).

4.3.4 Additional flanking region of *CEN6* from *S. cerevisiae* is sufficient in *S. uvarum* to properly segregate plasmid.

Due to high plasmid loss in both *S. paradoxus* and *S. uvarum,* we speculated that sequence divergence at the consensus sequence at the centromere DNA elements (CDE) regions might be responsible for reducing the binding of proteins and complexes required for complete centromeric function in the background of these other species. To test this hypothesis, we cloned a 125 bp fragment of *CEN4* from *S. uvarum,* replacing the *S. cerevisiae CEN6* sequence into the same ARS-containing plasmid and tested plasmid loss. We note that we used CEN4 instead of CEN6 because the sequence of CEN6 is not

determined. As seen in **Figure 4.4,** the percentage of plasmid loss was similar to the loss with the *ScCEN6* sequence, despite the use of *S. uvarum CEN4* sequence. We next considered another explanation, different minimal length requirements for proper segregation. Although 125 bp CEN sequence was tested, a fully functional CEN sequence in diverged species may require additional flanking sequence. To address this point, we tested a plasmid we obtained from the Brewer Lab that contained an extended 500 bp portion of the *ScCEN6* sequence. We found that extending the sequence outside the 125 bp consensus sequences was sufficient to reduce the percent plasmid loss per generation to a percentage similar to what is detected as typical plasmid loss in *S. cerevisiae* **(Figure 4.4)**.

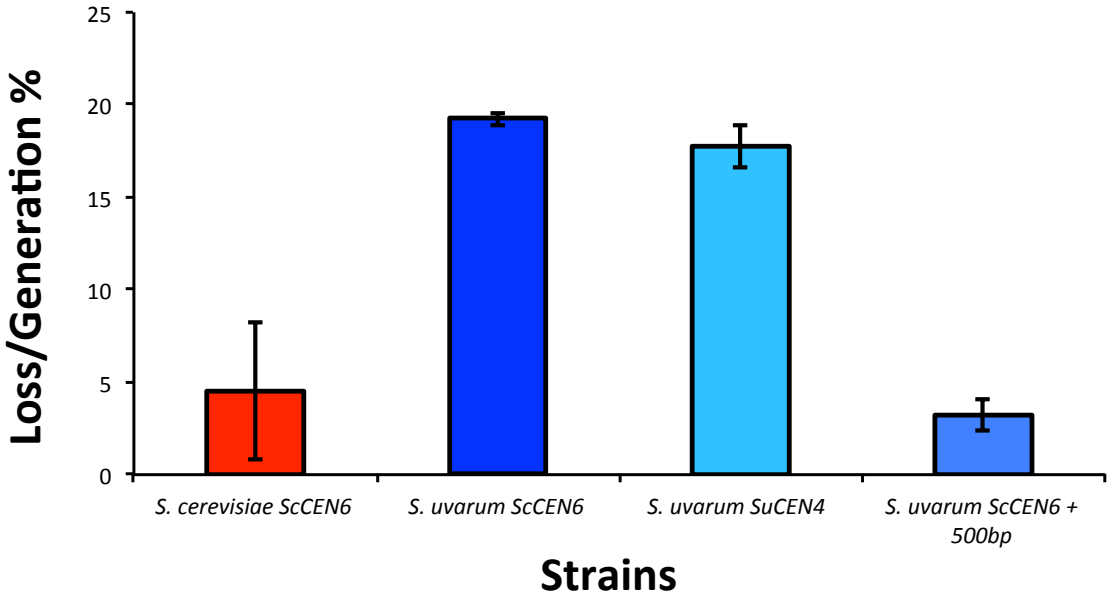


Figure 4.4 Additional flanking region of *CEN6* from *S. cerevisiae* is sufficient in *S. uvarum* to properly segregate plasmid. Bar chart illustrating percent plasmid loss per generation in *S. cerevisiae* (red bar) and *S. uvarum* (blue bars) backgrounds. All

plasmids tested contained an ARS and indicated CEN sequence (123 bp) unless otherwise noted.

4.3.5 min-CEN: identification of functional CEN elements across species.

Since extending the sequence length of the centromeric sequence was sufficient to reduce the percentage of plasmid loss, we were interested in simultaneously testing all centromeric sequences with varying size distributions to identify the most efficient portion of the centromeric sequence that is necessary for proper plasmid propagation. To achieve this goal, we constructed a plasmid library containing random lengths of CEN fragments from *S. cerevisiae* and *S. uvarum*. The random fragments were cloned into a plasmid containing an ARS and a dosage sensitive gene, *HTB2*. This dosage dependent gene is toxic in yeast at a high copy number, which is used as a selection against plasmids that accumulate multiple copies due to missegregation (**Figure 4.5**). We pooled a final

library of ~7,800 bacterial clones to make the final plasmid library, with an average insert size of ~300 bp.

We extracted plasmid DNA from the bacterial pool and transformed *S. cerevisiae* and *S. uvarum* with the plasmid library to ~10x fold coverage. Each pool was grown in batch in medium lacking uracil to select for the plasmid library. Centromeric sequences that cannot maintain the plasmid (loss of *URA3* selectable marker) or missegregate (accumulate high copies of lethal dosage gene) decrease in frequency in the population, while centromeric sequences that promote proper segregation increase in frequency (**Figure 4.5**). Using sequencing, we tracked the relative frequency of each centromeric fragment over time and calculated a relative fitness measurement for each fragment in the background of *S. cerevisiae* and *S. uvarum* (**Figure 4.6**).

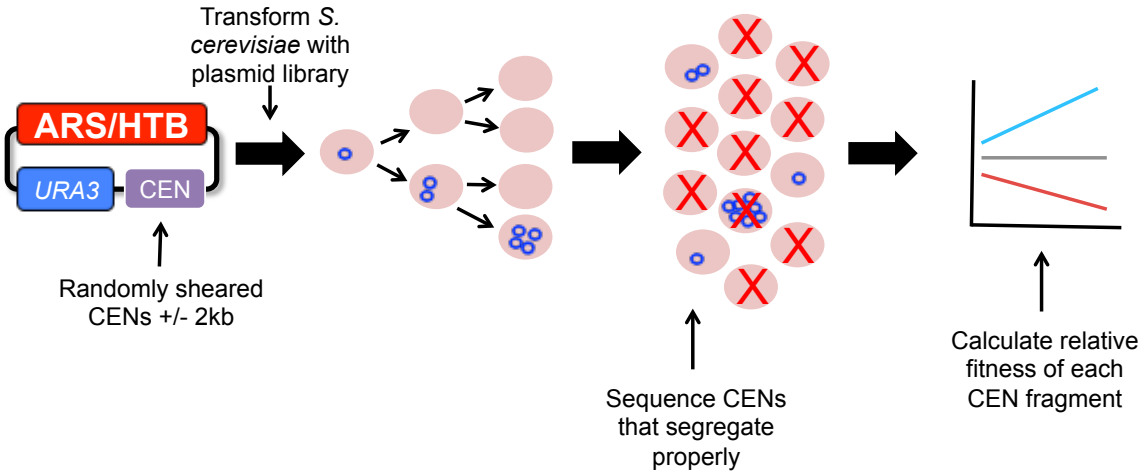


Figure 4.5 min-CEN: identification of functional CEN elements across species. Schematic of functional assay used to test CEN fragments' ability to properly segregate plasmids. Centromeric sequences are randomly fragmented from *S. cerevisiae* and *S. uvarum* and cloned into a plasmid containing a selectable marker (*URA3*) and a dosage sensitive gene (*HTB2*). *S. cerevisiae* and *S. uvarum* are transformed with the plasmid library and grown in batch culture. Samples are taken throughout the growth phase and sequenced to determine the frequency of the centromeric sequences.

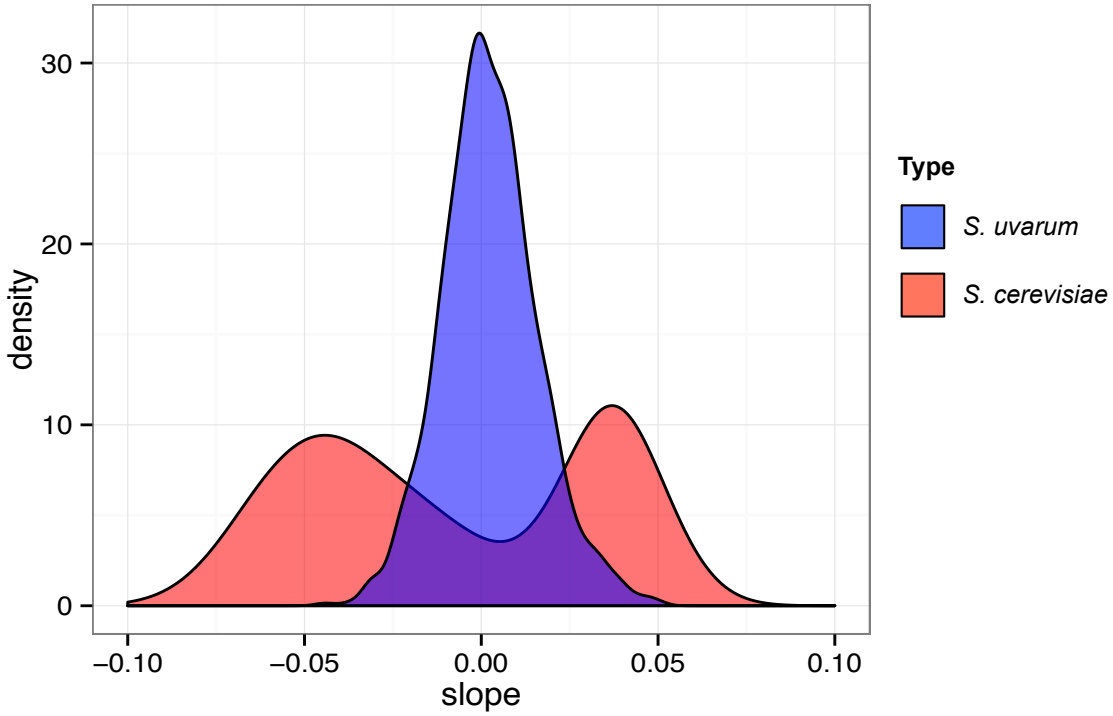


Figure 4.6 Distribution of CEN fragment fitness values in *S. uvarum* and *S. cerevisiae.* Density plots of relative fitness measurements for each centromeric fragment (one replicate). Blue density plot represents the fitness distribution of CEN fragments tested in *S. uvarum.* The red density plot represents the fitness measurements performed in *S. cerevisiae.*

While a clear difference of optimal fitness measurements is detectable in *S. cerevisiae,* most fitness values in *S. uvarum* followed a normalized distribution centered on zero, with very little variance (**Figure 4.6**). Furthermore, when we tested how well the fitness measurements correlated between replicates, the measurements in *S. uvarum* did not correlate (R^2 =0.0006) (**Figure 4.7A**). However, fitness measurements modestly correlated in the *S. cerevisiae* library (R^2 =0.637) (**Figure 4.7B**). Given the lack of correlation and the limited distribution of fitness values that were determined in *S. uvarum,* we conclude that the library either contained cloned fragments that were too small to promote segregation, or the toxic histone gene is not functional in *S. uvarum* to select against missegregating plasmids. Further work will determine why the plasmid library failed in *S. uvarum.* The data set I presented in the *S. cerevisiae* section was obtained from a limited amount of sequencing reads and will require more sequencing coverage and further analysis to make any further conclusions. However, the modest correlation between replicates in the fitness data from *S. cerevisiae* is promising and suggests that we will have the ability to find optimal sequences and lengths that promote properly segregating plasmids.

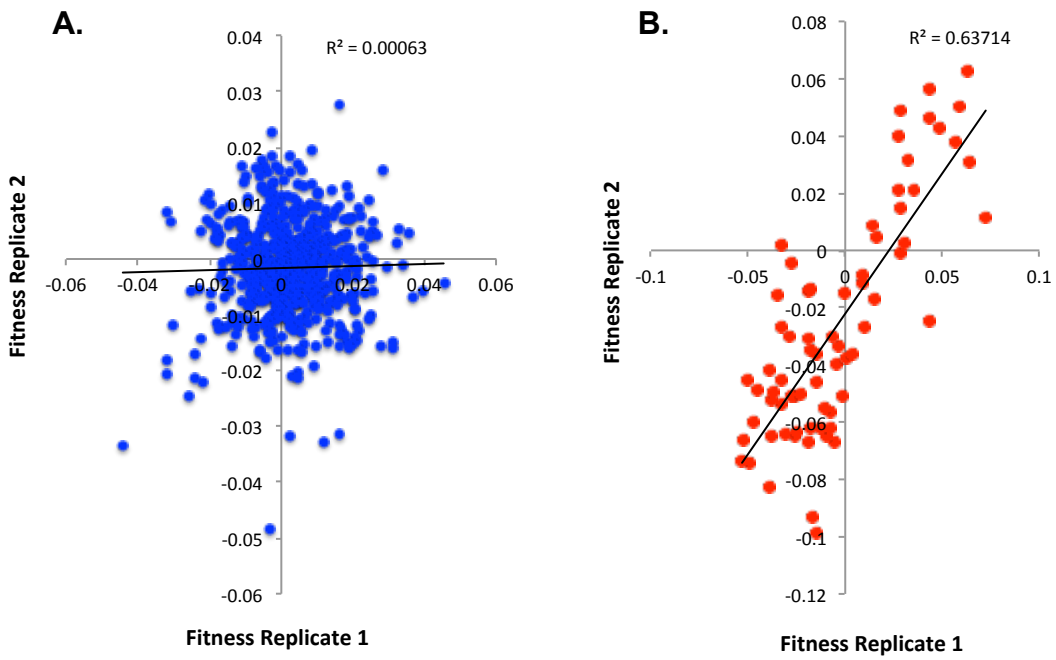


Figure 4.7 Scatter plots of fitness values between replicates A) Scatter plot of fitness values between replicates in blue represent data from *S. uvarum* (R^2=0.0006). B) Scatter plot of fitness values between replicates in red represent data from *S. cerevisiae* (R^2=0.637).

