Species-scale high-throughput functional analysis of natural variants in yeast

Chiann-Ling Yeh

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Maitreya J. Dunham, Chair

Celeste A. Berg

Stanley Fields

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Species-scale high-throughput functional analysis of natural variants in yeast

Chiann-Ling Yeh

Chair of the Supervisory Committee:

Maitreya J. Dunham

Department of Genome Sciences

The impact of natural genetic variation on phenotype is difficult to measure because we only partially understand how polymorphisms present in a population affect gene function. Understanding the relationship between genetic variation and phenotype has important implications for human therapeutics, but on a broader level is crucial for predicting evolutionary outcomes and disentangling adaptations that occurred in the past. In this work, I describe a high-throughput, cost-effective approach for assaying natural allelic variation on a species-wide level in the budding yeast, *Saccharomyces cerevisiae*. In the first chapter, I describe the many aspects of quantitative traits and current approaches used to understand natural allelic variation on a high-throughput level. I highlight approaches that have been developed particularly in yeast, as

this model system continues to be an amazing tool for genetics and genomics. Following with the second chapter, I describe a high-throughput functional assay that I developed that can measure the fitness of all natural alleles of a gene, in this case the high-affinity sulfate transporter *SUL1,* at the population level. I show that this approach can categorize alleles into functional, intermediate, or nonfunctional groups, and tying these results to ecological origins reveals patterns of the evolutionary history of *SUL1* in *S. cerevisiae*. In chapter three, I elaborate on a computational approach called PacRAT, a PacBio long-read sequencing algorithm with novel error-correcting properties, that improves the accuracy of barcode-allele pairs. I verified the success of this approach using simulated libraries and show that the method maximizes the number of reads that can be utilized from each PacBio SMRT cell, especially as gene length increases. Success of the aforementioned assay combined with this computational approach highlights the numerous other questions we can answer about natural variation and evolution. In the last part of this work (Chapter 4), I show an example of this approach to study phenotypic differences in paralogs in the maltose utilization pathway in *S. cerevisiae* and show that further examination can reveal more about paralog functional divergence and how strains have adapted to maltose-rich environments. The results here will also help deconvolute the genetic basis of adaptation to domesticated environments, as maltose-utilizing strains are typically isolated from beer samples. The final chapter concludes my dissertation where I summarize my work and discuss potential questions that can be further answered with these results. All in all, my work on natural allelic variation improves our understanding of how genotypes affect phenotypes and informs our understanding of how selection gave rise to the existing polymorphisms that affect populations today.

TABLE OF CONTENTS

List of Figures v

List of Tables vii

Chapter 1. Introduction 1

1.1 Scaling approaches for identifying genes underlying quantitative traits in yeast 2

1.1.1 The significance of high-throughput functional approaches 3

1.1.2 Common approaches for identifying quantitative trait loci 4

1.1.3 High-throughput genome-wide identification of genes contributing to trait differences 6

1.1.4 High-throughput interrogation of variant effects at a causal locus 9

1.1.5 Future directions for high-throughput functional approaches for identifying QTLs 12

1.2 Technologies used for linking barcodes to variants in pooled libraries 14

1.3 Curation of *S. cerevisiae* natural isolates for understanding genetic diversity 17

1.4 Role of gene duplications in evolution 18

1.5 Dissertation summary 20

Chapter 2. Developing a high-throughput assay for functional analysis of natural *SUL1* alleles 22

2.1 Background 23

2.2 Results 26

2.2.1 Allele library curation and characterization 26

2.2.2 Fitness distribution across natural SUL1 alleles 31

2.2.3 Effects of promoter mutations in natural SUL1 variants 34

2.2.4 Comparing competitive fitness with mutfunc 37

2.2.5 Phylogenetics and sequence analysis of natural SUL1 alleles 37
2.3 Discussion 42
2.4 Materials and methods 46
2.4.1 Strains and plasmids 46
2.4.2 Plasmid and yeast library generation 46
2.4.3 Linking barcodes with full-length variants 47
2.4.4 Pooled library competition in chemostats 49
2.4.5 Barcode sequencing and analysis 49
2.4.6 Pairwise fitness assays in chemostats 50
2.4.7 Measuring the growth of the 1,011 isolates on solid media 50
2.4.8 Phylogenetic tree generation and sequence analysis 51
2.4.9 Data availability 52
2.5 Acknowledgments 52
Chapter 3. Development of PacRAT: a program to improve barcode-variant mapping from PacBio long reads using multiple sequence alignments 53
3.1 Background 53
3.2 Results 57
3.2.1 Simulation of deep mutational scanning libraries to assess PacRAT performance 57
3.2.2 Assessing PacRAT performance through reanalysis of published libraries 58
3.3 Discussion 60
3.4 Materials and methods 61
3.4.1 PacRAT algorithm and development 61
3.4.2 Simulation deep mutational scanning (DMS) libraries 62

3.4.3 Reanalysis of PTEN/TPMT libraries 63
3.5 Acknowledgements 63
Chapter 4. High-throughput approach for understanding functional differences between paralogous genes in the maltose utilization pathway 64
4.1 Background 64
4.2 Results 70
4.2.1 Validation that a functional MALx3 can rescue growth in maltose 70
4.2.2 Determining the function of MAL33 on a species-wide scale in the chemostat 71
4.2.3 Phylogenetic and sequence analysis of MALx3 alleles 72
4.3 Discussion 79
4.4 Materials and methods 81
4.4.1 Strains, primers, and plasmids 81
4.4.2 Plasmid/yeast library generation and tests in maltose media 82
4.4.3 Pooled library competition in chemostats' 84
4.4.4 Barcode sequencing 85
4.4.5 Phylogenetic tree and sequence analysis 85
4.5 Acknowledgements 86
Chapter 5. Conclusions and future directions 87
5.1 Information gleaned from a high-throughput analysis of natural variants 87
5.2 Application of high-throughput analyses of natural variants to long-standing questions in evolution 91
5.3 Remaining optimizations to improve assay 95

5.4 Other applications and optimizations of PacRAT 98

5.5 Concluding remarks 100

Bibliography 101

Appendix A: Supplementary Figures 128

Appendix B: Supplementary Tables 134

LIST OF FIGURES

Figure 1.1. Traditional QTL mapping procedures (left panel) and recent advancements (right panel)..... 5

Figure 1.2. Summary of yeast library construction methods. 7

Figure 1.3. Summary of high-throughput functional assays in yeast. 10

Figure 2.1. Workflow for assaying natural variants in the 1,011 strain collection. 29

Figure 2.2. Species-level distribution of fitness effects of natural *SUL1* alleles. 30

Figure 2.3. Validation of pooled competition through direct competitions of selected natural *SUL1* alleles. 32

Figure 2.4. Coding polymorphisms are more useful for predicting deleterious effects compared to those in the promoter of *SUL1*. 35

Figure 2.5. Neighbor-joining gene cladogram generated through PHYLIP using unique genotypes of *SUL1* in the 1,011 strain collection. 38

Figure 2.6. Dairy and African beer subtree of the 1,011 *SUL1* genotypes. 40

Figure 3.1. Comparison of correct variant identification between ABP and PacRAT in simulated libraries with different lengths of variable region..... 56

Figure 4.1. The *MAL* locus permits cells to use maltose as the sole carbon source. 66

Figure 4.2. The *MAL* locus has duplicated several times in *S. cerevisiae*. 67

Figure 4.3. Functional differences can exist within a gene and between all of its paralogs.68

Figure 4.4. Paralogs of the *MAL* gene complement each other. 69

Figure 4.5. Saturation densities of cultures grown in flasks with differing maltose concentrations 71

Figure 4.6. Neighbor joining cladogram of unique *MAL33* alleles. 73

Figure 4.7. Polymorphisms across the *MAL33* locus. 74

Figure 4.8. Neighbor joining cladogram of unique *MAL13* alleles. 75

Figure 4.9. Polymorphisms across the *MAL13* locus. 76

Figure 4.10. Polymorphisms across the *MAL73* locus. 77

Figure 4.11. Neighbor joining cladogram of unique *MAL73* alleles. 78

Figure 4.12. Unrooted neighbor-joining tree of *MALx3* paralogs. 80

Supplementary Figure 1. Percentage of strains for each genome-wide ploidy that matched to at least one PacBio read. 128

Supplementary Figure 2. Allele frequencies found in PacBio allele library reflect those found in the Illumina reference sequences (expected values). 128

Supplementary Figure 3. Competitive fitness values calculated using FitSeq are well-correlated across replicates. 129

Supplementary Figure 4. Fitness and variance among loss-of-function alleles..... 130

Supplementary Figure 5. mutfunc determines which mutations are deleterious. 131

Supplementary Figure 6. Growth rates of unmodified isolates on sulfate limited solid plates. 132

Supplementary Figure 7. Barplot of number of genes with premature stop codons per strain, grouped by ecological origins. Variation within an ecological group may be due to geographical factors (for instance, yeast isolated from fruit in South America will have a different genetic makeup than yeast isolated from Southeast Asia). 133

LIST OF TABLES

Table 3.1. Comparison of mean passes through SimLoRD simulated and real libraries. 57

Table 3.2. *CYP2C9* and *MSH2* library statistics. 59

Supplementary Table 1. Primers used in *SUL1* natural allele study..... 134

Supplementary Table 2. Plasmids and strains used in *SUL1* allele study..... 136

Supplementary Table 3. Allele metadata for *SUL1* 139

Supplementary Table 4. One-way ANOVA test of simulated libraries to determine if there is a significant difference when using PacRAT. 158

Supplementary Table 5. Primers used for amplifying and cloning *MALx3* alleles. 159

Supplementary Table 6. Percent identity and similarity between *MALx3* sequences..... 161

ACKNOWLEDGEMENTS

With the culmination of this work also comes the culmination of many years of mentoring and support from all the advisors, labmates, family, and friends that I have had the privilege of having over this period. Dr. Graciela Unguez is my first science mentor from when I was an undergraduate at New Mexico State University and helped me fall in love with the scientific process, encouraged me to apply to graduate school, and challenged me to care about both my work and the world around me. I found an equally inspiring mentor here at the University of Washington with Dr. Maitreya Dunham, whose unending enthusiasm for outreach, molecular biology, and genetics invigorated me to continually ask questions and explore new areas that captivate my interests. Thank you for helping me develop as a scientist inside and outside the lab and always providing me with the space I needed to grow.

I would also like to thank members of the Dunham lab, particularly Dr. Bryce Taylor, who was my yeast genetics mentor at the bench, and Drs. Caiti Smukowski Heil and Monica Sanchez who both helped me with learning how to use the ministats. I am grateful for Drs. Abigail Keller, Christopher Ryan Livingston Large, and Clara Amorosi; it was incredibly fun to learn coding and science from and alongside my peers. Thank you to Emily Mitchell for always being available to help, making sure our lab ran smoothly, and for being a great friend. Thank you to Dr. Pengyao Jiang for being a thoughtful, humble, and easygoing scientist to work with, and thank you especially for helping me with complex population genetics concepts. Thanks to Dr. Matthew Rich for teaching me how to clone, barcode plasmids, and build variant libraries for an MPRA assay. Thank you to Dr. Anne Clark for help with analyzing *SUL1* across species – and identifying the introgression signatures in this gene. I thank Andrea Chang for your hard

work building and designing an assay for the *MAL33* library, even on days you had exams. Thank you to all the other members of the Dunham lab from past and present generations – every day felt like I was doing science with my best friends.

I extend my gratitude to my committee members (Drs. Celeste Berg, Stanley Fields, Kelley Harris, Wenying Shou, and Jennifer Nemhauser) for providing me with the advice and wisdom over the past five years to become a stronger scientist and for talking with me when I was uncertain about my future. I especially learned a lot from Celeste during my rotation in the Berg lab in terms of both molecular work and scientific writing/presentations, and I am grateful for those skills that I still use every day. Thank you to Dr. Joseph Schacherer and Andreas Tsouris for providing data and advice throughout the entirety of my project. Thank you to Jennifer Anderson for helping me through countless drafts of nearly every personal statement I wrote. Thank you to Andria Ellis for being wild enough to join me in creating Genome Hackers, our coding/genetics camp, and to Dr. Atom Lesiak and Joan Griswold for encouraging and helping me to continue its legacy. Finally, I would like to thank the graduate students in Genome Sciences, and all members of WiGS, GSAIMS, the antiracist book club, and Genomics Salon, for enriching our community and pushing our department to be better than the already spectacular place it is today.

On a personal level, I would like to thank my family and friends for the support they've provided me during my Ph.D. Thank you to Jane Kim, Noor Muhyi, and Hae-na Chung: it is hard to be a good friend from afar, but you did it so effortlessly. Because of you, I didn't struggle as much transitioning to a new city and I never had a day go by where I felt alone. Thank you to my Seattle friends, especially Emma De Neef, Zephyr McLaughlin, Jason Miles, and the other players on the Abiding Dudes, for all the carpool rides, for the soccer games, for all the camping

trips, and for keeping me physically active and mentally well. Thank you to both my parents and my brother Jerry for your support and love for me on this journey, and thank you to all my family back home in Taiwan who are always cheering me on. Thank you particularly to my grandmother who raised and gifted me with a strong sense of my Taiwanese identity that kept me grounded on days that were hard to get through. And finally, I would like to thank my very best friend and partner Jared Mohr for your kindness, humor, empathy, love, and companionship. We did everything in graduate school together, from doing homework to teaching to grading to science fair judging, and I am so elated to be writing my dissertation, defending, and celebrating alongside you.

Chapter 1. Introduction

Since the discovery of the role of DNA in coding proteins that lead to observable traits, the answer has remained elusive to the genotype-to-phenotype question: how do genetic changes give rise to trait differences? Most traits are not governed by one gene (also known as monogenic or Mendelian traits) but are governed by multiple genes (polygenic) with varying degrees of effect sizes (Boyle et al., 2017; MacKay et al., 2009; Morgante et al., 2018). These traits are called quantitative or complex traits. Diseases like cancer are an example of a complex trait, and the decades of research on cancer genetics illustrates that identifying the genetic basis of complex traits is particularly difficult. Understanding complex traits is also important in the agricultural sector, such as understanding how to produce more drought-tolerant crops to sustain human life. Climate change necessitating organisms to adapt to new environments is furthermore important for many animal populations that are experiencing rapid changes to their natural habitats. Several approaches have been developed for identifying regions of the genome responsible for quantitative traits, but most have limitations and do not capture the entire spectrum of phenotypic variation. For instance, common variants with small effect sizes are difficult to detect due to lack of statistical power, and variants with large effect sizes tend to be rare and difficult to capture due to insufficiently large sample sizes (Marouli et al., 2017; Visscher et al., 2012; Yang et al., 2010). Understanding the genetic basis of quantitative traits has great implications for medicine, industrial applications, agriculture, and evolution.

The budding yeast *Saccharomyces cerevisiae* has played a fundamental role in our understanding of quantitative traits. In this introduction, I discuss the major techniques used or developed in yeast that have moved the field forward and how they have scaled up in recent

years. I next discuss how sequencing of genomes from a variety of populations moves research in genetics beyond what we can glean from a single reference genome. Whole-genome sequencing of whole populations can be leveraged for understanding phenomena such as gene redundancy due to paralogs, and I cover the importance of such gene duplications in evolution in the following section.

1.1 Scaling approaches for identifying genes underlying quantitative traits in yeast

---

Figures and text in this section was adapted from a review written in collaboration with Dr. Pengyao Jiang submitted to *Current Opinion in Genetics & Development*.

Expansion of sequencing efforts to include thousands of genomes is providing a fundamental resource for determining the genetic diversity that exists in a population. Now, high-throughput approaches are necessary to begin to understand the role these genotypic changes play in affecting phenotypic variation. *Saccharomyces cerevisiae* maintains its position as an excellent model system to determine the function of unknown variants with its exceptional genetic diversity, phenotypic diversity, and reliable genetic manipulation tools. Here I review strategies and techniques developed in yeast that scale classic approaches of assessing variant function. These approaches improve our ability to better map quantitative trait loci at a higher resolution, even for rare variants, and are already providing greater insight into the role that different types of mutations play in phenotypic variation and evolution not just in yeast but across taxa.

1.1.1 *The significance of high-throughput functional approaches*

Genetic diversity present across the budding yeast *Saccharomyces cerevisiae* population has produced the extraordinary phenotypic diversity that we see in this species today (Fay, 2013; Peltier et al., 2019). With a variety of wild ecological origins and over 12,000 years of domestication across the globe, *S. cerevisiae* isolates in different clades have distinct polymorphisms that facilitate adaptation to specific environments (Bai et al., 2022; Gallone et al., 2016; Peter et al., 2018). The overarching question that still remains after decades of genetics and genomics work is which of the genotypic differences and changes in genetic architecture between individuals give rise to phenotypic variation. Changes in a single locus alone can be the cause for phenotypic differences and is known as a Mendelian trait. However, Mendelian traits are rare as most traits are complex and thus governed by multiple loci and gene-by-environment interactions. Understanding the genetic basis of complex traits has many implications for advancing therapeutics for disease, industrial applications, agricultural output for our changing climate, and our knowledge of evolution.

Curation, deep sequencing, and genomic analysis of 1,011 *S. cerevisiae* isolates collected from natural and domestic environments revealed the sheer number of variants present in the population (Peter et al., 2018). Containing over 1.6 million single nucleotide polymorphisms, this collection highlights that foundational approaches like QTL mapping for understanding the effects of these SNPs would be prohibitively labor-intensive, time-consuming, and oftentimes impossible given the current limitations of QTL mapping strategies (MacKay et al., 2009). Copy number variants (CNVs) further confound our understanding of the genetic architecture of traits; almost all open reading frames (ORFs) in the *S. cerevisiae* genome have a CNV in at least one of the 1,011 strains (Peter et al., 2018). The huge diversity in populations compared to the little that

is known about the effects of genotypic changes, with the added layer of intricacy that environmental factors play, necessitates high-throughput experiments that can confidently determine the impacts of mutations.

Even the huge number of variants as yet observed is dwarfed by the number of variants that could possibly exist. Obviously, this general problem is not limited to the yeast system: modern genetics is going to require high throughput approaches to understanding variation across taxa. Technological solutions developed in yeast can immediately be applied to other genomes either as a testbed for methods development to port to other systems, or by heterologously expressing genes in yeast.

While these challenges are daunting, the throughput of systems level approaches for determining and measuring variant function have increased considerably, particularly in *S. cerevisiae*, pointing to a path forward. The expansive genetic diversity of yeast, coupled with the extensive toolkit for dissecting genetic traits and engineering variants, provide an excellent foundation for understanding the impact of polymorphisms, learning fundamental principles underlying these effects, and ultimately predicting effects of potential mutations. Here we describe the latest technologies and strategies developed for addressing genotypic changes on a massive scale and what questions applications of these approaches can answer.

1.1.2 *Common approaches for identifying quantitative trait loci*

Genome-wide Association Study (GWAS) and Quantitative Trait Locus (QTL) mapping are the major approaches used to understand the effects of natural variants on complex traits (MacKay et al., 2009; Peter et al., 2018; Sardi et al., 2018). Even though tens of thousands of samples have been incorporated in human GWAS analysis, the missing heritability problem,

where mapped variants explain a small proportion of the total heritability (Manolio et al., 2009), is still not resolved. Previously in *S. cerevisiae*, due to the high degree of mosaicism and presence of population structure, GWAS required additional statistical corrections to avoid false positives (Connelly & Akey, 2012; Diao & Chen, 2012; Galardini et al., 2019; Peter et al., 2022; Strope et al., 2015). However, increases in sample size and diversity have improved the power of using GWAS for understanding the genetic basis of complex traits in yeast (Galardini et al., 2019; Maclean et al., 2017; Peter et al., 2018, 2022; Sardi et al., 2018; Strope et al., 2015).

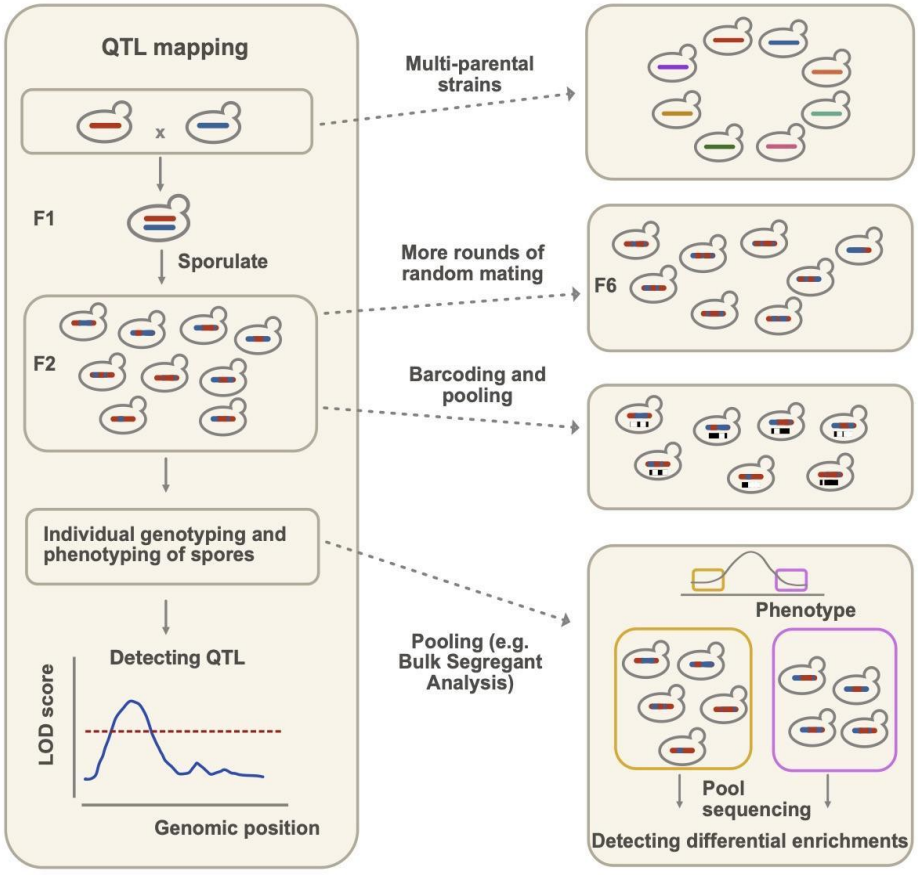


Figure 1.1. Traditional QTL mapping procedures (left panel) and recent advancements (right panel).

These approaches increase mapping precision, throughput, or diversity of natural variation for investigation.

QTL mapping, on the other hand, controls environmental factors and leverages the segregation of parental genotypes to pinpoint causal genetic variants leading to measurable phenotypes of interest (Bloom et al., 2013) (**Figure 1.1**, left panel). Phenotyping approaches for QTL mapping in yeast vary widely and include those that detect changes in molecular phenotypes such as gene expression or protein abundance (Albert et al., 2014; Brion et al., 2020; Duveau et al., 2021; Jackson et al., 2020; Renganaath et al., 2020; Shih & Fay, 2021; Sirr et al., 2020), colony or cell morphology (Nogami et al., 2007; Karin Voordeckers et al., 2012; Wilkening et al., 2014), flocculation patterns (Wilkening et al., 2014), growth rate (Bloom et al., 2019; Jakobson & Jarosz, 2019; Peter et al., 2018; Wilkening et al., 2014), enzymatic function (Collins et al., 2021), compound production (Eder et al., 2018), translation termination efficiency (Torabi & Kruglyak, 2011), and many more (Fay, 2013; Peltier et al., 2019).

1.1.3 *High-throughput genome-wide identification of genes contributing to trait differences*

Pooled approaches—such as bulk segregant analysis, which sequences pooled individuals with extreme phenotypes (Z. Wang et al., 2019), and barcoding multiple parental strains (Nguyen Ba et al., 2022) or individual spores (Matsui et al., 2022) in order to trace lineages of pooled segregants throughout a screening—improve the throughput of QTL mapping (**Figure 1.1**, **right panel**). Although these approaches can be enriched for false positives due to beneficial mutations, genomic instability, and diploidization events that may arise during the growth phase (Wilkening et al., 2014), advancements in automated workflows now allow phenotyping of a remarkable number of individual segregants. The largest QTL mapping study in yeast to date phenotyped an astounding 100,000 F2 barcoded segregants (Nguyen Ba et al., 2022). Collectively, QTL mapping studies have measured over 100 complex traits (Bloom et al., 2019;

Haas et al., 2019; Jakobson & Jarosz, 2019; Nguyen Ba et al., 2022; Z. Wang et al., 2019; Wilkening et al., 2014). Furthermore, deliberate selection of parent strains can survey most of the natural variation in the *S. cerevisiae* population, with 82% of biallelic SNPs captured by just 16 parental natural isolates (Bloom et al., 2019). Progeny of these 16 isolates, as well as hybrids from a diallel cross with 55 natural isolates, are also enriched for rare variants (Bloom et al., 2019; Fournier et al., 2019). Screening of these crosses confirmed that variants of large effects are usually rare in the total population, consistent with negative selection on these alleles (Bloom et al., 2019; Fournier et al., 2019; Nguyen Ba et al., 2022).

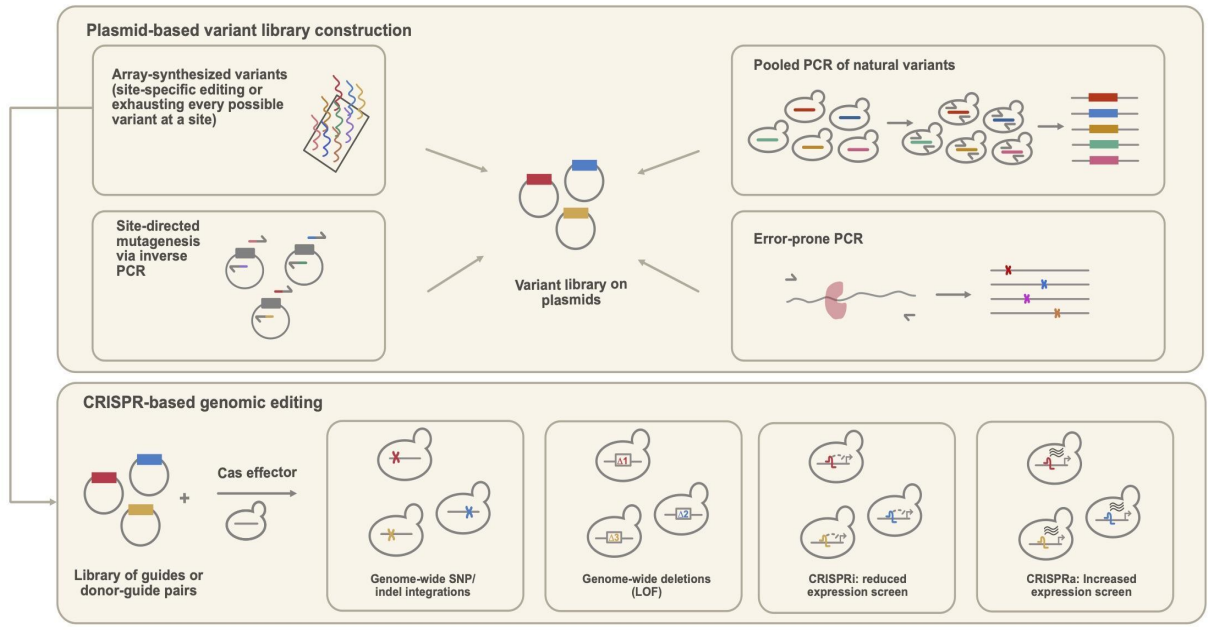


Figure 1.2. Summary of yeast library construction methods.
Top panel: Plasmid-based variant library construction. The effects of these variants can be determined by transforming yeast cells with this plasmid library. Bottom panel: CRISPR-based variant construction. Yeast cells are transformed with a plasmid library of guide or donor-guide pairs so that genetic changes are incorporated into the genome.

Ascertainment biases in QTL mapping still exist, resulting in variants of large effects being overrepresented in functional studies and variants of small effects being overshadowed. Identification of polymorphic differences between two strains and individually introducing the variants from one strain to another exposes the functional effects of those variants in an unbiased manner. Several approaches have leveraged advances made in CRISPR/Cas9 precise editing techniques to introduce SNPs identified between two strains in a high-throughput manner **Figure 1.2, bottom panel**) (Després et al., 2020; Guo et al., 2018; Roy et al., 2018; Sharon et al., 2018). Such studies have implemented improvements to increase homology-directed repair (HDR) efficiency over nonhomologous end joining (NHEJ) in yeast: efficient co-transformation of donor DNA and guide RNA to the same cell (Guo et al., 2018), increased copy number of the donor DNA strand (Sharon et al., 2018), and improved accessibility of the donor DNA to the Cas9-induced double-stranded break (Roy et al., 2018). These pooled variant studies have been able to interrogate the impacts of over 16,000 SNPs (in some cases over 30,000 SNPs) in one experiment, although variants of small effect tend to have low reproducibility and high false discovery rates (Després et al., 2020; Roy et al., 2018; Sharon et al., 2018). Once generated, libraries can be screened in various conditions to measure the interaction between each variant and the environment. Overcoming other limitations such as decreased editing accuracy with increased distance to PAM (protospacer-adjacent motifs, or the cut-site motif recognized by Cas9) sites or high oligonucleotide error rates will improve the power and accuracy of these assays.

Precise editing approaches have been useful for identifying expression QTLs (eQTLs), or *cis*-regulatory variants that alter expression and affect phenotype, making it possible to pinpoint causal *cis*- and *trans*-acting mutations in *S. cerevisiae* (Brion et al., 2020; Lutz et al., 2019;

Renganaath et al., 2020). Similarly, genome-wide perturbations that upregulate or downregulate nearly all genes in one assay can identify eQTLs as well, although not always at nucleotide resolution (Alford et al., 2021; Lian et al., 2019; McGlincy et al., 2021; Momen-Roknabadi et al., 2020). Scaled eQTL studies can now simultaneously measure expression and protein abundance with greater statistical power, providing a better understanding of how promoters and post-transcriptional processes may affect complex traits (Brion et al., 2020; Lutz et al., 2019; Renganaath et al., 2020). In addition to changes in expression and protein abundance patterns, complete loss of function (LOF) by full gene deletions, frameshift mutations, or introduction of premature stop codons can be investigated in a high-throughput manner to determine the impact of different types of LOF mutations (Guo et al., 2018; Sadhu et al., 2018).