4.4 DISCUSSION AND ONGOING WORK

The data presented in this chapter are a part of an ongoing project that aims to functional characterize centromeric sequences across the *Saccharomyces* clade of yeast species. This project utilized a variety of different functional tests that successfully identified the lack of proper plasmid segregation as a potential source of increased plasmid loss in *S. uvarum.* Furthermore, we identified that the length of centromeric sequence (or different portions of the consensus sequence region) may be important for proper plasmid segregation in *S. uvarum.* Finally, on-going experiments will identify the minimal functional elements required for efficient chromosome segregation, identify optimal sequences amongst all 32 chromosomes in the pool and probe for function in the context of different genetic backgrounds. One preliminary result from the screen in *S. cerevisiae* revealed that segments from the *S. uvarum CEN14* yielded the highest fitness score. However, more rigorous analysis and increased sequence coverage will validate this result. In summary, this study will not only shed light on the critical aspects of centromeric structure and functional sequence requirements, but will also lead to the improvement of universal cloning vectors that can be used for research and industrial applications.

4.5 MATERIALS AND METHODS

4.5.1 Strains, plasmids and primers

The strains, plasmids and primers used in this study are listed in **Supplementary Tables Appendix 4.1**, **4.2** and **4.3** respectively. Unless specified below, yeast strains were grown at 25°C for *S. uvarum* strains and 30°C for *S. cerevisiae* strains and standard

media recipes were used.

4.5.2 Plasmid loss assay (MCM assay)

Mini-chromosome maintenance (MCM) assays were as described in (Donato et al., 2006) using the plasmid pIL37. Specific modifications are described in (Liachko et al., 2013).

4.4.3 GFP plasmid analysis

S. cerevisiae and *S. uvarum* strains were transformed with pMS29 and grown in 3mL of –ura media for 24 hours. Samples were run on a flow cytometer (BD accuri) to determine the proportion of cells contained the GFP plasmid. From 3mL culture grown in –ura media, 3µl of cells were transferred to YPD and grown for 24 hours. Samples were diluted and ran on flow cytometer.

4.4.4 Construction of the min-CEN plasmid library

To make the min-CEN plasmid library, primers were designed (listed in **Supplementary Tables Appendix C.3)** 1kb upstream and downstream of CEN sequences in *S. cerevisiae* and *S. uvarum.* All centromeric sequences are represented except *S. uvarum* CEN10 and CEN7. PCR products were pooled together in equal molar ratios and fragmented using the Covaris. Random fragments were ligated into pIL49 and used to transform *E. coli.* Positive transformants (amp resistant) were scraped and pooled together (~7,800). Plasmid DNA was extracted using the Promega plasmid extraction kit.

4.4.5 min-CEN sequencing library preparation

The plasmid DNA extracted from the yeast population were subjected to analyses by deep sequencing as described (Liachko et al., 2013). Index primers are listed in **Supplementary Tables Appendix C.3**.

4.6 ACKNOWLEDGEMENTS

I would like to thank Tom Pohl and the Brewer lab for providing the CEN/ARS plasmid containing the extended 500 bp region around *CEN6*. I would also like to thank Celia Payen for her helpful suggestions about which natural isolates to test for the plasmid loss assays. I would also like to thank Rebecca Martin, Sarah Hilton, Azhar Khandekar, and Ivan Liachko for their involvement in this project.

CHAPTER 5: CONCLUSIONS AND FUTURE DIRECTIONS

While sequencing genomes is a solved problem, interpreting them is still difficult. One reason why it is still difficult is because we still don't understand how genetic variants interact with each other. That is, why might the same mutation be benign in one person and cause disease in another? Excluding environment effects, the answer must be that something in their genetic background. Discovering the heritability of many complex diseases will rely on a more comprehensive analysis of genetic background effects, ranging from large structure variants to understanding the joint effects of many loci of small effect (Eichler et al., 2010). Connecting genotype to phenotype amongst individuals is complicated by multiple genetic loci interactions observed as epistasis, buffering, or robustness, making it difficult to parse out causality (Hartman et al., 2001). Due to the complex nature of genetic background effects on phenotypes across human populations, we can turn to model organisms to study the effects of genetic interactions and modifying loci on general mechanisms of adaptation and gene function evolution. The following sections summarize the work that I have performed to address the impacts of genetic background on yeast evolution.

5.1 Genetic background and adaptation

In chapter two of this thesis, I discussed a study that utilized comparative experimental evolution to investigate the effects of genetic background differences between diverged species of yeast on adaptation to a nutrient limited environment. Using this approach, we identified how the genetic context of different species, specifically changes in non-coding regions, can shape multiple different evolutionary outcomes. First, we observed differential amplification events between sulfate transporter paralogs

and attributed this difference to beneficial fitness effects of each allele tested in the context of each species. Then we tested the fitness contributions of different combinations of non-coding and coding regions of each allele and identified reduced fitness effects and decreased expression associated with the non-coding region of the *S. uvarum SUL1* allele. While we did not identify the precise variant/variants that caused these differences, we presented several future experiments, such as deep mutation scanning of the *S. uvarum SUL1* promoter, to test for functionally important sequence variants and positions. To expand upon the conclusions I summarized in chapter two, I will focus this section on additional hypotheses that more fully explore the effects of the genomic context of *S. uvarum* and describe experiments that can address additional aspects of genetic background effects on adaptation more broadly.

Although the genomes of *S. cerevisiae* and *S. uvarum* are largely syntenic, differences in gene composition differ in the *SUL1* locus (Scannell et al., 2011). To determine if the genomic context of *S. uvarum* altered the ability for a *SUL1* amplification event to occur, we created a sensitized genetic background by creating a *SUL2* deletion and evolved this strain in sulfate limited conditions. We found evidence of *SUL1* amplification, ruling out one hypothesis that addressed the chromosomal context of the amplification event; however, other questions about different aspects of the genomic context remain.

For instance, to test our claim that expression differences between the paralogs are a sufficient driver of these different evolutionary trajectories, promoter swaps of all four alleles at the native *S. cerevisiae SUL1* and *S. uvarum SUL1* loci can be created and evolved in sulfate limited conditions to determine if the expression differences are

sufficient to select for amplification events at these chromosomal loci. Additional monitoring of the timing of the amplification events, what frequency they occur at and whether other beneficial mutations occur simultaneously can help track the dynamic events. Timing of events is particularly interesting in the context of *S. uvarum,* and it is worth noting that the *SUL2* amplification event was only observable after 500 generations. Several interesting aspects about this finding can be further explored to 1) determine what other beneficial mutations might arise and 2) test other genetic features that might affect the rate of amplification events in the *S. uvarum* genomic context.

Previous laboratory evolution experiments in several parallel populations of *S. cerevisiae* have identified a beneficial mutation in *SGF73,* in the context of clones both with and without *SUL1* amplification events (Araya et al., 2010; Gresham et al., 2008; Payen et al., ,2013). Loss of function of *SGF73* alone provides a 26% fitness advantage in sulfate-limited conditions. Performing whole genome sequencing on evolved clones every 50 generations while simultaneously testing the fitness may reveal other beneficial mutations in *S. uvarum* that lead to other adaptive routes rather than gene amplification events. Since the fitness benefit of the most highly adaptive route of amplifying *SUL1* is diminished in this species, this provides a different adaptive landscape that can be explored through alternative mutational spectrums. Examples of possible adaptive nonsynonymous mutations have been identified in *PIN4* and *MPS5 (Heil et al., 2016).*

One common mechanism of gene duplication relies on repetitive segments for sites of recombination. Ribosomal DNA (rDNA) arrays and telomeric loci are examples of repetitive DNA regions that have been shown to vary in copy number and contribute to dynamic changes within the genome. Other repetitive elements like Ty elements and long

terminal repeats are widely distributed across the *S. cerevisiae* genome (Demeke et al., 2015). It is known, however, that *S. uvarum* does not harbor any full-length Ty elements (Liti et al., 2005; Neuvéglise et al., 2002), which may contribute to a reduction in (1 CNV) the average number of CNV per clone versus 1.5 in *S. cerevisiae (Heil et al., 2016)*. Introducing full-length Ty elements in an *S. uvarum* background would therefore test if increasing repetitive elements would impact the frequency of detecting amplification events.

Furthermore, with the use of CRISPR/Cas9 technology, it would be feasible to engineer strains containing guide RNA targets to the conserved regions of Ty elements in various other species and natural isolates. Creating double stranded breaks in conserved regions of Ty sequences would provide template DNA to repair off multiple different Ty regions throughout the genome, creating vastly rearranged genomes. These diversely rearranged clones can be used to test how gross structural rearrangements affect adaptive potential, both within lab and other relevant environmental contexts, such as industrial fermentation environments that are known to select for specific traits in highly mosaic hybrid yeast strains (Dunn and Sherlock, 2008; Libkind et al., 2011).

Lastly, other experiments could also test overall adaptive potential of diverse wild strains of natural isolates to novel growth environments that span a variety of different ecological niches, such as sub-optimal temperatures. Preliminary work (Payen, unpublished) has explored the adaptability of natural strains of *S. cerevisiae* to explore genetic background effects on the adaptive potential to a strong selection pressure of sulfate limitation, identifying an anti-correlation between starting fitness and over-all fitness potential. Furthering these types of studies by pairing experimental evolution in a

variety of ecologically relevant conditions and testing the adaptability of different wild strain combinations can provide a model for directly testing diminishing returns epistasis. Because different strain backgrounds harbor varying degrees of genetic diversity, deleterious mutations hindering known mechanisms of gene amplifications may exist, and may also guide different adaptive routes that are directly dependent on the starting landscape. Overall, the principle of performing comparative experimental evolution across a diverse set of genetic backgrounds will prove to be useful in furthering our knowledge about the relationship between genome evolution and trait variation and their effects on evolutionary mechanisms of general adaptive processes.

5.2 Genetic background and gene function evolution

While comparative experimental evolution lets us bridge the gap between long and short evolutionary timescales, it isn't a substitute for thoroughly understanding these different species which allows us to better interpret the differences we're observing. Accurate functional annotation of genes is required for understanding how biological processes are formed and operate. Elucidation of gene function and functional tests of mutations remains to be one of the largest hurdles facing modern biology. In chapter three of this thesis, I addressed the need for more functional tests in non-standard genetic backgrounds and discussed a high-throughput method that was implemented in *S. uvarum* to characterize differential gene dispensability between *S. cerevisiae* and *S. uvarum*. The aim of this chapter was to begin to functionally characterize this non-standard species and to investigate gene function evolution between diverged species of yeast, using differences in gene dispensability as a proxy for gene function divergence. Since I

included conclusions of this project in the discussion of chapter three, I will limit this section to addressing specific remaining questions for future studies, such as the potential for applications of this large-scale library to identify genetic background effects on gene function evolution more broadly.

Using the library of random insertion mutants in *S. uvarum,* we identified 12% of genes to be differentially essential between the two species. While this category still requires further validation to confirm the remaining genes in this category, particularly given our relatively high false discovery rate, GO analysis revealed that genes that were essential only in *S. uvarum* are significantly enriched for structural constituents of the ribosome. Interestingly, a previous study in *Candida albicans* also identified a core group of 25 genes that were haploinsufficient across all conditions tested and were also significantly enriched for structural constituents of the ribosome GO enrichment category (Oh et al., 2010). Future studies should investigate these results further; however, identifying conserved functional categories across diverged species is a promising step to furthering our understanding of the pan genome that might be useful for designing broad-spectrum anti-fungal therapeutics.

One surprising result suggests that most differences in gene dispensability are likely due to *trans*-acting factors. An additional key point that remains to be addressed is identifying what specific loci are responsible for differences in gene essentiality between *S. cerevisiae* and *S. uvarum.* Given the results from our complementation tests where most genes can complement the inviable phenotype, these effects are most likely to be *trans*-acting. One method to pinpoint candidate loci involved in species-specific gene essentiality is quantitative trait loci (QTL) analysis in viable segregants from a hybrid

cross harboring gene knockouts of a species-specific allele. The identification of loci enriched in the viable segregants carrying the deletion can help identify particular regions of the genome that are responsible for differences in gene dispensability, highlighting specific genetic interactions. Although further optimizations of the random spore analysis and haploid selection protocols are required for their use in hybrids, I recently constructed auxin inducible, tagged degron *S. cerevisiae* strains that can be used to make hybrids to test this hypothesis.

The ability to perform hybrid genetics was an advantage of selecting *S. uvarum* as a conveniently placed species on the phylogenic tree to study. Further analysis of gene function in the context of hybrid evolution can also be studied using this library. Mating the *S. uvarum* transposon library to wild type or mutant *S. cerevisiae* can provide information about genes that might be involved in genome conflict resolution, hybrid-specific genetic interactions (both positive and negative) or even identify speciation genes. A recent study performed SAturated Transposon Analysis in Yeast (SATAY) in *S. cerevisiae* and would be an ideal library to pair with, given the highly dense coverage of the library (Michel et al., 2017b). Furthermore, a previous study investigated haploinsufficiency and haploproficiency in hybrids by mating the deletion collection to wild type *S. uvarum* (Lancaster, unpublished) and has already identified genes important for growth in nutrient limited conditions, serving as a useful dataset to compare against.

One major limitation of our library was the modest overall insertional coverage, which likely led to many of our false positives. Additionally, it would be difficult to attain any substantial information about essential protein domains by analyzing the specific gaps in insertion densities within the coding regions. However, future studies

may include applying the SATAY system in *S. uvarum* and a variety of other species to gain fine-scale resolution of the impacts of sequence level, natural variation on functional protein domains.