1.1.4 *High-throughput interrogation of variant effects at a causal locus*

When genes of interest are identified by any of these mapping methods, the question arises of how natural genetic variation or potential new mutations in these genes impact function. Multiplexed Assays of Variant Effects, or MAVEs, can deeply interrogate the effects of substitutions and *cis*-regulatory mutations in one locus by creating large libraries of variants and measuring *en masse* how they affect fitness, protein interactions, expression, or enzymatic function (Weile & Roth, 2018). Variants can be generated through a variety of methods (**Figure 1.2, top panel**): Error-prone PCR is a cost-effective approach that introduces random mutations into a region of interest and can explore both single and combinatorial effects of mutations, but specific mutations and variants of interest are not guaranteed to be generated (Rich et al., 2016). Inverse PCR allows for site-directed mutagenesis and has now been scaled to allow for saturation; this saturation mutagenesis approach has been useful for understanding the effects of

substitutions of all amino acids in one site but is quite labor intensive (Amorosi et al., 2021; Morton et al., 2020). Oligonucleotide synthesis also allows for deep interrogation into single substitutions and can include more complex variants, or variants with more than one mutation, but has a higher cost and error rate (Ollodart et al., 2021). Saturation mutagenesis can now be performed endogenously with CRISPR/Cas9 editing, namely by designing synonymous

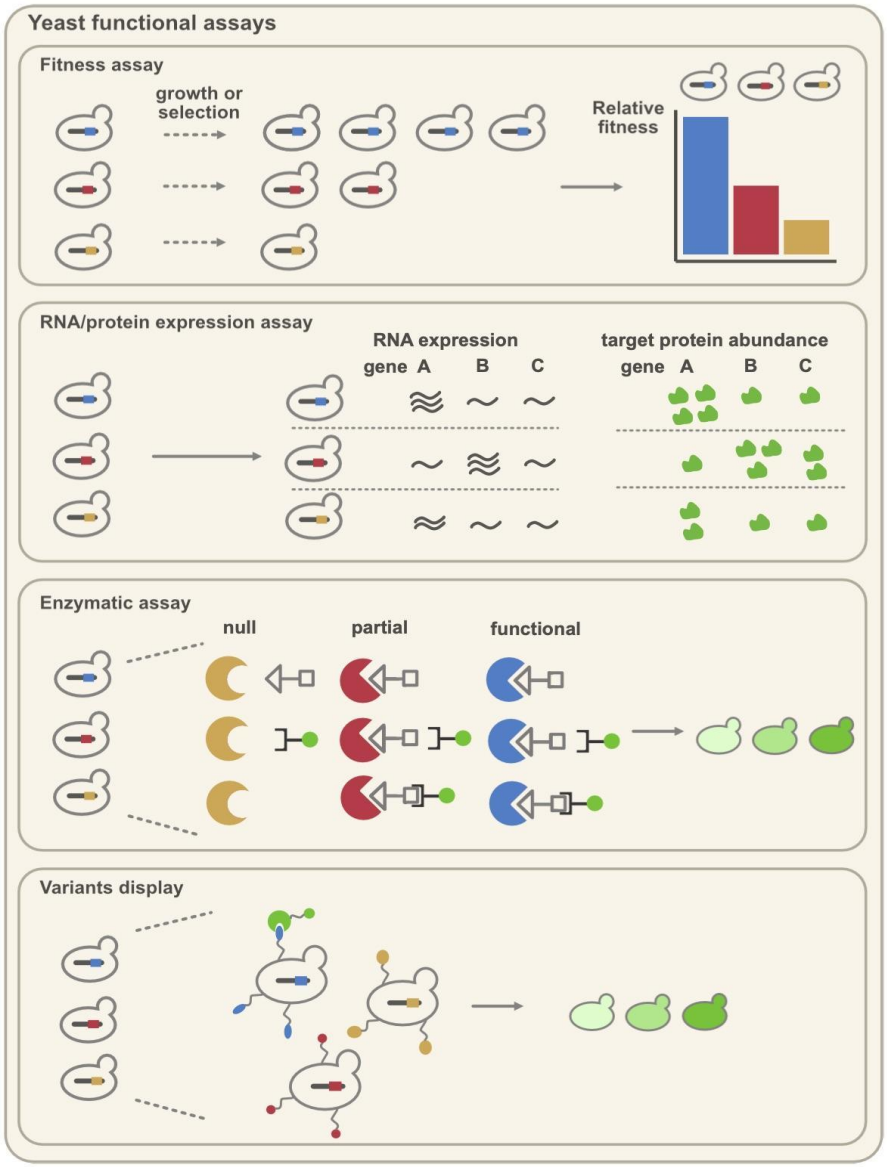


Figure 1.3. Summary of high-throughput functional assays in yeast.
The starting library can be developed using either plasmid-based or CRISPR-based methods.

mutations in the donor DNA strand to prevent unintentional recognition and cleavage due to strand complementarity to the guide RNA (Roy et al., 2018) (**Figure 1.2**, bottom panel). Sequencing full variants or short DNA tags that act as barcodes for each variant allows the tracking of these variants throughout an assay, which can be used to infer variant function (**Figure 1.3**).

MAVEs in yeast have been useful for predicting how new mutations alter phenotype, not just for yeast genes, but for genes from other systems as well (**Figure 1.3**). Remarkably, many human genes complement *S. cerevisiae* gene knockouts, smaller scale homology can be taken advantage of for those that don't, and even genes without homologs can be functional when heterologously expressed in yeast (Amorosi et al., 2021; Boonekamp et al., 2021; Cervelli et al., 2020; Hamza et al., 2020; Kapolka et al., 2020; Newberry et al., 2020; Ollodart et al., 2021; Sirr et al., 2020). A recent example is the combination of a yeast display assay and saturation mutagenesis on the receptor binding domain (RBD) of SARS-CoV-2 revealing the substitutions in the RBD that affect binding to human receptor proteins (Starr et al., 2020). These results have been phenomenally important for quickly anticipating the effects of mutations in SARS-CoV-2 and for developing effective vaccines (Starr et al., 2020).

This style of high-throughput, pooled phenotype testing can also be applied to more complex allele libraries. For example, we have generated and phenotyped a library of all natural alleles of one gene, including multiple SNPs and insertion/deletions in one haplotype (Yeh et al., 2021). This MAVE approach for understanding gene function on a species-wide scale reveals not only the natural fitness distribution of variants in this gene, but informs its evolutionary history as well. Decreases in cost along with increases in throughput and accuracy of long-read

sequencing enable the genotyping of large libraries containing complex variants that are longer than what can be fully captured from next-generation sequencing (301bp).

Finally, variant libraries need not be based on natural sequences at all. Generation of randomized sequences has been useful for understanding how translation and transcription factor-binding sites are utilized in promoter regions, with the effects of up to 100 million sequences determined in one experiment (Cuperus et al., 2017; de Boer et al., 2020; Vaishnav et al., 2022). The increasing number of studies focusing on exploring this space enhances our understanding of variant effects and is fundamental to computational programs more accurately predicting function (Gray et al., 2018; Jackson et al., 2020; Livesey & Marsh, 2020; Reeb et al., 2020; Weile et al., 2017; Zhou & Cai, 2021).

1.1.5 *Future directions for high-throughput functional approaches for identifying QTLs*

Advancements in approaches for identifying QTLs as well as the underlying causal genes and nucleotides have revealed an extraordinary amount of information about complex traits. High throughput methods such as those described here facilitate moving beyond individual example cases and allow for general patterns to surface that begin to reveal the categories of complex traits and what is required for a comprehensive understanding of their genetic basis. Complex traits vary in patterns themselves and can involve multiple common variants, multiple rare variants, or a combination of both (Bloom et al., 2019; Nguyen Ba et al., 2022), justifying the need for independent interrogations of variants under multiple environments. For some traits, variants clustered in one locus can have effects in the same direction (Sharon et al., 2018; She & Jarosz, 2018); for others, however, variants can have canceling effects that result in neither variant being detected in a QTL (Renganaath et al., 2020). These variants can be coding or

noncoding, and the effect sizes of these mutations are relatively similar (Jakobson & Jarosz, 2019; Sharon et al., 2018). Effects of variants can be additive, although greater sample sizes of segregants revealed that the terminal effects are more indicative of epistatic interactions (Bloom et al., 2019; Fournier et al., 2019; Matsui et al., 2022; Nguyen Ba et al., 2022). Identifying epistatic interactions is still challenging, and most high-throughput approaches for doing so have been through whole gene deletions or engineered mutant alleles (Costanzo et al., 2016, 2021; Diss & Lehner, 2018; Domingo et al., 2018; Kuzmin et al., 2020) or by investigating genetic network changes as a result of single gene perturbations (Caudal et al., 2021; Jackson et al., 2020). Thus, the effects of most pairwise and complex SNP interactions at the genomic scale remain to be determined.

The end goal of these studies is to predict how genotypic changes alter phenotype. Saturation-level analysis can be achieved by taking a gene-centric approach to understand how variation in a locus affects phenotype. Good candidates for such analysis are the causal genes identified by QTL studies. Certain genes (such as *MKT1, HAP1, IRA2*) continually resurface in QTL maps, indicating a high degree of pleiotropy (Demogines et al., 2008; Deutschbauer & Davis, 2005; Dimitrov et al., 2009; Gou et al., 2019; Jakobson & Jarosz, 2019; Kim & Fay, 2009; Nguyen Ba et al., 2022; Steinmetz et al., 2002; Wilkening et al., 2014). Yet, how regulatory and substitution changes alter their function under these various environments and across many more alleles and genetic backgrounds is still largely unknown. Measuring the effects of genes using standing variation can reveal patterns of evolution and signatures of selection (Yeh et al., 2021). Moreover, coupling growth phenotypes with molecular phenotypes like expression or protein function can lead to mechanistic understanding. They also illustrate a high-throughput way for studying distribution of fitness effects of genes, which is an important

and long-standing question in understanding evolution.

Even with the high-throughput approaches developed to date, many challenges and prospects in identifying causal variants persist. In order to measure variant effects on complex traits, the traits must have phenotypes that can be measured accurately in high throughput. Additionally, knowing only one of the multiple traits affected by a pleiotropic gene may confound interpretations of how variation affects fitness. The impact of intergenic regions remains comparatively understudied as well. Increases in whole-genome sequencing have also revealed that causal copy number variants explain a larger fraction of phenotypic variance when compared to SNPs (Peter et al., 2018). Thus, future studies will need to move beyond nucleotide-level variants; increasing the throughput for studying the effects of mutations such as copy number variation, translocation, and aneuploidies will provide a more exhaustive view of how genotype affects phenotype. Finally, the success of heterologously expressing human genes in yeast to understand gene function is evidence that this versatile model organism can test gene function across other organisms as well. Applications of these high-throughput methods to across taxa will inform the evolutionary history of selection, adaptation, and drift in spanning diverse populations.

1.2 TECHNOLOGIES USED FOR LINKING BARCODES TO VARIANTS IN POOLED LIBRARIES

Advances in sequencing technology supported the discovery of the magnitude of variant effects in pooled variant studies. In these pooled variant studies, genes (or regions of genes) are cloned into a plasmid and tagged with a unique barcode. In short-read sequencing, the length of the variable region of interest is the limiting factor. To date, the largest number of base pairs that

can be sequenced from a molecule in next-generation sequencing, without further manipulation, was 302 bp (151 bp paired-end reads). To overcome this limitation, a method called subassembly was developed to enable sequencing of a longer variable region (Hiatt et al., 2010). In this approach, variable regions tagged with unique DNA barcodes are randomly sheared and ligated with an adaptor for sequencing. Sequencing from the ends of the breakpoints ensures that the entirety of the variable region is covered, and sequencing results from each sheared molecule can be merged together (Hiatt et al., 2010). However, this approach becomes more laborious with increased regions of variable length.

Another approach developed for linking barcodes with variants using short-read sequencing is a droplet-based approach (Borgström et al., 2015; Lan et al., 2016). Droplet microfluidics has been instrumental in advancing single-cell sequencing, but it can also be used for high-resolution sequencing of long DNA fragments. In this approach, single DNA fragments containing the barcode and variant are isolated into a single droplet (Borgström et al., 2015; Lan et al., 2016). The molecule is amplified, fragmented, and tagged with a unique index barcode (in Borgström et al. (2015), indexing is a bead-based process). All fragments are freed from the droplet and then sequenced in parallel. Sequencing reads are demultiplexed and assembled to determine the barcode-variant map for the entire pool. Because short-read sequencing still has the highest accuracy among all sequencing platforms, this approach provides an affordable option for linking barcodes to variants in longer DNA molecules. Although short-read sequencing does not have a 100% accuracy rate, the depth by which these molecules are sequenced provides higher confidence in read assemblies.

Long-read sequencing technologies have recently become more affordable and more accurate, opening the possibility of using long-reads and bypassing the need to assemble short-

read fragments. Oxford Nanopore Technology (ONT) sequencing can be affordable, but has a low accuracy rate that can sometimes be as low as 69% (Logsdon et al., 2020). The accuracy rate in PacBio sequencing is low as well (85%), but recent chemistry and improvements have increased its throughput and accuracy (Wenger et al., 2019). Specifically, PacBio high-fidelity (HiFi) sequencing produces reads with an accuracy rate of up to 99% through the ligation of SMRTbell adaptors that circularize a DNA fragment (Logsdon et al., 2020; Wenger et al., 2019). The reason HiFi reads are more accurate is that the circularization of molecules allows the polymerase to traverse the molecule multiple times, increasing the depth by which each fragment is sequenced. Using PacBio HiFi sequencing for linking barcode to variants has been successful and is becoming increasingly common (Amorosi et al., 2021; Ollodart et al., 2021; Starita et al., 2015; Yeh et al., 2021). Highly accurate reads are especially important for correctly identifying variants, and improvements in both sequencing technology and computational assembly software will only improve the accuracy of long read sequencing. What is remarkable to me is that when I started graduated school, the average number of times the polymerase passed through molecules was ~13, and each SMRT cell produced ~200,000 HiFi reads. SMRT cells now produce upwards of ~2 million HiFi reads with >20 passes on average for a 4kb DNA fragment. These improvements in throughput and sequencing chemistry allow for greater sequencing depth of the library, and also allow multiplexing of libraries, which can greatly help alleviate the high price tag of PacBio sequencing.

1.3 CURATION OF *S. CEREVISIAE* NATURAL ISOLATES FOR UNDERSTANDING GENETIC DIVERSITY

A large sampling of genomes from a population is needed to properly represent genetic diversity and the role it plays in phenotypic diversity. The rapid decrease in cost of sequencing has permitted the curation of large collections in several organisms. These large collections include the 1000 Genome Project analyzing the human population (Altshuler et al., 2010), the 1001 Genome Project analyzing the *Arabidopsis thaliana* population (Alonso-Blanco et al., 2016), and the 1002 Genomes analyzing the *Saccharomyces cerevisiae* population (Peter et al., 2018). What stands out about *S. cerevisiae* is the remarkable diversity harbored in the compact genome of this microorganism. With powerful tools for understanding complex traits in yeast such as QTL mapping and GWAS, these data have enabled insight into how genome diversity (including ploidy, copy number changes, signatures of introgression, etc.) shaped evolutionary history (Peter et al., 2018).

Sequencing of over 1,000 genomes in a population across all these species was performed using short-read sequencing (Alonso-Blanco et al., 2016; Altshuler et al., 2010; Peter et al., 2018). Further analyses of these datasets underscored that while these datasets are good for high-level comparisons of individuals or strains, phasing predictions are often inaccurate (Belsare et al., 2019). Thus, resequencing with long-read sequencing technology can greatly improve confidence in the quality of data in terms of genome structure, duplication, and haplotyping (Kronenberg et al., 2021; Logsdon et al., 2020; Saada et al., 2022). Accuracy and throughput of long-read sequencing has drastically improved in the past few years. Advances in computational tools provide the opportunity to move beyond using a single strain as the reference sequence

towards using pan-genomes for understanding a species' genome through functional and sequence-based methods.

1.4 Role of gene duplications in evolution

The outcomes of gene duplications have great potential for driving changes in the genome that shape evolution (Magadum et al., 2013; Ohno, 1970; Jianzhi Zhang, 2003). Alterations in copy number can result in a multitude of fates for homologs. Due to the redundancy, one potential fate of the gene pairs is that one retains normal function and the other acquires mutations that result in loss of function (Jianzhi Zhang, 2003). This loss of function due to redundancy is known as pseudogenization, and pseudogenes are often found in gene-dense regions (Torrents et al., 2003). Old pseudogenes may be undetectable since homology between the gene pairs decreases over time, but pseudogenes may also retain some purpose, such as providing a template for gene conversion in the immune system of chickens (Ota & Nei, 1995). Another potential fate of gene duplications is subfunctionalization, where the role of the original gene is divided into the two gene duplicates. These can be particularly important for controlling differential expression, like tissue-specific expression in an osteoglycin gene in fish (Costa et al., 2018). Neofunctionalization, or emergence of a new function, may occur as a result of gene duplication as well. In this process, one gene duplicate retains its original function, and the relaxed selection on the other copy allows it to acquire mutations that result in a new function (Jianzhi Zhang, 2003). For example, duplication of a cytochrome P450 gene resulted in insecticide resistance in planthoppers (Zimmer et al., 2018). Especially for populations experiencing rapid changes in environments, neofunctionalization provides an important role in adaptation.

Duplications can arise from a variety of events: unequal crossing over during recombination, transposon-mediated duplications like retrotranspositions, duplicative transposition, or polyploidization (Magadum et al., 2013; Jianzhi Zhang, 2003). Unequal crossing over can occur when two sister chromatids are misaligned during recombination and results in novel tandem repeats (Magadum et al., 2013; Jianzhi Zhang, 2003). Transposon-mediated duplications, like retrotranspositions, occur when mRNA is reverse transcribed to cDNA and reinserted into the genome. Unlike unequal crossing over events, genes duplicated from retrotranspositions are usually not linked (Magadum et al., 2013; Jianzhi Zhang, 2003). Duplicative transposition occurs during nonallelic homologous recombination or nonhomologous end joining and can result in repeats like segmental duplications, which are the most common type of gene duplication events in humans (Magadum et al., 2013; Samonte & Eichler, 2002). Finally, polyploidization can also result in gene duplications. These can be chromosomal, but whole-genome duplication events have occurred, typically thought to be due to hybridization, chromosome doubling, and then gene loss (Magadum et al., 2013).

Based on sequence and knowledge of how gene duplications arise, we can use the context around where gene duplicates, or paralogs, exist to determine the age and potential function of each paralog. Genome assembly and comparative genomics confirmed the hypothesis of an ancient whole genome duplication event in *S. cerevisiae* (Kellis et al., 2004). Vertebrate genomes also show signatures of whole-genome duplications and chromosome-scale duplications (Holland & Ocampo Daza, 2018). Kellis et al., 2004 discovered that typically one gene in the paralog pair showed faster evolution than the other, supporting the idea of neofunctionalization or subfunctionalization. The idea of differentially evolving paralogs was supported by studies such as those of genes in the maltase family that encode the enzyme that

can process maltose or other carbon sources (Brown et al., 2010; K Voordeckers et al., 2012). What was discovered was that each paralog in the multigene family supported subfunctionalization and neofunctionalization models due to substrate-specific activities across the paralogous enzymes. In yeast, the loci where these maltase genes are located are subtelomeric, and thus have higher rates of recombination and mutations (Brown et al., 2010). The higher rates of recombination and mutations in subtelomeric maltase genes permitted the rapid divergence of these genes that resulted in paralogs with differing function (Brown et al., 2010; K Voordeckers et al., 2012).

1.5 DISSERTATION SUMMARY

In this dissertation, I describe my approach for determining the function natural variants of *Saccharomyces cerevisiae* in a high-throughput manner. The approach I developed was piloted using the gene *SUL1*, a high-affinity sulfate permease, that imports sulfate molecules into the cell (Chapter 2). In addition to confirming the loss-of-function phenotype of alleles with premature stop codons, I can assign function of alleles matched to strains in the 1,011 yeast (Peter et al., 2018). I also show what this information provides on the species level in terms of function and evolution. In Chapter 3, I describe an approach developed using multiple sequence alignments to improve the accuracy by which barcodes and variants are mapped using PacBio long-read sequencing. With co-authors, I validated the performance of this program using simulations of PacBio deep mutational scanning libraries and furthermore show the sequencing depth needed to achieve a high percentage of accurately mapped reads. Following the development of both the experimental pipeline and computational corrections, I show how this approach can be applied to a biological question, specifically in understanding paralog

differences in yeast (Chapter 4). Application to understanding paralog differences was done in *MALx3*, which is a multigene family that includes a transcription factor required for activating expression of the maltose transport and maltase genes for utilization of maltose when it is the sole carbon source. Comparative analysis on a population scale reveals significant differences in patterns among the three *MALx3* genes investigated, suggesting differential function of these paralogs. Finally, in Chapter 5, I discuss other questions in evolution that can be resolved using this method. I also elaborate on pitfalls in this method and what optimizations can be made to improve its utility.

Chapter 2. Developing a high-throughput assay for functional analysis of natural *SUL1* alleles

---

The work in this chapter is derived from a manuscript currently posted on *bioRxiv* titled "High-throughput functional analysis of natural variants in yeast." Andreas Tsouris conducted the whole isolate growth assay on sulfate-limited agar plates. PacBio sequencing was performed by the UW PacBio sequencing core, and barcode sequencing was performed by Noah Hanson.

How natural variation affects phenotype is difficult to determine given our incomplete ability to deduce the functional impact of the polymorphisms detected in a population. Although current computational tools can predict allele function and experimental tools can measure it, there has previously been no assay that does so in a high-throughput manner while also representing haplotypes derived from wild populations. In this work, I present such an assay that measures the fitness of hundreds of natural alleles of a given gene without site-directed mutagenesis or DNA synthesis. In this assay, I amplify a gene of interest, the high-affinity sulfate transporter *SUL1*, from genomic DNA pooled from a collection of 1,011 *S. cerevisiae* strains and clone alleles of the genes *en masse* onto a low-copy, barcoded plasmid. I show that this approach allows us to measure the fitness conferred by each allele and to stratify functional and nonfunctional alleles. Additionally, I pinpoint the polymorphisms in both coding and noncoding regions that are detrimental to fitness or are of small effect and result in intermediate phenotypes. Integrating these results with a phylogenetic tree, I observe the frequency that a loss-of-function mutation occurs and whether or not there is an evolutionary pattern to the observable phenotypic results. This approach is easily applicable to other genes. These results, which

complement classical genotype-phenotype mapping strategies and demonstrate a high-throughput approach for understanding the effects of polymorphisms across an entire species, can greatly propel future investigations into quantitative traits.

2.1 Background

Quantitative traits, or traits that vary on a continuous distribution rather than in discrete categories, are responsible for most phenotypic differences across all organisms (MacKay et al., 2009; Morgante et al., 2018). Despite decades of efforts investigating how genotype informs phenotype, the molecular underpinnings of quantitative traits are still largely unknown, especially on a species-wide scale. Understanding how sequence changes lead to phenotypic changes is difficult to disentangle as these traits often involve interactions between multiple loci that each in themselves have genetic variation among natural populations. Even for a single locus, an exhaustive population-scale determination of the impact of genetic variants remains out of sight. Improved approaches for investigating the function of genetic variation in a cost-effective and high-throughput manner will broaden insights into the genetic basis of trait variation, ranging from deleterious diseases to adaptive evolution.

Due to the rapid advancement and decreased cost of high-throughput sequencing, forward genetics approaches have boosted our ability to pinpoint loci underlying traits of interest. For instance, quantitative trait loci (QTL) and linkage mapping provide avenues for identifying loci responsible for phenotypic differences between individuals and even in some instances can result in determining which polymorphisms are integral for certain phenotypes (Ehrenreich et al., 2012; Treusch et al., 2015). However, QTL mapping method relies on pairwise crosses between a small subset of genetic backgrounds and is difficult to scale to investigate phenotypic variation on the

population level (Stinchcombe & Hoekstra, 2008). Approaches like genome-wide association studies do investigate loci on the population scale, but they lack the ability to infer and experimentally ascertain the functional effects of rare or low-frequency variants (Morgante et al., 2018). Other computational approaches deduce function based on collected data describing metrics such as conservation, statistical analyses of genomic architecture, allele frequency, predicted changes in protein stability, known sites of protein-protein interactions, and transcription factor-binding motifs, but all still require experimental validation (Adzhubei et al., 2010; Mitchell-Olds et al., 2007; Schymkowitz et al., 2005; She & Jarosz, 2018; Wagih et al., 2018; Wray et al., 2013).

Recently, multiplexed assays of variant effects (MAVE) studies have provided an approach for determining the function of thousands of variants in a high-throughput manner (Starita et al., 2017; Weile & Roth, 2018). MAVEs have been extremely useful in understanding how missense mutations or nucleotide changes alter gene function and/or expression (Duveau et al., 2017; Fowler & Fields, 2014; Matreyek et al., 2018). However, most MAVEs have been limited to studying single nucleotide or amino acid substitutions that differ from a reference sequence, as technologies don't yet exist to generate and measure the consequences of the large libraries that would be necessary to explore combinatorial variation. Additionally, the majority of variants assayed are rarely reflective of those in natural populations. For instance, natural alleles can have more than one polymorphism, not all of which are seen exclusively in coding or exclusively in noncoding regions, and thus are not surveyed completely in many MAVE studies. The ability to directly test the function of natural variants of whole populations provides context for how polymorphisms and combinations of polymorphisms alter phenotype. Furthermore, such an approach would provide deeper insight into the evolutionary history of a gene and how both

weak and strong selection or drift have resulted in natural variation (T. Johnson & Barton, 2005; Mitchell-Olds et al., 2007). Thus, developing a method for testing natural variants in a high-throughput manner is of high interest.

Here, we developed such an assay to determine the function of natural variants on a species-wide scale using the budding yeast *Saccharomyces cerevisiae*. With the rapid advances in high-throughput whole-genome sequencing, we now have large collections of natural *S. cerevisiae* strains that contain genomic data as well as geographical and ecological information (Bergström et al., 2014; Liti et al., 2009; Peter et al., 2018; Schacherer et al., 2009; Strope et al., 2015; Zhu et al., 2016). Although much research has been done on laboratory strains for understanding biology, curation of these natural collections revealed the striking diversity within this popular model organism: *S. cerevisiae* has been isolated from a variety of countries all over the globe and from environments like humans clinical samples, domesticated products like beer and bread, and tree and fruit samples. Sequencing of these genomes has revealed an extraordinary depth of genetic variation, but still little is known about how these genetic changes impact phenotypic variation outside of a handful of association studies and QTL mapping efforts (Ehrenreich et al., 2012; Kim et al., 2012; Peltier et al., 2019; Wilkening et al., 2014). With the large genome sequencing efforts and strain collections, in addition to the wealth of molecular tools developed for yeast, *S. cerevisiae* is the ideal system to develop this assay and investigate the effects of natural polymorphisms for whole populations.

For piloting and developing our approach, we used the natural alleles of *SUL1* from a collection of 1,011 isolates to test whether we can deconvolute how variation affects cell growth under sulfate-limiting conditions (Peter et al., 2018). *SUL1* encodes a high-affinity sulfate permease and is expressed under sulfate limitation. When different strains of *S. cerevisiae* are

evolved under sulfate limitation in the chemostat, cells with amplifications of *SUL1* have high fitness and rise in frequency in the population (Gresham et al., 2008; Payen et al., 2014; Sanchez et al., 2017). Strong selection for amplification of this locus in sulfate-limiting conditions allows for a reliable functional assay in which we can mimic amplifications by transforming cells with a low-copy plasmid containing *SUL1*. Additionally, we have performed a deep mutational scan on the promoter of *SUL1*, giving us a dataset measuring the functional consequences of single mutations for comparison (Rich et al., 2016). By co-culturing a population of cells transformed with a barcoded library of natural alleles, we can measure competitive fitness via barcode sequencing, thereby determining *SUL1* functionality *en masse*. Our results show that this assay is accurate in predicting function and useful in understanding which genetic changes affect phenotype. These data allow for insight into the evolutionary history of *SUL1* function and possible evidence for selection of loss-of-function mutations. This approach, especially when combined with established forward genetics approaches in identifying causal loci, will greatly strengthen our understanding of quantitative traits on a species-wide scale.

2.2 RESULTS

2.2.1 *Allele library curation and characterization*

When sulfate is a limiting nutrient, *S. cerevisiae* increases expression of *SUL1*, which encodes a high-affinity sulfate permease that increases the uptake of sulfate molecules into the cell. Previously, we measured the competitive fitness of *SUL1* alleles isolated from 10 different wild yeast isolates (Payen et al., in preparation). We found that these alleles confer a wide range of fitness: some had loss-of-function phenotypes while others performed better than the allele found in the reference strain S288C. In order to determine whether this wide variation was

representative of the entire species and whether it correlated with features such as the environment from which each strain was isolated, we set out to survey *SUL1* functionality across a bigger sample of natural isolates. For this study, we used the 1,011 *S. cerevisiae* strain collection, which was curated from a variety of geographical and ecological origins (Peter et al., 2018). In addition, the collection contains at least 250 unique alleles of *SUL1,* with 354 variable sites in the gene. Alleles contain 11 polymorphisms on average vs. the reference allele, with the most polymorphic allele having 79 mutations. Therefore, these factors make *SUL1* a powerful tool for us to better understand the natural variation of a single gene in *S. cerevisiae* populations.

In our previous studies, the fitness of individual *SUL1* alleles was measured by transforming the reference strain with an additional copy of a *SUL1* allele on a low-copy plasmid, and the resulting transformant was competed against the lab strain without a plasmid containing a GFP cassette in the HO locus, but is otherwise isogenic, under sulfate limitation (Sanchez et al., 2017, Payen et al., in preparation). While this assay is reliable and consistent, it would be difficult and unrealistic to scale to measure hundreds of alleles, as each competition is labor-intensive and requires extensive resources. Thus, we developed a high-throughput, multiplexed approach that allows us to simultaneously measure these fitness values directly (**Figure 2.1**). To do this, we pooled the 1,011 isolates together, extracted genomic DNA, and used barcoded primers binding to conserved regions to isolate and amplify all natural alleles of the *SUL1* gene. These sequences were cloned *en masse* into low-copy CEN/ARS plasmids to create an allele library that was used to transform the reference strain (FY). The resulting library in the FY background contained approximately 6,000 barcodes for an estimated 250 unique alleles (24X coverage) to ensure complete coverage and internal replicates (**Figure 2.1A**).

We used PacBio long-read circular consensus sequencing (CCS) to pair barcodes with their respective alleles (**Figure 2.1.B**). Although PacBio CCS has drastically decreased sequencing errors over the past few years, we found that many reads still contained errors that were especially noticeable in the form of insertions and deletions. To further eliminate these sequencing artifacts, we performed multiple sequencing alignments on CCS reads that shared the same barcode, and we used those alignments to derive new consensus sequences. In total, our analysis produced 8,386 barcode-variant pairs, which we determined was still an overestimate given our library size of ~6,000 barcodes. We removed consensus reads that appeared only once to eliminate false positives or negatives in our downstream analysis, with 3,787 barcodes remaining.

Among these 3,787 barcodes, we identified 407 unique alleles in our library. Of these variants, we were able to match 228 alleles to at least one strain in the 1,011 strain collection, with a total of 880 strains that had at least one matched allele in the library (**Supplementary Figure 1**). To determine how well this library reflected the polymorphisms in the strain collection, we plotted the correlation of polymorphism frequency in both the variant reference sequences and library sequences and found that these values were highly correlated (Pearson's correlation, r=0.978, **Supplementary Figure 2**). Correlation values were similar for polymorphisms in all regions of the gene: the 5'-UTR, coding region, and 3'-UTR were all well-correlated (Pearson's correlation, r=0.956, 0.980, and 0.993, respectively). Of the 354 variable sites found in the reference sequences only 45 of them were not detected in the allele library, nine of which were rare polymorphisms. Our pipeline did not reveal any *de novo* mutations that could have resulted from PCR or sequencing artifacts. Capturing the remaining missing

polymorphisms may require deeper sequencing. Furthermore, errors in variant calling may be a factor in being able to accurately match alleles to the strains of origin.