In conclusion, the largest challenge and ultimate goal of genome research is to predict phenotype from genotype. A systems level approach of understanding how natural genetic variation influences allelic specific effects on biological systems as a whole is required to approach this challenge. The scope of the work presented in this thesis begins to tackle this challenge by taking comparative experimental evolution and systems level approached to studying genetic background effects of gene function evolution to provide new insights into mechanisms of molecular adaptation.

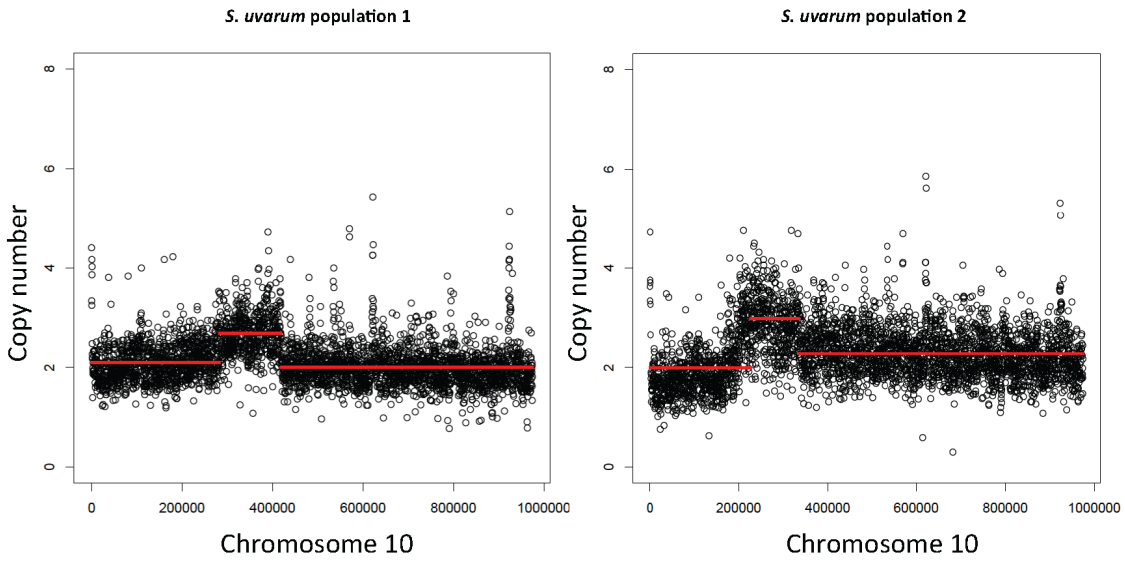
APPENDICES

APPENDIX A: SUPPLEMENTARY MATERIAL FOR CHAPTER 2

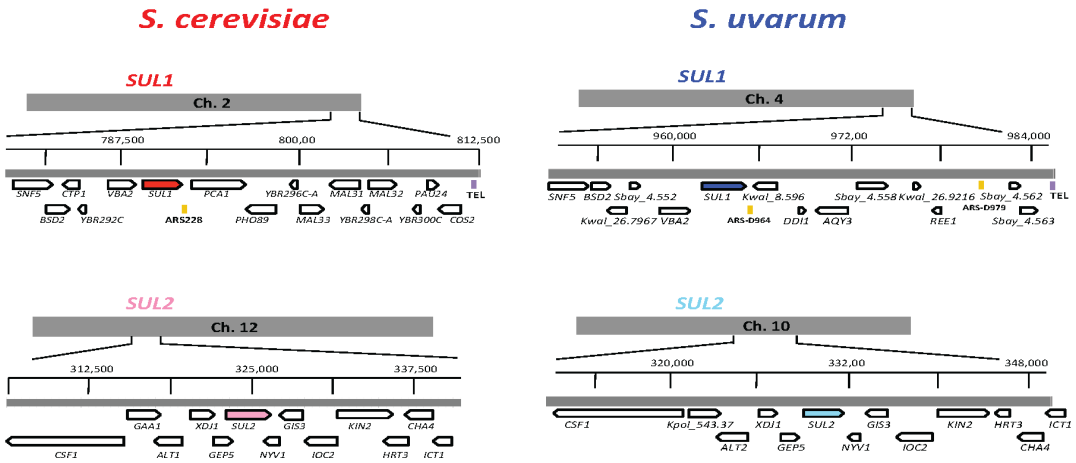
Supplemental Table 2.1 List of strains

Supplemental Table 2.2 List of plasmids

Supplemental Table 2.3 List of primers

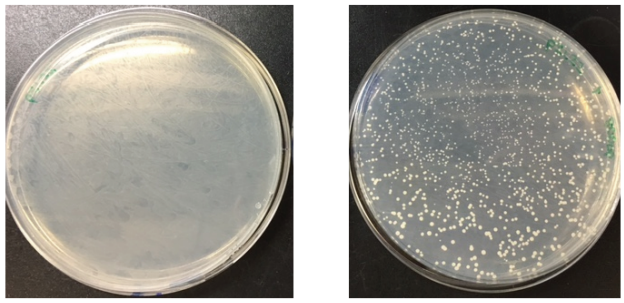


Supplemental Figure 2.1 Chromosome X copy number plots of two evolved populations of *S. uvarum*.

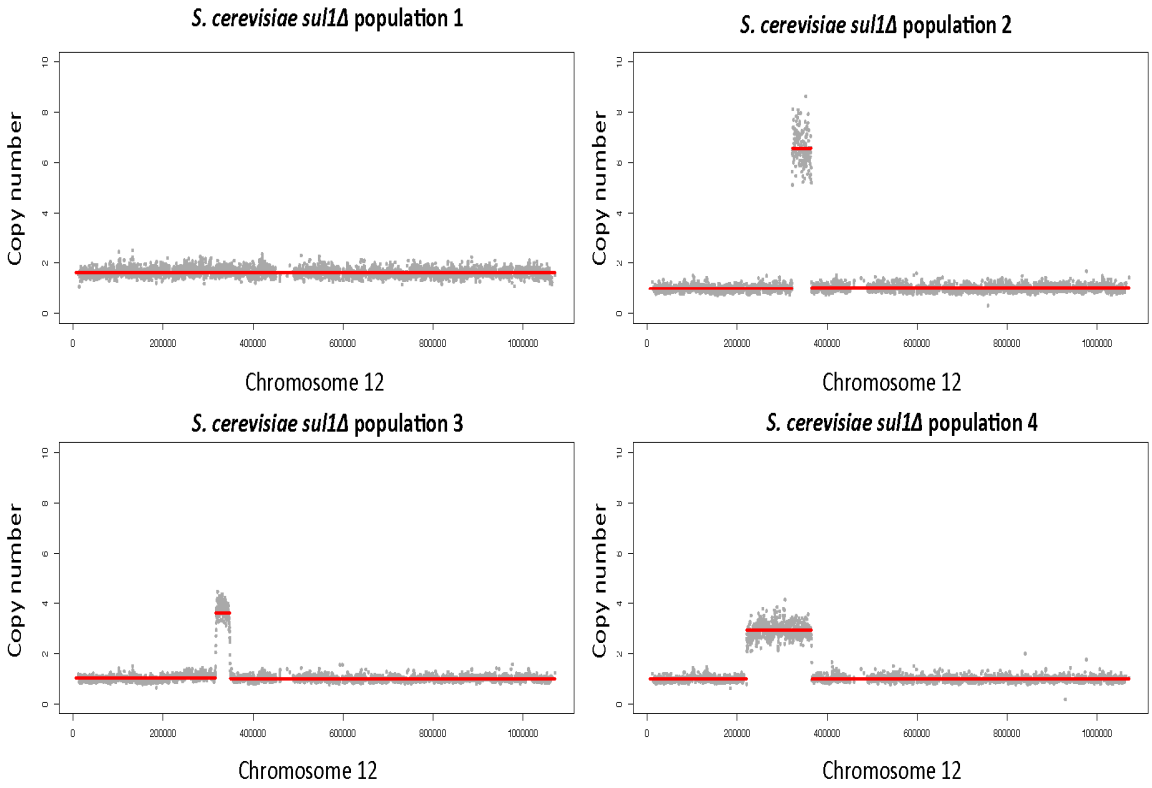


Supplemental Figure 2.2 Genetic context of the *SUL*

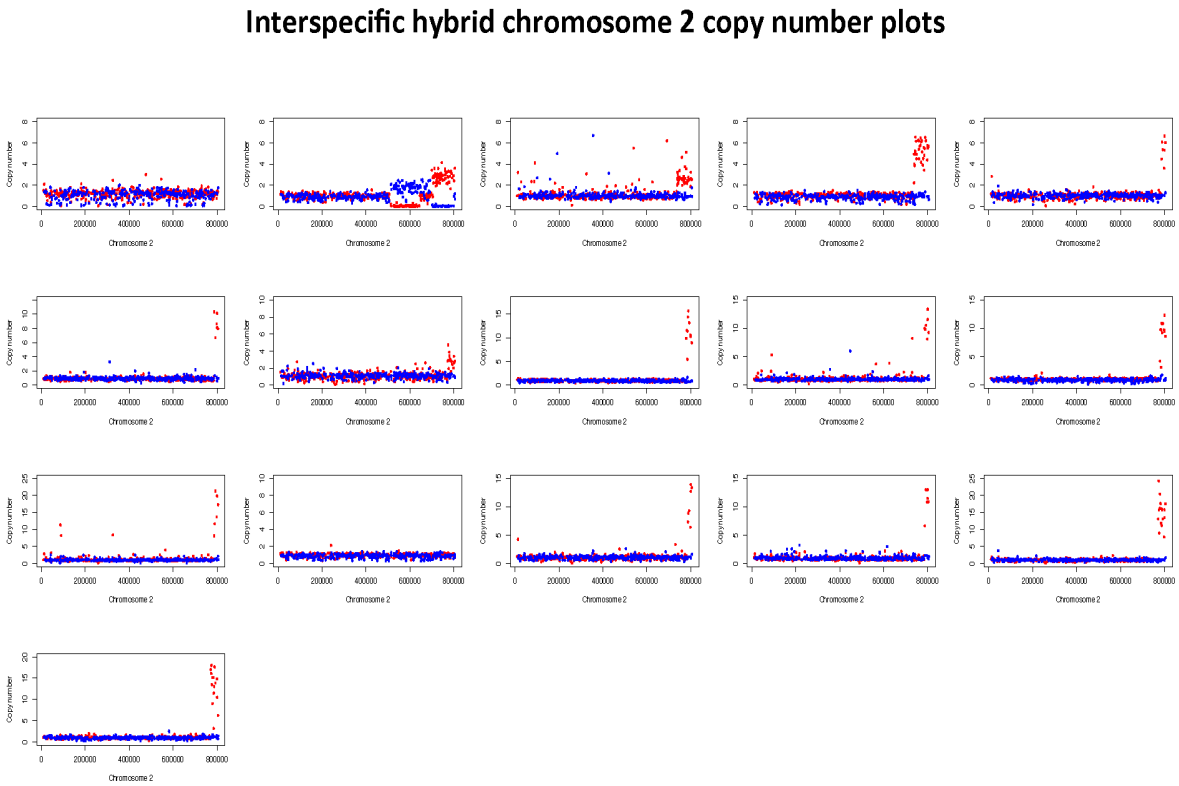
alleles in *S. cerevisiae* and *S. uvarum*



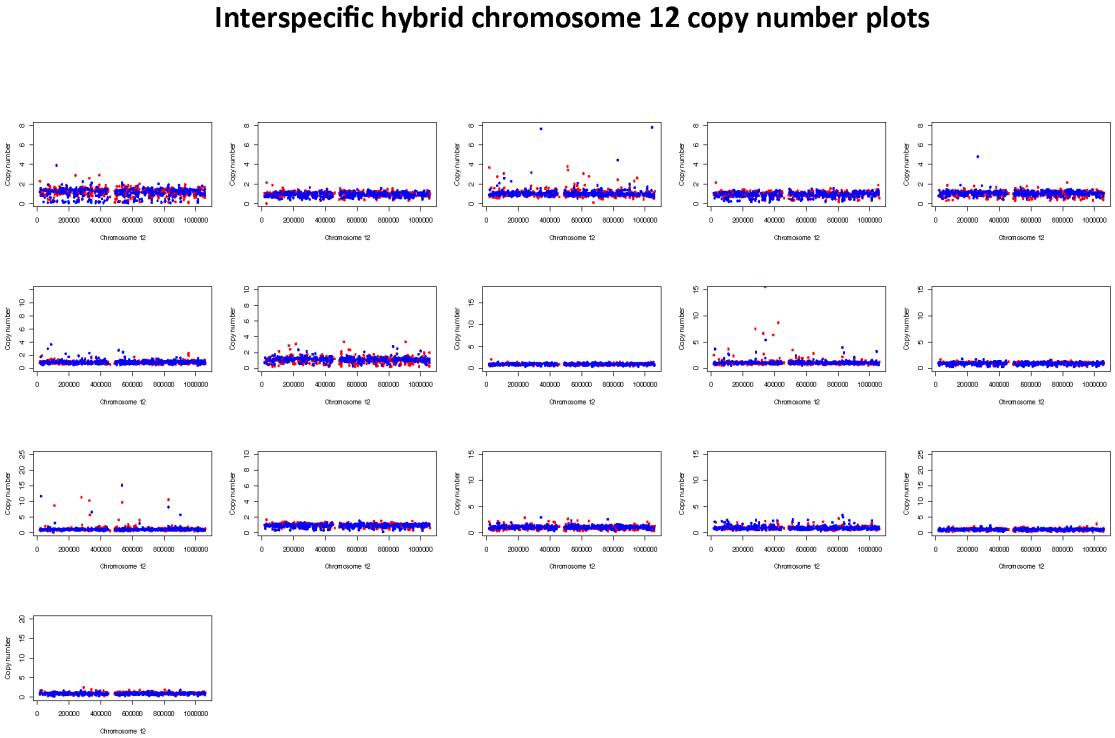
Supplemental Figure 2.3 Plasmid replication containing 800 bp downstream of *SUL1* from *S. uvarum*



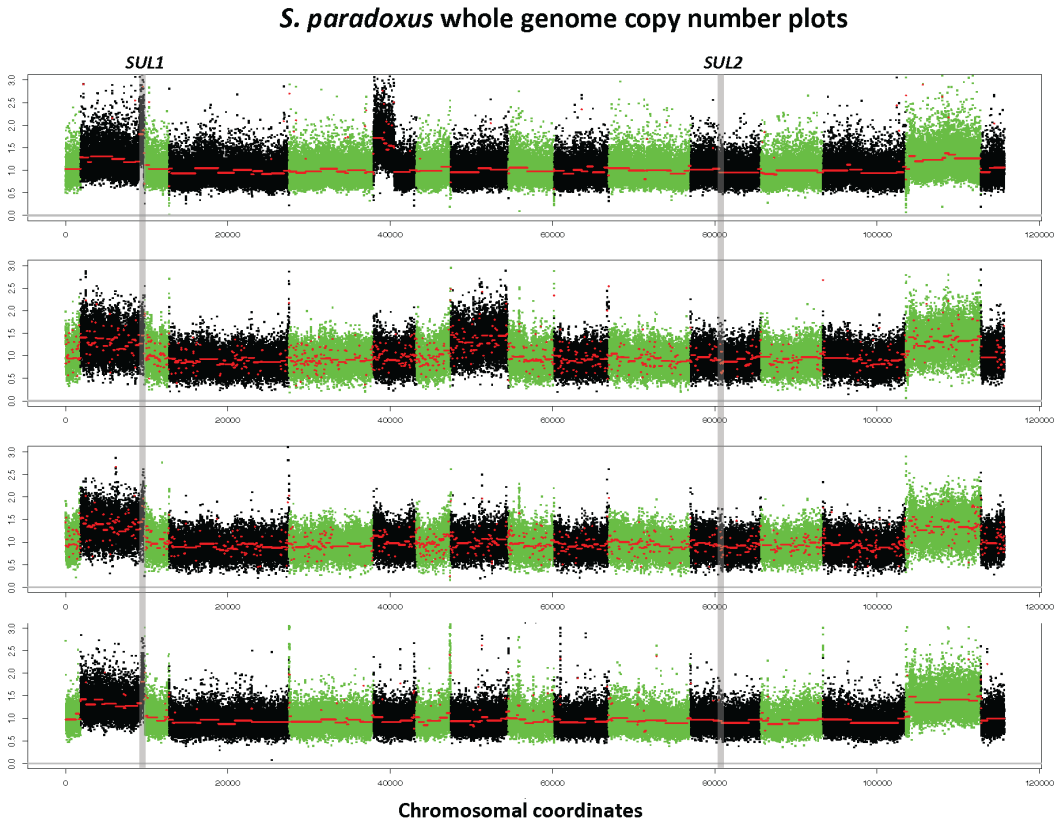
Supplemental Figure 2.4 Chromosome XII copy number plots of four evolved *sul1*Δ *S. cerevisiae* populations



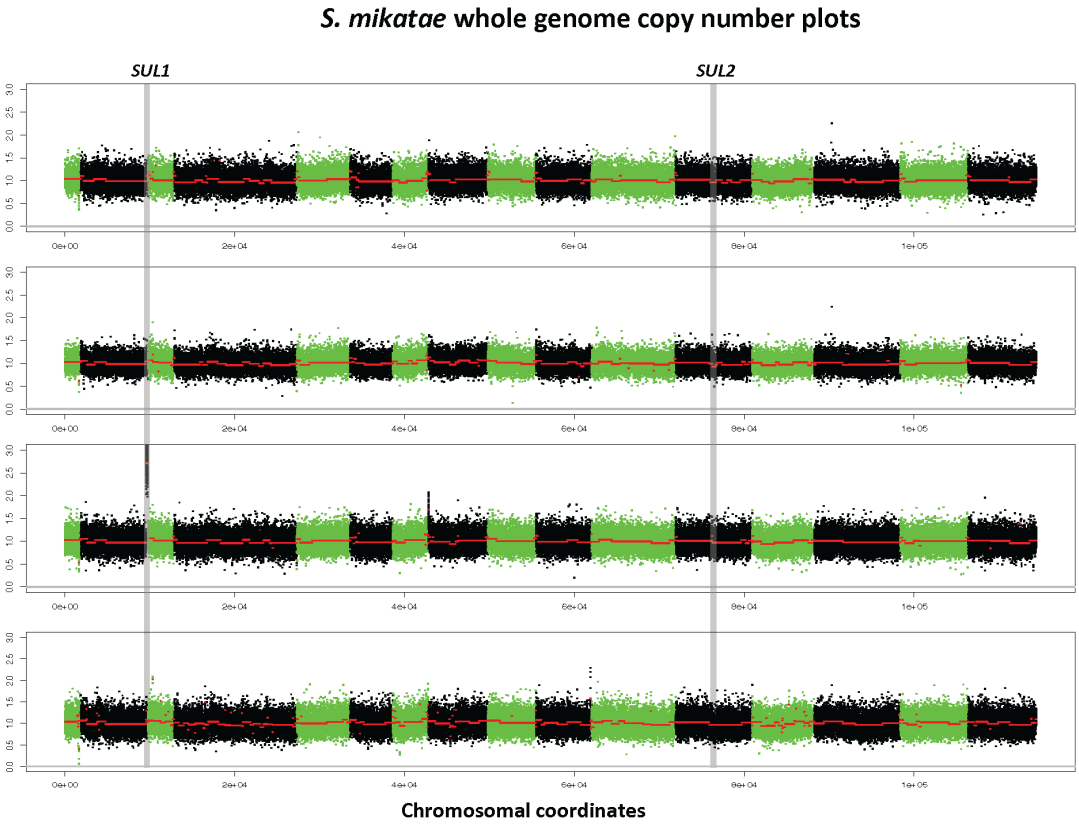
Supplemental Figure 2.5 Chromosome II copy number plots of 16 evolved hybrid clones. Note that each panel is scaled according to the range of values for that individual experiment.



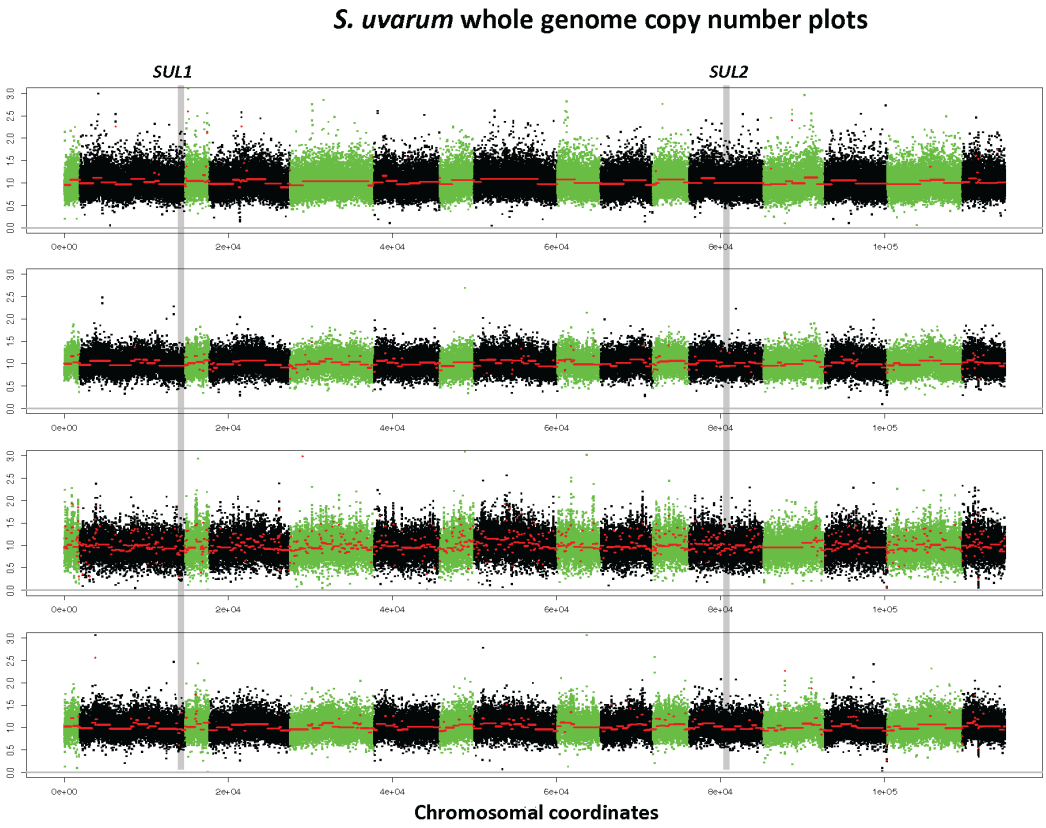
Supplemental Figure 2.6 Chromosome XII copy number plots of 16 evolved hybrid clones. Note that each panel is scaled according to the range of values for that individual experiment.



Supplemental Figure 2.7 Whole genome copy number plots of four evolved *S. paradoxus* populations



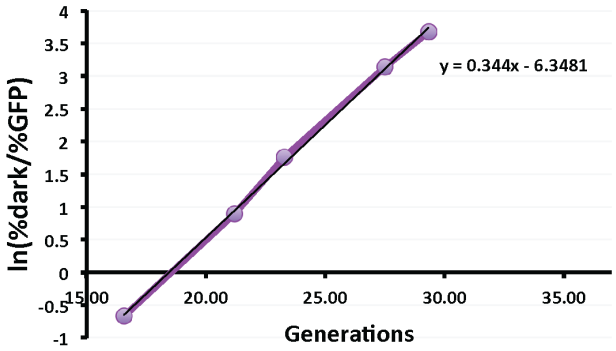
Supplemental Figure 2.8 Whole genome copy number plots of four evolved *S. mikatae* populations



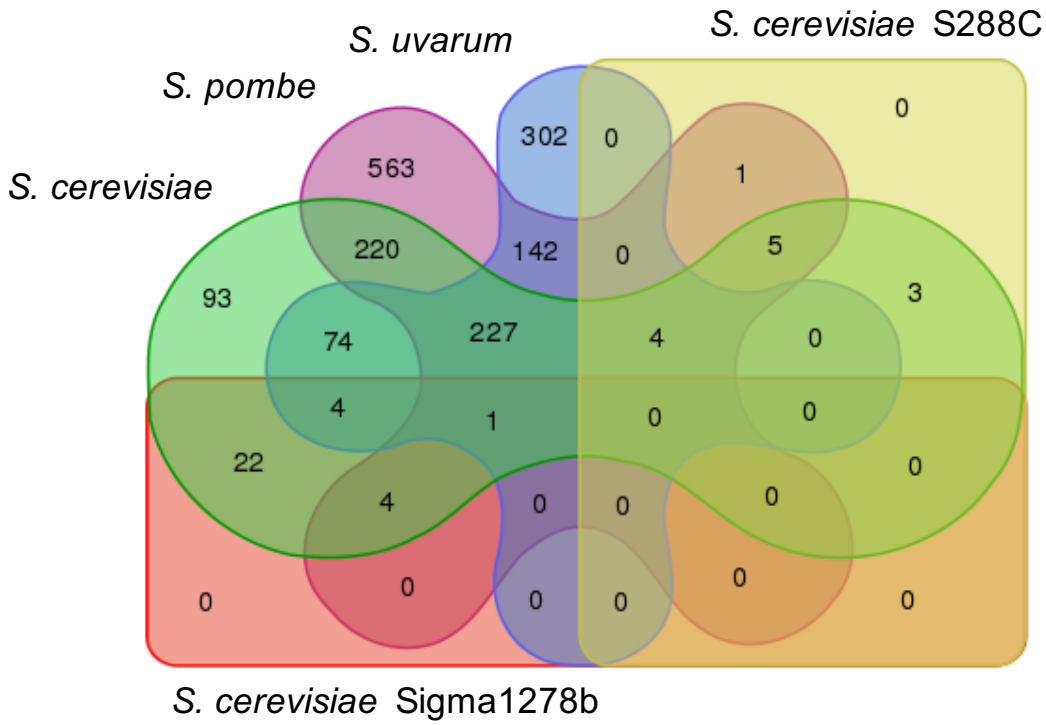
Supplemental Figure 2.9 Whole genome copy number plots of four evolved *S. uvarum* populations

Example of fitness coefficient calculations

% GFP	Generation	Ln(dark/GFP)	Fitness (slope)
63.67%	19.88	-0.561069648	0.344
23.52%	23.48	1.179178146	
11.98%	25.98	1.994325469	
4.14%	29.76	3.142193006	

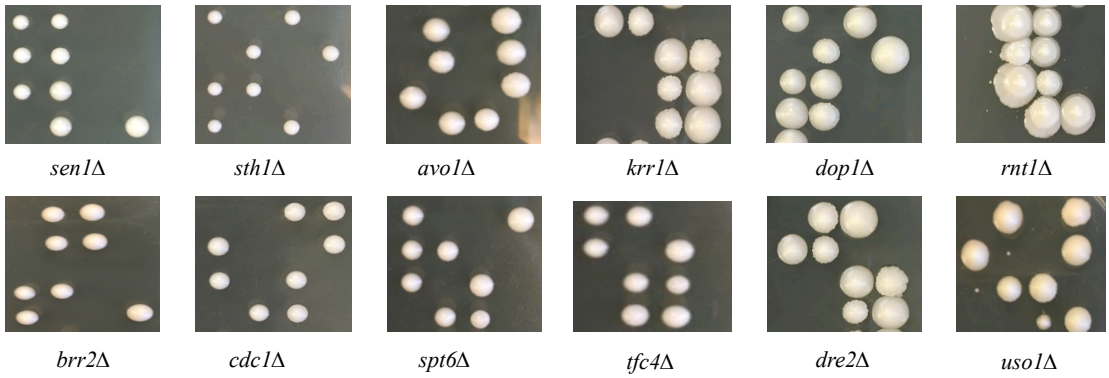


Supplemental Figure 2.10 Example of fitness coefficient calculations using data from a *S. cerevisiae* strain transformed with a *ScSUL1* containing plasmid