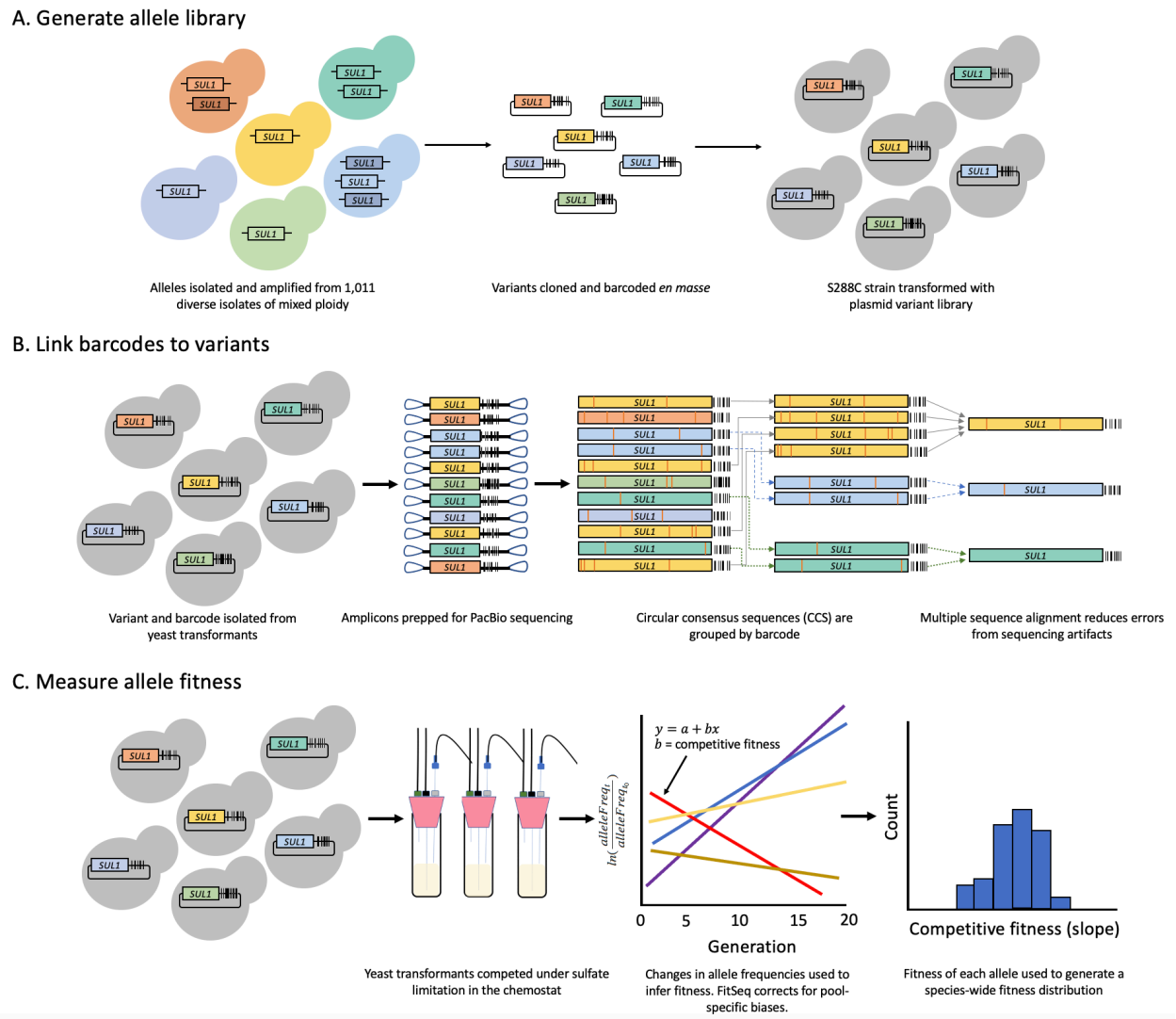


Figure 2.1. Workflow for assaying natural variants in the 1,011 strain collection.

A) The S288C lab strain is transformed with *SUL1* natural allele barcoded plasmid library.

B) PacBio long-read sequencing is used to link barcodes with variants.

C) Transformants are competed together in one flask under sulfate limitation. Each culture represents a biological replicate. Barcode sequencing every 3-4 generations is used to calculate the abundance of each variant and its respective competitive fitness.

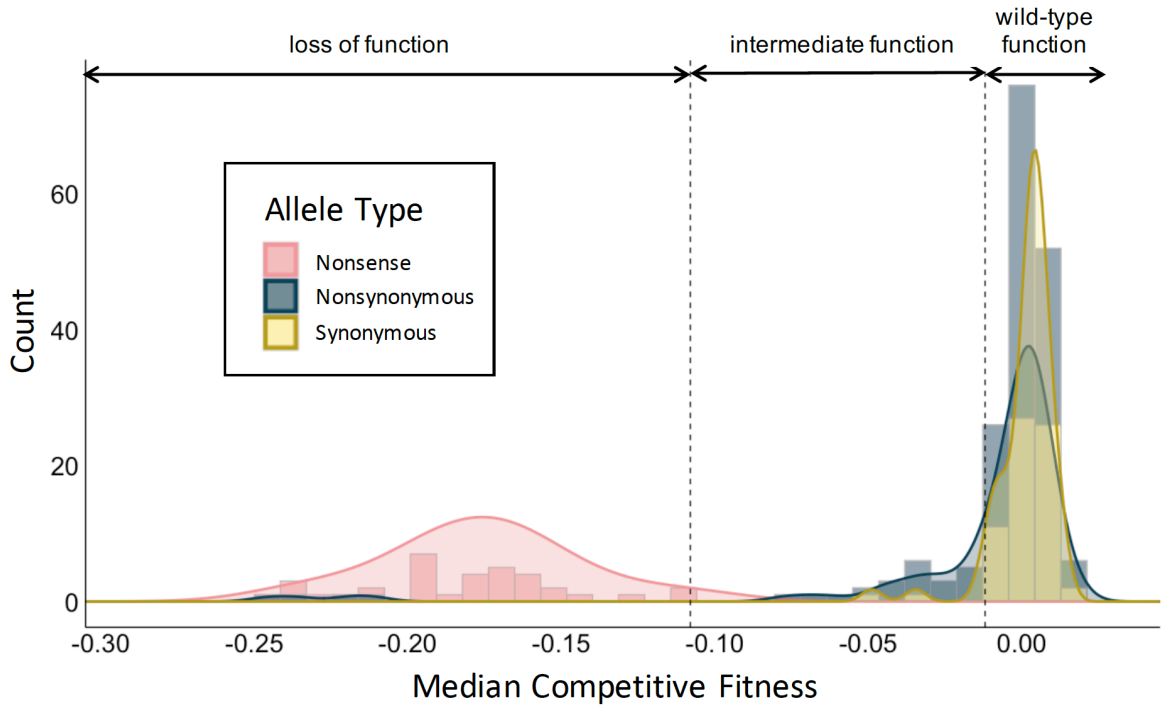


Figure 2.2. Species-level distribution of fitness effects of natural *SUL1* alleles.
Yeast transformants of an allele library of *SUL1* cloned onto a low-copy plasmid were competed in sulfate-limited media in the chemostat. The log-fold change in proportions of each barcode across 12 timepoints was measured through barcode sequencing and used to calculate competitive fitness. Alleles categorized as nonsense alleles may also contain synonymous and nonsynonymous polymorphisms. Those grouped as nonsynonymous alleles may contain synonymous polymorphisms, but do not have premature stop codons. Synonymous alleles do not have nonsynonymous or nonsense polymorphisms. All alleles may contain polymorphisms in the promoter or 3'UTR. Loss-of-function alleles were defined as having a fitness lower than the highest-fitness allele with a premature stop codon. Wild-type function alleles have a fitness higher than the lowest-fitness synonymous allele.

We were also unable to capture the alleles from 23 strains that were identified to have *SUL1* introgressed from *Saccharomyces paradoxus*. The inability to capture these sequences was likely due to these sequences being more highly diverged and therefore unable to hybridize with the primers that were designed. However, for completeness, we were still able to measure the

functionality of the introgressed *SUL1* alleles using our lower throughput method of direct competitions, as described below.

2.2.2 *Fitness distribution across natural SUL1 alleles*

To determine the fitness landscape of all the *SUL1* alleles present in our allele library, we competed the library of yeast transformants in a continuous culture system under a sulfate limitation condition. Samples from 12 timepoints across four replicates were collected every 3-4 generations.

For each sample, we extracted the plasmids from sampled cultures and sequenced the barcodes using Illumina short-read sequencing. By tracking the change in barcode frequencies over the 12 timepoints, we determined the competitive fitness values for strains carrying each allele (**Figure 2.2**). The calculated competitive fitness of the three replicates showed strong correlation and reproducibility (**Supplementary Figure 3**).

In our barcoded library, 863 of 3,787 barcodes were associated with alleles identical to that of the S288C reference strain. We normalized all fitness values to the average fitness of these wild-type alleles (0.0097, standard deviation 0.0698). Reassuringly, we found that many barcodes with lower fitness values (fitness < -0.03) were associated with alleles containing natural premature stop codons (**Figure 2.2**). In fact, upon analyzing the sequences in each strain, we found 74 strains that are homozygous for premature stop codons in their *SUL1* alleles. Among the 31 alleles with premature stop codons, fifteen occur in amino acid positions 155 and 184 (Y155* and Q184**)*, where amino acid sequences are compared to the S288C protein sequence (859 amino acids total). The prevalence of the premature stop codons at positions 155 and 184 may be due to sampling bias, but may indicate functional significance for *SUL1*.

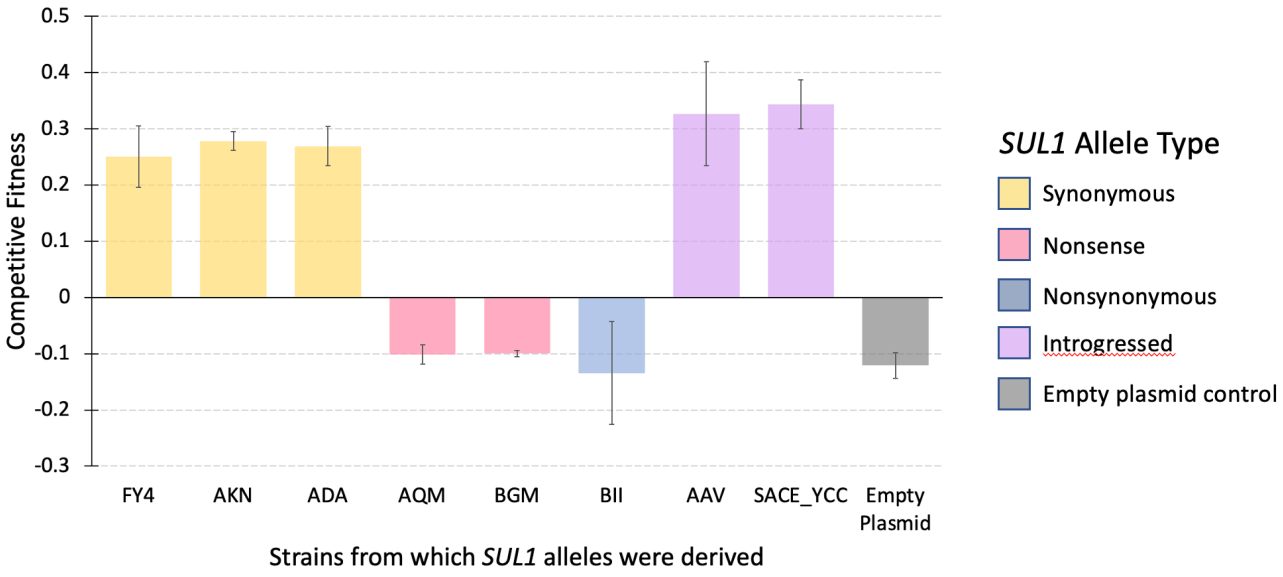


Figure 2.3. Validation of pooled competition through direct competitions of selected natural *SUL1* alleles.

S288C strains transformed with a specific *SUL1* allele on a low-copy plasmid were individually competed against an isogenic GFP-marked strain in the chemostat. Fitness values were calculated by tracking the log-fold change in proportion of non-fluorescent strains and fluorescent strains over 20 generations. These values were used to validate select alleles and their phenotypes observed in the pooled competition. Alleles were selected based on definitive categorization in wild-type-like (pooled competitive fitness close to 0) or loss-of-function (pooled competitive fitness less than -0.10) phenotypes. Of the loss-of-function alleles, AQM and BGM have premature stop codons while BII is loss of function due to nonsynonymous polymorphisms. *SUL1* alleles in AKN, ADA, and BII were done in a prior experiment (Payen et al., in preparation).

Due to the wide range in fitness of alleles with premature stop codons, we investigated whether stop codons that occurred earlier in *SUL1* have a greater impact on function. We found that the location of stop codons in *SUL1* did not dictate the deleterious effects of containing a nonsense mutation (**Supplementary Figure 4A**). However, the fitness of alleles with premature

stop codons at amino acid position 671 consistently have much lower fitness compared to others with premature stop codons elsewhere. This stop codon occurs in the predicted extracellular STAS (sulfate transporter and anti-sigma factor antagonist) domain, which is thought to be crucial for metabolism sensing, and may be further impacting sulfate transport under sulfate-limiting conditions (Sharma et al., 2011).

We compared the standard deviations among barcodes that shared the same loss-of-function alleles to that of barcodes that shared the same wild-type alleles (**Supplementary Figure 4B**). The barcodes linked to loss-of-function alleles varied more in fitness (Welch two sample t-test, $p < 0.005$), although we attribute this variance to increased errors that occur when measuring fitness on a log scale. In regard to magnitude, the barcode counts were reliable, but the barcode counts tended to be less accurate when frequencies were low and continued to decrease through later time points.

In addition to stratifying alleles with premature stop codons and alleles with wild-type phenotypes, we identified alleles with nonsynonymous polymorphisms that also resulted in a loss of function. For instance, alleles that have a single polymorphism resulting in a T669K amino acid substitution showed a loss of function. We also found that alleles with A454P and D483N and alleles with S699L substitutions (and no additional nonsense or promoter polymorphisms) had a loss of function phenotype in our pooled library. Alleles with their polymorphism information, corresponding strain information, and measured fitness values can be found in **Supplementary Table 3**.

We assessed how well the fitness values are reflected in direct competitions by selecting *SUL1* alleles from seven isolates and cloning them individually on the same low-copy plasmid. We transformed S288C haploid yeast with these individual plasmids and competed each allele

directly against an isogenic GFP strain with no plasmid (**Figure 2.3**). Three of the alleles were selected to validate a wild-type-like phenotype and corresponded to the values calculated in the pooled competition. Three other alleles selected showed a loss-of-function phenotype in the pooled competition, which was reflected in the direct competitions. Two of these alleles contained a deletion that resulted in a frameshift (from strains BGM and AQM), and the third allele had nonsynonymous mutations (from strain BII). The BII strain has previously been evolved through sulfate limitation for 150 generations, and it was found that a natural polymorphism that results in a P296L change is responsible for the loss-of-function phenotype (Payen et al., in preparation). In each case, we found the results of the direct competitions recapitulated those found in our pooled competition.

Since we were unable to measure functionality of introgressed alleles in our library, we used the same approach of a direct competition to assay introgressed allele functionality. After validating the fitness of the *SUL1* orthologue from *S. paradoxus* in the *S. cerevisiae* background, which has previously shown high fitness (Sanchez et al., 2017), we also tested the fitness of two alleles that show signatures of introgression from *S. paradoxus*. The two introgressed alleles, despite having over 40 amino acid differences compared to the reference allele, also had a wild-type phenotype (**Figure 2.3**).

2.2.3 *Effects of promoter mutations in natural SUL1 variants*

The fitness distribution across the natural alleles shows alleles with only synonymous site changes in the coding region that nevertheless had a lower competitive fitness compared to strains carrying the wild-type coding sequence from the reference strain (**Figure 2.2**). We reasoned that one explanation is that these alleles may instead carry functional differences in the

noncoding sequences. We found that these alleles share the n.-456G>A polymorphism, and upon further inspection discovered that this SNP is present only in alleles (including those with additional nonsynonymous SNPs) with lower competitive fitness values under sulfate limitation (median competitive fitness = -0.04). Since this competitive fitness value is not as low as alleles with premature stop codons (median competitive fitness = -0.17), it is indicative of an intermediate phenotype. This SNP occurs in a putative Cbf1-binding motif, and binding of the Cbf1 transcription factor is important for growth in sulfate-limiting conditions (Rich et al., 2016; Siggers et al., 2011). The SNP also decreased fitness in a *SUL1*-promoter mutagenesis study, further supporting the functional effects of changes in this motif (Rich et al., 2016).

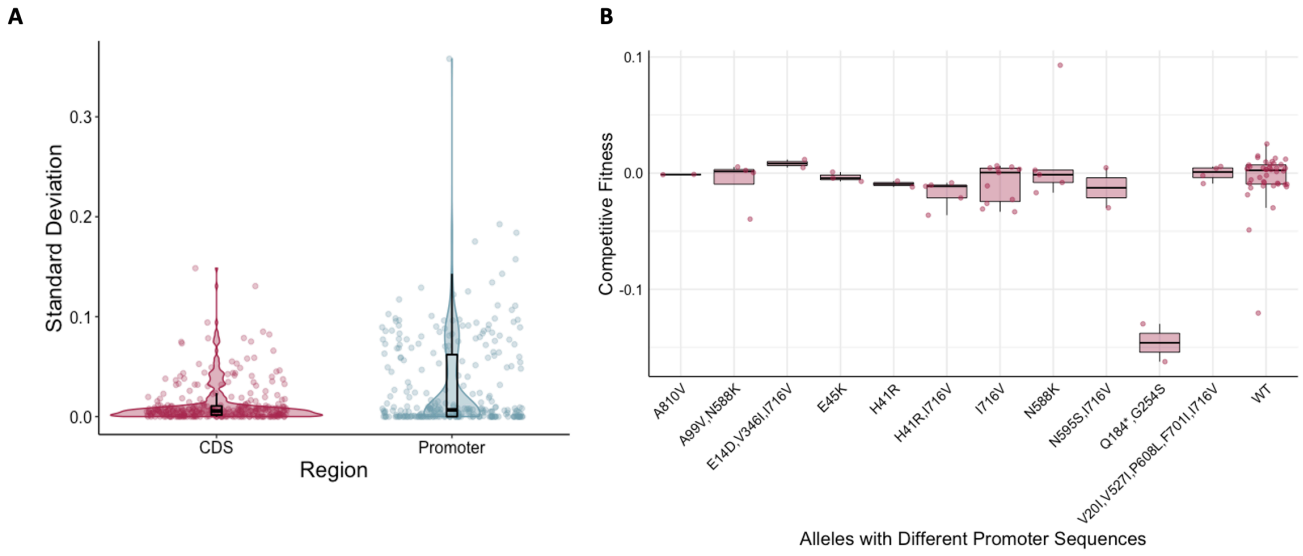


Figure 2.4. Coding polymorphisms are more useful for predicting deleterious effects compared to those in the promoter of *SUL1*.

A) Violin plots of the standard deviations of the competitive fitness for barcodes grouped by those that share the same coding sequence compared with the standard deviation of those that share the same promoter sequence.

B) Boxplots of competitive fitness of the sequences that share the same coding sequence but differ in the promoter sequences.

We used the highest fitness of an allele that contains a premature codon (median competitive fitness = -0.108) and the lowest fitness of alleles without promoter or nonsynonymous polymorphisms (median competitive fitness = -0.0120) to establish a range for other alleles with intermediate phenotypes. Twenty-two unique alleles show an intermediate phenotype, and 9/20 alleles with nonsynonymous polymorphisms also have the n.-456G>A polymorphism. Using these benchmarks, we also identified nonsynonymous changes that did not confer a complete loss of function.

The observation of promoter mutations affecting phenotype in sulfate limitation led us to inspect how much promoter polymorphisms in general contribute to the fitness values observed across the entire allele library. We compared the standard deviation in fitness for sequences that share the same coding sequence to the standard deviation in fitness for sequences that share the same promoter sequences. We found that the coding sequences seemed to determine fitness of a strain more consistently under sulfate limitation (**Figure 2.4A**). That is, alleles with the same promoter sequences had a greater variance in fitness values. Furthermore, alleles that shared the same coding sequences but differed in promoter sequences showed few significant differences in fitness (**Figure 2.4B**). Finally, despite the fact that the promoter mutagenesis study found mutations that could improve fitness under sulfate limitation, we did not identify such polymorphisms among our natural variants. This result may be because the benefit due to regulatory mutations is not strong enough to affect fitness significantly in the natural habitats of yeast.

2.2.4 *Comparing competitive fitness with mutfunc*

With nonsense mutations, loss of function can be predicted based on sequence alone. However, predicting the functional effects of other mutations based on sequence alone is much more challenging. To determine how well these fitness values are reflected in functional computational predictors, we used mutfunc to compare our results to predicted functional effects (Wagih et al., 2018). For each variant, we took the most putatively detrimental mutation and compared its SIFT (Sorting Intolerant From Tolerant) score, which is based on amino acid biochemistry and sequence homology, to the fitness values calculated in our pooled competition assay. While the SIFT scores and our fitness values themselves showed very little correlation (**Supplementary Figure 5A**), we found that most alleles with a loss-of-function phenotype had a low SIFT score (**Supplementary Figure 5B**). Interestingly, many mutations that SIFT predicted would be detrimental actually had a wild-type-like phenotype under sulfate limitation. The differences between the predicted and measured scores highlight the value of experimentally measuring the function of variants, especially in cases for which we need to consider the functional impacts of multiple polymorphisms on the same haplotype.

2.2.5 *Phylogenetics and sequence analysis of natural SUL1 alleles*

To assess phenotypic patterns of *SUL1* on the population level, we annotated a distance-based gene tree of *SUL1* (**Figure 2.5**) with the competitive fitness values we calculated from our pooled competition assay. In our gene tree, we used the *SUL1* allele of *Saccharomyces paradoxus* (CBS432) as the outgroup. We removed branch lengths from these trees to simplify interpretations. Using these annotated trees, we can interpret phenotype in relation to ecological origins and phylogenetic relationships (**Figure 2.5**). We firstly looked at the strains

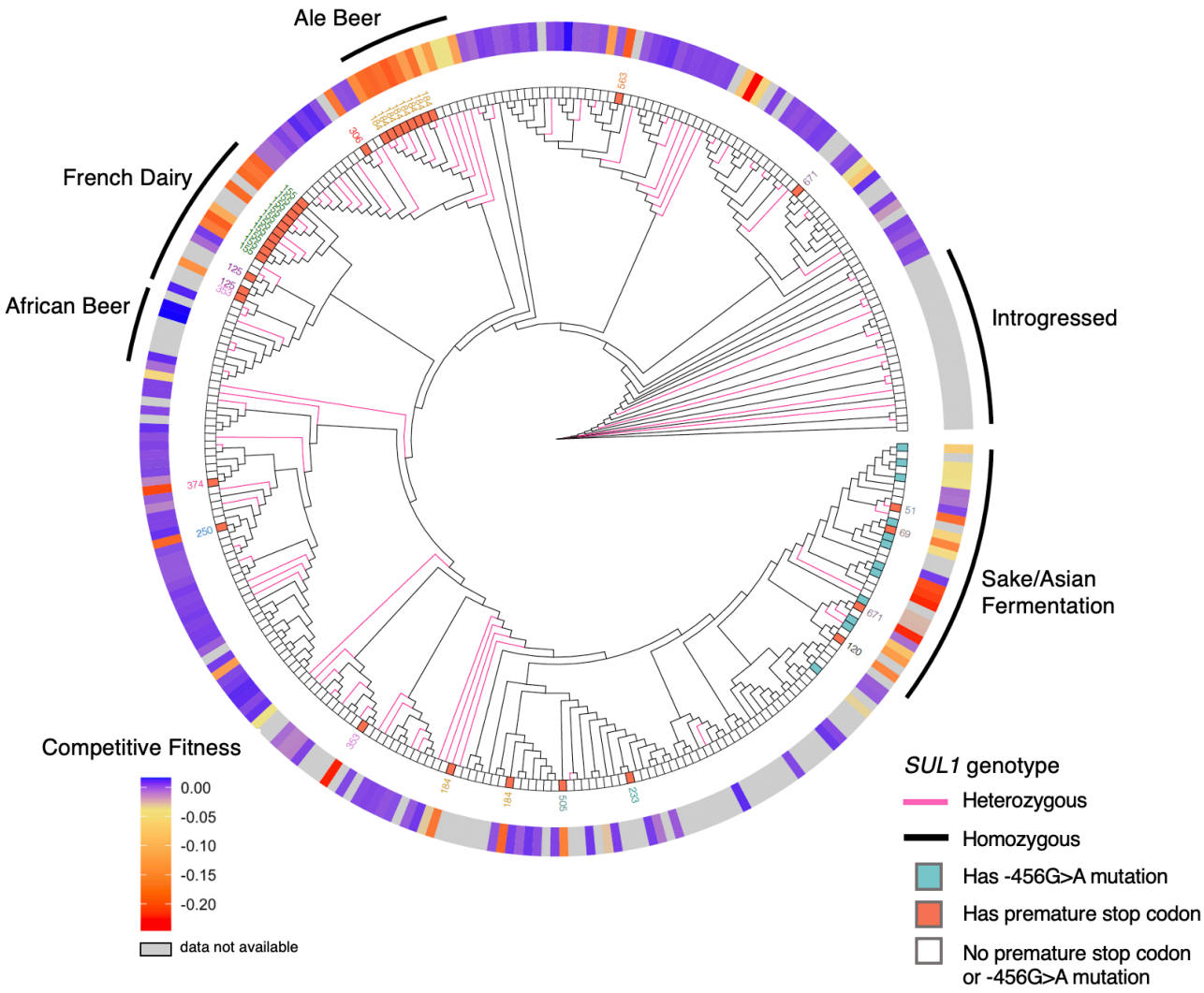
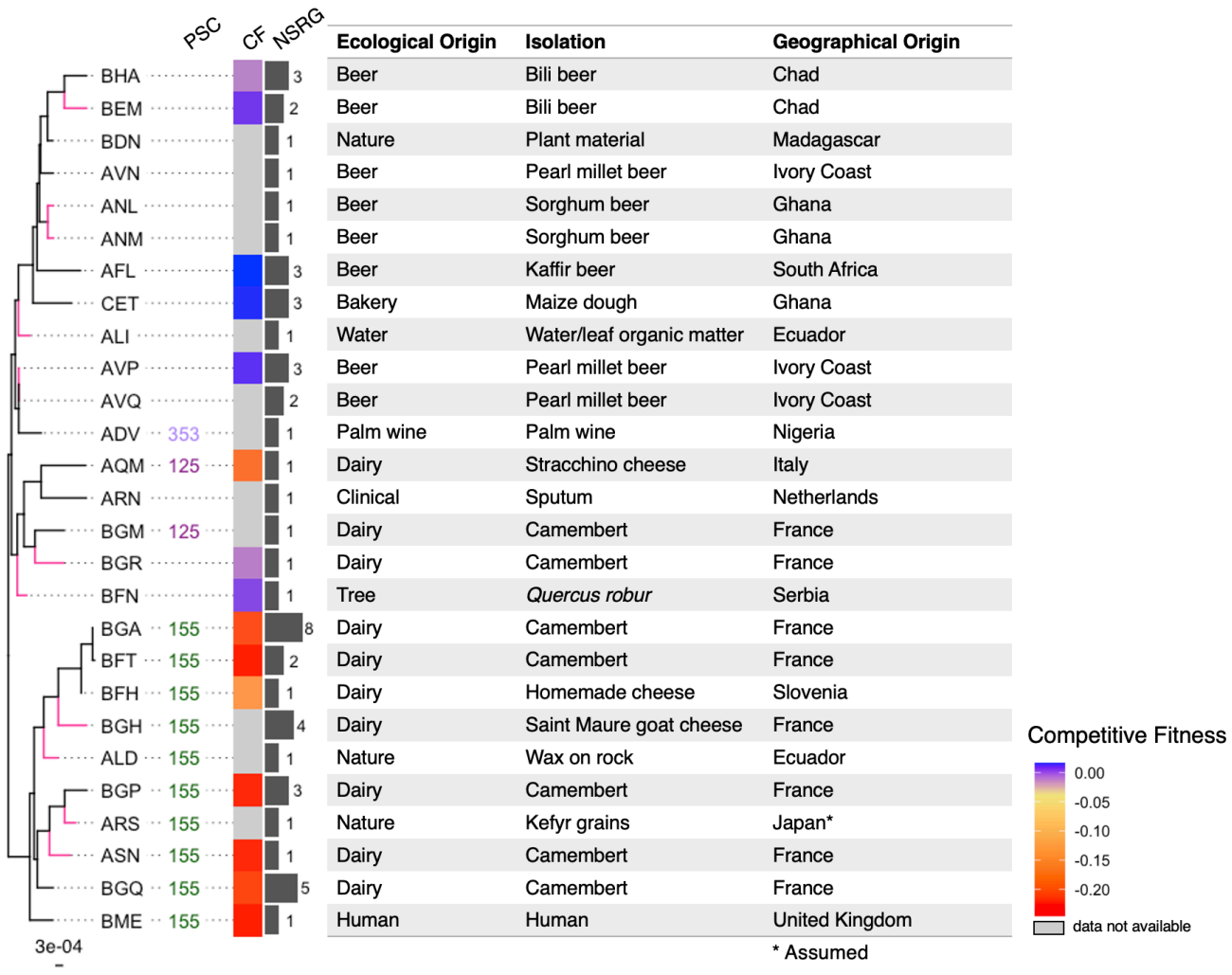


Figure 2.5. **Neighbor-joining gene cladogram generated through PHYLIP using unique genotypes of *SUL1* in the 1,011 strain collection.**
French dairy and sake/Asian fermentation clades both show multiple independent instances of loss-of-function mutations. A stop codon at amino acid position 184 occurs independently in different strains. Color of edges (pink or black) indicates whether the genotypes for those terminal nodes is homozygous or heterozygous. Heterozygous alleles can be derived from diploid, triploid, tetraploid, or even pentaploid strains. Boxes directly adjacent to terminal nodes indicate the genotypes that are homozygous for a premature stop codon (red) or a -456G>A mutation (cyan). Flanking boxes of genotypes with premature stop codons are numbers indicating where in the amino acid sequence the premature stop codon occurred. The ring surrounding the tree denotes the mean *SUL1* competitive fitness values for a given strain's allele on a purple (wild-type-like fitness) to red (loss-of-function fitness) gradient. Labeled regions are generalizations for what comprises most of those clades.

homozygous for premature stop codons in *SUL1*. The polymorphism that results in Q184* does not occur in a singular clade, reducing the possibility that this premature stop codon arose in prevalence as a result of drift or identity by descent. Alleles with Y155* are primarily present in strains isolated from dairy environments in Normandy, France; however, not all dairy strains share the same nonsense mutation (**Figure 2.6**). Two other strains derived from dairy, AQM and BGM, instead have the L125* frameshift mutation. This pattern suggests that a loss-of-function mutation could be beneficial in a dairy environment.

The majority of strains with the detrimental promoter mutation n.-456G>A were isolated from sake or Asian fermentation strains. Additionally, many strains in this clade have a premature stop codon and or nonsynonymous polymorphisms that result in loss of function, which again supports the idea that there may be a trade-off for having a loss-of-function *SUL1* allele since more than one loss-of-function allele sequence exists among these strains.

Based on the distribution of deleterious alleles over the phylogeny, we wondered if these allele differences lead to phenotype differences when the alleles are in their native strain context. We grew all isolates (unmodified) from the 1,011 strain collection on solid minimal media agar plates under sulfate limitation and compared the growth rates to that of the strains pinned on sulfate-abundant minimal media. Interestingly, we found little to no correlation between the growth rates of strains and the competitive fitness values of their *SUL1* alleles (**Supplementary Figure 6A,B**). We additionally looked for growth patterns among ploidy, geographical origins, and clade and found no patterns related to these groupings (**Supplementary Figure 6C**). These results argue that additional background effects beyond the *SUL1 locus* matter for determining fitness in sulfate limitation. Measuring the fitness of the alleles in the library in additional strain backgrounds may help further characterize this genetic complexity.



Ecological Origin	Isolation	Geographical Origin
Beer	Bili beer	Chad
Beer	Bili beer	Chad
Nature	Plant material	Madagascar
Beer	Pearl millet beer	Ivory Coast
Beer	Sorghum beer	Ghana
Beer	Sorghum beer	Ghana
Beer	Kaffir beer	South Africa
Bakery	Maize dough	Ghana
Water	Water/leaf organic matter	Ecuador
Beer	Pearl millet beer	Ivory Coast
Beer	Pearl millet beer	Ivory Coast
Palm wine	Palm wine	Nigeria
Dairy	Stracchino cheese	Italy
Clinical	Sputum	Netherlands
Dairy	Camembert	France
Dairy	Camembert	France
Tree	*Quercus robur*	Serbia
Dairy	Camembert	France
Dairy	Camembert	France
Dairy	Homemade cheese	Slovenia
Dairy	Saint Maure goat cheese	France
Nature	Wax on rock	Ecuador
Dairy	Camembert	France
Nature	Kefyr grains	Japan*
Dairy	Camembert	France
Dairy	Camembert	France
Human	Human	United Kingdom

* Assumed

Figure 2.6. Dairy and African beer subtree of the 1,011 *SUL1* genotypes.
Although dairy strains AQM and BGM share a more recent common ancestor to African beer strains, they show different but independent and homozygous loss-of-function polymorphisms. Color of edges (pink or black) indicates whether genotype for those terminal nodes is homozygous or heterozygous. PSC, amino acid site with premature stop codon (homozygous); CF, competitive fitness; NSRG, number of strains represented by genotype. Boxes around terminal nodes indicate the genotypes that are homozygous for a premature stop codon (red) or a -456G>A mutation (cyan). Scale (bottom left) indicates number of nucleotide substitutions per site.

We calculated the average *dN/dS* value of *SUL1* across all 1,011 strains and found that the value was low (*dN/dS* < 0.2), suggesting that there may be purifying selection on *SUL1*. Additionally, Tajima's *D* statistic suggests that *SUL1* is unlikely to be evolving neutrally (*D* = -2.85). This statistic may indicate that this locus has not reached equilibrium after a bottleneck in the past and is still undergoing expansion. The neutrality index calculated from the McDonald-Kreitman test indicated no evidence of selection (NI = 1.117, Fisher's exact two-tailed test, p = 0.625); however, there are mutations in the *S. cerevisiae* population that are slightly and fully deleterious which have been shown to cause errors in predictions of adaptive evolution using this test (Charlesworth & Eyre-Walker, 2008).

In order to determine whether *SUL1* is exceptional in the prevalence of loss-of-function mutations, we determined the frequency of likely deleterious premature stop codons at all loci in the 1,011 strain collection sequences. Using the sequencing data curated in the 1,011 *S. cerevisiae* strains, we analyzed the coding sequences of genes in the pangenome for premature stop codons that occurred in the first 90% of the gene. We excluded genes that either did not appear in the pan-genome or contained premature stop codons in the pan-genome reference sequences. Grouping these genes enriched in premature stop codons by ecological origins, we found that dairy strains tended to have a consistently higher number of genes that are homozygous for premature stop codons compared to strains isolated from other ecological origins (**Supplementary Figure 7**). Enrichment of dairy strains with premature stop codons is consistent with previous studies that identified enriched loss-of-function alleles among dairy strains that were a result of drift and are important for trait variation (Legras et al., 2018; Zorgo et al., 2012). Of all the genes in the pangenome, 2,465 genes contain a premature stop codon in at least two strains, with 862 of these genes containing premature stop codons in more than 20

strains. Gene Ontology (GO) term analysis revealed that 158 of these 862 genes are involved in ion and/or transmembrane transport. This GO analysis corresponds with previous analyses that found that genes encoding transmembrane proteins tended to be closer to telomeric ends of chromosomes and were more likely to acquire loss-of-function mutations (Bergström et al., 2014). Of the 1601 genes that have premature stop codons in fewer than 20 strains, 284 are involved in catabolic processes (Holm-Bonferroni test/Benjamini Hochberg p-value < 3e-4) and 385 are involved in responses to stimuli (p-value < 6e-5). The number of genes with loss-of-function variants is much greater than the number found in previous studies, likely due to the fact that this dataset has a greater number of strains and much more diversity among strains in regards to factors such as ploidy and isolation origin (Bergström et al., 2014; Jelier et al., 2011).

2.3 Discussion

Assessing the phenotype of alleles on a species-wide scale is crucial for understanding how quantitative traits vary in a population. Previous approaches for experimentally identifying causal variants use DNA synthesis or mutagenesis, and in many cases do not reflect the alleles found in natural populations. We developed here a high-throughput and low-cost functional approach that can measure the fitness of nearly all alleles present in a population. Specifically in our study, we investigated the function of 228 natural variants of *SUL1*, a high-affinity sulfate transporter gene, present in the 1,011 *S. cerevisiae* strain collection. Our assay identified instances of functional, intermediate, and loss-of-function phenotypes. Using these data, as well as gene and whole genome sequencing data, we related *SUL1* fitness to its evolutionary history. *SUL1* acquired multiple independent instances of loss of function, the majority of which were due to premature stop codons. Other alleles suffered from frameshift, nonsynonymous, and

promoter polymorphisms that negatively affect fitness. These multiple independent instances provide evidence that there may be a fitness trade-off for yeast to have a loss-of-function *SUL1* allele. The strains carrying these loss-of-function alleles were largely isolated from dairy, beer, and sake clades. Because not all loss-of-function polymorphisms were identical in each clade (for instance, there are three different premature stop codons among dairy strains), these events were likely not due to drift but have a functional benefit instead. An alternative explanation is that some strains, including those from dairy environments, naturally carry a high burden of loss-of-function polymorphisms, and *SUL1* could simply represent an easily tolerated loss that is recurrent by chance. As speculated in previous studies, enriched loss-of-function events in specific populations are thought to arise as a result of genetic drift and play an important role in maintaining genetic variation (Legras et al., 2018; Zorgo et al., 2012).

Nevertheless, there is some evidence that a loss-of-function *SUL1* allele may confer a trade-off and be beneficial under particular environments. There are toxic analogues of sulfate, such as chromate and selenate, that can be transported into the cell through the Sul1 permease (Cherest et al., 1997; A. J. Johnson et al., 2016). Other compounds, such as cadmium, that affect cell function and growth are toxic due to the uptake of sulfate by Sul1 (X. Zhang et al., 2020). These results show instances in which having a functional copy of *SUL1* would be detrimental and suggest that *SUL1* may have some antagonistic pleiotropic effects. Having multiple functions may also explain the lack of gain-of-function alleles in our library, as having a higher-affinity *SUL1* may not be beneficial in natural environments. Despite the results from previous studies, many of which investigated the effects of toxic compounds in lab strain backgrounds similar to what we used here, we have been unable to recapitulate these trade-offs.

Identifying loss-of-function alleles by searching for premature stop codons is relatively straightforward. Additionally, we found that many of the nonsynonymous polymorphisms were predicted by mutfunc to have a deleterious effect, although many of these predicted deleterious polymorphisms were false positives. Moreover, the effects of polymorphisms in regulatory regions are more challenging to predict computationally. Using natural variation, we have identified instances where a single polymorphism (n.-456G>A) in a predicted transcription factor-binding site affects fitness of cells under sulfate limitation, a result that was also apparent in our prior promoter mutagenesis study (Rich et al., 2016).

Our approach also identifies intermediate phenotypes, many of which in our pool were likely a result of a natural promoter polymorphism that affects expression. In studying variants, it is challenging to identify deleterious mutations in a population. Here, we illustrate an example showing the importance of studying both coding and noncoding polymorphisms, as both normal expression and protein structure affect phenotype and thus how selection acts on a population.

While some *SUL1* alleles have single polymorphisms that can result in a total loss of function, there were also alleles with several nonsynonymous mutations that had wild-type-like fitness under sulfate limitation. Notable examples include the two *SUL1* alleles found across 21 unique isolates that had signatures of introgression from *S. paradoxus*; these alleles had over 40 amino acid differences, yet they functioned normally in the S288C background. These results support our previous findings that *SUL1*'s high affinity has been maintained across *S. paradoxus* and *S. cerevisiae* (Sanchez et al., 2017), and the fitness measurements of the introgressed alleles support the idea that these sequences maintain their function even in a new genetic background context. The wide variation in *SUL1* function under sulfate limitation is stark, and using these natural variants has provided further evidence for non-neutral evolution.

In this study and our prior study, we found no correlation between *SUL1* function and yeast growth on sulfate-limited media (Payen et al., in preparation). Despite the fact that *SUL1* copy number increases in evolution experiments under sulfate limitation, the fitness of endogenous copies of *SUL1* did not necessarily dictate cell performance under sulfate limitation. One possible reason for this observation is that these strains contain functional copies of the *SUL1* paralog, *SUL2*. Despite being a lower affinity sulfate permease compared to *SUL1*, we found no strains that were homozygous for obvious loss-of-function *SUL2* alleles. A possible explanation for this lack of loss-of-function mutations in *SUL2* is that between the two paralogous genes, *SUL1* is evolving new or diverging function while *SUL2* maintains the original function of transporting sulfate. The alleles of *SUL2* and other transporters like *SOA1* likely also play an important role in growth under sulfate-limiting conditions. Alternatively, small growth rate changes may not be observable in our solid media growth rate assays compared to what is possible to measure in chemostat culture.

In sum, leveraging the technologies available in high-throughput Illumina and PacBio sequencing, we present here a widely applicable and affordable approach for assaying hundreds of natural variants in high-throughput. Assaying natural variants in this manner is especially useful when coupled with whole-genome sequencing data, as it allows us to better understand function in relation to molecular evolution. Furthermore, our method compares many alleles of a gene in isolation in an otherwise isogenic background away from the complexities of genetic background interactions. This approach complements methods like QTL mapping, providing a more thorough investigation of phenotypic patterns across an entire species, which can also contribute to our understanding of how pleiotropic a gene is. Further application of this approach in other genes and other genetic backgrounds will be greatly beneficial to our understanding of

how selection acts on natural populations and how multiple polymorphisms contribute to function and ultimately phenotype.

2.4 MATERIALS AND METHODS

2.4.1 *Strains and plasmids*

Natural isolates from the 1,011 Saccharomyces cerevisiae collection were used to isolate natural variants of SUL1 (Peter et al., 2018). Strains pinned on yeast extract peptone dextrose (YPD) agar plates were transferred to liquid YPD in 96-well plates, grown overnight at 30°C, and stored in 30% glycerol at -80°C. The FY3 S288C strain DBY7284 (MATa ura3-52) was used for transformation and competition experiments (described below). A GFP-marked strain YMD1214 (MATa hoΔ::GFP-KANMX) that has neutral fitness under sulfate limitation was used for validation competition assays. Prototrophic FY3 (DBY11069), YMD4321 (MATa ura3-52 sul1Δ::URA3-KANMX), YMD4322 (MATa ura3-52 sul2Δ::URA3-KANMX), and YMD4323 (MATa ura3-52 sul1Δ::URA3-KanMX sul2Δ::URA3-KanMX) were used to validate growth rates on sulfate-limited and sulfate-abundant agar plates. A pRS316 vector with an NruI site inserted in the BamHI site (YMD2307) was used in this study for molecular cloning and competitions described below. A complete list of strains and plasmids can be found in **Supplementary Table 2**.

2.4.2 *Plasmid and yeast library generation*

Strains from the 1,011 *S. cerevisiae* collection were pooled together from colonies on a solid agar plate. Genomic DNA was then extracted using the QIAGEN Genomic-tip 100/G kit.

Natural variants of *SUL1* were amplified with primers designed to hybridize to conserved regions 844 bp upstream of the translation start site and 262 bp downstream of the stop codon (oligos 1 and 2, **Supplementary Table 1**). The promoter region contains 39 bp that overlap with the coding region of the upstream gene *VBA2.* Oligo 2 also contained an 8 bp randomized sequence to serve as a barcode. PCR was performed using KAPA HiFi Hotstart Readymix with the following cycling conditions: 95°C for 3 min, then 19 cycles of 98°C for 20 seconds, 60°C for 15 seconds, and 72°C for 4 minutes. Final extension was at 72°C for 4 minutes, and then the reaction was cooled to 4°C. The barcoded product was purified using the DNA Clean and Concentrator kit from Zymo Research and assembled into an NruI-digested plasmid via Gibson assembly. Chemically competent *E. coli* cells were transformed with the product using heat shock at 42°C, and >20,000 transformants were collected and pooled. Plasmids were extracted from the pooled transformants using Wizard® *Plus SV* Miniprep DNA Purification Kit and then used to transform yeast (DBY7284) using 100 µL of 2 M lithium acetate, 800 µL of 50% 4000 polyethylene glycol, 100 µL of 1M dithiothreitol, and 50 µL of 10 mg/mL of carrier DNA. Approximately 6,000 Ura+ yeast transformants were collected for pooled competition experiments and for PacBio sequencing (**Figure 2.1A**).

2.4.3 *Linking barcodes with full-length variants*

Plasmids were extracted from the yeast transformant pool using Zymoprep Yeast Plasmid Miniprep II (Zymo Research). Plasmid fragments containing the barcode and variant were isolated using M13/pUC primers with KAPA HiFi Hotstart Readymix and the following cycling conditions: 95°C for 3 min, then 13 cycles of 98°C for 20 seconds, 60°C for 15 seconds, and 72°C for 4 minutes. The final product was extracted from a 0.5% agar gel using Qiagen's Gel

Extraction kit and cleaned using Ampure PB beads (Pacific Biosciences). Two PacBio libraries were made using the SMRTbell™ Template Prep Kit 1.0 (Pacific Biosciences) and sent to University of Washington PacBio Sequencing Services for sequencing and Sequel II circular consensus sequence (CCS) analysis.

BAM files of CCS reads were aligned to the plasmid reference file using BWA/0.7.13 mem (Heng Li, 2013). Reads that were aligned to the reference were piped to a new BAM file with Samtools/1.9 (H. Li et al., 2009). These reads were also analyzed with cigar strings to validate alignment of PacBio reads. From there, the barcodes were extracted, and a barcode-variant map was generated that contained a file with all of the barcode-variant reads and all of the highest quality reads for each barcode, as previously described (Matreyek et al., 2018). Since the resulting barcode-variant map still showed a considerable number of insertion and deletion errors, we used a multiple sequence alignment of all the reads that shared the same barcodes to eliminate additional sequencing errors. Alignments were done using MUSCLE (v.3.8.31) (Edgar, 2004). Any further ambiguous nucleotides were resolved by performing a pairwise alignment against the highest quality read (EMBOSS Needle v. 6.4.0) (Needleman & Wunsch, 1970).

To match PacBio reads to strains in the 1,011 collection, reference sequences were first extracted from the GVCF in the 1,011 collection genome data using BCFtools consensus. We then used regular expressions to search for reads that were putatively derived from these reference sequences. We removed barcodes that contained only one CCS read or were not represented in our barcode sequencing analysis (**Figure 2.1B**).

2.4.4 *Pooled library competition in chemostats*

Sulfate-limited media (3mg/L ammonium sulfate) was prepared as previously described (Gresham et al., 2008; Payen et al., 2014). Four 50-mL chemostat culture vessels were filled with 20 mL of media at 30°C and inoculated with 1 mL of the yeast transformant pool. This culture was grown for 24 hours, after which the pumps were turned on and the culture switched to a continuous culture system at a dilution rate of about 0.17 volumes per hour (~3.4 mL/hour in a 20 mL culture). Samples were taken twice a day for 5 days, or about 25 generations. For each sample, 1 mL was stored in 25% glycerol at -80°C, and another 1 mL was used for plasmid extraction (**Figure 2.1C**).

2.4.5 *Barcode sequencing and analysis*

For each time point and replicate from the pooled library competition, plasmids were again extracted using the Zymoprep Yeast Plasmid Miniprep kit. One replicate was discarded due to technical errors. Barcodes were isolated and amplified using forward oligo 25 and indexed reverse oligos 26-40 and 120-128 that included Illumina Nextera sequencing adaptors. KAPA HiFi Hotstart Readymix was used with 1 µL of 1X SYBR™ Green I and the following PCR cycles: 95°C for 3 min, then 17-19 cycles of 98°C for 20 seconds, 60°C for 15 seconds, and 72°C for 15 seconds. The reaction was run on a Bio-Rad MiniOpticon (Bio-Rad) to avoid overamplification. PCR products were cleaned using Ampure XP Beads (Agencourt) and quantified using the KAPA Library Quantification Kit for Illumina® Platforms (Roche). Libraries were sequenced on a NextSeq sequencer (Illumina) with sequencing oligos 41 (Read 1), 44 (Read 2), and 100 (Index). Paired-end reads were merged using PEAR/0.9.5 (Jiajie Zhang et al., 2014). Using FitSeq, we calculated the fitness of each barcode for each given replicate (F.

Li et al., 2018). FitSeq normalizes each pool to account for experimental error between replicates, providing a more accurate readout of fitness. The fitness values were then normalized by the average fitness of barcodes associated with the wild-type (S288C) alleles. The effects of single mutations were also compared with predicted consequences of mutations from mutfunc (Wagih et al., 2018) (**Figure 2.1C**).

2.4.6 *Pairwise fitness assays in chemostats*

To assess fitness of strains carrying individual alleles, 300 µL of a liquid culture of each strain was inoculated into a chemostat containing 20 mL of sulfate-limited media at 30°C. For each of the competitors, one competitor strain contained a plasmid with an extra copy of *SUL1* and the other isogenic strain contained a neutral GFP marker. Additionally, competition experiments of strains carrying each allele being assayed were conducted in at least two biological replicates. Cultures were grown for 24 hours before switching to a continuous culture system. Once cultures achieved steady state, the competing cultures were mixed at a 1:1 ratio (around 72 hours). Cultures were competed for 15 generations after mixing and sampled twice daily (approximately every 3-6 generations). For each sample, cultures were assayed for percent GFP cells with a BD Accuri C6 flow cytometer (BD Biosciences). Competitive fitness values were calculated by plotting ln(number of dark cells/number of GFP+ cells) over about 25 generations and taking the linear slope of the linear regression from this data.

2.4.7 *Measuring the growth of the 1,011 isolates on solid media*

Solid sulfate-limited media (3mg/L ammonium sulfate) was prepared by adding 2% agarose to liquid sulfate-limited media and poured in PlusPlates (Singer Instruments). Solid

sulfate abundant media was prepared by adding ammonium sulfate (to 5g/L) to the sulfate-limited media. To ensure the depletion of sulfate in the cells, all 1,011 natural isolates were grown overnight (~14 hours) on solid sulfate-limited media. The isolates were then replicated in quadruplicate on solid sulfate-limited and sulfate-abundant media. Photos of the colonies were taken every 12 hours for 3 days and the R package gitter was used to calculate the size of each colony in the photos (Wagih & Parts, 2014). For each time point on both limited and abundant conditions, we subtracted the colony size at the first time point from the colony size at subsequent time points (colony size = $size_t$ - $size_{t=0}$). Growth rates were calculated by taking the average of the ratio of the colony size in limited media over the colony size in abundant media across 72 hours.

2.4.8 *Phylogenetic tree generation and sequence analysis*

To generate our phylogenetic trees, *SUL1* sequences from the 1,011 strains and from *S. paradoxus* strain CBS432 were aligned using MUSCLE (Edgar, 2004). The genetic distances for *SUL1* alleles were calculated using the maximum-likelihood-based distances through DNADIST in the PHYLIP package (Felsenstein, 2005). A gene tree for *SUL1* was then generated using the NEIGHBOR program, and the final tree was visualized and annotated using R/ggtree (Yu et al., 2017). *dN/dS* was calculated using the yn00 function in PAML.

To determine the prevalence of loss-of-function mutations across all genes in the 1,011 strains, we used sequences from the core ORFs in the pangenome as reference sequences and identified which strains were homozygous for premature stop codons in each of the core ORFs (Peter et al., 2018). Premature stop codons that occurred in the last 90% of an ORF were not included, as previous studies have shown that these mutations would not necessarily cause a

significant loss of function (Bergström et al., 2014). Gene Ontology (GO) analysis was conducted using Yeastmine (accessed May 22, 2020) and both Benjamini-Hochberg and Bonferroni test corrections were used to account for multiple testing (Balakrishnan et al., 2012).

2.4.9 *Data availability*

Raw sequencing data can be found in the Sequencing Read Archive (BioProject Accession PRJNA681436 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA681436). Scripts and full version of the Supplemental Tables used for this paper can be found at https://github.com/dunhamlab/SUL1_natural_variants. All alleles, matched strains, barcodes, fitness, coding mutations, and noncoding mutations can be found in **Supplementary Table 3.**

2.5 ACKNOWLEDGMENTS

Thank you to Andreas Tsouris for help phenotyping natural isolates of sulfate limited media. Thanks to Noah Hanson for performing the barcode sequencing on my samples. Thank you to Dr. Anne Clark for providing the analysis on introgressed *SUL1* alleles, Dr. Mary Kuhner for consultation and expertise on the best approach for creating and interpreting phylogenetic trees, and Dr. Pengyao Jiang and Dr. Kelley Harris for assisting with visualization of phylogenetic trees. I would also like to thank Fangfei Li and Dr. Sasha Levy for their assistance with FitSeq used to calculate fitness values, and thank you to the UW PacBio sequencing core for help with amplicon sequencing.

Chapter 3. Development of PacRAT: a program to improve barcode-variant mapping from PacBio long reads using multiple sequence alignments

---

Work described in this chapter is derived from "PacRAT: a program to improve barcode-variant mapping from PacBio long reads using multiple sequence alignments" published in *Bioinformatics*. Simulations and figures were done in collaboration with Clara Amorosi and Soyeon Showman. Reanalysis of the *CYP2C9* library was performed by Clara Amorosi.

Use of PacBio sequencing for characterizing barcoded libraries of genetic variants is on the rise. However, current approaches in resolving PacBio sequencing artifacts can result in a high number of incorrectly identified or unusable reads. Here, we developed a **Pac**Bio **R**ead **A**lignment **T**ool (PacRAT) that improves the accuracy of barcode-variant mapping through several steps of read alignment and consensus calling. To quantify the performance of our approach, we simulated PacBio reads from eight variant libraries of various lengths and showed that PacRAT improves the accuracy in pairing barcodes and variants across these libraries. Analysis of real (non-simulated) libraries also showed an increase in the number of reads that can be used for downstream analyses when using PacRAT.

3.1 Background

Improvements in sequencing technology have greatly expanded our capacity not just for detecting genetic variants, but also for assessing their function. Approaches such as deep mutational scanning (DMS) use a library of variants to characterize the impact of genetic variation

on protein structure, gene expression, and gene function across many fields including protein science (McLaughlin et al., 2012; Newberry et al., 2020; Thompson et al., 2020; Tinberg et al., 2013), evolutionary biology (Hietpas et al., 2013; Hoggard et al., 2016; Rich et al., 2016)), and pathology (Chiasson et al., 2019; Gray et al., 2019; Matreyek et al., 2018; Starita et al., 2015; Suiter et al., 2020). Such variant libraries are commonly "barcoded" with short random DNA tags, and this barcoding allows for short-read sequencing of multiple time points and removes the step of directly sequencing entire variants at every time point. Barcodes are associated with library variants; this association is a simple process if the variable region and the linked barcode can be sequenced together with short-read methods. Hierarchical tag-directed sequencing, known as subassembly, converts longer sequences (>600 bp) into short-read-accessible fragments that ultimately get assembled into the full-length sequence. However, with longer variable regions and/or distance to the barcode, subassembly becomes increasingly costly, challenging, and time-consuming (Hiatt et al., 2010). To overcome the labor-intensive and expensive approach of subassembly, many studies are turning to PacBio sequencing to generate high-quality long reads that span the entirety of the barcode and variable region (Amorosi et al., 2021; Matreyek et al., 2018; Ollodart et al., 2021; Suiter et al., 2020).

Compared to Illumina short-read sequencing, PacBio generates much longer continuous reads. Whereas Illumina sequencing using clonal populations of molecules to generate highly accurate reads, PacBio sequencing directly retrieves sequence information from each individual molecule. Molecules are immobilized at the bottom of a zero-mode waveguide sequencing cell, where a polymerase then traverses the molecule and emits a unique fluorophore for each nucleotide added (Buck et al., 2017). Longer reads overcome previous limitations, though they have the disadvantage of an increased error rate, which can be as high as 15% (Wenger et al., 2019). This

is due to a low signal-to-noise ratio from fluorophores cleaved off from the zero-mode waveguide present in the sequencing cell. Errors are not prone to clustering in AT- or GC-rich regions, like they are in next-generation sequencing. Rather, errors are dispersed fairly evenly throughout a molecule (Buck et al., 2017). To reduce errors, highly accurate reads are generated from circular consensus sequencing (CCS), which circularizes molecules by ligating single-stranded SMRTbell™ adaptors and allowing a single DNA molecule to be sequenced multiple times. Each pass through the molecule produces a subread, and these subreads are collapsed to form a consensus sequence that improves the quality score of a sequence (Rhoads & Au, 2015; Wenger et al., 2019).

Although PacBio chemistry has improved, errors are still pervasive in CCS reads. Despite the fact that errors are unbiased, CCS errors still tend to cluster as insertion/deletion (indel) errors in homopolymer regions due to suboptimal sequence alignments that are performed when retrieving consensus sequences. Compared to Illumina NovaSeq, the PacBio mismatch rate is 17X lower, but indel rates are 181X higher (Wenger et al., 2019). Such indel errors are more prominent in CCS reads with lower numbers of subreads, often those from longer target sequences, as fewer complete passes are contained within the average read length. In a recent study that performed PacBio barcode-variant mapping, more than 30% of barcodes were associated with indels and were discarded during analysis due to variation in length (Matreyek et al., 2018). Some libraries include variants that have real indels, which cannot always be ignored or discarded. Therefore, additional computational error correction steps are required to improve reliability and maximize information gained from PacBio consensus reads.

In this work, we present **Pac**Bio **R**ead **A**lignment **T**ool ("PacRAT") which maximizes the number of usable reads for barcode-variant mapping while reducing sequencing errors found in

CCS reads. We tested the utility of PacRAT by simulating eight libraries of various read lengths and analyzing two real libraries. Our approach reduces costs of sequencing by utilizing reads that otherwise were assigned incorrect sequences in previous methods. Pac-RAT also provides researchers with metrics for estimating the necessary read coverage for different types of libraries, an approach that allows for more efficient multiplexing of variant libraries.

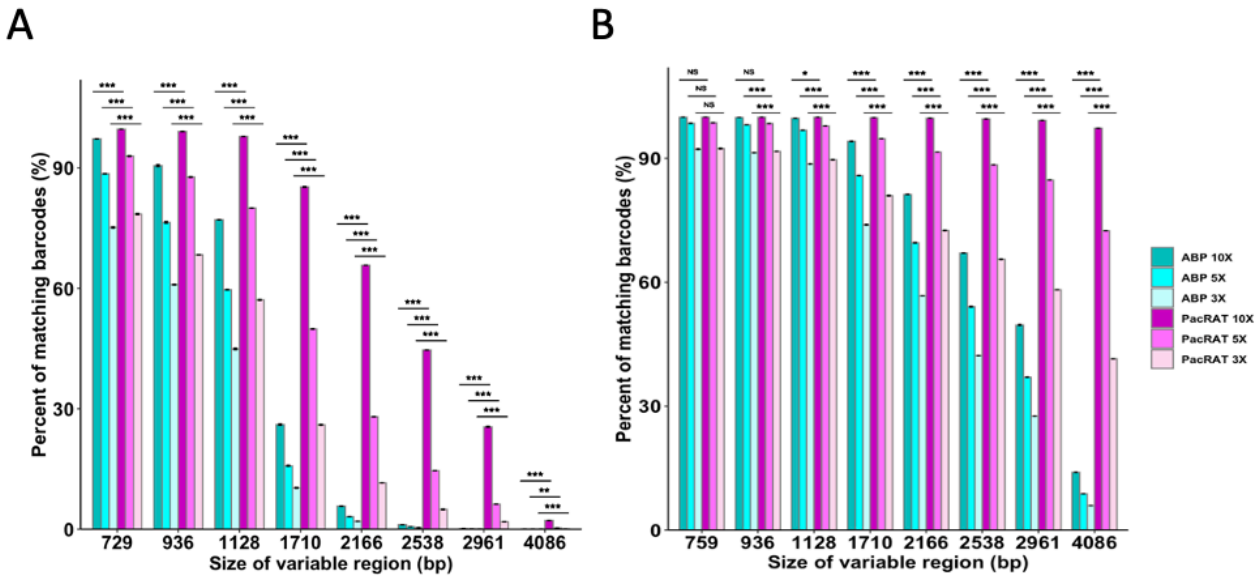


Figure 3.1. Comparison of correct variant identification between ABP and PacRAT in simulated libraries with different lengths of variable region.

Variant libraries of eight different variable region lengths (759, 936, 1128, 1710, 2166, 2538, 2961, and 4086 bp) were generated in triplicate using SimLoRD (Stöcker et al., 2016) at 10X, 5X, and 3X read coverage. For both plots, the number of correctly identified variants using ABP and PacRAT were compared. **A)** Simulations were performed with default SimLoRD Chi-squared parameters (-xn 1.8923e-03 2.5394e+00 5500 -xs 0.0121 -5.12 675 48303.073 1.469). **B)** Simulations performed with adjusted Chi-squared parameters (-xn 8e-03 2.54 5500 -xs 2.78e-03 -0.48 825 48303.073 1.469), chosen to better simulate longer variable regions (**Table 3.1**). Significance values are derived from a one-way ANOVA test and Tukey *post hoc* pairwise comparisons (**Supplementary Table 4**): Not significant (NS), $p < 0.05$ (**), $p<0.0005$(***).

3.2 RESULTS

3.2.1 *Simulation of deep mutational scanning libraries to assess PacRAT performance*

To assess the accuracy of PacRAT, we computationally generated eight DMS libraries with a range of sizes (759-4086 bp). These DMS libraries were created from cDNA sequences found in the human genome that span a range of lengths. Using SimLoRD, a PacBio SMRT read simulator (Stöcker et al., 2016), we simulated CCS reads in triplicate from these DMS libraries with 10X, 5X, and 3X read coverage (**Figure 3.1**). Using SimLoRD's default Chi-squared

Gene	**length (bp)**	**mean number of passes**	**real or simulated library**	**default or adjusted SimLoRD parameters**
MSH2	4675	11.55	real	NA
4086 bp simulated	4086	4.34	simulated	default
4086 bp simulated	4086	9.65	simulated	adjusted
CYP2C9	1625	19.45	real	NA
1710-bp simulated	1710	6.51	simulated	default
1710-bp simulated	1710	14.86	simulated	adjusted

Table 3.1. Comparison of mean passes through SimLoRD simulated and real libraries.
We determined that the average pass numbers produced using SimLoRD's default Chi-squared (-xn 1.8923e-03 2.5394e+00 5500 -xs 0.0121 -5.12 675 48303.073 1.469) did not accurately reflect those of real PacBio libraries. Thus, we adjusted these parameters (-xn 8e-03 2.54 5500 -xs 2.78e-03 -0.48 825 48303.073 1.469) to better reflect the number of passes found in Sequel II data.
as variable region lengths increase since MSA is able to correct many of the errors present in longer libraries. 5X and 3X downsamples of our libraries correctly matched 72% and 41% of barcodes from the longest simulated library, respectively. In a smaller 2.1 kb library, we observed that 99.7%, 91.5%, and 72.7% of barcodes were correctly identified with PacRAT at 10X, 5X, and 3X coverage, respectively. These results indicate that PacRAT can accommodate reduced sequencing read requirements while maintaining accuracy of data, which could reduce library sequencing costs and improve throughput.

parameters (-xn and -xs), which are designed to model P4 chemistry (RS II), PacRAT increased the number of correctly identified variants in libraries sized between 1700-2500 bp, but its effectiveness dropped rapidly after 2500 bp. (**Figure 3.1A**). The average pass numbers produced using these default parameters did not accurately reflect those of real PacBio libraries (**Table 3.1**). Thus, we adjusted these parameters (described in Stöcker et al., 2016) to better reflect the number of passes found in Sequel II data (**Table 3.1**). A one-way ANOVA test and *post hoc* Tukey test showed that barcode-variant mapping significantly improves with PacRAT in nearly all libraries (**Figure 3.1**, **Supplementary Table 4**). For the largest simulated library (4086 bp), PacRAT correctly identified the variant for 97% of barcodes, whereas ABP only identified 14% (**Figure 3.1**). Efficiency decrease is amplified as variable region lengths increase since MSA is able to correct many of the errors present in longer libraries. 5X and 3X downsamples of our libraries correctly matched 72% and 41% of barcodes from the longest simulated library, respectively. In a smaller 2.1 kb library, we observed that 99.7%, 91.5%, and 72.7% of barcodes were correctly identified with PacRAT at 10X, 5X, and 3X coverage, respectively. These results indicate that PacRAT can accommodate reduced sequencing read requirements while maintaining accuracy of data, which could reduce library sequencing costs and improve throughput.