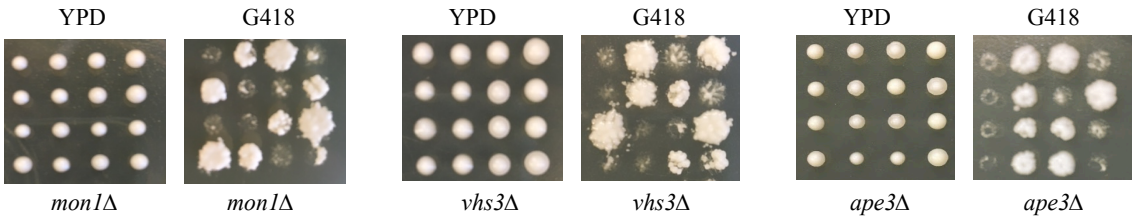
Supplemental Figure 3.1 Conservation comparison between *S. cerevisiae, S. pombe* and predicted *S. uvarum* essential genes

Conserved Essential Genes: ScE:SuE



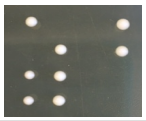
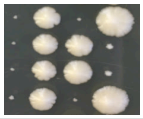
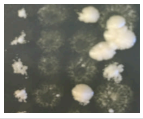
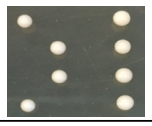
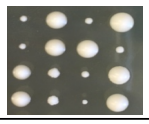
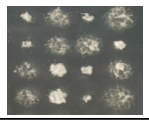
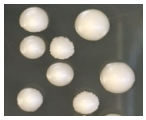
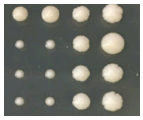
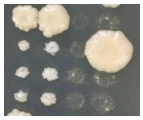
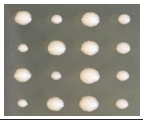
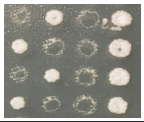
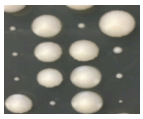
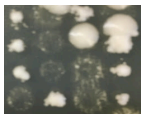
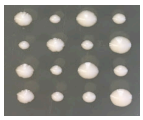
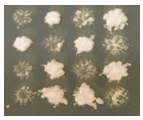
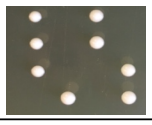
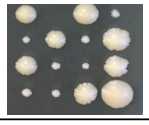
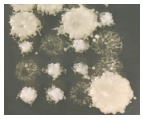
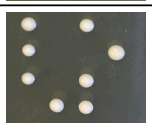
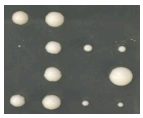
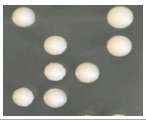
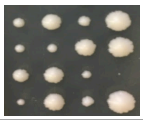
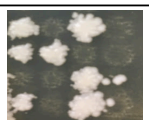
Supplemental Figure 3.2 Confirmed tetrad analysis for conserved, predicted essential genes

Conserved Nonessential Genes: ScNE:SuNE



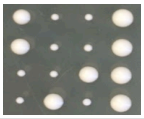
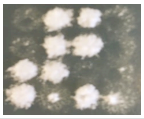
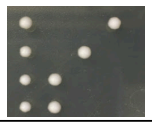
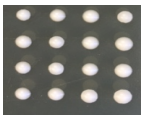
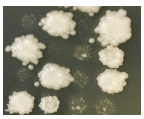
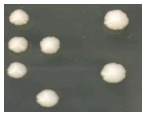
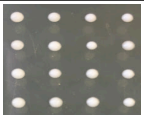
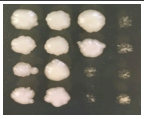
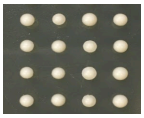
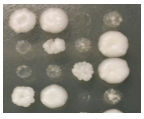
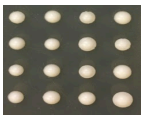
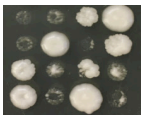
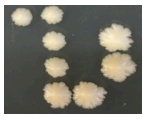
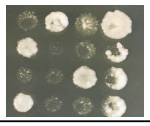
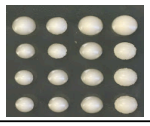
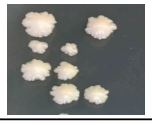
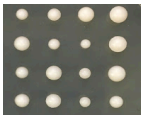
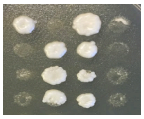
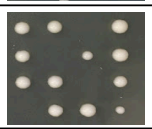
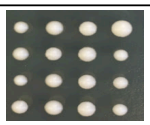
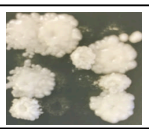
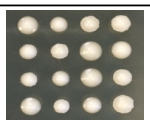
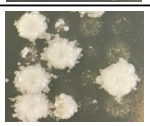
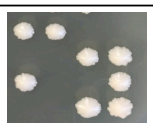
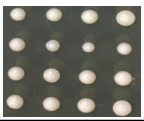
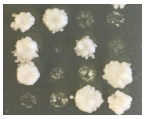
Supplemental Figure 3.3 Confirmed tetrad analysis for conserved, predicted non-essential genes

S. uvarum Specific Essential Genes:

Gene	*S. uvarum* Tetrads YPD	*S. cerevisiae* Tetrads YPD	*S. cerevisiae* Tetrads G418
*kap104*Δ			
*ssq1*Δ			
*tup1*Δ			
*aro7*Δ			
*mdm10*Δ			
*sac3*Δ			
*ccm1*Δ			
*vma5*Δ			
*aft1*Δ			

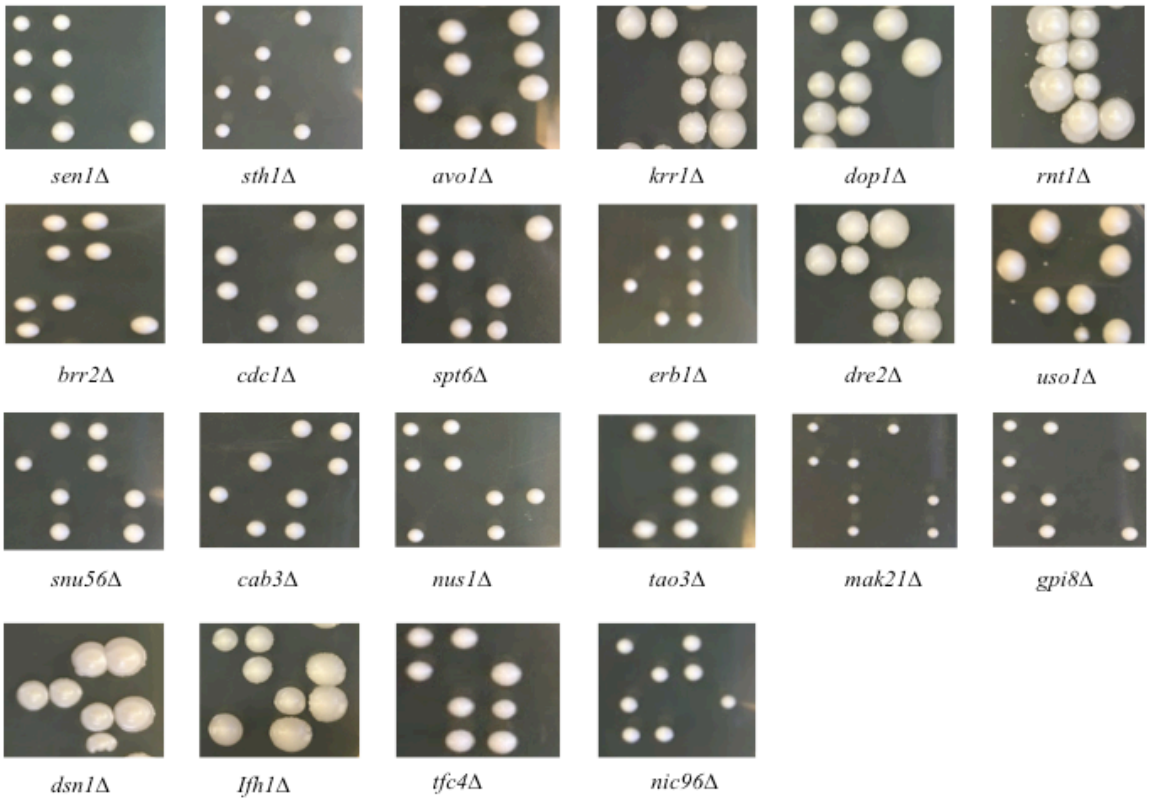
Supplemental Figure 3.4 Confirmed tetrad analysis for predicted *S. uvarum*-specific essential genes

S. cerevisiae Specific Essential Genes:

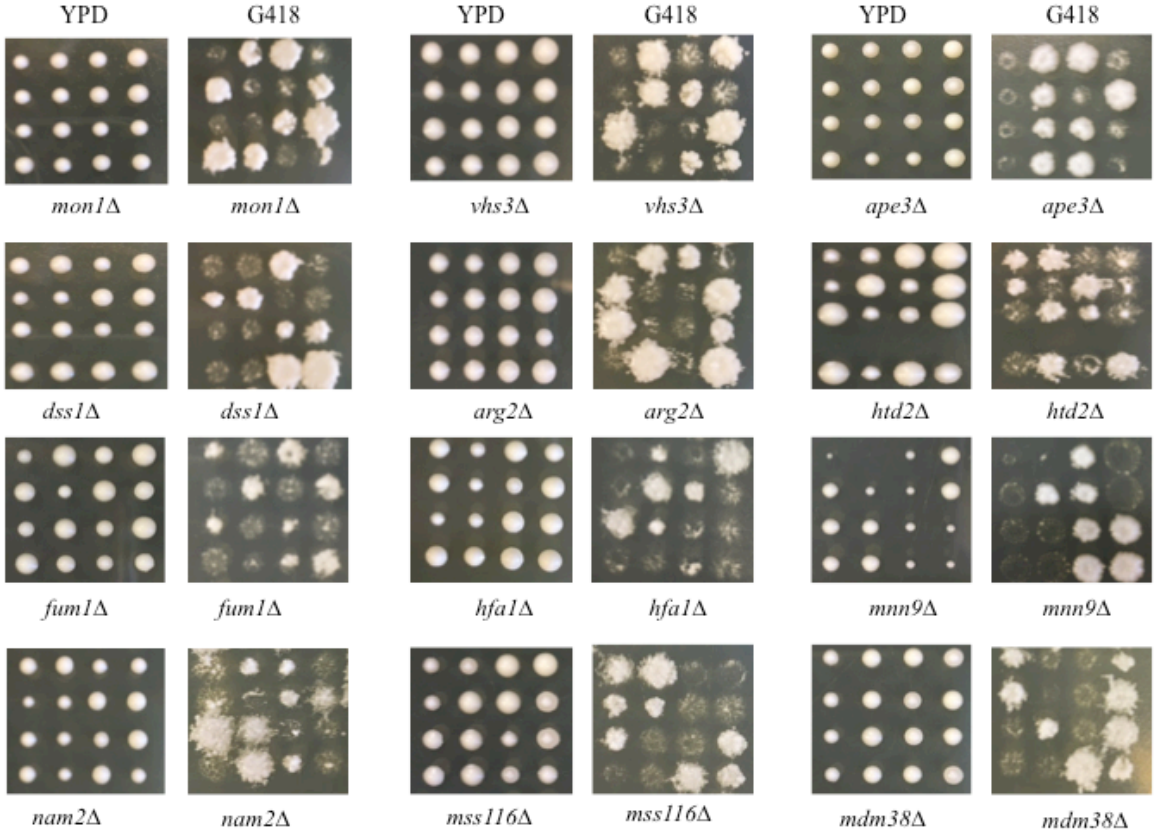
Gene	*S. cerevisiae* Tetrads YPD	*S. uvarum* Tetrads YPD	*S. uvarum* Tetrads G418
cdc25Δ			
sec24Δ			
mcm10Δ			
alr1Δ ch15			
alr1Δ ch7	NA		
inn1Δ			
vtc4Δ			
shr3Δ			
tfc3Δ			
myo2Δ			