3.2.2 *Assessing PacRAT performance through reanalysis of published libraries*

To verify that our program improves barcode variant mapping with real PacBio data, we reanalyzed the PTEN and TPMT libraries sequenced from Matreyek et al., 2018. These older libraries were sequenced with an RS II instrument, which has a much lower read quality than newer chemistry (Sequel and Sequel II). The software for converting raw RS II sequencing files to .bam

Library	***CYP2C9* without PacRAT**	***CYP2C9* with PacRAT**	***MSH2* without PacRAT**	***MSH2* with PacRAT**
Total CCS reads	432,500		273,308	
Total unique barcodes recovered from PacBio	118,896		12,081	
# barcodes associated with on-target mutations	36,108	39,847	2,687	3,441
# barcodes associated with indels	66,761	61,551	7,004	6,209
# barcodes associated with WT sequences	1,565	1,723	1,632	1,919
# barcodes associated with off-target mutations (excluding indels)	14,462	15,775	854	615
# of unique on-target variants (observed/expected)	8,980/15,648	9,261/15,648	236/241	241/241
Coverage of on-target variants	4.0	4.3	18.3	22.2

Table 3.2. *CYP2C9* and *MSH2* library statistics.

The *MSH2* library was synthesized from Twist Biosciences and sequenced on a Sequel II machine (Sequel 3.0) using one flow cell (Ollodart et al., 2021). The *CYP2C9* library is the activity library described in Amorosi et al., 2021, and was sequenced on a Sequel II machine (Sequel 2.1) using two flow cells. For all PacRAT analyses, parameters --cutoff 1 and --threshold 0.6 were used.

files (bax2bam) and for generating CCS reads (ccs2) now exist as legacy softwares and are no longer maintained by PacBio. Despite version and compatibility issues, we were able to reanalyze the libraries with more recent versions of the software. For the PTEN library (variable region length 1398 bp), we found that PacRAT decreased the number of reads discarded due to indels by 9.2% and increased the number of usable reads by 11.6%. PacRAT ultimately increased the number of unique variants in the library by 11.3%. For the TPMT library (variable region length

1016 bp), the number of indels decreased by 9.1%. The number of usable reads increased by 14% and the number of unique variants in the pool increased by 15.1%.

We also re-analyzed two published variant libraries created for genes *MSH2* (Ollodart et al., 2021) and *CYP2C9* (Amorosi et al., 2021), both of which were sequenced on a Sequel II machine. Additional information on the generation and characteristics of these libraries can be found in **Table 3.1** and **Table 3.2**. The *MSH2* library was generated at 22X coverage via DNA synthesis, where specific single amino acid variants were requested from Twist Biosciences. PacRAT decreased the number of unexpected variants from 16.8% of barcodes analyzed with ABP to 10.5% with PacRAT (**Table 3.2**). We additionally analyzed smaller subsets of this library and observed similar results. We are unable to distinguish if these errors are sequencing or synthesis errors.

For the ~118,000-barcode *CYP2C9* library, which was generated by NNK mutagenesis at each codon via inverse PCR, PacRAT increased the number of barcodes associated with on-target variants from 36,108 barcodes to 39,847 barcodes, while reducing the number of barcodes associated with indels from 66,761 barcodes to 61,551 barcodes. Barcodes associated with indels are expected to have low functional scores, but with ABP only, we observed 3,043 barcodes associated with indels that had scores similar to wild-type variants, likely indicating a variant misclassification. PacRAT was able to reduce the number of these misclassifications by 48.4%.

3.3 Discussion

PacRAT improves upon existing PacBio library barcode-variant mapping pipelines by implementing multiple sequence alignments to resolve the variant sequence associated with each barcode. Our method is reference-free and resolves sequence artifacts, commonly indels. While

not all reads will map to an expected variant, users can adjust thresholds to determine the ideal parameters for each specific library.

Importantly, because PacBio sequencing is still much more costly than next-generation sequencing, this approach provides users with an idea for how to multiplex libraries for one single PacBio run. Each library sequenced with current PacBio sequencing chemistry (Sequel II) can return nearly 3 million CCS reads. The simulations performed here with PacRAT provide users with an idea for how much sequencing coverage an individual library needs for highly accurate barcode-variant mapping and how many libraries can be multiplexed in one run. To reduce sequencing costs further, PacRAT utilizes reads that were assigned incorrect sequences or otherwise discarded in previous methods, thereby maximizing the number of usable reads in each library.

3.4 MATERIALS AND METHODS

3.4.1 *PacRAT algorithm and development*

PacRAT is a reference-free, highly reliable approach for linking barcodes to variants and improves upon the output of a previously used method, Assembly By PacBio (ABP) (Matreyek et al., 2018). ABP assigns barcodes to variants by taking either the highest-frequency sequence associated with each barcode or highest average quality sequence associated with each barcode. Crucially, this assigned variant often retains sequencing errors (most commonly 1 bp indels), depending on the coverage and quality of reads associated with each barcode.

To fix this issue, our program assigns barcodes to variants by aligning CCS reads with the same barcode using multiple sequence alignment (MSA) (Edgar, 2004). If a sequencing error

occurs in a barcode, it will not be includd in the alignment process. A consensus sequence is then generated from these alignments. If no ambiguous nucleotides are present, this consensus sequence is used as the final, error-corrected sequence. However, if ambiguities in these alignments persist, a second pairwise alignment is performed with EMBOSS Needle (Rice et al., 2000) between the highest average quality CCS read and the derived consensus sequence. Sites with ambiguities are resolved by taking the nucleotide sequence from the highest quality read. Thresholds for determining consensus sequences can be specified by the user, with the default requiring that the majority of sequences share the same base at each position.

3.4.2 *Simulation deep mutational scanning (DMS) libraries*

The simulated DMS libraries were generated using custom scripts (https://github.com/dunhamlab/PacRAT) by generating all possible amino acid substitutions at each site. Once these libraries were generated, SimLoRD version 1.0.4 was used to simulate a PacBio sequencing run. For each library, we specified the number of reads with the parameter –num-reads, and these numbers were selected based on desired sequencing coverage. For default parameters, -xn and -xs parameters were not included. For adjusted parameters, we used -xn 8e-03 2.54 5500 -xs 2.78e-03 -0.48 825 48303.073 1.469. We aligned using the Burrows-Wheeler Aligner (version 0.7.17) with options -C -M and -L 80. Extra flags in bam file were removed and CIGAR string files were generated using samtools (version 1.9). Barcode-variant maps were generated using scripts from AssemblyByPacBio (https://github.com/shendurelab/AssemblyByPacBio). Comparisons between reference sequences and *in silico* PacBio reads returns the error correction rates between ABP and PacRAT.

3.4.3 *Reanalysis of PTEN/TPMT libraries*

The libraries from Matreyek et al., 2018 were downloaded from SRA and converted from raw sequencing files to .bam files using the program bax2bam version 0.08. Circular consensus sequences (CCS) were generated using the ccs function in SMRT Link version 6.0. The resulting libraries were aligned to a reference sequence using bwa version 0.7.10 with options -C -M -L 80 and flagged with samtools version 1.9. Output files were processed through Assembly by PacBio and the subsequent results were compared with results from PacRAT using custom scripts.

3.5 ACKNOWLEDGEMENTS

Thank you to Phil Green for his suggestion of this approach. Thank you to Jochen Weil, Atina Cole, and Fritz Roth for the helpful suggestions to improve PacRAT. Thanks to Nick Popp for beta testing PacRAT, and thank you to William Noble for his advice on what platform this program should be developed on.

Chapter 4. High-throughput approach for understanding functional differences between paralogous genes in the maltose utilization pathway

Experiments performed in this chapter were done with the assistance of Andrea Chang.

In Chapter 2, I described an approach for assessing allele function on a species-wide scale. This chapter describes an application of that approach to answer a different biological question: How do paralogous differences affect phenotype throughout different clades in the *Saccharomyces cerevisiae* population. This approach is applied to *MAL33*, a gene that encodes an activator protein that facilitates the import and utilization of maltose as a carbon source in *S. cerevisiae*. Not only are the functional differences among the alleles of *MAL33* unknown, but the functional differences among the five other paralogs of *MAL33* (*MAL13, MAL23, MAL43, MAL63,* and *MAL73*) are largely unknown as well. By determining the function of *MAL33* on a species-wide scale, I begin to understand how maltose utilization is varies throughout different clades in the population. Comparative sequence and phylogenetic analysis among *MAL13, MAL33*, and *MAL73* portray how vastly different the paralogs are in terms of number of unique alleles, prevalence of premature stop codons, and range in copy number. This chapter highlights an exciting application to understand genes on a population level and the potential to comprehensively compare paralog function.

4.1 Background

Paralogs, or genes that arose from a duplication event in an ancestral genome, play a critical role in evolution. Functional redundancy of paralogs permits genes to diverge and acquire

mutations, resulting in pseudogenization or neofunctionalization that may affect fitness in new or changing environments (Guan et al., 2007). While paralogous genes share a significant amount of sequence identity, natural allelic differences within even just one gene in a paralogous group contribute to phenotypic variation in a population. Although some studies have pinpointed polymorphisms that contribute to functional differences among paralogs, few have investigated how phenotypic variation among multiple paralogs affects fitness of an organism, especially on a species-wide scale (Leh-Louis et al., 2004; Noree et al., 2019; Will et al., 2010).

The human genome is highly redundant as more than 13,000 of about 19,000 protein-coding genes have at least one paralog (T. Wang et al., 2015). Most studies of the human genome do not account for gene redundancy, and understanding functional differences between redundant regions becomes particularly important to understand when determining how cells respond to therapeutics. Due to recent developments in high-throughput genotype-to-phenotype approaches, we now have the capacity to study not just allelic differences within a gene, but allelic variation among paralogous groups of genes as well.

The work herein describes the beginnings of an application of the high-throughput approach detailed in Chapter 2 to analyze natural allelic function among paralogs, namely from activator genes in the *MAL* loci. *S. cerevisiae* has several *MAL* loci, each of which consists of a group of three genes that are essential for the transport (*MALx1*), regulation (*MALx3*), and hydrolysis (*MALx2*) of maltose (**Figure 4.1**). Several *MAL* loci exist (*MAL1, MAL2, MAL3, MAL4, MAL6*), all of which have diverged in sequence but still retain similar and complementary functions (**Figure 4.2A**) (Charron et al., 1989). Genes in each locus are named based on the locus in which they exist. For instance, the *MAL* genes in the *MAL1* locus are termed *MAL11, MAL12,* and *MAL13.* Copy number differences within each paralogous locus may also exist

(Duan et al., 2018). Allelic variation in these loci are present primarily in domesticated strains, hypothesized to have formed due to outcrossing of diverse wild isolates (Duan et al., 2018). Maltose utilization is a major characteristic of domestication in yeast; most domesticated strains—including yeast isolated from sources without maltose like milk and honey—can use maltose as the sole carbon source, but wild isolates typically cannot (Bai et al., 2022). Although wild isolates do contain *MAL* genes, the reason behind why these isolates largely cannot use maltose is still unknown.

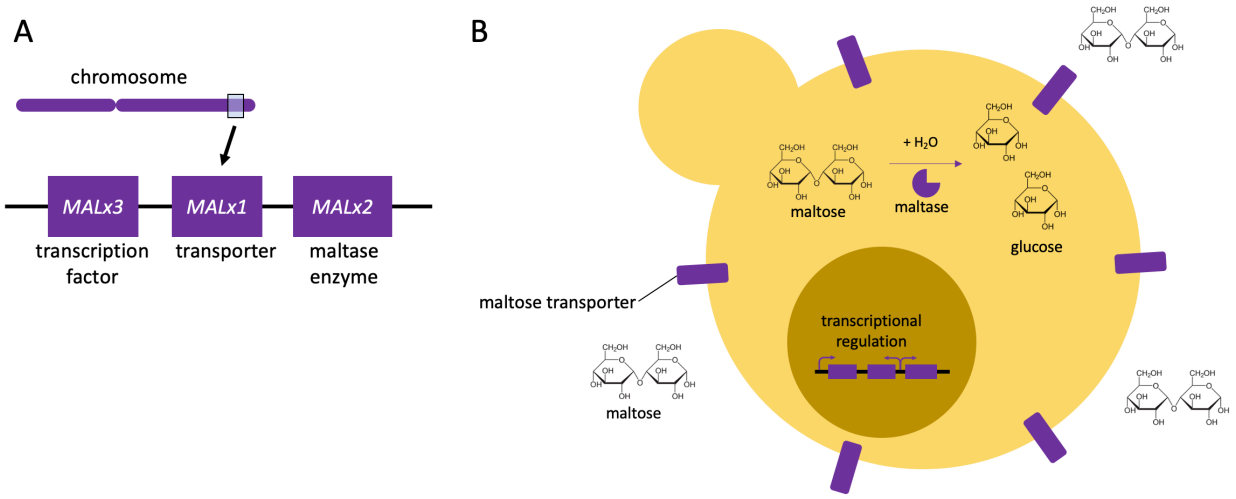


Figure 4.1. The *MAL* locus permits cells to use maltose as the sole carbon source.
A) Each *MAL* locus contains three genes that encode a transporter (*MALx1*), an enzyme that catalyzes the hydrolysis of maltose (*MALx2*), and a transcription factor that regulates the expression of genes in the locus (*MALx3*). **B)** Visual representation of the three genes necessary for growth in maltose.

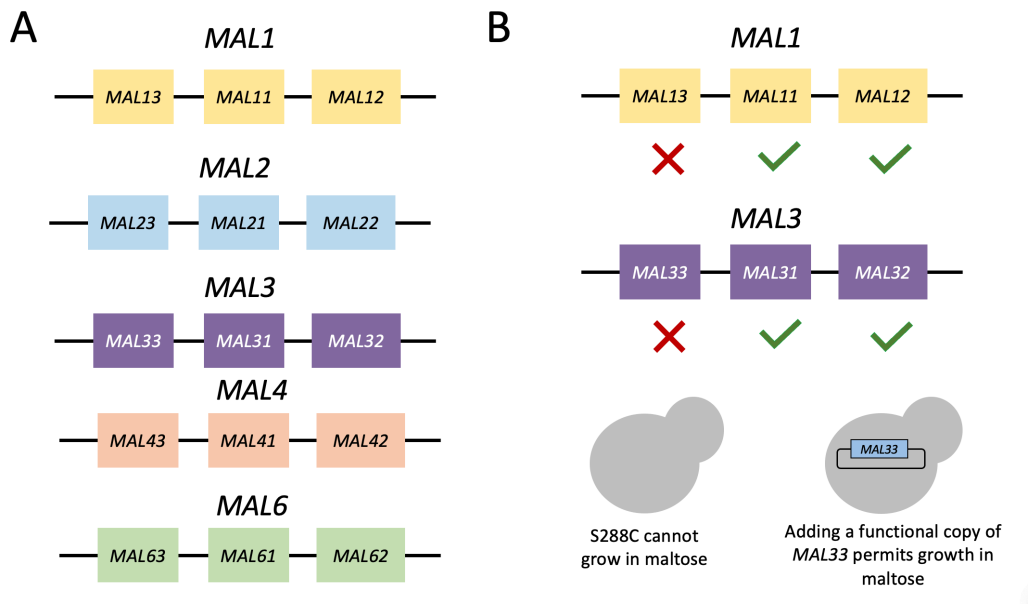


Figure 4.2. The *MAL* locus has duplicated several times in *S. cerevisiae*.
A) All known *MAL* loci across all strains of *S. cerevisiae*. Each locus contains a *MALx1, MALx2*, and *MALx3* gene. **B)** The lab strain S288C contains two of the known loci, *MAL1* and *MAL3*. The transporter (*MALx1*) and maltase (*MALx2*) genes are functional; however, the transcription factor-encoding gene (*MALx3*) is not, and thus S288C cannot grow using maltose as its sole carbon source. An addition of a functional *MALx3* allele can rescue growth.

MAL loci are approximately 9.0-kb in length and are typically located near telomeric regions (**Figure 4.1**) (Charron et al., 1989; Naumov et al., 1994). The location of these genes strongly suggests that duplication of these loci occurred through translocation of the genes to different chromosomes. Not all strains contain the same duplications or number of *MAL* loci. Furthermore, not all strains are able to utilize maltose due to loss of function in one, two, or all three genes in the *MAL* loci. For instance, the lab reference strain S288C has two *MAL* (*MAL1* and *MAL3*) loci, both of which have functional maltose transporter (*MALx1*) and maltase (*MALx2*) genes, but nonfunctional alleles of the *MAL*-activator protein (*MALx3*) (**Figure 4.2B**). Despite decades of work understanding the structure of maltose loci, determining functionality of

alleles in these paralogs is still difficult, especially given the diversity across the entire *S. cerevisiae* population. Variation exists between and within paralogs, adding an additional layer of complexity in understanding allele function (**Figure 4.3**).

Because paralogs are complementary, the loss of function of *MALx3* genes (which encode the activator protein) in S288C makes the *MALx3* paralogs ideal candidates for testing function across paralogous alleles (**Figure 4.4**). The S288C reference strain only contains copies of *MAL13, MAL33*, or *MAL73* (which is also known as *YPR196W,* see footnote[*]). In order to sample all natural *MALx3* alleles in a species-wide manner, we used genomic DNA from a

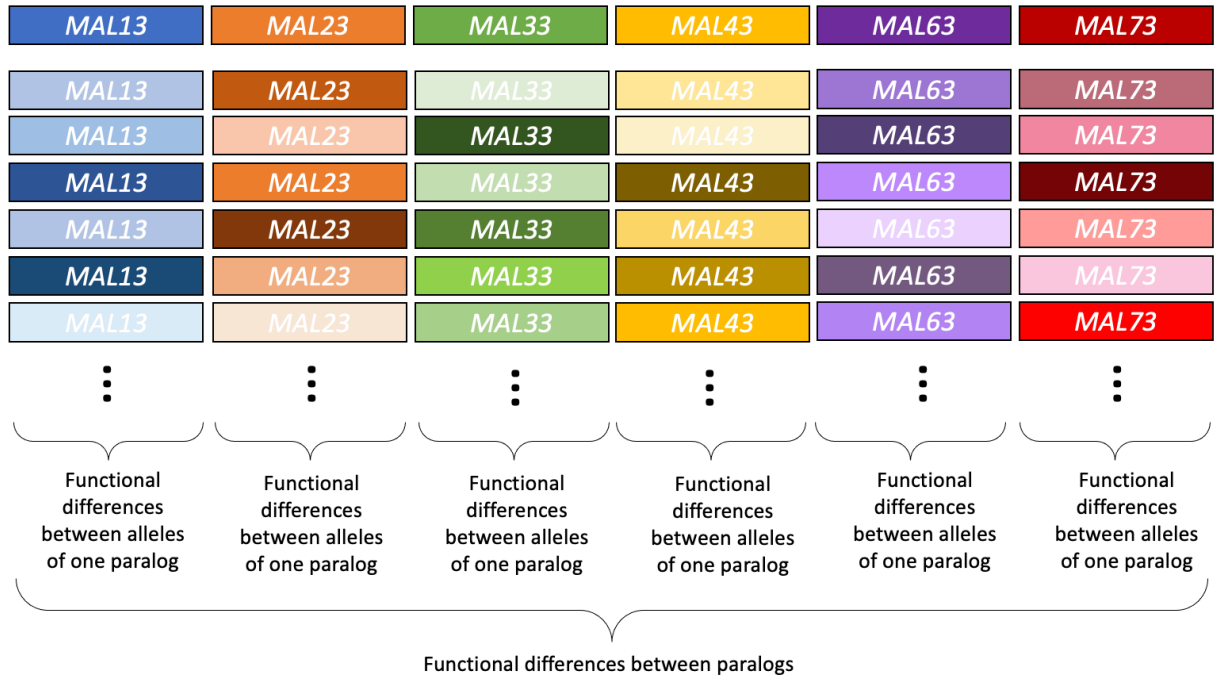


Figure 4.3. Functional differences can exist within a gene and between all of its paralogs. Six different paralogs of *MALx3* exist across the *S. cerevisiae* population. Not all *S. cerevisiae* have all six paralogs, and the copy number of each paralog can vary (Brown et al., 2010; Duan et al., 2018). The function of each paralog can differ between isolates (Brown et al., 2010).

[*] *MAL7* is not a known or characterized *MAL* locus comprising all three genes in a canonical *MAL* locus. Prior to the characterization study done in Ohdate et al., 2018, the gene was known only as a putative maltose-response transcription factor located on chromosome XVI.

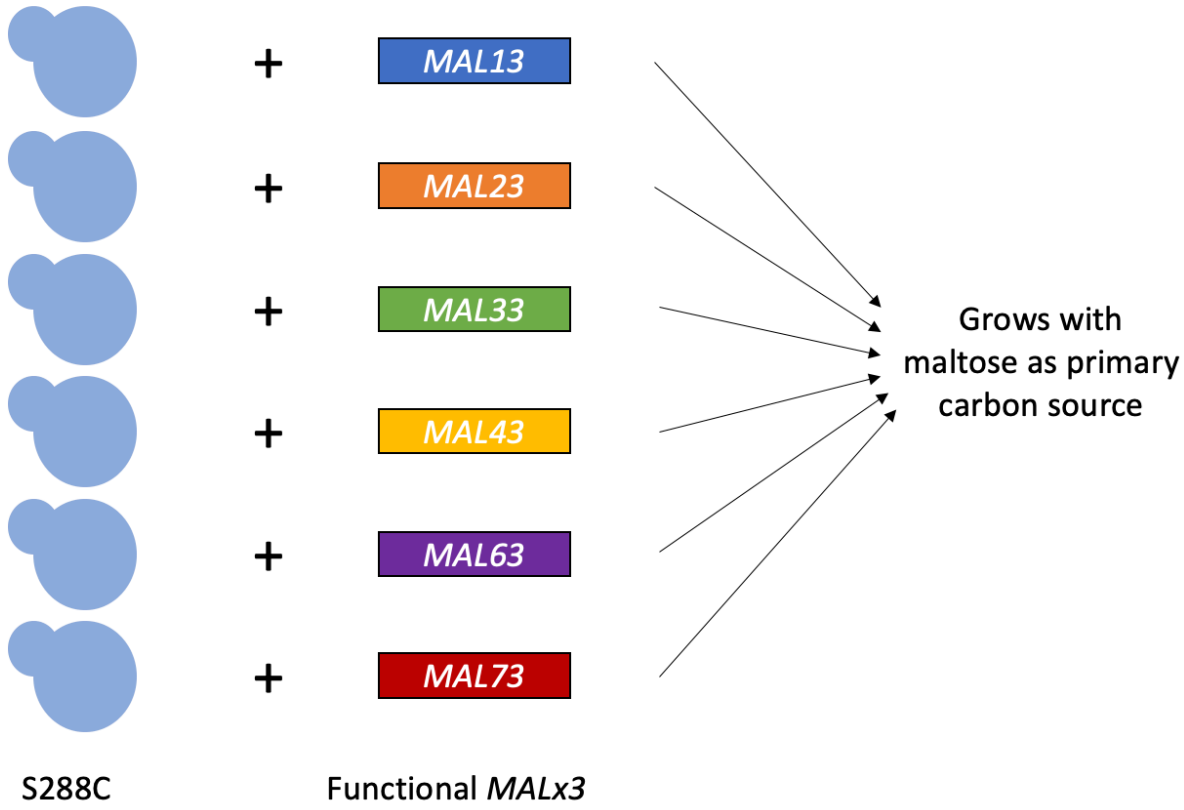


Figure 4.4. Paralogs of the *MAL* gene complement each other.

S288C does not have a functional copy of *MALx3*, but has functional copies of *MAL11, MAL12, MAL31, and MAL32.* Because the paralogs of *MAL* genes complement each other, we can use *MALx3* to test the functional differences of paralogs on a species-wide scale by competing allele libraries in maltose limiting media. Any functional *MALx3* should rescue growth, including ones from loci that are not present in S288C (*MAL23, MAL43, MAL63)* (Charron et al., 1989).

collection of 1,011 strains isolated from diverse geographical and ecological origins. We started our assay by first validating that the lab strain transformed with a plasmid containing a functional *MALx3* allele from the *MAL3* locus (*MAL33*) rescues growth in maltose. Next, we tested the growth of strains under different maltose conditions to determine the maltose-limiting concentration. We then developed a library of *MAL33* and competed this library in maltose-limited chemostats. Although these experiments are still underway, PacBio long-read sequencing

and barcode sequencing will reveal a lot about *MAL33* function across the *S. cerevisiae* population. Insights into *MALx3* paralog function across all alleles will not only reveal how maltose utilization has evolved in *S. cerevisiae*, but they will also emphasize the importance of investigating paralog allelic differences within the human genome for understanding phenotypic variation and responses to therapeutics.

4.2 RESULTS

4.2.1 *Validation that a functional MALx3 can rescue growth in maltose*

Before proceeding with building a barcoded allele library, we first validated that 1) the reference strain S288C (FY) cannot grow using maltose as its sole carbon source, and 2) a functional *MALx3* allele on a low-copy plasmid in the background of the reference strain can rescue growth in maltose for the reference strain. We grew the S288C in 2% maltose for 72 hours and saw no changes or increases in optical density. Next, we created a small library for three *MALx3* genes (*MAL13, MAL33,* and *MAL73*), transforming the lab strain with the library, and observing whether cells grew. We designed primers in conserved regions that flank *MAL13*, *MAL33*, and *MAL73*. Using the 1,011 natural isolate collection (Peter et al., 2018), we isolated and amplified the three genes from pooled genomic DNA, creating a small library of ~250 *E. coli* transformants and ~2,000 yeast transformants per paralogous gene. The yeast mini libraries were grown in 2% maltose flasks for seven days. After 48 hours, we observed growth in the *MAL33* library, but did not see any growth in the *MAL13* and *MAL73* libraries. Lack of growth in the *MAL13* and *MAL73* libraries may have been due to insufficient sampling of the 1,011 natural isolates. Alternatively, lack of growth could have been the inability of *MAL13* and *MAL73* to complement the *MAL33* in the S288C background.

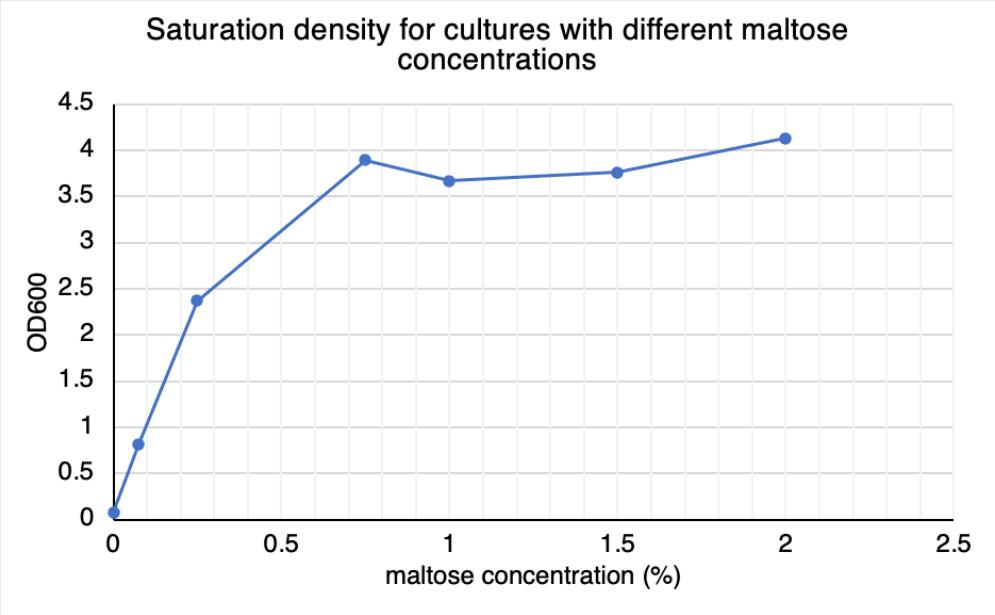


Figure 4.5. Saturation densities of cultures grown in flasks with differing maltose concentrations

Cultures with greater than 0.25% maltose had approximately the same saturation density, indicating that those cultures had a sufficient amount of maltose for cells to utilize as their sole carbon source, and thus the carbon source was not limiting.

4.2.2 *Determining the function of MAL33 on a species-wide scale in the chemostat*

In order to compare functions of *MALx3* alleles through competitions in the chemostat, we determined which concentration of maltose is limiting to facilitate competition between cells carrying different *MALx3* alleles. Although prior studies have tested the growth of cells in maltose limitation in the chemostat, we performed control tests to determine the optimal maltose concentration (Jansen et al., 2004). We first tested growth in flasks of the reference strain transformed with a functional *MAL33* on a low-copy plasmid. Despite prior studies using 0.75% maltose at a dilution rate of 0.1 hr^{-1} (Jansen et al., 2004), our results support that 0.75% maltose is not limiting (**Figure 4.5**). Thus, we opted to use 0.25% maltose with carbon-limiting

chemostat media. We also tested cultures in the chemostat to determine the dilution rate and found that cells were able to remain at steady-state at a dilution rate of 0.15 hr^{-1}.

We next created a more complex library of *MAL33* using the same approach described in Section 4.2.1, but instead with barcoded primers. These primers captured 802 bp upstream and 598 bp downstream of the *MAL33* coding sequence. *MAL33* is substantially more polymorphic than *SUL1*, with 840 unique genotypes among the 1,011 natural isolates. We collected ~11,000 yeast transformants and competed them together for ~20 generations. Although results are still being collected, we found ~20,000 unique barcodes in the starting pool. Barcode sequencing over these 20 generations will reveal a fitness distribution, and PacBio long read sequencing will provide information about each haplotype contributing to growth in maltose limited media.

4.2.3 *Phylogenetic and sequence analysis of MALx3 alleles*

The reference strain does not contain a premature stop codon in its *MAL33* allele, yet the allele is nonfunctional. Loss of function across multiple alleles may be the case for many *MAL33* alleles (**Figure 4.6**). Notably, not many clades labeled as "Not domesticated" have premature stop codons, despite the fact that it is uncommon for wild strains to have the capacity to grow using maltose as the sole carbon source. It is particularly striking that the number of unique *MAL33* alleles is so high, yet the polymorphisms that render the gene nonfunctional are generally unknown (apart from premature stop codons). In addition to having a high number of polymorphisms across the entire population, *MAL33* also varies in copy number (**Figure 4.6, Figure 4.7**). Variation in copy number further complicates our ability to predict which strains can utilize maltose, as the different copies of *MAL33* in one genome can all have multiple alleles with varying function.

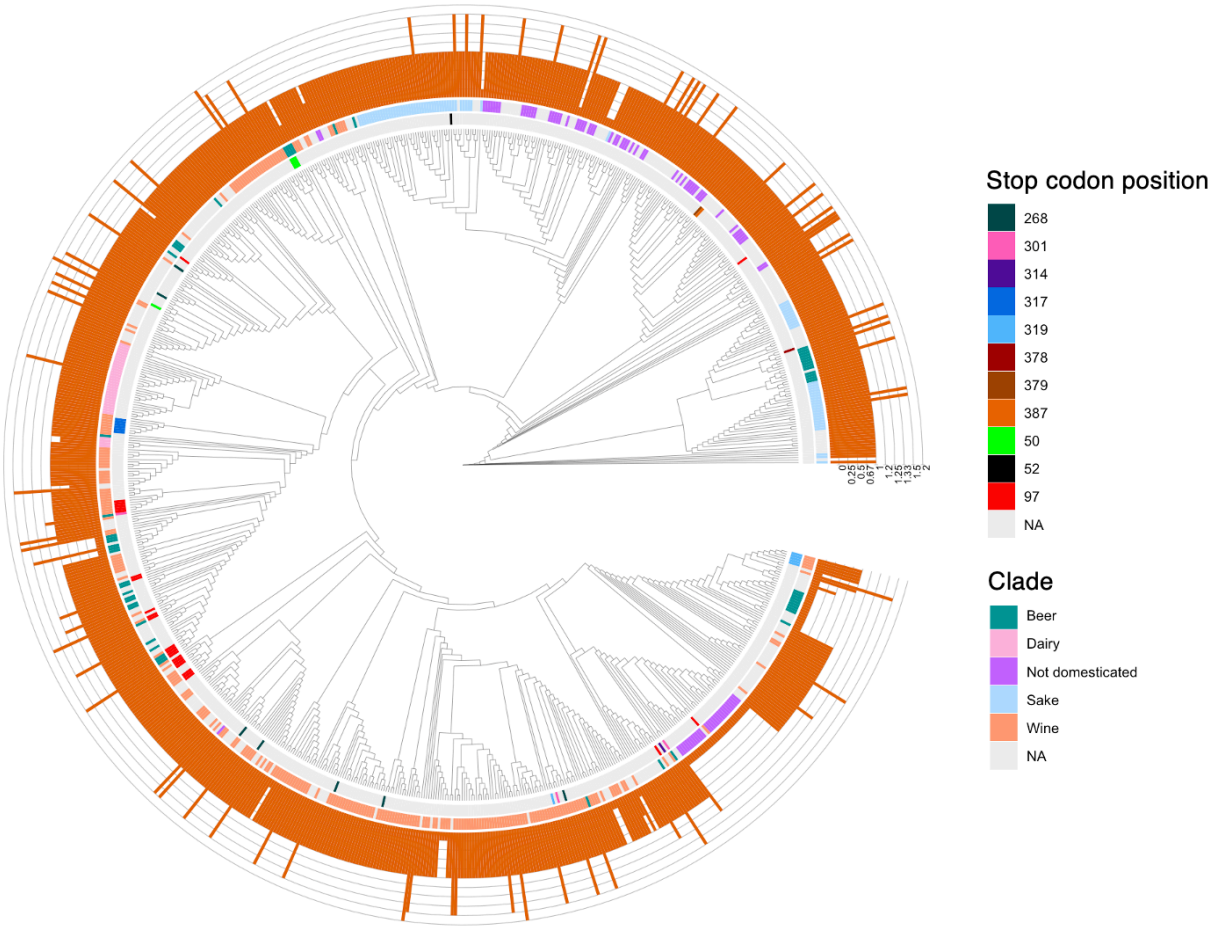


Figure 4.6. Neighbor joining cladogram of unique *MAL33* alleles.
The innermost ring adjacent to the cladogram indicates whether a homozygous premature stop codon exists in the *MAL33* allele for a particular genotype. Premature stop codons are categorized by color; light gray indicates genotypes that are not homozygous for premature stop codons. The second innermost ring labels the clade from which the genotype is derived. The outermost ring with gridlines indicates an approximate copy number for strains with those genotypes.

Further confounding these results is the fact that some sequences for *MAL33* were called for strains that were reported to have 0 copies of *MAL33* in the genome (**Figure 4.6**). This discordance is likely a result of paralogs aligning to the *MAL33* locus: the S288C reference strain does not have copies of *MAL23, MAL43,* or *MAL63* and therefore does not contain these

sequences in the reference genome. Thus, if other isolates contain sequences from these three paralogs, they were likely to have been mapped to the *MAL33, MAL13*, or *MAL73 (YPR196W)* locus.

Although *MAL33* is relatively polymorphic, the number of polymorphisms across the *MAL33* gene is particularly higher at about 777 bp into the gene through the end of the 3'UTR (**Figure 4.7**). This base pair position is at amino acid position 260 in the MAL33 protein, which is the initial location where many of the observed premature stop codons occur (**Figure 4.6**). The nuclear localization signal (amino acid position 41-49) for MAL33 is also relatively well conserved. Other predicted domains, such as the DNA-binding domain, occur early in the protein

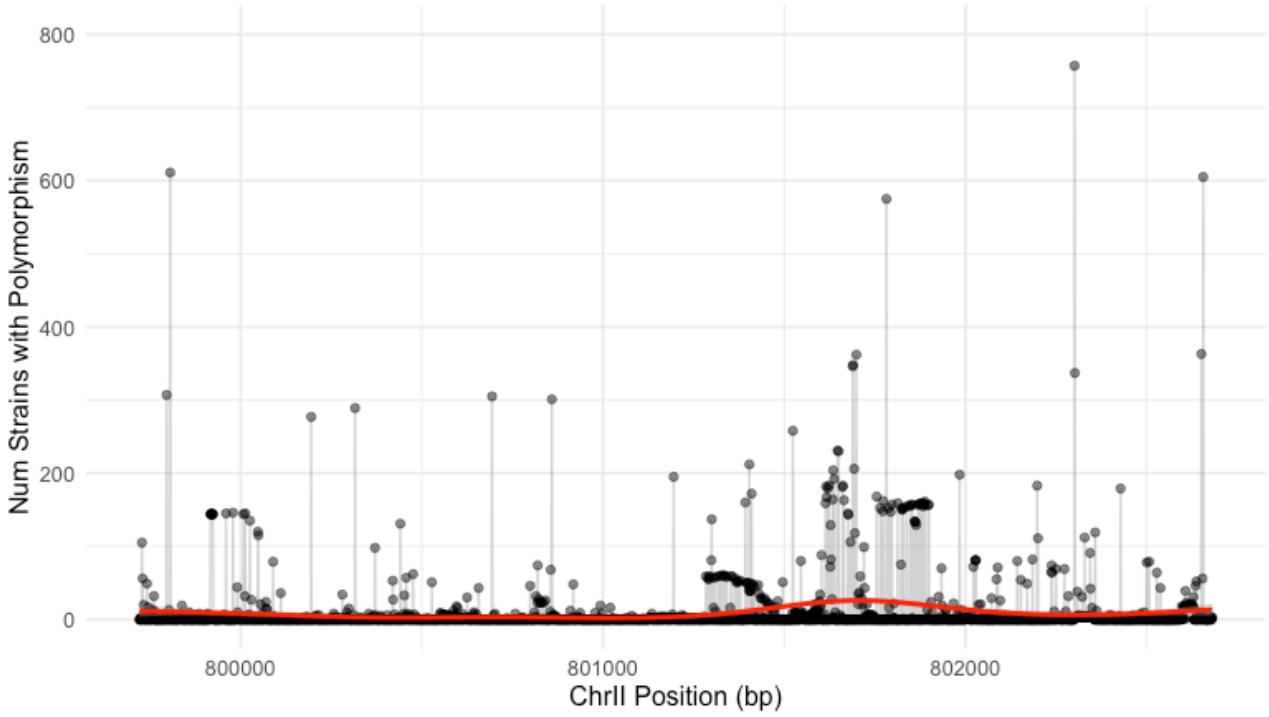


Figure 4.7. Polymorphisms across the *MAL33* locus.
The red trendline is the locally weighted smoothing (LOESS) regression analysis and shows a general pattern between the polymorphisms across the entire locus. The DNA-binding domain is located within the first 60 amino acids of the transcription factor.

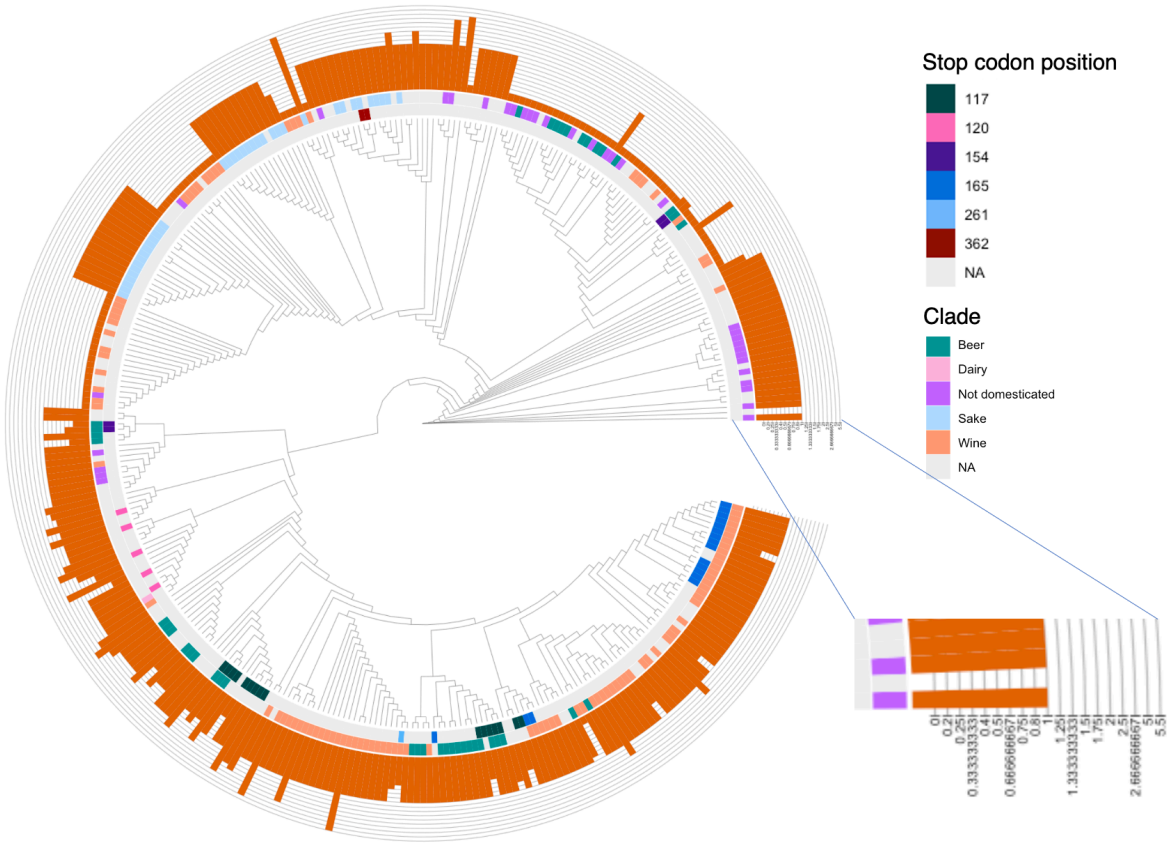


Figure 4.8. Neighbor joining cladogram of unique *MAL13* alleles.
The innermost ring adjacent to the cladogram indicates whether a homozygous premature stop codon exists in the *MAL13* allele for a particular genotype. Premature stop codons are categorized by color; light gray indicates genotypes that are not homozygous for premature stop codons. The second innermost ring labels the clade from which the genotype is derived. The outermost ring with gridlines indicates an approximate copy number for strains with those genotypes.

as well (between amino acids 2-58). The location of these important domains suggests that despite having 467 amino acids, the transcription factor maintains the most crucial domains towards the beginning of the protein.

To determine if all *MALx3* alleles have this same pattern, I performed the same analysis on *MAL13*. Interestingly, *MAL13* had a greater variety in copy number, ranging from 0 to ~5

copies in some strains (**Figure 4.8**). The number of unique premature stop codons is smaller, and these tend to originate from the same clade. *MAL13* is less polymorphic, with 340 unique alleles across the population (compared to 840 in *MAL33*) (**Figure 4.9.**). Wine strains have fewer unique premature stop codons in *MAL13* compared to *MAL33*, and the fact that so many strains do not have any copies of *MAL13* (including those of beer, wine, and wild isolates) suggests that other *MALx3* paralogs may act as the functional transcription factor permitting growth using maltose as the sole carbon source. Future investigations will include determining when these duplications occurred, and if the lower number of polymorphisms in *MAL13* indicate purifying selection or less time to diverge.

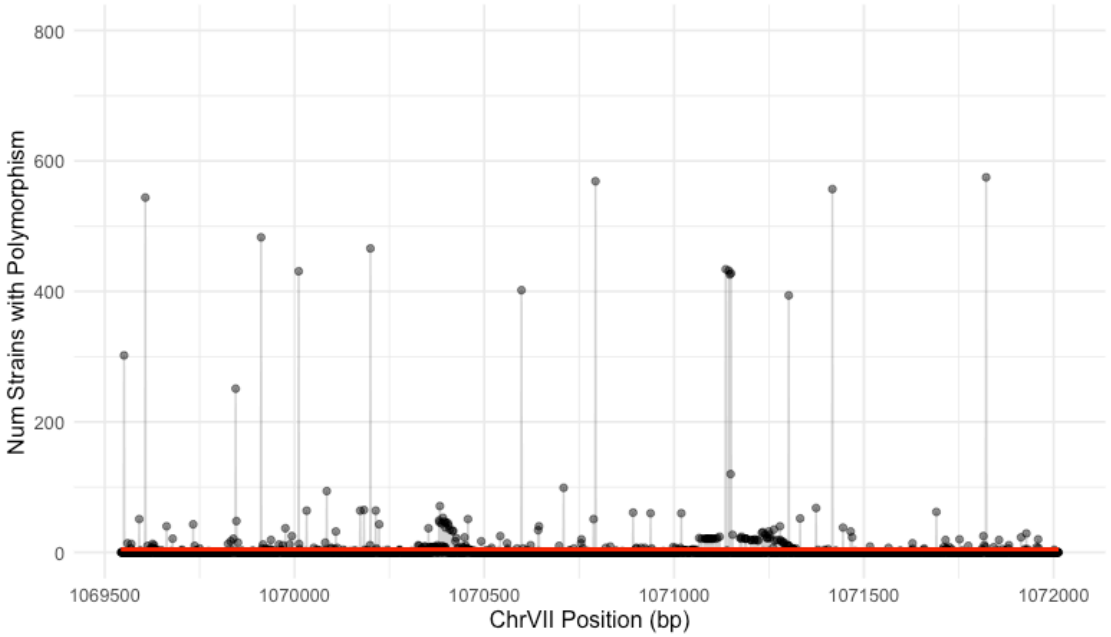


Figure 4.9. Polymorphisms across the *MAL13* locus.
The red trendline is the locally weighted smoothing (LOESS) regression analysis and shows a general pattern between the polymorphisms across the entire locus.

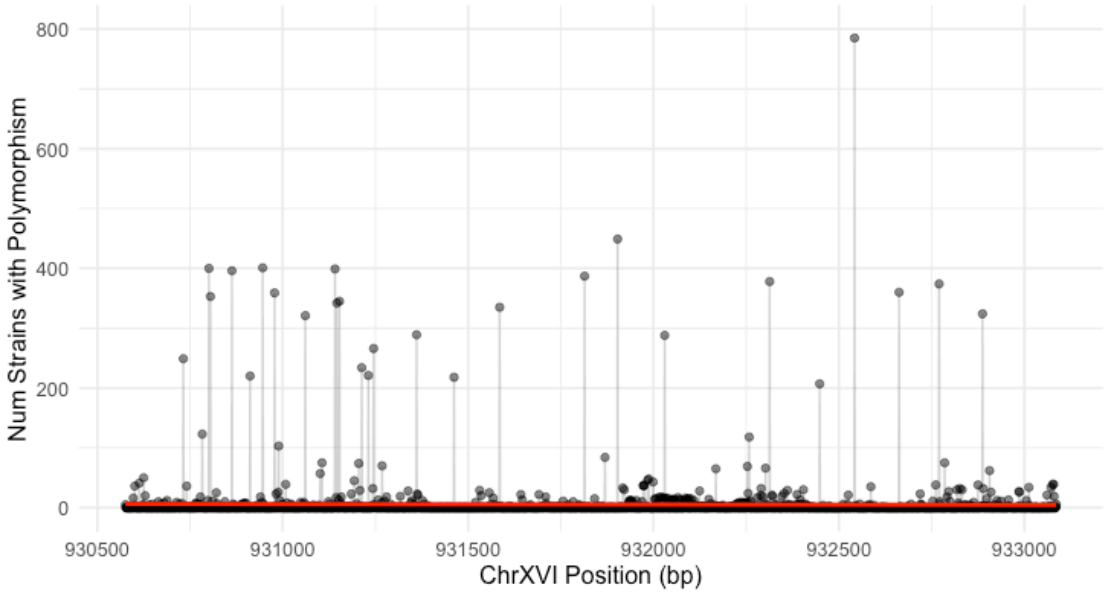


Figure 4.10. Polymorphisms across the *MAL73* locus.
The red trendline is the locally weighted smoothing (LOESS) regression analysis and shows a general pattern between the polymorphism across the entire *MAL73* gene.

The same analysis done on *MAL73 (YPR196W)* also produced patterns different than those of *MAL13* and *MAL33. MAL73* showed a very similar polymorphic pattern as *MAL13* (**Figure 4.10**). Very few of the 467 unique *MAL73* genotypes had a premature stop codon, at least not compared to the reference S288C sequence (**Figure 4.11**). The number of unique premature stop codons is also very small, with only 22 strains exhibiting these nonsense mutations. Copy number varies further in this gene, with one strain even having 7 copies in its genome (**Figure 4.11**). The polymorphic differences between these alleles may play an important role in maltose utilization, although the extent to which any of these alleles were incorrectly mapped reads from *MAL23*, *MAL43*, and *MAL63* is unknown.

Comparison of the three *dN/dS* values of *MAL13*, *MAL33*, and *MAL73* provided a glimpse into the substitution patterns of these three paralogs (*dN/dS* = 0.23, 0.226, and 0.42, respectively). The *dN/dS* values of *MAL13* and *MAL33* are lower than that of *MAL73*, suggesting

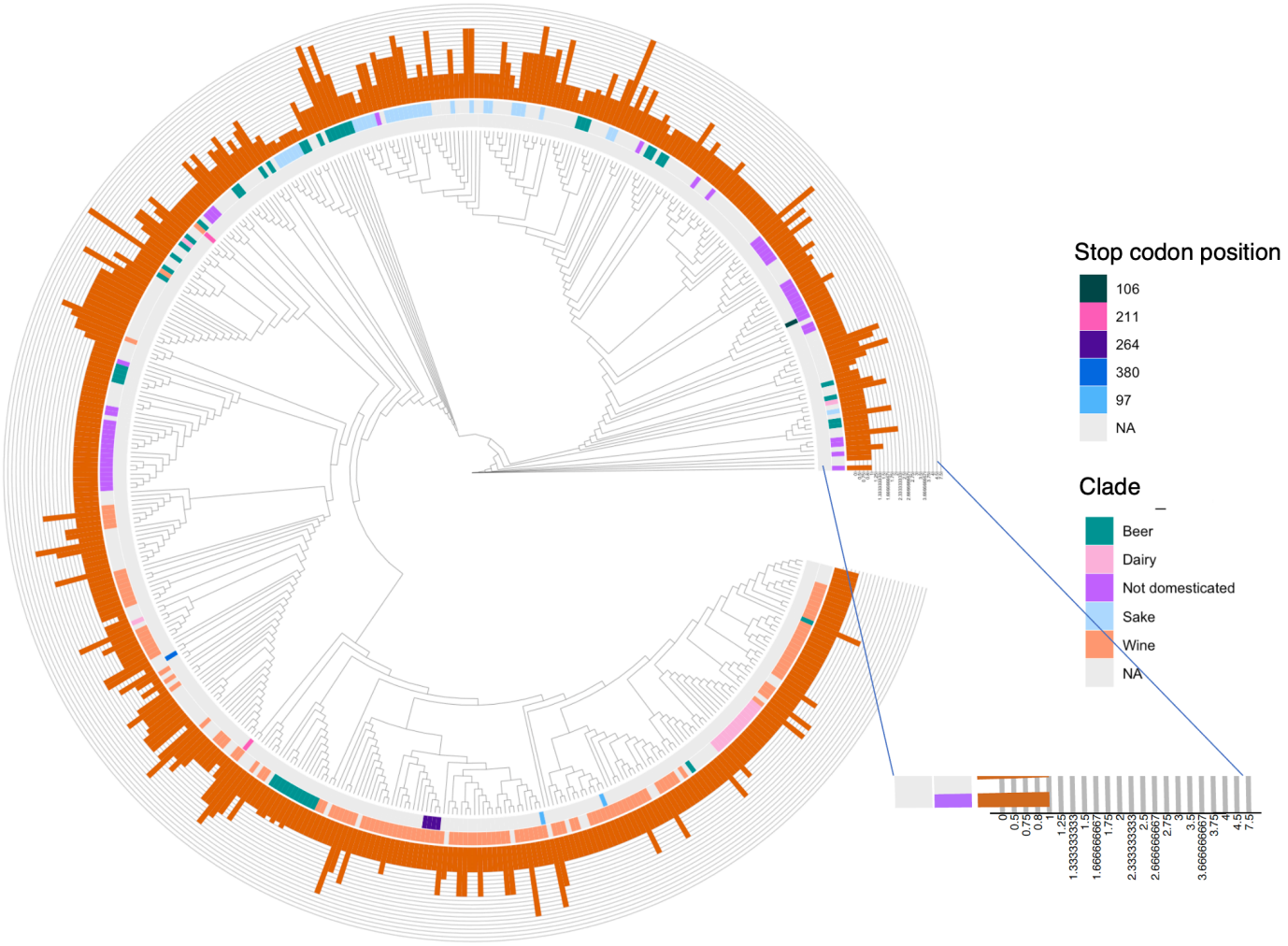


Figure 4.11. Neighbor joining cladogram of unique *MAL73* alleles.
Similar to the other tree figures, the centermost ring flanking the cladogram indicates the location of a homozygous premature stop codon (if present) in the particular *MAL73* genotype. The middle ring indicates the clade from which the genotype is derived, and the outermost ring with gridlines indicates an approximate copy number for strains with those genotypes. An inlet is provided for reading the axis label on the copy number ring.

that they may be under greater purifying selection that *MAL73*. Like *SUL1*, the low *dN/dS* values may represent balancing selection, where multiple alleles are being maintained in the population. The high functioning, nonfunctional, and intermediary alleles all seem to play an important role: the function of the weak *MAL73* allele is important for sake fermentation (Ohdate et al., 2018),

inability to utilize maltose is common in wild isolates (Bai et al., 2022), and high-functioning alleles are common in beer strains where maltose is abundant and sometimes the sole carbon source (Gibson et al., 1997).

4.3 Discussion

We show here an application of an inexpensive high-throughput approach to functionalize natural variants present in a population in order to understand the role of paralogs in evolution. This approach can demonstrate how the genotype of a cell contributes to its phenotype in certain environments. Being able to disentangle the function of paralogs on a large scale is important for understanding their roles in complex traits and evolution. Specifically, the experiments and analyses described here provide evidence for what *MAL33* alleles are functional across the 1,011 *S. cerevisiae* strains. The paralogs of *MALx3* can exist in a strain in a variety of combinations and copy numbers. The results from the phylogenetic trees show that some paralogs (such as *MAL33*) are more prevalent among all strains; however, some paralogs (such as *MAL13* and *MAL73*) can vary more in terms of copy number. Differences in species-wide patterns between paralogs can reveal a lot about paralog function, subfunctionalization, and neofunctionalization. The use of a reference genome and mapping software that can distinguish between the five *MALx3* paralogs will further strengthen the results from whole-genome sequencing.

Allelic differences have been observed before in several *MALx3* paralogs (Gibson et al., 1997; Ohdate et al., 2018). Some paralogs have alleles that can be split into the categories of inducible, noninducible, and constitutive alleles, which may explain why some *MALx3* genes do not rescue growth in S288C. Interestingly, Gibson et al. (1997) observed such a pattern in the

paralogs *MAL23, MAL43*, and *MAL63*, and found that the domain that causes inducibility is located in the C-terminal region of the *MALx3* proteins. Complete truncation of this region did not affect the ability of strains to utilize maltose, but only affected the their ability to respond to inducible factors. Although *MAL33* is more diverged from *MAL23, MAL43*, and *MAL63* (**Figure 4.12, Supplementary Table 6**), it is possible that polymorphisms in the same homologous, C-terminal region of MAL33 can also cause differences like inducibility in allelic function. Similarly, allelic differences in *MAL73* were observed to be due to a frameshift mutation that resulted not in loss of function, but in weaker maltose fermentation. Future work functionalizing different paralogs of *MALx3* will demonstrate if this pattern is unique to *MAL73* and will identify loss-of-function alleles that are not detectable simply through sequence alone. The results from both these papers—as well as the results from *MAL13* and *MAL33* shown here—indicate that identification of frameshift or nonsense mutations may not be good indicators of *MALx3* function, and that experimental validations are the optimal way to determine how these paralogs all contribute to a cell's ability to grow in maltose.

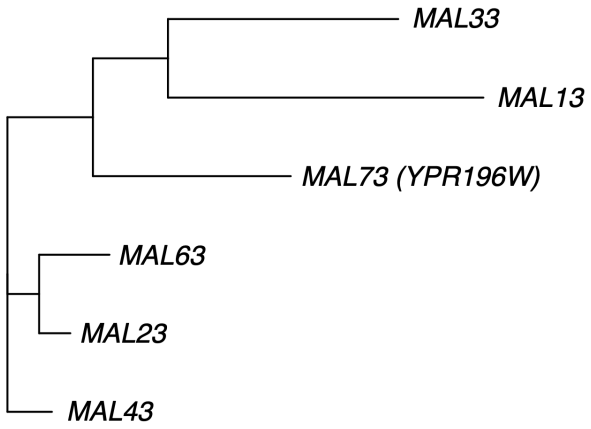


Figure 4.12. Unrooted neighbor-joining tree of *MALx3* paralogs.
This neighbor joining tree was generated by entering the DNA sequences of *MALx3* alleles derived from the pangenome of the 1,011 strains into Clustal Omega.

Interestingly, *MAL13*, *MAL33*, and *MAL73* appear to be evolving more rapidly compared to *MAL23*, *MAL43*, and *MAL63* (**Figure 4.12**). Additionally, the percent identity and similarity between *MAL23, MAL43,* and *MAL63* (both DNA and amino acid sequences) are above 92%; these values are greater than the percent identity and similarity of *MAL13*, *MAL33*, and *MAL73* (**Supplementary Table 6**). Alignment of the 1,011 strains against a reference sequence that contains *MAL23*, *MAL43*, and *MAL63* would help our understanding of 1) paralog function; 2) when translocation mutations may have occurred; and 3) if specific clades retain certain *MALx3* paralogs due to functional benefits. The same analysis can be performed using an ancestrally reconstructed sequence. Few studies have investigated these paralogous genes on a species-wide level, and investigation of the functional differences between these genes will provide insight into why some paralogs in this gene family are diverging more rapidly than others.

Although this project still has much work left to be done, analyses reveal here provide the start to an interesting story for how the paralogous *MALx3* genes diverged in sequence, function, copy number, and prevalence in *Saccharomyces cerevisiae*. Future work creating allele libraries for all genes in the *MAL*-activator family will assist in understanding functional differences between paralogous genes. The *MAL* activator proteins are a complex set of paralogs; application of this approach to another set of paralogs – perhaps *SUL2*, a paralog of *SUL1* – would be interesting to complement results from Chapter 2.

4.4 MATERIALS AND METHODS

4.4.1 *Strains, primers, and plasmids*

Primers for this chapter were designed by extracting the number of strains with polymorphisms at each position where the primer would bind. Optimally, primers would have

fewer than two sites with a polymorphism, and at most 500 strains with polymorphisms at either of those positions. The most conserved regions were selected for primer binding. Primers used in this chapter are listed in Supplementary Table 5.

Natural isolates from the 1,011 *S. cerevisiae* collection were pinned on YPD agar plates, transferred to liquid YPD in 96-well plates, grown overnight at 30°C, and stored in 30% glycerol at -80°C. Prototrophic FY3 (DBY11069) was used to test growth in 2% maltose. The FY3 S288C strain DBY7284 (*MATa ura3-52*) was used for transformation and competition experiments. Vector pRS316 with an NruI site inserted in the BamHI site (YMD2307) was used for molecular cloning and competitions described below.

4.4.2 *Plasmid/yeast library generation and tests in maltose media*

Strains from the 1,011 *S. cerevisiae* collection were pooled together from colonies on a solid agar plate. Genomic DNA was then extracted using the QIAGEN Genomic-tip 100/G kit.

For creating the mini library, Gibson primers of *MAL13*, *MAL33*, and *MAL73* were used to amplify regions ~800bp upstream and ~300 bp downstream of the coding sequence. PCR was performed using KAPA HiFi Hotstart Readymix with the following cycling conditions: 95°C for 3 min, then 23 cycles of 98°C for 20 seconds, 58°C/64°C/67°C for 15 seconds (respectively), and 72°C for 4 minutes. Final extension was at 72°C for 4 minutes, and the reaction was cooled to 4°C. The product was purified using the DNA Clean and Concentrator kit from Zymo Research and assembled into the NruI-digested plasmid YMD2307 via Gibson assembly. Chemically competent *E. coli* cells were transformed with the product using heat shock at 42°C, and ~250 transformants were collected per library. Plasmids were extracted from bacterial transformants using Wizard® *Plus SV* Miniprep DNA Purification Kit and used to transform

DBY7284 using 100 µL of 2 M lithium acetate, 800 µL of 50% 4000 polyethylene glycol, 100 µL of 1M dithiothreitol, and 50 µL of 10 mg/mL of carrier DNA. Approximately 2,000 Ura+ yeast transformants were collected for testing growth in C-Ura +2% maltose media. Starting cultures in 2% maltose media for each library were at OD600=0.05 and allowed to grow for 7 days at 30°C.

Single colonies were isolated from the *MAL33* mini library that grew in 2% maltose and grown overnight in C-Ura (glucose) media. 0.125%, 0.25%, 0.75%, 1%, 1.5% and 2% maltose was added to carbon limiting media (0.1g calcium chloride dihydrate, 0.1g sodium chloride, 0.5g magnesium sulfate heptahydrate, 1g potassium phosphate monobasic, 5g ammonium sulfate, vitamins and minerals as previously described (Gresham et al., 2008), water to 1L). Culture started at OD=0.09 and density was measured on the 3rd, 5th, and 7th days (until saturation).