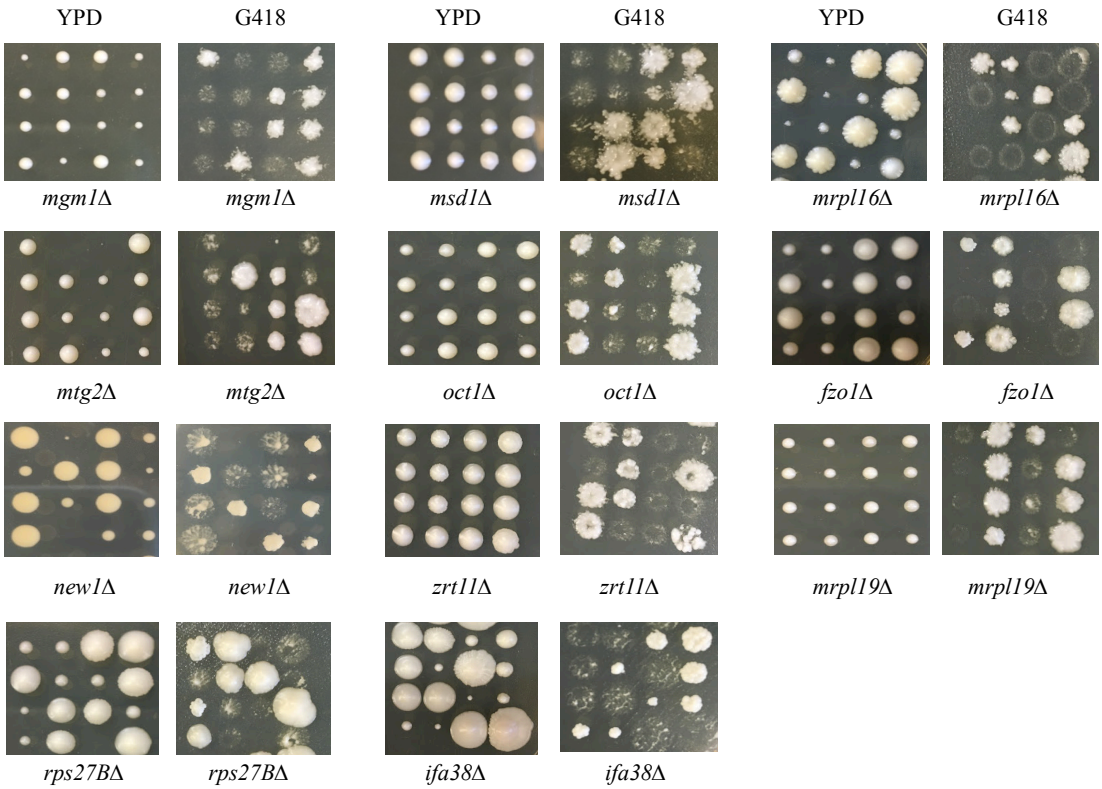
Supplemental Figure 3.5 Confirmed tetrad analysis for predicted *S. cerevisiae*-specific essential genes

Conserved Essential Genes: ScE:SuE

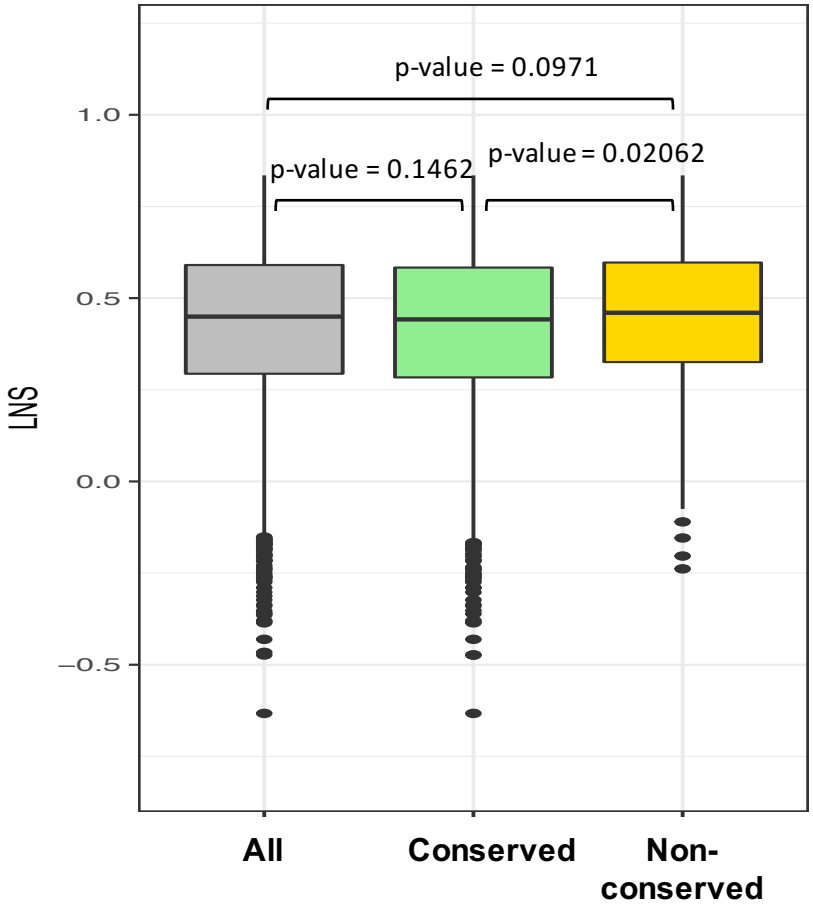


Conserved Nonessential Genes: ScNE:SuNE

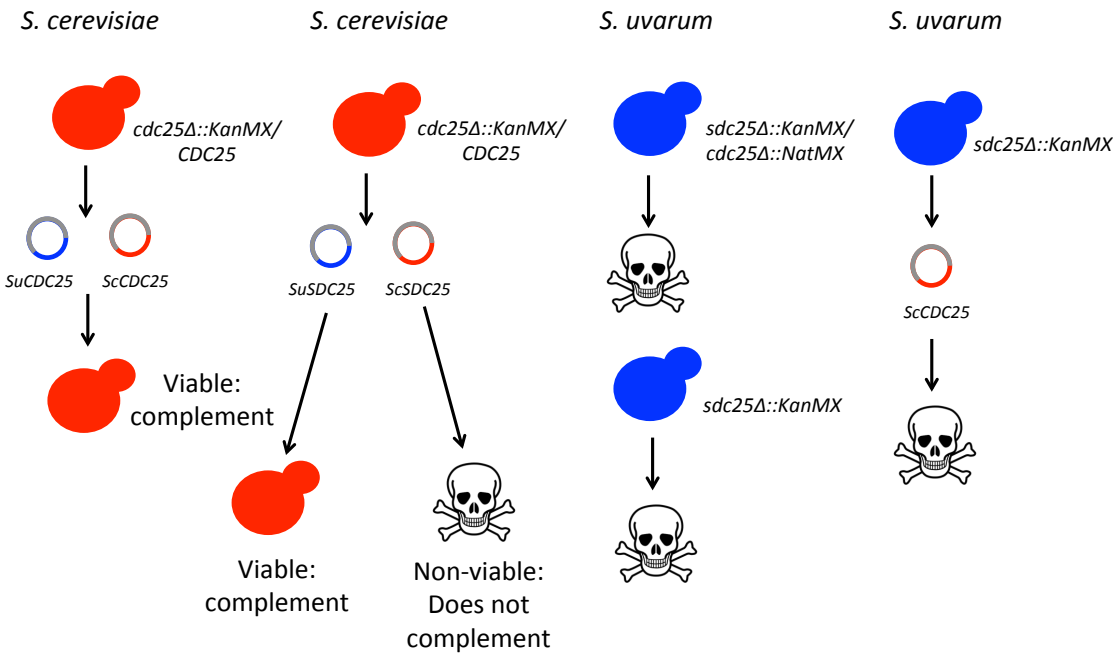




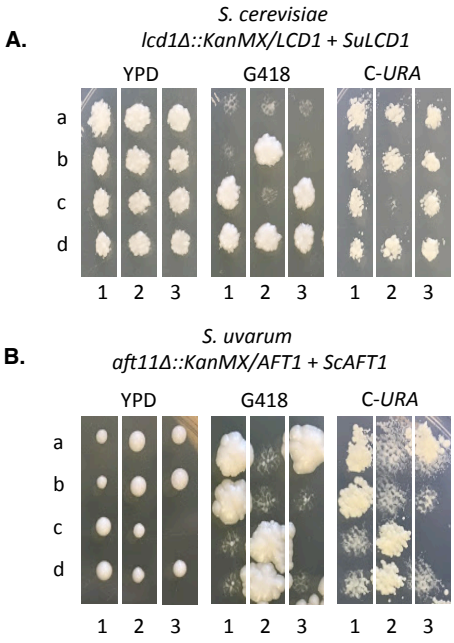
Supplemental Figure 3.6 Confirmed tetrad analysis for all confirmed *S. uvarum* conserved essential and non-essential genes



Supplemental Figure 3.7 Differential gene expression alone cannot explain difference in gene dispensability



Supplemental Figure 3.8 Schematic summarizing complementation assays and double mutant tetrad analysis



Supplemental Figure 3.9 Complementation assay confirming two examples of genes that differ in essentiality but complement the viability phenotype in both genetic backgrounds

APPENDIX C: SEQUENCING ANALSIS FOR A COLLABORATIVE PROJECT ON LOW-TEMPERATURE FERMENTATION

In addition to the sequencing analysis I described in this thesis, I also analyzed sequencing data for a collaborative project titled: "Evolutionary engineering of a wine yeast strain revealed a key role of inositol and mannoprotein metabolism during low-temperature fermentation" published in BMC Genomics in 2015 (López-Malo et al., 2015).

APPENDIX D: COMPETITION EXPERIMENTS FOR A COLLABORATIVE PROJECT ON THE EFFECTS OF CIS-REGULATORY MUTATIONS IN THE *SUL1* GENE

To test the correlation between fitness measurements between variants determined in a pooled approach on plasmids, I tested the fitness of individual, integrated *SUL1* promoter variants in sulfate-limited conditions using chemostats. The values were highly correlated and published as a supplemental figure in a study titled: "Comprehensive analysis of the *SUL1* promoter of *Saccharomyces cerevisiae*" published in Genetics in 2016 (Rich et al., 2016).

REFERENCE

Adams, J., and Rosenzweig, F. (2014). Experimental microbial evolution: history and conceptual underpinnings. Genomics *104*, 393–398.

Adams, M.D., and Sekelsky, J.J. (2002). From sequence to phenotype: reverse genetics in drosophila melanogaster. Nat. Rev. Genet. *3*, 189–198.

Alföldi, J., and Lindblad-Toh, K. (2013). Comparative genomics as a tool to understand evolution and disease. Genome Res. *23*, 1063–1068.

Araya, C.L., Payen, C., Dunham, M.J., and Fields, S. (2010). Whole-genome sequencing of a laboratory-evolved yeast strain. BMC Genomics *11*, 88.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. *2*, 2006.0008.

Bachmann, A., and Knust, E. (2008). The use of P-element transposons to generate transgenic flies. Methods Mol. Biol. Clifton NJ *420*, 61–77.

Barrick, J.E., and Lenski, R.E. (2013). Genome dynamics during experimental evolution. Nat. Rev. Genet. *14*, 827–839.

Barrick, J.E., Yu, D.S., Yoon, S.H., Jeong, H., Oh, T.K., Schneider, D., Lenski, R.E., and Kim, J.F. (2009). Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature *461*, 1243–1247.

Beadle, G.W., and Tatum, E.L. (1941). Genetic Control of Biochemical Reactions in Neurospora. Proc. Natl. Acad. Sci. *27*, 499–506.

Bellí, G., Garí, E., Piedrafita, L., Aldea, M., and Herrero, E. (1998). An activator/repressor dual system allows tight tetracycline-regulated gene expression in budding yeast. Nucleic Acids Res. *26*, 942–947.

Bensasson, D., Zarowiecki, M., Burt, A., and Koufopanou, V. (2008). Rapid Evolution of Yeast Centromeres in the Absence of Drive. Genetics *178*, 2161–2167.

Berardinis, V. de, Vallenet, D., Castelli, V., Besnard, M., Pinet, A., Cruaud, C., Samair, S., Lechaplais, C., Gyapay, G., Richez, C., et al. (2008). A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. Mol. Syst. Biol. *4*, 174.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Breen, M.S., Kemena, C., Vlasov, P.K., Notredame, C., and Kondrashov, F.A. (2012). Epistasis as the primary factor in molecular evolution. Nature *490*, 535–538.

Brewer, B.J., Payen, C., Raghuraman, M.K., and Dunham, M.J. (2011). Origin-Dependent Inverted-Repeat Amplification: A Replication-Based Model for Generating Palindromic Amplicons. PLoS Genet. *7*, e1002016.

Brewer, B.J., Payen, C., Rienzi, S.C.D., Higgins, M.M., Ong, G., Dunham, M.J., and Raghuraman, M.K. (2015). Origin-Dependent Inverted-Repeat Amplification: Tests of a Model for Inverted DNA Amplification. PLOS Genet *11*, e1005699.

Brown, C.J., Todd, K.M., and Rosenzweig, R.F. (1998). Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. Mol. Biol. Evol. *15*, 931–942.

Byrne, K.P., and Wolfe, K.H. (2005). The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. *15*, 1456–1461.

Carroll, S.B. (2005). Evolution at Two Levels: On Genes and Form. PLOS Biol. *3*, e245.

Caudy, A.A., Guan, Y., Jia, Y., Hansen, C., DeSevo, C., Hayes, A.P., Agee, J., Alvarez-Dominguez, J.R., Arellano, H., Barrett, D., et al. (2013). A New System for Comparative Functional Genomics of Saccharomyces Yeasts. Genetics *195*, 275–287.

Chang, S.-L., and Leu, J.-Y. (2011). A Tradeoff Drives the Evolution of Reduced Metal Resistance in Natural Populations of Yeast. PLOS Genet *7*, e1002034.

Chen, Z., Cheng, C.-H.C., Zhang, J., Cao, L., Chen, L., Zhou, L., Jin, Y., Ye, H., Deng, C., Dai, Z., et al. (2008). Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. Proc. Natl. Acad. Sci. *105*, 12944–12949.

Chénais, B., Caruso, A., Hiard, S., and Casse, N. (2012). The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. Gene *509*, 7–15.

Cherest, H., Davidian, J.C., Thomas, D., Benes, V., Ansorge, W., and Surdin-Kerjan, Y. (1997). Molecular characterization of two high affinity sulfate transporters in Saccharomyces cerevisiae. Genetics *145*, 627–635.