For creating the full barcoded *MAL33* library, oligos P172 and P194 were used to amplify the gene. P194 contains a 10-bp randomized sequence to serve as a barcode. PCR was performed using KAPA HiFi Hotstart Readymix with the following conditions: 95°C for 3 min, then 20 cycles of 98°C for 20 seconds, 64°C for 15 seconds, and 72°C for 4 minutes. Final extension was at 72°C for 4 minutes, and then the reaction was cooled to 4°C. The reaction was done in 8 replicates, and final PCR reactions were pooled, cleaned, and eluted in 10uL of water using the DNA Clean and Concentrator kit from Zymo Research. The barcoded product was assembled into YMD2307 via Gibson assembly. NEB® 10-beta electrocompetent *E. coli* were transformed with the product, and 20,000 transformants were collected and pooled by plating 10% of the entire transformation (the rest was discarded). Plasmids were extracted from the pooled transformants using Wizard® *Plus SV* Miniprep DNA Purification Kit and then used to transform yeast (DBY7284). To achieve high efficiency transformation, yeast culture was grown

overnight in 20mL of YPD. Cells were diluted to OD600=0.45 the following day in 50mL YPD, and regrown for 5 hours until the OD600 reached ~1.8. Cells were pelleted at 25°C for 3 minutes at 3000 rpm in 50mL falcon tubes and resuspended in 30 mL LiSORB (100mM lithium acetate, 1M sorbitol in TE). Cells were pelleted again at room temperature for 3 minutes at 3000 rpm. This wash was repeated again. LiSORB was decanted and cells were resuspended in 2mL of LiSORB solution, allowed to rotate at room temperature for 30 minutes. 50uL of boiled (10 minutes) and cooled salmon sperm DNA (2mg/mL) was added to 1ug of the *MAL33* plasmid DNA library. 200uL of cells still in LiSORB were added to the DNA mix and vortexed. 1 mL of LiPEG (100mM lithium acetate, 40% PEG 3350 in TE) was added to the solution, and mix was allowed to rotate at room temperature for 30 minutes. Next, 100uL of DMSO was added, and the mixture was put in 42°C for 10 minutes. Solution was centrifuged at 5000 rpm at room temperature. Supernatant was aspirated, and final solution was resuspended in 3.75 mL of YPD and 1.25 mL of 2M sorbitol (0.5M final concentration). Cells were allowed to recover for 1 hour at 30°C in a roller drum. Two C-Ura plates that had been plated with 10% of cell culture were used to collect transformants, and approximately 11,000 Ura+ yeast transformants were collected (in both glycerol stocks and frozen pellets) for pooled competition experiments and further analysis.

4.4.3 *Pooled library competition in chemostats'*

Maltose-limited media (0.25%) was added to carbon-limiting media (1g calcium chloride dihydrate, 1g sodium chloride, 5g magnesium sulfate heptahydrate, 10g potassium phosphate monobasic, 50g ammonium sulfate, vitamins and minerals as previously described (Gresham et al., 2008), water to 10L). Two 230 mL chemostat culture vessels were filled with media in a

30°C water jacket. 1 mL of yeast transformant pool was added to the culture, grown for 30 hours, and then diluted at a rate of about 0.15 volumes per hour. Samples were taken twice a day for 5 days, or about 25 generations. For each sample, 2mL were stored in 25% glycerol at -80°C and another 2mL pelleted for plasmid extraction.

4.4.4 *Barcode sequencing*

For each time point and replicate from the pooled library competition, plasmids were again extracted using the Zymoprep Yeast Plasmid Miniprep kit. Barcodes were isolated and amplified using forward primer P199 and indexed reverse primer P200, which included Illumina Nextera sequencing adaptors. KAPA HiFi Hotstart Readymix was used with 1 µL of 1X SYBR™ Green I and the following PCR cycles: 95°C for 3 min, then 20 cycles of 98°C for 20 seconds, 66°C for 30 seconds, and 72°C for 30 seconds. PCR products were cleaned using Sample Purification Beads (Illumina). Libraries were sequenced on a NextSeq sequencer (Illumina) with sequencing P197 (Read 1), P198 (Read 2), and P202 (Index).

4.4.5 *Phylogenetic tree and sequence analysis*

To generate phylogenetic trees, *MAL33* sequences from the 1,011 strains and from *S. paradoxus* strain CBS432 were aligned using MUSCLE (Edgar, 2004). The genetic distances for *MAL33* alleles were calculated using the maximum-likelihood-based distances through DNADIST in the PHYLIP package (Felsenstein, 2005). A gene tree for *MAL33* was then generated using the NEIGHBOR program, and the final tree was visualized and annotated using R/ggtree (Yu et al., 2017).

Mutation and copy number data were retrieved from Peter et al., 2018. *dN/dS* values were calculated using the yn00 median value from PAML. Phylogenetic trees between all the known *MALx3* genes and amino acid sequences were generated using Clustal Omega. Percent identity and similarity were calculated using the SMS (Stothard, 2000).

4.5 ACKNOWLEDGEMENTS

Thank you to my mentee, Andrea Chang, for performing the majority of these experiments under my supervision. Thank you to Dr. Barbara Dunn for providing useful literature and background on the *MAL* genes. Thank you to Noah Hanson for performing the barcode sequencing for this project.

Chapter 5. Conclusions and future directions

5.1 Information gleaned from a high-throughput analysis of natural variants

Predicting phenotype based on genotype is a long-standing question that has been investigated for decades. Accurate predictions have great implications for many fields including evolution, medicine, industrial settings, and agriculture. Since the start of the work described in this dissertation (2017), many technological advances have been made to improve our understanding of the effects of polymorphisms. The ability to identify quantitative trait loci and causal polymorphisms that affect phenotype has scaled now to assays that survey whole genomes or thousands of single nucleotide polymorphisms (SNPs) in one or few experiments. The work here fills the niche of surveying entire haplotypes of genes that are essential under certain environmental conditions on a population level. It also provides a bigger picture of the fitness distribution as a result of changes to that one gene. It does so in a cost-effective manner and avoids having to use DNA oligonucleotide synthesis, which has an error rate that is difficult to determine/validate and can be expensive as the length of the locus and the number of unique variants increase.

In Chapter 2, I describe a proof-of-concept experiment that portrays how competing natural variants together provides a comprehensive fitness distribution of alleles. Using sequences inferred from whole-genome sequencing and mapped to the reference S288C strain genome, I am able to identify which strains contain alleles that have premature stop codons and validate that those alleles confer a lower fitness compared to the wild-type allele. These results reaffirmed that both coding and noncoding SNPs affect fitness and function. With phylogenetics, I could interpret the evolutionary history of *SUL1*; multiple independent instances of loss-of-

function mutations suggest that the phenotype conferred by a loss-of-function allele may be beneficial under certain conditions. Of particular interest is that many of the loss-of-function mutations occurred in domesticated strains (isolated from beer, dairy, sake, or wine); these results raise the question of). It is also thought-provoking to consider whether these environments select for nonfunctional alleles, or whether the functional alleles were lost long before domestication occurred. One possibility is that some environments (such as dairy and bakeries) provide protein-rich nutrients, so cells are able to use the protein as their sulfur source. Losing the necessity of using the sulfate transporter due to these protein-rich environments would be an example where loss of function due to drift may have taken place. Alternatively, sulfate is abundant in the grape juice used to make wine, and under nitrogen limitation, a high-affinity sulfate transporter can produce hydrogen sulfide and impart to the wine an undesirable flavor (Walker et al., 2021). An intentional selection of low-affinity sulfate transporters may be a result of domestication in the winery. The fitness of cells transformed with the *SUL1* allele library grown in these domesticated environments may provide some interesting insights into validating whether either of these hypotheses is true.

I was surprised to find that the *SUL1* allele that contained signatures of introgression from *Saccharomyces paradoxus* had a fitness in the *S. cerevisiae* background that was similar to the reference allele in S288C. However, this finding is perhaps not surprising since the *SUL1* allele of *S. paradoxus* functions nearly as well as the *S. cerevisiae* S288C allele of *SUL1* (Sanchez et al., 2017). Because *S. paradoxus* is more often isolated from wild environments, similarity in *SUL1* function may further support that *SUL1* function is important in wild environments where protein content is not as accessible as protein content is in domesticated environments. The question now left unanswered is why signatures of introgression from *S.*

paradoxus still remain in the *S. cerevisiae* genome if these signatures do not significantly alter ability to transport sulfate into the cell; this transporter may be pleiotropic, meaning that it can influence more than one trait.

S. paradoxus has higher nucleotide diversity compared to *S. cerevisiae*, but has not been as extensively domesticated (L. J. Johnson et al., 2004). The genetic diversity in *S. paradoxus* provides another excellent resource for surveying natural genetic variation without the influence of thousands of years of domestication by humans. Because many of the features of *S. paradoxus* are the same as *S. cerevisiae* – for instance, the species have the same number of chromosomes and have a high level of synteny – we can generate natural allele libraries from *S. paradoxus* to gain a better understanding of gene function across the two species to understand gene function divergence.

Another key feature of domestication in *S. cerevisiae* is the utilization of maltose as the sole carbon source. Despite the *MAL* locus being essential for maltose uptake and utilization, these loci are also present in *S. paradoxus* strains, where the influence of domestication has not occurred. Some alleles are functional in *S. paradoxus*, and the orthologs of *MAL13*, *MAL33*, and *MAL73* analyzed in this dissertation are distinguishable in the genome of the reference CBS432 *S. paradoxus* isolate (Naumov et al., 1994). A functional *S. cerevisiae MAL13* allele can serve as an activator in *S. paradoxus* for the maltase gene that results in the hydrolysis of maltose (Naumov et al., 1994). The Naumov et al. (1994) study hypothesizes that maltose utilization is weak in *S. paradoxus,* but a survey of *MAL* alleles in *S. paradoxus* isolates, or even *S. cerevisiae* x *S. paradoxus* hybrids, will undoubtedly reveal information about paralog divergence and maltose utilization in different species and ecological environments. Investigation into the

equivalent orthologs in even more diverged species, such as *Saccharomyces uvarum*, would be interesting as well.

Applying this high-throughput approach of assaying natural variants of *MALx3* alleles will also provide great insight into domestication and *MALx3* functional diversity. I have already begun to test the possibility of building these libraries through my experiments with *MAL33*. With just sequence analysis alone, great differences are already observable in copy number, SNPs, and function. The fact that some strains can have up to 7 copies of *MAL73* is surprising in itself, and this paralog tends to have fewer premature stop codons compared to the *MAL13* and *MAL33* alleles. Subfunctionalization or neofunctionalization may be occurring throughout these paralogs. Once *MAL13* and *MAL73* libraries are built, the libraries of these three paralogs can be pooled and competed together. It would be especially interesting to determine if any paralog (regardless of allele) is more effective at activating expression of *MAL* genes, or if the paralogs all have alleles that can be categorized as high-functioning, loss-of-function, or intermediary.

This approach is preferable over those like deep mutational scanning or random mutagenesis. All alleles assayed in these experiments are present in the natural *S. cerevisiae* population and provide information on how cells respond to environmental conditions. Some alleles contain only one mutation; however, many have multiple mutations. While we did not find mutations in combination that affect gene function, investigation of other genes may reveal examples where polymorphisms of a gene are epistatic and only impact function in combination with other polymorphisms. Deep mutational scanning investigates only single substitutions, and combinatorial mutations become exponentially more complex to study. Random mutagenesis does not always reflect variants present in the wild; however, it is useful in exploring a fitness landscape that may not be achieved from simply using natural alleles.

I was surprised that no *SUL1* alleles showed any signatures of gain of function. The results described here reinforce the hypothesis that the reference allele of *SUL1* in the *S. cerevisiae* background confers the highest fitness. Perhaps additional work on other genes, such as a *MALx3* gene, would reveal instances of gain-of-function SNPs.

All in all, the results from this work underscore the persistent utility of yeast as a model organism for studying questions about domestication, loss of function, and paralog evolution. Natural allele libraries of *SUL1* and *MAL33* already provide a better picture of gene function than what was previously known. By leveraging the diversity present among these strains, we can better understand how genotypic changes give rise to phenotypic differences between isolates.

5.2 Application of high-throughput analyses of natural variants to long-standing questions in evolution

The benefit of having an allele library is its use to test how genotypic changes on a species-wide scale alter phenotype under changing environments. Many prior studies have demonstrated that selection against a functional sulfate transporter can be achieved by using toxic sulfate analogues like potassium dichromate and sodium selenate (Cherest et al., 1997; Walker et al., 2021). Although I was unable to replicate the antagonistic effects of these compounds with functional *SUL1* alleles, there may be alternative media that I did not test but can result in selection against a high-affinity sulfate transporter.

The concept of one gene affecting multiple traits is called pleiotropy. The concept of an allele being beneficial in one environment but detrimental in another is called antagonistic pleiotropy. *SUL1* has this trade-off phenotype, and understanding how allelic differences in

SUL1 affect fitness distribution under antagonistic environments is another question that can be answered through these libraries of natural alleles. Several studies have identified genes with antagonistic pleiotropy, but whether or not these trade-off phenotypes are consistent across the population remains to be determined. Excellent examples of genes with trade-offs under different environments are the genes *AQY1* and *AQY2*, where a functional copy of an *AQY* gene provides higher tolerance to cycles of freeze-thaw with a trade-off of lower tolerance to high osmolarity (Will et al., 2010). Loss of function results in the opposite phenotype (low freeze-thaw tolerance and improved tolerance to high osmolarity). Both functional and nonfunctional alleles exist in the *S. cerevisiae* population, but apart from frameshift and nonsense mutations, loss of function is challenging to identify. Some phenotypic information may be gleaned from ecological origins, but these allelic functions may still need to be validated experimentally.

Another gene I identified with characteristics of antagonistic pleiotropy is *RIM15*. This gene encodes a protein kinase involved in cell proliferation in response to certain nutrients, and prior experiments have shown that a functional copy improves growth under toxic environments like cadmium chloride and lead nitrate (Kessi-Pérez et al., 2016). The antagonistic condition for *RIM15* is when the sole carbon source is ethanol and is only available in a limiting concentration (N. Zhang et al., 2009). Although this gene illustrates a perfect example of antagonistic pleiotropy, the gene is large by yeast standards (nearly 5 kb) and thus it may be difficult to accurately assemble its full haplotype. However, with improvements in PacBio sequencing, the feasibility of investigating this gene accurately is not too far away.

Other genes that may exhibit signs of antagonistic pleiotropy have been identified before. In a study published in 2012, researchers were able to identify instances of antagonistic pleiotropy throughout the whole genome by analyzing nonessential genes (Qian et al., 2012).

Allele libraries are validating and identifying how frequently such patterns appear across the entire population. A question that remains unanswered is if all alleles of a gene exhibit this trade-off, or if these results are only due to surveying a handful of alleles. Understanding allele function on a species-wide level can provide an ample amount of information about evolutionary history and the relationship of a particular isolate to environmental factors.

This population-wide survey of natural alleles can also be applied to understanding epistasis or protein-protein interactions. This approach can be combined with a yeast display assay to investigate agglutinin proteins such as *AGA1, AGA2*, *SAG1*, and even the *AGA1* paralog known as *FIG2*. By building two individual libraries, we can investigate variation in the interaction between two interacting proteins can inform how well natural alleles agglutinate and how much these interactions promote mating in yeast. This approach would also be a way to survey what isolates are compatible with one another for mating and if other isolates are less likely to agglutinate and have less of an affinity for mating. Similar to work done using *MAL* libraries, it may be interesting to build allele libraries in *S. paradoxus* and may reveal information about what isolates are more likely to hybridize with *S. cerevisiae.*

Another set of genes that can be used to investigate epistasis are *STE5* and *STE7.* These two genes interact to activate an important kinase pathway in yeast and are also important for yeast mating. In a project not detailed in this dissertation, I have begun to identify instances of epistasis between variants of these two genes. Generating libraries of natural variants of *STE5* and *STE7* could help in understanding if epistatic effects between these alleles play an important role in this kinase pathway and ultimately in mating evolution.

As mentioned before, the genotypic and phenotypic diversity in *S. cerevisiae* makes it an ideal model organism for investigating genetic changes. Therefore, the genetic backgrounds that

these alleles are in undoubtedly plays an important role in gene function. In investigating *SUL1* allelic differences, I was unable to find instances where genetic background dramatically altered the phenotype of alleles, even in different backgrounds. A notable example of this is that the introgressed allele – and even the *S. paradoxus* allele – when moved to the background of the reference S288C strain, did not show significant fitness differences. Although I was unable to find these differences in *SUL1*, usage of other genes for allele libraries can be used to investigate how genetic background alters allelic function and fitness distribution of a gene. Differences in function due to changes in genetic backgrounds can also help identify epistatic and *trans*-acting factors that were previously unknown. One gene that could potentially be used for allele libraries is *MLH1*, which confers differences in mutation rate as a result of changes in genetic background (Wanat et al., 2007). Traditional fluctuation assays make it difficult to determine mutation rate in a high-throughput manner, although a recent study multiplexing mutation rate assays in a turbidostat streamlines this approach (Ollodart et al., 2021).

Application of this approach can also be used to investigate differences in expression. Work described here confirmed studies that showed that SNPs in the promoter region of *SUL1* can affect fitness (Rich et al., 2016). Numerous studies investigating diversity in quantitative traits found that expression differences are responsible for a significant amount of phenotypic diversity (Salinas et al., 2016), and this focus on natural variation in transcriptional changes has increased in recent years. A gene that could be used for identifying these natural expression differences is *SGE1*, a gene encoding an efflux pump, which shows differences in imidazolium ionic liquids (IIL) tolerance based on changes in transcript levels (Higgins et al., 2018).

The applications of studying natural variation on a species-wide scale are endless. The examples listed above are a handful of the few genes that can be investigated to investigate a

number of biological questions, including how paralog functions differ, how gene function changes under altered environmental conditions, how genetic background changes fitness distributions, and how expression differences underlie phenotypic diversity.

5.3 Remaining optimizations to improve assay

The current approach of studying natural variants in a high-throughput manner still has many aspects that could be improved. After I retrieved PacBio sequences from the *SUL1* allele library, an obstacle that I immediately encountered was matching these alleles to their strain of origin. This obstacle arose because currently most of the strains in the collection are diploid, and resolved haplotypes are not complete in many of these strains. With decreasing costs and improved confidence in long-read sequencing, resequencing of these isolates with longer reads could improve haplotype information. Without long-read sequencing, one approach to improve mapping of alleles back to the strain of origin would be using a Hidden-Markov Model (HMM), based on allele frequency both in the strain collection and in the PacBio sequenced alleles. The frequency of these alleles would serve as the basis for the "start" probability of each allele in terms of which strain it matches up with. Clade information could help improve confidence of this information, especially if these clades contain both haploid and polyploid strains and the unique alleles in these strains can be used to infer haplotypes for the polyploid strains. An increasing number of strains are being sequenced with Oxford Nanopore Technology (ONT) or PacBio, and reanalysis with these new data could warrant a reanalysis of the libraries generated here.

While the phenotypic data overlaid on the phylogenetic data reveal information about allele function, the majority of these results were interpreted in a single genetic background.

How fitness distribution and allele function change as a result of changes in background would provide more information on phenotypic variation. Even having the allele library on a plasmid, rather than integrated into the genome, with a selectable marker also present on the plasmid could cause a misinterpretation of gene function. For instance, if a gene of interest has an epistatic interaction with the selectable marker used in an assay, the allele could affect function of this gene of interest. In future libraries, establishing a similar system that would result in integration into the native locus in the genome would allow us to better understand how genotypic changes affect phenotype when the differences exist at the native location in the genome. Advances in CRISPR/Cas9 technology can facilitate this integration, although integrating multiple edits throughout a ~4 kb locus may be an obstacle.

One complication of many of these high-throughput studies is that measuring gene function is not always easy. In the libraries created and assayed here, fitness was measured as a result of the ability of a yeast strain to utilize or transport a specific nutrient. For cells that may have phenotypic differences, such as morphology, separating distinct types of cells may be more difficult. For genes without a measurable phenotype, or even with only subtle changes in phenotype, performing a study on natural variants may be more difficult.

The curation and extensive analysis of the 1,011 *S. cerevisiae* strains facilitated the possibility of developing this assay. Numerous parts of this assay depended on the accuracy of this strain collection and its associated dataset. However, during analysis, I discovered that some alleles were incorrectly called. For instance, the genotype for *SUL1* in the strain DBVPG1106 was missing the crucial SNP that results in a loss-of-function phenotype under sulfate limitation. While I was able to correct this error and validate that the SNP exists through Sanger sequencing, I am unable to determine how prevalent these errors are. This type of error also

existed in my analysis of the *MALx3* alleles, where an allele sequence would exist in the database for a strain that has no copies of that particular gene. I discussed in Chapter 4 that this discordance may be as a result of a mapping error, but the extent to which this mismapping exists and which alleles are actually mismapped reads is difficult to deconvolute. In general, identifying true mismapping errors is a pitfall in current genome assemblies, but improvements long-read sequencing technologies and algorithms are reducing these issues (Logsdon et al., 2020).

The ability to capture all alleles of a gene is also an integral part of this assay. I show in the *SUL1* library that none of the introgressed alleles were captured. This failure was a result of designing primers based on a conserved region in a subset of strains. In subsequent libraries, I developed an approach for designing primers based on all 1,011 strains in the collection. Interestingly, the strains with signatures of introgression were incorrectly called in the sequencing results, likely due to mapping errors as well. Thus, while this new approach for designing primers may be able to capture more variants, it is possible that it would still not be able to capture sequences as diverged as introgressed sequences.

Another factor of concern is template switching during PCR. Some groups have observed this from their PacBio sequencing specifically when using the Kapa HiFi polymerase; however, I did not identify any heteroduplexes in my libraries, nor did I observe any barcodes with heteroduplexes. To test if this template-switching issue is unique to Kapa HiFi, different polymerases could be used on the same library and sequenced together to determine if the same barcode-sequences are returned.

In sum, the multitude of approaches for improving this technique will optimize the accuracy of matching allele sequences to the strain of origin and its resulting phenotype.

Although many drawbacks exist in both the experimental and computational design of this approach, these data can be reanalyzed to provide much higher accuracy in matching alleles to strain of origin, especially as technologies continue to improve.

5.4 Other applications and optimizations of PacRAT

This high-throughput approach relies on the accuracy of PacBio sequencing reads. Since the first sequencing run of *SUL1* was performed, PacBio sequencing technology has greatly improved. Using polymerases with higher fidelity, which increases the number of passes through each circular consensus sequence (CCS), provided higher quality reads. Furthermore, one SMRT cell can now sequence up to 3 million molecules, an output that is much greater than the 200,000 reads/cell obtained in previous models. The simulations provided in this dissertation showed how much coverage was necessary for accurately matching barcodes to variants. Since PacBio sequencing quality has improved since doing those simulations, it is likely that the required coverage is now even lower. Lower required coverage combined with higher sequence quality provides room for multiplexing in one SMRT cell, a combination that would greatly reduce cost and labor for preparing libraries for sequencing.

One optimization of PacRAT would be utilizing minimap2 to perform pairwise alignment (Heng Li, 2018). This software is much faster than current aligners and is integrated with Python. It also outperforms other aligners when aligning long reads, such as those retrieved from ONT or PacBio. Since it does not have a significant difference in output compared to other aligners for PacBio SMRT reads, the current approach using EMBOSS Needle for pairwise alignment works well (Heng Li, 2018; Needleman & Wunsch, 1970; Rice et al., 2000).

Additional features added to PacRAT would be useful for determining the accuracy of alignments as well. For instance, adding a metric that reports the concordance between variants with the same barcode sequenced from individual molecules would help with reducing the false discovery rate. Being able to identify whether subclusters of barcode-variant molecules exist may also help, either in the unlikely event that one barcode contains multiple variants or if an error occurred early during PCR and resulted in a jackpot event. Knowing how many of these subclusters exist and if a minor allele is present for a specific barcode may be useful as well. Furthermore, it may be helpful to understand if the errors were due to discordant reads, barcode swaps, or sequencing errors alone.

Sequence aligners can inaccurately align homopolymer tracks between two reads. We found that many of the errors between alignments occurred in such homopolymer regions. Although PacBio polymerases do not exhibit nucleotide bias in errors, they do have a higher likelihood of insertion/deletion errors that, when they occur in these homopolymer tracks, are difficult to remediate. These issues will likely resolve themselves as CCS accuracy continues to improve.

Correspondence with other researchers using PacBio for linking barcodes to variants also revealed that heteroduplexes can form with PCR products. That is, in some cases, due to the high degree of identity between sequences, template switching may occur between barcodes and their respective variants. Future analysis will need to address this issue. In addition, PacBio sequencing errors can occur in the barcode, and deeper investigation into the barcode sequencing errors will reveal how often these types of errors occur.

5.5 CONCLUDING REMARKS

The outcomes of this dissertation foreshadow a thrilling future for the characterization of genes. Assaying hundreds or thousands of variants in one experiment raises our standards for understanding gene function in yeast, and these assays can now be done on a population level. Although this approach is imperfect, its current form already reveals new patterns and questions for investigating the impact of SNPs under certain environments. I am excited for future optimizations and work using this approach, and I am inspired by what questions it has the capacity to answer about evolution and population genetics.

BIBLIOGRAPHY

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. In *Nature Methods* (Vol. 7, Issue 4, pp. 248–249). Cambridge Univ. Press. https://doi.org/10.1038/nmeth0410-248

Albert, F. W., Treusch, S., Shockley, A. H., Bloom, J. S., & Kruglyak, L. (2014). Genetics of single-cell protein abundance variation in large yeast populations. *Nature*, *506*(7489), 494–497. https://doi.org/10.1038/nature12904

Alford, B. D., Tassoni-Tsuchida, E., Khan, D., Work, J. J., Valiant, G., & Brandman, O. (2021). ReporterSeq reveals genome-wide dynamic modulators of the heat shock response across diverse stressors. *ELife*, *10*. https://doi.org/10.7554/eLife.57376

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M. M., Cao, J., Chae, E., Dezwaan, T. M. M., Ding, W., Ecker, J. R. R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G. G., Hancock, A. M. M., Henz, S. R. R., Holm, S., … Zhou, X. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell*, *166*(2), 481–491. https://doi.org/10.1016/j.cell.2016.05.063

Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., Nickerson, D. A., … Peterson, J. L. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Amorosi, C. J., Chiasson, M. A., McDonald, M. G., Wong, L. H., Sitko, K. A., Boyle, G.,

Kowalski, J. P., Rettie, A. E., Fowler, D. M., & Dunham, M. J. (2021). Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. *American Journal of Human Genetics*, *108*(9), 1735–1751. https://doi.org/10.1016/j.ajhg.2021.07.001

Bai, F. Y., Han, D. Y., Duan, S. F., & Wang, Q. M. (2022). The Ecology and Evolution of the Baker's Yeast *Saccharomyces cerevisiae*. In *Genes* (Vol. 13, Issue 2, p. 230). MDPI. https://doi.org/10.3390/genes13020230

Balakrishnan, R., Park, J., Karra, K., Hitz, B. C., Binkley, G., Hong, E. L., Sullivan, J., Micklem, G., & Cherry, J. M. (2012). YeastMine-An integrated data warehouse for *Saccharomyces cerevisiae* data as a multipurpose tool-kit. *Database*, *2012*. https://doi.org/10.1093/database/bar062

Belsare, S., Levy-Sakin, M., Mostovoy, Y., Durinck, S., Chaudhuri, S., Xiao, M., Peterson, A. S., Kwok, P. Y., Seshagiri, S., & Wall, J. D. (2019). Evaluating the quality of the 1000 genomes project data. *BMC Genomics*, *20*(1), 620. https://doi.org/10.1186/s12864-019-5957-x

Bergström, A., Simpson, J. T., Salinas, F., Barré, B., Parts, L., Zia, A., Nguyen Ba, A. N., Moses, A. M., Louis, E. J., Mustonen, V., Warringer, J., Durbin, R., & Liti, G. (2014). A high-definition view of functional genetic variation from natural yeast genomes. *Molecular Biology and Evolution*, *31*(4), 872–888. https://doi.org/10.1093/molbev/msu037

Bloom, J. S., Boocock, J., Treusch, S., Sadhu, M. J., Day, L., Oates-Barker, H., & Kruglyak, L. (2019). Rare variants contribute disproportionately to quantitative trait variation in yeast. *ELife*, *8*. https://doi.org/10.7554/eLife.49212

Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V., & Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature*, *494*(7436), 234–237.

https://doi.org/10.1038/nature11867

Boonekamp, F. J., Knibbe, E., Vieira-Lara, M. A., Wijsman, M., Luttik, M. A. H., Eunen, K. van, Ridder, M. den, Bron, R., Suarez, A. M. A., Rijn, P. van, Wolters, J. C., Pabst, M., Daran, J.-M., Bakker, B., & Daran-Lapujade, P. (2021). A yeast with muscle doesn't run faster: full humanization of the glycolytic pathway in *Saccharomyces cerevisiae*. *bioRxiv*, 2021.09.28.462164. https://doi.org/10.1101/2021.09.28.462164

Borgström, E., Redin, D., Lundin, S., Berglund, E., Andersson, A. F., & Ahmadian, A. (2015). Phasing of single DNA molecules by massively parallel barcoding. *Nature Communications*, *6*(1), 1–6. https://doi.org/10.1038/ncomms8173

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. In *Cell* (Vol. 169, Issue 7, pp. 1177–1186). Cell Press. https://doi.org/10.1016/j.cell.2017.05.038

Brion, C., Lutz, S. M., & Albert, F. W. (2020). Simultaneous quantification of mrna and protein in single cells reveals post-transcriptional effects of genetic variation. *ELife*, *9*, 1–34. https://doi.org/10.7554/eLife.60645

Brown, C. A., Murray, A. W., & Verstrepen, K. J. (2010). *Rapid Expansion and Functional Divergence of Subtelomeric Gene Families in Yeasts | Elsevier Enhanced Reader*. Current Biology.

Buck, D., Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X. J., & Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, *6*. https://doi.org/10.12688/f1000research.10571.2

Caudal, E., Friedrich, A., Jallet, A., Garin, M., Hou, J., & Schacherer, J. (2021). Population-level

survey of loss-of-function mutations revealed that background dependent fitness genes are rare and functionally related in yeast. *BioRxiv*, 2021.08.25.457624. https://doi.org/10.1101/2021.08.25.457624

Cervelli, T., Lodovichi, S., Bellè, F., & Galli, A. (2020). Yeast-based assays for the functional characterization of cancer-associated variants of human DNA repair genes. In *Microbial Cell* (Vol. 7, Issue 7, pp. 162–174). Shared Science Publishers OG. https://doi.org/10.15698/mic2020.07.721

Charlesworth, J., & Eyre-Walker, A. (2008). The McDonald-Kreitman Test and Slightly Deleterious Mutations. *Molecular Biology and Evolution*, *25*(6), 1007–1015. https://doi.org/10.1093/molbev/msn005

Charron, M. J., Read, E., Haut, S. R., Michels, C. A., Dubin, R. A., Perkins, E., Michels, C. A., & Needleman, R. B. (1989). *Molecular Evolution of the Telomere-Associated MAL Loci of Saccharomyces*.