Christ, D., and Chin, J.W. (2008). Engineering Escherichia coli heat-resistance by synthetic gene amplification. Protein Eng. Des. Sel. *21*, 121–125.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. (2003). Finding Functional Features in Saccharomyces Genomes by Phylogenetic Footprinting. Science *301*, 71–76.

Cliften, P.F., Hillier, L.W., Fulton, L., Graves, T., Miner, T., Gish, W.R., Waterston, R.H., and Johnston, M. (2001). Surveying Saccharomyces Genomes to Identify Functional Elements by Comparative DNA Sequence Analysis. Genome Res. *11*, 1175–1186.

Coelho, P.S., Kumar, A., and Snyder, M. (2000). Genome-wide mutant collections: toolboxes for functional genomics. Curr. Opin. Microbiol. *3*, 309–315.

Colegrave, N., and Collins, S. (2008). Experimental evolution: experimental evolution and evolvability. Heredity *100*, 464–470.

Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: How duplicated genes find new functions. Nat. Rev. Genet. *9*, 938–950.

Costanzo, M. (2016). Systems biology: A yeast global genetic interaction map. Nat. Methods *13*, 904–904.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., et al. (2010). The Genetic Landscape of a Cell. Science *327*, 425–431.

Csink, A.K., and Henikoff, S. (1998). Something from nothing: the evolution and utility of satellite repeats. Trends Genet. TIG *14*, 200–204.

DeBolt, S. (2010). Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales. Genome Biol. Evol. *2*, 441–453.

Demeke, M.M., Foulquié-Moreno, M.R., Dumortier, F., and Thevelein, J.M. (2015). Rapid Evolution of Recombinant Saccharomyces cerevisiae for Xylose Fermentation through Formation of Extra-chromosomal Circular DNA. PLoS Genet. *11*.

DeNicola, G.M., Karreth, F.A., Adams, D.J., and Wong, C.C. (2015). The utility of transposon mutagenesis for cancer studies in the era of genome editing. Genome Biol. *16*.

Dettman, J.R., Rodrigue, N., Melnyk, A.H., Wong, A., Bailey, S.F., and Kassen, R. (2012). Evolutionary insight from whole-genome sequencing of experimentally evolved microbes. Mol. Ecol. *21*, 2058–2077.

Donalies, U.E.B., and Stahl, U. (2002). Increasing sulphite formation in Saccharomyces cerevisiae by overexpression of MET14 and SSU1. Yeast *19*, 475–484.

Donato, J.J., Chung, S.C.C., and Tye, B.K. (2006). Genome-Wide Hierarchy of Replication Origin Usage in Saccharomyces cerevisiae. PLOS Genet. *2*, e141.

Dowell, R.D., Ryan, O., Jansen, A., Cheung, D., Agarwala, S., Danford, T., Bernstein, D.A., Rolfe, P.A., Heisler, L.E., Chin, B., et al. (2010). Genotype to phenotype: a complex problem. Science *328*, 469.

Dujon, B. (2010). Yeast evolutionary genomics. Nat. Rev. Genet. *11*, 512–524.

Dunham, M.J., Badrane, H., Ferea, T., Adams, J., Brown, P.O., Rosenzweig, F., and Botstein, D. (2002). Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S. A. *99*, 16144–16149.

Dunn, B., and Sherlock, G. (2008). Reconstruction of the genome origins and evolution of the hybrid lager yeast Saccharomyces pastorianus. Genome Res. *18*, 1610–1623.

Duraisingh, M.T., and Cowman, A.F. (2005). Contribution of the pfmdr1 gene to antimalarial drug-resistance. Acta Trop. *94*, 181–190.

Dykhuizen, D.E., and Hartl, D.L. (1983). Selection in chemostats. Microbiol. Rev. *47*, 150–168.

Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.

Eisen, J.A. (1998). Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. Genome Res. *8*, 163–167.

Ellegren, H. (2008). Comparative genomics and the study of evolution by natural selection. Mol. Ecol. *17*, 4586–4596.

Engle, E.K., and Fay, J.C. (2012). Divergence of the Yeast Transcription Factor FZF1 Affects Sulfite Resistance. PLOS Genet *8*, e1002763.

Felsenstein, J. (1985). Phylogenies and the Comparative Method. Am. Nat. *125*, 1–15.

Fischer, G., Neuvéglise, C., Durrens, P., Gaillardin, C., and Dujon, B. (2001). Evolution of Gene Order in the Genomes of Two Related Yeast Species. Genome Res. *11*, 2009–2019.

Fitzgerald-Hayes, M., Clarke, L., and Carbon, J. (1982). Nucleotide sequence comparisons and functional analysis of yeast centromere DNAs. Cell *29*, 235–244.

Fleig, U., Beinhauer, J.D., and Hegemann, J.H. (1995). Functional selection for the centromere DNA from yeast chromosome VIII. Nucleic Acids Res. *23*, 922–924.

Foote, S.J., Thompson, J.K., Cowman, A.F., and Kemp, D.J. (1989). Amplification of the multidrug resistance gene in some chloroquine-resistant isolates of P. falciparum. Cell *57*, 921–930.

Forsburg, S.L. (2001). The art and design of genetic screens: yeast. Nat. Rev. Genet. *2*, 659–668.

Fraser, H.B., Moses, A.M., and Schadt, E.E. (2010). Evidence for widespread adaptive evolution of gene expression in budding yeast. Proc. Natl. Acad. Sci. *107*, 2977–2982.

Gagnon-Arsenault, I., Marois Blanchet, F.-C., Rochette, S., Diss, G., Dubé, A.K., and Landry, C.R. (2013). Transcriptional divergence plays a role in the rewiring of protein interaction networks after gene duplication. J. Proteomics *81*, 112–125.

Gangadharan, S., Mularoni, L., Fain-Thornton, J., Wheelan, S.J., and Craig, N.L. (2010). DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo. Proc. Natl. Acad. Sci. *107*, 21966–21972.

Garland, T., Midford, P.E., and Ives, A.R. (1999). An Introduction to Phylogenetically Based Statistical Methods, with a New Method for Confidence Intervals on Ancestral Values. Integr. Comp. Biol. *39*, 374–388.

Gout, J.-F., and Lynch, M. (2015). Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization. Mol. Biol. Evol. msv095.

Gresham, D., Desai, M.M., Tucker, C.M., Jenq, H.T., Pai, D.A., Ward, A., DeSevo, C.G., Botstein, D., and Dunham, M.J. (2008). The Repertoire and Dynamics of Evolutionary Adaptations to Controlled Nutrient-Limited Environments in Yeast. PLoS Genet *4*, e1000303.

Gu, Z., Steinmetz, L.M., Gu, X., Scharfe, C., Davis, R.W., and Li, W.-H. (2003). Role of duplicate genes in genetic robustness against null mutations. Nature *421*, 63–66.

Guan, Y., Dunham, M.J., and Troyanskaya, O.G. (2007). Functional Analysis of Gene Duplications in Saccharomyces cerevisiae. Genetics *175*, 933–943.

Guan, Y., Dunham, M., Caudy, A., and Troyanskaya, O. (2010). Systematic Planning of Genome-Scale Experiments in Poorly Studied Species. PLOS Comput. Biol. *6*, e1000698.

Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D.C., and Shyr, Y. (2013). Large Scale Comparison of Gene Expression Levels by Microarrays and RNAseq Using TCGA Data. PLOS ONE *8*, e71462.

Haaf, T., and Willard, H.F. (1997). Chromosome-specific alpha-satellite DNA from the centromere of chimpanzee chromosome 4. Chromosoma *106*, 226–232.

Hardy, S., Legagneux, V., Audic, Y., and Paillard, L. (2010). Reverse genetics in eukaryotes. Biol. Cell *102*, 561–580.

Hartman, J.L., Garvik, B., and Hartwell, L. (2001). Principles for the Buffering of Genetic Variation. Science *291*, 1001–1004.

Heil, C.S.S., DeSevo, C.G., Pai, D.A., Tucker, C.M., Hoang, M.L., and Dunham, M.J. (2016). Selection on heterozygosity drives adaptation in intra- and interspecific hybrids. bioRxiv 73007.

Henikoff, S., Ahmad, K., and Malik, H.S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. Science *293*, 1098–1102.

Hensel, M., Shea, J.E., Gleeson, C., Jones, M.D., Dalton, E., and Holden, D.W. (1995). Simultaneous identification of bacterial virulence genes by negative selection. Science *269*, 400–403.

Hittinger, C.T., and Carroll, S.B. (2007). Gene duplication and the adaptive evolution of a classic genetic switch. Nature *449*, 677–681.

Ho, C.H., Magtanong, L., Barker, S.L., Gresham, D., Nishimura, S., Natarajan, P., Koh, J.L.Y., Porter, J., Gray, C.A., Andersen, R.J., et al. (2009). A molecular barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds. Nat. Biotechnol. *27*.

Hoffman, C.S., and Winston, F. (1987). A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of Escherichia coli. Gene *57*, 267–272.

Hoffmann, R.D., and Palmgren, M. (2016). Purifying selection acts on coding and non-coding sequences of paralogous genes in Arabidopsis thaliana. BMC Genomics *17*, 456.

Hsiao, C.L., and Carbon, J. (1981). Characterization of a yeast replication origin (ars2) and construction of stable minichromosomes containing cloned yeast centromere DNA (CEN3). Gene *15*, 157–166.

Hughes, A.L. (1994). The Evolution of Functionally Novel Proteins after Gene Duplication. Proc. R. Soc. Lond. B Biol. Sci. *256*, 119–124.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., et al. (2000). Functional Discovery via a Compendium of Expression Profiles. Cell *102*, 109–126.

Kawecki, T.J., Lenski, R.E., Ebert, D., Hollis, B., Olivieri, I., and Whitlock, M.C. (2012). Experimental evolution. Trends Ecol. Evol. *27*, 547–560.

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

Khan, A.I., Dinh, D.M., Schneider, D., Lenski, R.E., and Cooper, T.F. (2011). Negative Epistasis Between Beneficial Mutations in an Evolving Bacterial Population. Science *332*, 1193–1196.

Kieliszek, M., Błażejak, S., Gientka, I., and Bzducha-Wróbel, A. (2015). Accumulation and metabolism of selenium by yeast cells. Appl. Microbiol. Biotechnol. *99*, 5373–5382.

Kim, D.-U., Hayles, J., Kim, D., Wood, V., Park, H.-O., Won, M., Yoo, H.-S., Duhig, T., Nam, M., Palmer, G., et al. (2010a). Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nat. Biotechnol. *28*, 617–623.

Kim, D.-U., Hayles, J., Kim, D., Wood, V., Park, H.-O., Won, M., Yoo, H.-S., Duhig, T., Nam, M., Palmer, G., et al. (2010b). Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nat. Biotechnol. *28*, 617–623.

Kumar, A., Seringhaus, M., Biery, M.C., Sarnovsky, R.J., Umansky, L., Piccirillo, S., Heidtman, M., Cheung, K.-H., Dobry, C.J., Gerstein, M.B., et al. (2004). Large-Scale Mutagenesis of the Yeast Genome Using a Tn7-Derived Multipurpose Transposon. Genome Res. *14*, 1975–1986.

Kvitek, D.J., Will, J.L., and Gasch, A.P. (2008). Variations in Stress Sensitivity and Genomic Expression in Diverse S. cerevisiae Isolates. PLOS Genet *4*, e1000223.

Langridge, G.C., Phan, M.-D., Turner, D.J., Perkins, T.T., Parts, L., Haase, J., Charles, I., Maskell, D.J., Peters, S.E., Dougan, G., et al. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res. *19*, 2308–2316.

Lau, G.W., Haataja, S., Lonetto, M., Kensit, S.E., Marra, A., Bryant, A.P., McDevitt, D., Morrison, D.A., and Holden, D.W. (2001). A functional genomic analysis of type 3 Streptococcus pneumoniae virulence. Mol. Microbiol. *40*, 555–571.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

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

Libkind, D., Hittinger, C.T., Valério, E., Gonçalves, C., Dover, J., Johnston, M., Gonçalves, P., and Sampaio, J.P. (2011). Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. Proc. Natl. Acad. Sci. *108*, 14539–14544.

Lieberman, T.D., Michel, J.-B., Aingaran, M., Potter-Bynoe, G., Roux, D., Jr, M.R.D., Skurnik, D., Leiby, N., LiPuma, J.J., Goldberg, J.B., et al. (2011). Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat. Genet. *43*, 1275–1280.

Liti, G., Peruffo, A., James, S.A., Roberts, I.N., and Louis, E.J. (2005). Inferences of evolutionary relationships from a population survey of LTR-retrotransposons and telomeric-associated sequences in the Saccharomyces sensu stricto complex. Yeast Chichester Engl. *22*, 177–192.

Liti, G., Haricharan, S., Cubillos, F.A., Tierney, A.L., Sharp, S., Bertuch, A.A., Parts, L., Bailes, E., and Louis, E.J. (2009). Segregating YKU80 and TLC1 Alleles Underlying Natural Variation in Telomere Properties in Wild Yeast. PLOS Genet. *5*, e1000659.

López-Igual, R., Wilson, A., Leverenz, R.L., Melnicki, M.R., Bourcier de Carbon, C., Sutter, M., Turmo, A., Perreau, F., Kerfeld, C.A., and Kirilovsky, D. (2016). Different Functions of the Paralogs to the N-Terminal Domain of the Orange Carotenoid Protein in the Cyanobacterium Anabaena sp. PCC 71201[OPEN]. Plant Physiol. *171*, 1852–1866.

López-Malo, M., García-Rios, E., Melgar, B., Sanchez, M.R., Dunham, M.J., and Guillamón, J.M. (2015). Evolutionary engineering of a wine yeast strain revealed a key role of inositol and mannoprotein metabolism during low-temperature fermentation. BMC Genomics *16*, 537.

Lynch, M., and Force, A. (2000). The Probability of Duplicate Gene Preservation by Subfunctionalization. Genetics *154*, 459–473.

Malik, H.S., and Henikoff, S. (2009). Major Evolutionary Transitions in Centromere Complexity. Cell *138*, 1067–1082.

Marcet-Houben, M., and Gabaldón, T. (2015). Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. PLOS Biol. *13*, e1002220.

Martin, O.C., DeSevo, C.G., Guo, B.Z., Koshland, D.E., Dunham, M.J., and Zheng, Y. (2009). Telomere behavior in a hybrid yeast. Cell Res. *19*, 910–912.

McClintock, B. (1950). The Origin and Behavior of Mutable Loci in Maize. Proc. Natl. Acad. Sci. U. S. A. *36*, 344–355.

Mei, J.-M., Nourbakhsh, F., Ford, C.W., and Holden, D.W. (1997). Identification of Staphylococcus aureus virulence genes in a murine model of bacteraemia using signature-tagged mutagenesis. Mol. Microbiol. *26*, 399–407.

Michel, A.H., Hatakeyama, R., Kimmig, P., Arter, M., Peter, M., Matos, J., Virgilio, C.D., and Kornmann, B. (2017a). Functional mapping of yeast genomes by saturated transposition. eLife *6*, e23570.

Michel, A.H., Hatakeyama, R., Kimmig, P., Arter, M., Peter, M., Matos, J., Virgilio, C.D., and Kornmann, B. (2017b). Functional mapping of yeast genomes by saturated transposition. eLife *6*, e23570.

Miller, A.W., Befort, C., Kerr, E.O., and Dunham, M.J. (2013). Design and Use of Multiplexed Chemostat Arrays. J. Vis. Exp.

Miller, W., Makova, K.D., Nekrutenko, A., and Hardison, R.C. (2004). Comparative Genomics. Annu. Rev. Genomics Hum. Genet. *5*, 15–56.

Müller, C.A., and Nieduszynski, C.A. (2012). Conservation of replication timing reveals global and local regulation of replication origin activity. Genome Res. *22*, 1953–1962.

Murphy, T.D., and Karpen, G.H. (1998). Centromeres take flight: alpha satellite and the quest for the human centromere. Cell *93*, 317–320.

Nakao, Y., Kanamori, T., Itoh, T., Kodama, Y., Rainieri, S., Nakamura, N., Shimonaga, T., Hattori, M., and Ashikari, T. (2009b). Genome Sequence of the Lager Brewing Yeast, an Interspecies Hybrid. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes *16*, 115–129.

Nakao, Y., Kodama, Y., and Shimonaga, T. (2009a). Sulfate Ion Transporter Gene and Use Thereof.

Neuvéglise, C., Feldmann, H., Bon, E., Gaillardin, C., and Casaregola, and S. (2002). Genomic Evolution of the Long Terminal Repeat Retrotransposons in Hemiascomycetous Yeasts. Genome Res. *12*, 930–943.

Nieduszynski, C.A., Knox, Y., and Donaldson, A.D. (2006). Genome-wide identification of replication origins in yeast by comparative genomics. Genes Dev. *20*, 1874–1879.

Noble, S.M., French, S., Kohn, L.A., Chen, V., and Johnson, A.D. (2010). Systematic screens of a Candida albicans homozygous deletion library decouple morphogenetic switching and pathogenicity. Nat. Genet. *42*, 590–598.

Novick, A., and Szilard, L. (1950). Description of the Chemostat. Science *112*, 715–716.

Oh, J., Fung, E., Schlecht, U., Davis, R.W., Giaever, G., St. Onge, R.P., Deutschbauer, A., and Nislow, C. (2010). Gene Annotation and Drug Target Discovery in Candida albicans with a Tagged Transposon Mutant Collection. PLoS Pathog. *6*.

Ohno S. (1970). Evolution by Gene duplication (Berlin: Springer-Verlag: Olson KA).

van Opijnen, T., and Camilli, A. (2013). Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat. Rev. Microbiol. *11*, 435–442.

van Opijnen, T., Bodi, K.L., and Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat. Methods *6*, 767–772.

Ostrowski, E.A., Woods, R.J., and Lenski, R.E. (2008). The genetic basis of parallel and divergent phenotypic responses in evolving populations of Escherichia coli. Proc. R. Soc. Lond. B Biol. Sci. *275*, 277–284.

Paquin, C., and Adams, J. (1983). Frequency of fixation of adaptive mutations is higher in evolving diploid than haploid yeast populations. Nature *302*, 495–500.

Parts, L. (2014). Genome-wide mapping of cellular traits using yeast. Yeast *31*, 197–205.

Payen, C., Di Rienzi, S.C., Ong, G.T., Pogachar, J.L., Sanchez, J.C., Sunshine, A.B., Raghuraman, M.K., Brewer, B.J., and Dunham, M.J. (2013). The Dynamics of Diverse Segmental Amplifications in Populations of Saccharomyces cerevisiae Adapting to Strong Selection. G3 GenesGenomesGenetics *4*, 399–409.

Pedram, M., Sprung, C.N., Gao, Q., Lo, A.W.I., Reynolds, G.E., and Murnane, J.P. (2006). Telomere Position Effect and Silencing of Transgenes near Telomeres in the Mouse. Mol. Cell. Biol. *26*, 1865–1878.

Porwollik, S., Santiviago, C.A., Cheng, P., Long, F., Desai, P., Fredlund, J., Srikumar, S., Silva, C.A., Chu, W., Chen, X., et al. (2014). Defined Single-Gene and Multi-Gene Deletion Mutant Collections in Salmonella enterica sv Typhimurium. PLOS ONE *9*, e99820.

Qian, W., and Zhang, J. (2014). Genomic evidence for adaptation by gene duplication. Genome Res. *24*, 1356–1362.

Reilly, M.T., Faulkner, G.J., Dubnau, J., Ponomarev, I., and Gage, F.H. (2013). The Role of Transposable Elements in Health and Diseases of the Central Nervous System. J. Neurosci. *33*, 17577–17586.

Rich, M.S., Payen, C., Rubin, A.F., Ong, G.T., Sanchez, M.R., Yachie, N., Dunham, M.J., and Fields, S. (2016). Comprehensive Analysis of the SUL1 Promoter of Saccharomyces cerevisiae. Genetics *203*, 191–202.

Rienzi, S.C.D., Lindstrom, K.C., Mann, T., Noble, W.S., Raghuraman, M.K., and Brewer, B.J. (2012). Maintaining replication origins in the face of genomic change. Genome Res. *22*, 1940–1952.

Rokas, A., and Carroll, S.B. (2008). Frequent and Widespread Parallel Evolution of Protein Sequences. Mol. Biol. Evol. *25*, 1943–1953.

de la Rosa, J., Weber, J., Friedrich, M.J., Li, Y., Rad, L., Ponstingl, H., Liang, Q., de Quirós, S.B., Noorani, I., Metzakopian, E., et al. (2017). A single-copy Sleeping Beauty transposon mutagenesis screen identifies new PTEN-cooperating tumor suppressor genes. Nat. Genet. *49*, 730–741.

Ross-Macdonald, P., Coelho, P.S.R., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K.-H., Sheehan, A., Symoniatis, D., Umansky, L., et al. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption. Nature *402*, 413–418.

Ryu, S.-L., Murooka, Y., and Kaneko, Y. Reciprocal translocation at duplicated RPL2 loci might cause speciation of Saccharomyces bayanus and Saccharomyces cerevisiae. Curr. Genet. *33*, 345–351.

Salama, N.R., Shepherd, B., and Falkow, S. (2004). Global Transposon Mutagenesis and Essential Gene Analysis of Helicobacter pylori. J. Bacteriol. *186*, 7926–7935.

Scannell, D.R., Butler, G., and Wolfe, K.H. (2007). Yeast genome evolution—the origin of the species. Yeast *24*, 929–942.

Scannell, D.R., Zill, O.A., Rokas, A., Payen, C., Dunham, M.J., Eisen, M.B., Rine, J., Johnston, M., and Hittinger, C.T. (2011). The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. G3 GenesGenomesGenetics *1*, 11–25.

Schwarzmüller, T., Ma, B., Hiller, E., Istel, F., Tscherner, M., Brunke, S., Ames, L., Firon, A., Green, B., Cabral, V., et al. (2014). Systematic Phenotyping of a Large-Scale Candida glabrata Deletion Collection Reveals Novel Antifungal Tolerance Genes. PLOS Pathog. *10*, e1004211.

Selmecki, A.M., Maruvka, Y.E., Richmond, P.A., Guillet, M., Shoresh, N., Sorenson, A.L., De, S., Kishony, R., Michor, F., Dowell, R., et al. (2015). Polyploidy can drive rapid adaptation in yeast. Nature *519*, 349–352.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., et al. (2014). Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. Science *343*, 84–87.

Stapley, J., Reger, J., Feulner, P.G.D., Smadja, C., Galindo, J., Ekblom, R., Bennison, C., Ball, A.D., Beckerman, A.P., and Slate, J. (2010). Adaptation genomics: the next generation. Trends Ecol. Evol. *25*, 705–712.

Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science *302*, 249–255.

Teichmann, S.A., and Babu, M.M. (2004). Gene regulatory network growth by duplication. Nat. Genet. *36*, 492–496.

Thomas, S., Maynard, N.D., and Gill, J. (2015). DNA library construction using Gibson Assembly®. Nat. Methods *12*.

Thorsen, M., Lagniel, G., Kristiansson, E., Junot, C., Nerman, O., Labarre, J., and Tamás, M.J. (2007). Quantitative transcriptome, proteome, and sulfur metabolite profiling of the Saccharomyces cerevisiae response to arsenite. Physiol. Genomics *30*, 35–43.

Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Pagé, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H., et al. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science *294*, 2364–2368.

Triglia, T., Foote, S.J., Kemp, D.J., and Cowman, A.F. (1991). Amplification of the multidrug resistance gene pfmdr1 in Plasmodium falciparum has arisen as multiple independent events. Mol. Cell. Biol. *11*, 5244–5250.

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F.M., Bassel, G.W., Tanimoto, M., Chow, A., Steinhauser, D., Persson, S., and Provart, N.J. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ. *32*, 1633–1651.

VanderSluis, B., Bellay, J., Musso, G., Costanzo, M., Papp, B., Vizeacoumar, F.J., Baryshnikova, A., Andrews, B., Boone, C., and Myers, C.L. (2010). Genetic interactions reveal the evolutionary trajectories of duplicate genes. Mol. Syst. Biol. *6*, 429.

Varshney, G.K., Lu, J., Gildea, D.E., Huang, H., Pei, W., Yang, Z., Huang, S.C., Schoenfeld, D., Pho, N.H., Casero, D., et al. (2013). A large-scale zebrafish gene knockout resource for the genome-wide study of gene function. Genome Res. *23*, 727–735.

Verdaasdonk, J.S., and Bloom, K. (2011). Centromeres: unique chromatin structures that drive chromosome segregation. Nat. Rev. Mol. Cell Biol. *12*, 320–332.

Vogwill, T., Kojadinovic, M., Furió, V., and MacLean, R.C. (2014). Testing the Role of Genetic Background in Parallel Evolution Using the Comparative Experimental Evolution of Antibiotic Resistance. Mol. Biol. Evol. msu262.

Voordeckers, K., Brown, C.A., Vanneste, K., van der Zande, E., Voet, A., Maere, S., and Verstrepen, K.J. (2012). Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication. PLoS Biol. *10*, e1001446.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., et al. (2008). The diploid genome sequence of an Asian individual. Nature *456*, 60–65.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi. Nature *449*, 54–61.

Weerdenburg, E.M., Abdallah, A.M., Rangkuti, F., Ghany, M.A.E., Otto, T.D., Adroub, S.A., Molenaar, D., Ummels, R., Veen, K. ter, Stempvoort, G. van, et al. (2015). Genome-Wide Transposon Mutagenesis Indicates that Mycobacterium marinum Customizes Its Virulence Mechanisms for Survival and Replication in Different Hosts. Infect. Immun. *83*, 1778–1788.

Wetmore, K.M., Price, M.N., Waters, R.J., Lamson, J.S., He, J., Hoover, C.A., Blow, M.J., Bristow, J., Butland, G., Arkin, A.P., et al. (2015). Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons. mBio *6*, e00306-15.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science *285*, 901–906.

Wood, T.E., Burke, J.M., and Rieseberg, L.H. (2005). Parallel genotypic adaptation: when evolution repeats itself. Genetica *123*, 157–170.

Woods, R., Schneider, D., Winkworth, C.L., Riley, M.A., and Lenski, R.E. (2006). Tests of parallel molecular evolution in a long-term experiment with Escherichia coli. Proc. Natl. Acad. Sci. *103*, 9107–9112.

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. *8*, 206–216.

Yung, M.C., Park, D.M., Overton, K.W., Blow, M.J., Hoover, C.A., Smit, J., Murray, S.R., Ricci, D.P., Christen, B., Bowman, G.R., et al. (2015). Transposon Mutagenesis Paired with Deep Sequencing of Caulobacter crescentus under Uranium Stress Reveals Genes Essential for Detoxification and Stress Tolerance. J. Bacteriol. *197*, 3160–3172.

VITA

Monica (Mascareñas) Sanchez was born and raised in Albuquerque, New Mexico where she attended the University of New Mexico and attained a B.S. in Biology. While attending UNM, Monica worked in the lab of Gabriel Lopez, Jeremy Edwards and Maggie Werner-Washburne studying sequencing technology development and yeast genetics. Monica spent two summers away at undergraduate research programs at Harvard University and Genome Sciences where she worked with George Whitesides and Jay Shendure. After graduating from UNM in 2011, Monica entered the Molecular and Cellular PhD program at the University of Washington and joined the Department of Genome Sciences. Monica enjoys running, attending concerts and spending time with family, especially her husband Joe.