Cherest, H., Davidian, J. C., Thomas, D., Benes, V., Ansorge, W., & Surdin-Kerjan, Y. (1997). Molecular characterization of two high affinity sulfate transporters in *Saccharomyces cerevisiae*. *Genetics*, *145*(3), 627–635. http://www.ncbi.nlm.nih.gov/pubmed/9055073

Chiasson, M., Dunham, M. J., Rettie, A. E., & Fowler, D. M. (2019). Applying Multiplex Assays to Understand Variation in Pharmacogenes. *Clinical Pharmacology & Therapeutics*, *106*(2), 290–294. https://doi.org/10.1002/cpt.1468

Collins, M. A., Avery, R. R., & Albert, F. W. (2021). Substrate-Specific Effects of Natural Genetic Variation on Proteasome Activity 2. *BioRxiv*, 2021.11.23.469794. https://doi.org/10.1101/2021.11.23.469794

Connelly, C. F., & Akey, J. M. (2012). On the prospects of whole-genome association mapping

in *Saccharomyces cerevisiae*. *Genetics*, *191*(4), 1345–1353. https://doi.org/10.1534/genetics.112.141168

Costa, R. A., Martins, R. S. T., Capilla, E., Anjos, L., & Power, D. M. (2018). Vertebrate *SLRP* family evolution and the subfunctionalization of osteoglycin gene duplicates in teleost fish. *BMC Evolutionary Biology*, *18*(1), 191. https://doi.org/10.1186/s12862-018-1310-2

Costanzo, M., Hou, J., Messier, V., Nelson, J., Rahman, M., VanderSluis, B., Wang, W., Pons, C., Ross, C., Ušaj, M., San Luis, B. J., Shuteriqi, E., Koch, E. N., Aloy, P., Myers, C. L., Boone, C., & Andrews, B. (2021). Environmental robustness of the global yeast genetic interaction network. *Science*, *372*(6542). https://doi.org/10.1126/science.abf8424

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., Pelechano, V., Styles, E. B., Billmann, M., Van Leeuwen, J., Van Dyk, N., Lin, Z. Y., Kuzmin, E., Nelson, J., Piotrowski, J. S., … Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, *353*(6306). https://doi.org/10.1126/science.aaf1420

Cuperus, J. T., Groves, B., Kuchina, A., Rosenberg, A. B., Jojic, N., Fields, S., & Seelig, G. (2017). Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res.*, *27*(12), 2015–2024. https://doi.org/10.1101/gr.224964.117

de Boer, C. G., Vaishnav, E. D., Sadeh, R., Abeyta, E. L., Friedman, N., & Regev, A. (2020). Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nature Biotechnology*, *38*(1), 56–65. https://doi.org/10.1038/s41587-019-0315-8

Demogines, A., Smith, E., Kruglyak, L., & Alani, E. (2008). Identification and dissection of a complex DNA repair sensitivity phenotype in Baker's yeast. *PLoS Genet.*, *4*(7), e1000123.

https://doi.org/10.1371/journal.pgen.1000123

Després, P. C., Dubé, A. K., Seki, M., Yachie, N., & Landry, C. R. (2020). Perturbing proteomes at single residue resolution using base editing. *Nature Communications*, *11*(1), 1–13. https://doi.org/10.1038/s41467-020-15796-7

Deutschbauer, A. M., & Davis, R. W. (2005). Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nat. Genet.*, *37*(12), 1333–1340. https://doi.org/10.1038/ng1674

Diao, L., & Chen, K. C. (2012). Local ancestry corrects for population structure in *Saccharomyces cerevisiae* genome-wide association studies. *Genetics*, *192*(4), 1503–1511. https://doi.org/10.1534/genetics.112.144790

Dimitrov, L. N., Brem, R. B., Kruglyak, L., & Gottschling, D. E. (2009). Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of *Saccharomyces cerevisiae* S288C strains. *Genetics*, *183*(1), 365–383. https://doi.org/10.1534/genetics.109.104497

Diss, G., & Lehner, B. (2018). The genetic landscape of a physical interaction. *ELife*, *7*. https://doi.org/10.7554/eLife.32472

Domingo, J., Diss, G., & Lehner, B. (2018). Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature 2018 558:7708*, *558*(7708), 117–121. https://doi.org/10.1038/s41586-018-0170-7

Duan, S. F., Han, P. J., Wang, Q. M., Liu, W. Q., Shi, J. Y., Li, K., Zhang, X. L., & Bai, F. Y. (2018). The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. *Nature Communications*, *9*(1), 1–13. https://doi.org/10.1038/s41467-018-05106-7

Duveau, F., Yuan, D. C., Metzger, B. P. H., Hodgins-Davis, A., & Wittkopp, P. J. (2017).

Effects of mutation and selection on plasticity of a promoter activity in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(52), E11218–E11227. https://doi.org/10.1073/pnas.1713960115

Duveau, F., Zande, P. Vande, Metzger, B. P. H., Diaz, C. J., Walker, E. A., Tryban, S., Siddiq, M. A., Yang, B., & Wittkopp, P. J. (2021). Mutational sources of trans-regulatory variation affecting gene expression in *Saccharomyces cerevisiae*. *ELife*, *10*. https://doi.org/10.7554/eLife.67806

Eder, M., Sanchez, I., Brice, C., Camarasa, C., Legras, J. L., & Dequin, S. (2018). QTL mapping of volatile compound production in *Saccharomyces cerevisiae* during alcoholic fermentation. *BMC Genomics*, *19*(1), 166. https://doi.org/10.1186/s12864-018-4562-8

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Ehrenreich, I. M., Bloom, J., Torabi, N., Wang, X., Jia, Y., & Kruglyak, L. (2012). Genetic architecture of highly complex chemical resistance traits across four yeast strains. *PLoS Genetics*, *8*(3). https://doi.org/10.1371/journal.pgen.1002570

Fay, J. C. (2013). The molecular basis of phenotypic variation in yeast. In *Current Opinion in Genetics and Development* (Vol. 23, Issue 6, pp. 672–677). NIH Public Access. https://doi.org/10.1016/j.gde.2013.10.005

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the Author. Department of Genome Sciences, University of Washington, Seattle.*

Fournier, T., Saada, O. A., Hou, J., Peter, J., Caudal, E., & Schacherer, J. (2019). Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. *ELife*, *8*. https://doi.org/10.7554/eLife.49258

Fowler, D. M., & Fields, S. (2014). Deep mutational scanning: A new style of protein science. In *Nature Methods* (Vol. 11, Issue 8, pp. 801–807). Nature Publishing Group. https://doi.org/10.1038/nmeth.3027

Galardini, M., Busby, B. P., Vieitez, C., Dunham, A. S., Typas, A., & Beltrao, P. (2019). The impact of the genetic background on gene deletion phenotypes in *Saccharomyces cerevisiae*. *Molecular Systems Biology*, *15*(12), e8831. https://doi.org/10.15252/msb.20198831

Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., Merlevede, A., Roncoroni, M., Voordeckers, K., Miraglia, L., Teiling, C., Steffy, B., Taylor, M., Schwartz, A., Richardson, T., White, C., Baele, G., Maere, S., & Verstrepen, K. J. (2016). Domestication and Divergence of *Saccharomyces cerevisiae* Beer Yeasts. *Cell*, *166*(6), 1397-1410.e16. https://doi.org/10.1016/j.cell.2016.08.020

Gibson, A. W., Wojciechowicz, L. A., Danzi, S. E., Zhang, B., K, J. H., Hut, Z., & Michels, C. A. (1997). *Constitutive Mutations of the Saccharomyces cerevisiae MAL-Activator Genes MAL23, MAL43, MAL63, and mal64*.

Gou, L., Bloom, J. S., & Kruglyak, L. (2019). The Genetic Basis of Mutation Rate Variation in Yeast. *Genetics*, *211*(2), 731–740. https://doi.org/10.1534/genetics.118.301609

Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J., & Fowler, D. M. (2018). Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. In *Cell Systems* (Vol. 6, Issue 1, pp. 116-124.e3). https://doi.org/10.1016/j.cels.2017.11.003

Gray, V. E., Sitko, K., Ngako Kameni, F. Z., Williamson, M., Stephany, J. J., Hasle, N., & Fowler, D. M. (2019). Elucidating the molecular determinants of Ab aggregation with deep mutational scanning. *G3: Genes, Genomes, Genetics*, *9*(11), 3683–3689.

https://doi.org/10.1534/g3.119.400535

Gresham, D., Desai, M. M., Tucker, C. M., Jenq, H. T., Pai, D. A., Ward, A., DeSevo, C. G., Botstein, D., & Dunham, M. J. (2008). The Repertoire and Dynamics of Evolutionary Adaptations to Controlled Nutrient-Limited Environments in Yeast. *PLoS Genetics*, *4*(12), e1000303. https://doi.org/10.1371/journal.pgen.1000303

Guan, Y., Dunham, M. J., & Troyanskaya, O. G. (2007). Functional Analysis of Gene Duplications in *Saccharomyces cerevisiae*. *Genetics*, *175*(2), 933–943. https://doi.org/10.1534/genetics.106.064329

Guo, X., Chavez, A., Tung, A., Chan, Y., Kaas, C., Yin, Y., Cecchi, R., Garnier, S. L., Kelsic, E. D., Schubert, M., DiCarlo, J. E., Collins, J. J., & Church, G. M. (2018). High-throughput creation and functional profiling of DNA sequence variant libraries using CRISPR–Cas9 in yeast. *Nature Biotechnology*. https://doi.org/10.1038/nbt.4147

Haas, R., Horev, G., Lipkin, E., Kesten, I., Portnoy, M., Buhnik-Rosenblau, K., Soller, M., & Kashi, Y. (2019). Mapping Ethanol Tolerance in Budding Yeast Reveals High Genetic Variation in a Wild Isolate. *Front. Genet.*, *10*, 998. https://doi.org/10.3389/fgene.2019.00998

Hamza, A., Driessen, M. R. M., Tammpere, E., O'Neil, N. J., & Hieter, P. (2020). Cross-Species Complementation of Nonessential Yeast Genes Establishes Platforms for Testing Inhibitors of Human Proteins. *Genetics*, *214*(3), 735–747. https://doi.org/10.1534/genetics.119.302971

Hiatt, J. B., Patwardhan, R. P., Turner, E. H., Lee, C., & Shendure, J. (2010). Parallel, tag-directed assembly of locally derived short sequence reads. *Nature Methods*, *7*(2), 119–122. https://doi.org/10.1038/nmeth.1416

Hietpas, R. T., Bank, C., Jensen, J. D., & Bolon, D. N. A. (2013). Shifting fitness landscapesin

response to altered environments. *Evolution*, *67*(12), 3512–3522. https://doi.org/10.1111/evo.12207

Higgins, D. A., Young, M. K. M., Tremaine, M., Sardi, M., Fletcher, J. M., Agnew, M., Liu, L., Dickinson, Q., Peris, D., Wrobel, R. L., Hittinger, C. T., Gasch, A. P., Singer, S. W., Simmons, B. A., Landick, R., Thelen, M. P., & Sato, T. K. (2018). Natural Variation in the Multidrug Efflux Pump *SGE1* Underlies Ionic Liquid Tolerance in Yeast. *Genetics*, *210*(1), 219–234. https://doi.org/10.1534/genetics.118.301161

Hoggard, T., Liachko, I., Burt, C., Meikle, T., Jiang, K., Craciun, G., Dunham, M. J., & Fox, C. A. (2016). *High Throughput Analyses of Budding Yeast ARSs Reveal New DNA Elements Capable of Conferring Centromere-Independent Plasmid Propagation*. https://doi.org/10.1534/g3.116.027904

Holland, L. Z., & Ocampo Daza, D. (2018). A new look at an old question: When did the second whole genome duplication occur in vertebrate evolution? *Genome Biology*, *19*(1), 1–4. https://doi.org/10.1186/s13059-018-1592-0

Jackson, C. A., Castro, D. M., Saldi, G.-A., Bonneau, R., & Gresham, D. (2020). Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *ELife*, *9*. https://doi.org/10.7554/eLife.51254

Jakobson, C. M., & Jarosz, D. F. (2019). Molecular Origins of Complex Heritability in Natural Genotype-to-Phenotype Relationships. *Cell Syst*, *8*(5), 363-379.e3. https://doi.org/10.1016/j.cels.2019.04.002

Jansen, M. L. A., Daran-Lapujade, P., De Winde, J. H., Piper, M. D. W., & Pronk, J. T. (2004). Prolonged Maltose-Limited Cultivation of *Saccharomyces cerevisiae* Selects for Cells with Improved Maltose Affinity and Hypersensitivity. *Applied and Environmental Microbiology*,

70(4), 1956–1963. https://doi.org/10.1128/AEM.70.4.1956-1963.2004

Jelier, R., Semple, J. I., Garcia-Verdugo, R., & Lehner, B. (2011). Predicting phenotypic variation in yeast from individual genome sequences. *Nature Genetics*, *43*. https://doi.org/10.1038/ng.1007

Johnson, A. J., Veljanoski, F., O 'doherty, P. J., Zaman, M. S., Petersingham, G., Bailey, T. D., Mü, G., Kersaitis, C., & Wu, M. J. (2016). Revelation of molecular basis for chromium toxicity by phenotypes of *Saccharomyces cerevisiae* gene deletion mutants. *Metallomics*, *8*(8), 542–550. https://doi.org/10.1039/c6mt00039h

Johnson, L. J., Koufopanou, V., Goddard, M. R., Hetherington, R., Schäfer, S. M., & Burt, A. (2004). Population Genetics of the Wild Yeast *Saccharomyces paradoxus*. *Genetics*, *166*(1), 43–52. https://doi.org/10.1534/genetics.166.1.43

Johnson, T., & Barton, N. (2005). Theoretical models of selection and mutation on quantitative traits. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1459), 1411–1425. https://doi.org/10.1098/rstb.2005.1667

Kapolka, N. J., Taghon, G. J., Rowe, J. B., Morgan, W. M., Enten, J. F., Lambert, N. A., & Isom, D. G. (2020). Dcyfir: A high-throughput CRISPR platform for multiplexed G protein-coupled receptor profiling and ligand discovery. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(23), 13117–13126. https://doi.org/10.1073/pnas.2000430117

Kellis, M., Birren, B. W., & Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, *428*(6983), 617–624. https://doi.org/10.1038/nature02424

Kessi-Pérez, E. I., Araos, S., García, V., Salinas, F., Abarca, V., Larrondo, L. F., Martínez, C., &

Cubillos, F. A. (2016). *RIM15* antagonistic pleiotropy is responsible for differences in fermentation and stress response kinetics in budding yeast. *FEMS Yeast Research*, *16*(3), fow021. https://doi.org/10.1093/femsyr/fow021

Kim, H. S., & Fay, J. C. (2009). A combined-cross analysis reveals genes with drug-specific and background-dependent effects on drug sensitivity in *Saccharomyces cerevisiae*. *Genetics*, *183*(3), 1141–1151. https://doi.org/10.1534/genetics.109.108068

Kim, H. S., Huh, J., Riles, L., Reyes, A., & Fay, J. C. (2012). A noncomplementation screen for quantitative trait alleles in *Saccharomyces cerevisiae*. *G3: Genes, Genomes, Genetics*, *2*(7), 753–760. https://doi.org/10.1534/g3.112.002550

Kronenberg, Z. N., Rhie, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., Porubsky, D., Kuhn, K., Mueller, K. A., Low, W. Y., Hiendleder, S., Fedrigo, O., Liachko, I., Hall, R. J., Phillippy, A. M., Eichler, E. E., Williams, J. L., Smith, T. P. L., Jarvis, E. D., … Kingan, S. B. (2021). Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nature Communications*, *12*(1), 1–10. https://doi.org/10.1038/s41467-020-20536-y

Kuzmin, E., Vandersluis, B., Ba, A. N. N., Wang, W., Koch, E. N., Usaj, M., Khmelinskii, A., Usaj, M. M., Leeuwen, J. Van, Kraus, O., Tresenrider, A., Pryszlak, M., Hu, M. C., Varriano, B., Costanzo, M., Knop, M., Moses, A., Myers, C. L., Andrews, B. J., & Boone, C. (2020). Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. *Science*, *368*(6498). https://doi.org/10.1126/science.aaz5667

Lan, F., Haliburton, J. R., Yuan, A., & Abate, A. R. (2016). Droplet barcoding for massively parallel single-molecule deep sequencing. *Nature Communications*, *7*(1), 1–10. https://doi.org/10.1038/ncomms11784

Legras, J.-L., Galeote, V., Bigey, F., Camarasa, C., Marsit, S., Nidelet, T., Sanchez, I., Couloux, A., Guy, J., Franco-Duarte, R., Marcet-Houben, M., Gabaldon, T., Schuller, D., Sampaio, J. P., Dequin, S., & Wittkopp, P. (2018). Adaptation of *S. cerevisiae* to Fermented Food Environments Reveals Remarkable Genome Plasticity and the Footprints of Domestication. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msy066

Leh-Louis, V., Wirth, B., Despons, L., Wain-Hobson, S., Potier, S., & Souciet, J. L. (2004). Differential evolution of the *Saccharomyces cerevisiae DUP240* paralogs and implication of recombination in phylogeny. *Nucleic Acids Research*, *32*(7), 2069–2078. https://doi.org/10.1093/nar/gkh529

Li, F., Salit, M. L., & Levy, S. F. (2018). Unbiased Fitness Estimation of Pooled Barcode or Amplicon Sequencing Studies. *Cell Systems*, *7*(5), 521-525.e4. https://doi.org/10.1016/j.cels.2018.09.004

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Heng. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. http://arxiv.org/abs/1303.3997

Li, Heng. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Lian, J., Schultz, C., Cao, M., HamediRad, M., & Zhao, H. (2019). Multi-functional genome-wide CRISPR system for high throughput genotype–phenotype mapping. *Nature Communications*, *10*(1), 1–10. https://doi.org/10.1038/s41467-019-13621-4

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P.,

Roberts, I. N., Burt, A., Koufopanou, V., Tsai, I. J., Bergman, C. M., Bensasson, D., O'Kelly, M. J. T., van Oudenaarden, A., Barton, D. B. H., Bailes, E., Nguyen, A. N., Jones, M., … Louis, E. J. (2009). Population genomics of domestic and wild yeasts. *Nature*, *458*(7236), 337–341. https://doi.org/10.1038/nature07743

Livesey, B. J., & Marsh, J. A. (2020). Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.*, *16*(7), e9380. https://doi.org/10.15252/msb.20199380

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. In *Nature Reviews Genetics* (Vol. 21, Issue 10, pp. 597–614). Nature Research. https://doi.org/10.1038/s41576-020-0236-x

Lutz, S., Brion, C., Kliebhan, M., & Albert, F. W. (2019). DNA variants affecting the expression of numerous genes in trans have diverse mechanisms of action and evolutionary histories. *PLoS Genetics*, *15*(11), e1008375. https://doi.org/10.1371/journal.pgen.1008375

MacKay, T. F. C., Stone, E. A., & Ayroles, J. F. (2009). The genetics of quantitative traits: Challenges and prospects. In *Nature Reviews Genetics* (Vol. 10, Issue 8, pp. 565–577). Nature Publishing Group. https://doi.org/10.1038/nrg2612

Maclean, C. J., Metzger, B. P. H., Yang, J.-R., Ho, W.-C., Moyers, B., & Zhang, J. (2017). Deciphering the Genic Basis of Yeast Fitness Variation by Simultaneous Forward and Reverse Genetics. *Mol. Biol. Evol.*, *34*(10), 2486–2502. https://doi.org/10.1093/molbev/msx151

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., & Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *Journal of Genetics*, *92*(1), 155–161. https://doi.org/10.1007/s12041-013-0212-8

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R., Kjaer, T. R., Fine, R. S., Lu, Y., Schurmann, C., Highland, H. M., Rüeger, S., Thorleifsson, G., Justice, A. E., Lamparter, D., Stirrups, K. E., Turcot, V., Young, K. L., Winkler, T. W., Esko, T., … Lettre, G. (2017). Rare and low-frequency coding variants alter human adult height. *Nature*, *542*(7640), 186–190. https://doi.org/10.1038/nature21039

Matreyek, K. A., Starita, L. M., Stephany, J. J., Martin, B., Chiasson, M. A., Gray, V. E., Kircher, M., Khechaduri, A., Dines, J. N., Hause, R. J., Bhatia, S., Evans, W. E., Relling, M. V., Yang, W., Shendure, J., & Fowler, D. M. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genetics*, *50*(6), 874–882. https://doi.org/10.1038/s41588-018-0122-z

Matsui, T., Mullis, M. N., Roy, K. R., Hale, J. J., Schell, R., Levy, S. F., & Ehrenreich, I. M. (2022). The interplay of additivity, dominance, and epistasis on fitness in a diploid yeast cross. *Nature Communications*, *13*(1), 1–14. https://doi.org/10.1038/s41467-022-29111-z

McGlincy, N. J., Meacham, Z. A., Reynaud, K. K., Muller, R., Baum, R., & Ingolia, N. T. (2021). A genome-scale CRISPR interference guide library enables comprehensive phenotypic profiling in yeast. *BMC Genomics*, *22*(1), 205. https://doi.org/10.1186/s12864-021-07518-0

McLaughlin, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S., & Ranganathan, R. (2012). The

spatial architecture of protein function and adaptation. *Nature*, *491*(7422), 138–142. https://doi.org/10.1038/nature11500

Mitchell-Olds, T., Willis, J. H., & Goldstein, D. B. (2007). Which evolutionary processes influence natural genetic variation for phenotypic traits? In *Nature Reviews Genetics* (Vol. 8, Issue 11, pp. 845–856). Nature Publishing Group. https://doi.org/10.1038/nrg2207

Momen-Roknabadi, A., Oikonomou, P., Zegans, M., & Tavazoie, S. (2020). An inducible CRISPR interference library for genetic interrogation of *Saccharomyces cerevisiae* biology. *Communications Biology*, *3*(1), 1–12. https://doi.org/10.1038/s42003-020-01452-9

Morgante, F., Huang, W., Maltecca, C., & Mackay, T. F. C. (2018). Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals. *Heredity*, *120*(6), 500–514. https://doi.org/10.1038/s41437-017-0043-0

Morton, E. A., Dorrity, M. W., Zhou, W., Fields, S., & Queitsch, C. (2020). Transcriptional re-wiring by mutation of the yeast Hsf1 oligomerization domain 1. *BioRxiv*, 2020.05.23.112250. https://doi.org/10.1101/2020.05.23.112250

Naumov, G. I., Naumova, E. S., & Michelst, C. A. (1994). *Genetic Variation of the Repeated MAL Loci i n Natural Populations of Saccharomyces cereuisiae and Saccharomyces paradoxus*.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Newberry, R. W., Leong, J. T., Chow, E. D., Kampmann, M., & DeGrado, W. F. (2020). Deep mutational scanning reveals the structural basis for α-synuclein activity. *Nature Chemical Biology*, *16*(6), 653–659. https://doi.org/10.1038/s41589-020-0480-6

Nguyen Ba, A. N., Lawrence, K. R., Rego-Costa, A., Gopalakrishnan, S., Temko, D., Michor, F., & Desai, M. M. (2022). Barcoded Bulk QTL mapping reveals highly polygenic and epistatic architecture of complex traits in yeast. *ELife*, *11*. https://doi.org/10.7554/eLife.73983

Nogami, S., Ohya, Y., & Yvert, G. (2007). Genetic Complexity and Quantitative Trait Loci Mapping of Yeast Morphological Traits. *PLoS Genetics*, *3*(2), e31. https://doi.org/10.1371/journal.pgen.0030031

Noree, C., Sirinonthanawech, N., & Wilhelm, J. E. (2019). *Saccharomyces cerevisiae ASN1* and *ASN2* are asparagine synthetase paralogs that have diverged in their ability to polymerize in response to nutrient stress. *Scientific Reports*, *9*(1), 1–8. https://doi.org/10.1038/s41598-018-36719-z

Ohdate, T., Omura, F., Hatanaka, H., Zhou, Y., Takagi, M., Goshima, T., Akao, T., & Ono, E. (2018). *MAL73*, a novel regulator of maltose fermentation, is functionally impaired by single nucleotide polymorphism in sake brewing yeast. *PLOS ONE*, *13*(6), e0198744. https://doi.org/10.1371/JOURNAL.PONE.0198744

Ohno, S. (1970). *Evolution by Gene Duplication*. Springer Verlag.

Ollodart, A. R., Yeh, C.-L. C., Miller, A. W., Shirts, B. H., Gordon, A. S., & Dunham, M. J. (2021). Multiplexing mutation rate assessment: determining pathogenicity of Msh2 variants in *Saccharomyces cerevisiae*. *Genetics*, *218*(2), iyab058. https://doi.org/10.1093/genetics/iyab058

Ota, T., & Nei, M. (1995). Evolution of immunoglobulin *VH* pseudogenes in chickens. *Molecular Biology and Evolution*, *12*(1), 94–102. https://doi.org/10.1093/oxfordjournals.molbev.a040194

Payen, C., Di Rienzi, S. C., Ong, G. T., Pogachar, J. L., Sanchez, J. C., Sunshine, A. B., Raghuraman, M. K., Brewer, B. J., & Dunham, M. J. (2014). The dynamics of diverse segmental amplifications in populations of *Saccharomyces cerevisiae* adapting to strong selection. *G3 (Bethesda, Md.)*, *4*(3), 399–409. https://doi.org/10.1534/g3.113.009365

Peltier, E., Friedrich, A., Schacherer, J., & Marullo, P. (2019). Quantitative trait nucleotides impacting the technological performances of industrial *Saccharomyces cerevisiae* strains. *Frontiers in Genetics*, *10*(JUL), 683. https://doi.org/10.3389/fgene.2019.00683

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J.-M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., … Schacherer, J. (2018). Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature*, *556*(7701), 339–344. https://doi.org/10.1038/s41586-018-0030-5

Peter, J., Friedrich, A., Liti, G., & Schacherer, J. (2022). Extensive simulations assess the performance of genome-wide association mapping in various *Saccharomyces cerevisiae* subpopulations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *10.1098/rstb.2020.0514*.

Qian, W., Ma, D., Xiao, C., Wang, Z., & Zhang, J. (2012). The Genomic Landscape and Evolutionary Resolution of Antagonistic Pleiotropy in Yeast. *Cell Reports*, *2*(5), 1399–1410. https://doi.org/10.1016/J.CELREP.2012.09.017

Reeb, J., Wirth, T., & Rost, B. (2020). Variant effect predictions capture some aspects of deep mutational scanning experiments. *BMC Bioinformatics*, *21*(1), 107. https://doi.org/10.1186/s12859-020-3439-4

Renganaath, K., Cheung, R., Day, L., Kosuri, S., Kruglyak, L., & Albert, F. W. (2020).

Systematic identification of *cis*-regulatory variants that cause gene expression differences in a yeast cross. *ELife*, *9*. https://doi.org/10.7554/eLife.62669

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/J.GPB.2015.08.002

Rice, P., Longden, L., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. In *Trends in Genetics* (Vol. 16, Issue 6, pp. 276–277). Elsevier Ltd. https://doi.org/10.1016/S0168-9525(00)02024-2

Rich, M. S., Payen, C., Rubin, A. F., Ong, G. T., Sanchez, M. R., Yachie, N., Dunham, M. J., & Fields, S. (2016). Comprehensive Analysis of the *SUL1* Promoter of *Saccharomyces cerevisiae*. *Genetics*, *203*(1), 191–202. https://doi.org/10.1534/genetics.116.188037

Roy, K. R., Smith, J. D., Vonesch, S. C., Lin, G., Tu, C. S., Lederer, A. R., Chu, A., Suresh, S., Nguyen, M., Horecka, J., Tripathi, A., Burnett, W. T., Morgan, M. A., Schulz, J., Orsley, K. M., Wei, W., Aiyar, R. S., Davis, R. W., Bankaitis, V. A., … Steinmetz, L. M. (2018). Multiplexed precision genome editing with trackable genomic barcodes in yeast. *Nat. Biotechnol.*, *36*(6), 512–520. https://doi.org/10.1038/nbt.4137

Saada, O. A., Tsouris, A., Large, C., Friedrich, A., Dunham, M. J., & Schacherer, J. (2022). Phased polyploid genomes provide deeper insight into the multiple origins of domesticated *Saccharomyces cerevisiae* beer yeasts. *Current Biology*. https://doi.org/10.1016/j.cub.2022.01.068

Sadhu, M. J., Bloom, J. S., Day, L., Siegel, J. J., Kosuri, S., & Kruglyak, L. (2018). Highly parallel genome variant engineering with CRISPR–Cas9. *Nat. Genet.*, *50*(4), 510–514. https://doi.org/10.1038/s41588-018-0087-y

Salinas, F., de Boer, C. G., Abarca, V., García, V., Cuevas, M., Araos, S., Larrondo, L. F.,

Martínez, C., & Cubillos, F. A. (2016). Natural variation in non-coding regions underlying phenotypic diversity in budding yeast. *Scientific Reports*, *6*(1), 21849. https://doi.org/10.1038/srep21849

Samonte, R. V., & Eichler, E. E. (2002). Segmental duplications and the evolution of the primate genome. In *Nature Reviews Genetics* (Vol. 3, Issue 1, pp. 65–72). Nat Rev Genet. https://doi.org/10.1038/nrg705

Sanchez, M. R., Miller, A. W., Liachko, I., Sunshine, A. B., Lynch, B., Huang, M., Alcantara, E., DeSevo, C. G., Pai, D. A., Tucker, C. M., Hoang, M. L., & Dunham, M. J. (2017). Differential paralog divergence modulates genome evolution across yeast species. *PLoS Genetics*, *13*(2). https://doi.org/10.1371/journal.pgen.1006585

Sardi, M., Paithane, V., Place, M., Robinson, D. E., Hose, J., Wohlbach, D. J., & Gasch, A. P. (2018). Genome-wide association across *Saccharomyces cerevisiae* strains reveals substantial variation in underlying gene requirements for toxin tolerance. *PLoS Genet.*, *14*(2), e1007217. https://doi.org/10.1371/journal.pgen.1007217

Schacherer, J., Shapiro, J. A., Ruderfer, D. M., & Kruglyak, L. (2009). Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature*, *458*(7236), 342–345. https://doi.org/10.1038/nature07670

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Research*, *33*(SUPPL. 2), W382–W388. https://doi.org/10.1093/nar/gki387

Sharon, E., Chen, S. A. A., Khosla, N. M., Smith, J. D., Pritchard, J. K., & Fraser, H. B. (2018). Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell*, *175*(2), 544-557.e16. https://doi.org/10.1016/j.cell.2018.08.057

She, R., & Jarosz, D. F. (2018). Mapping Causal Variants with Single-Nucleotide Resolution Reveals Biochemical Drivers of Phenotypic Change. *Cell*, *172*(3), 478-490.e15. https://doi.org/10.1016/j.cell.2017.12.015

Shih, C.-H., & Fay, J. (2021). *Cis*-regulatory variants affect gene expression dynamics in yeast. *Elife*, *10*. https://doi.org/10.7554/eLife.68469

Siggers, T., Duyzend, M. H., Reddy, J., Khan, S., & Bulyk, M. L. (2011). Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Molecular Systems Biology*, *7*. https://doi.org/10.1038/msb.2011.89

Sirr, A., Lo, R. S., Cromie, G. A., Scott, A. C., Ashmead, J., Heyesus, M., & Dudley, A. M. (2020). A yeast-based complementation assay elucidates the functional impact of 200 missense variants in human *PSAT1*. *Journal of Inherited Metabolic Disease*, *43*(4), 758–769. https://doi.org/10.1002/jimd.12227

Starita, L. M., Ahituv, N., Dunham, M. J., Kitzman, J. O., Roth, F. P., Seelig, G., Shendure, J., & Fowler, D. M. (2017). Variant Interpretation: Functional Assays to the Rescue. *The American Journal of Human Genetics*, *101*, 315–325. https://doi.org/10.1016/j.ajhg.2017.07.014

Starita, L. M., Young, D. L., Islam, M., Kitzman, J. O., Gullingsrud, J., Hause, R. J., Fowler, D. M., Parvin, J. D., Shendure, J., & Fields, S. (2015). Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics*, *200*(2), 413–422. https://doi.org/10.1534/genetics.115.175802

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veesler, D., & Bloom, J. D. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals

Constraints on Folding and ACE2 Binding. *Cell*, *182*(5), 1295-1310.e20. https://doi.org/10.1016/j.cell.2020.08.012

Steinmetz, L. M., Sinha, H., Richards, D. R., Spiegelman, J. I., Oefner, P. J., McCusker, J. H., & Davis, R. W. (2002). Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, *416*(6878), 326–330. https://doi.org/10.1038/416326a

Stinchcombe, J. R., & Hoekstra, H. E. (2008). Combining population genomics and quantitative genetics: Finding the genes underlying ecologically important traits. In *Heredity* (Vol. 100, Issue 2, pp. 158–170). Nature Publishing Group. https://doi.org/10.1038/sj.hdy.6800937

Stöcker, B. K., Köster, J., & Rahmann, S. (2016). SimLoRD: Simulation of Long Read Data. *Bioinformatics*, *32*(17), 2704–2706. https://doi.org/10.1093/bioinformatics/btw286

Stothard, P. (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques*, *28*(6). https://doi.org/10.2144/00286ir01

Strope, P. K., Skelly, D. A., Kozmin, S. G., Mahadevan, G., Stone, E. A., Magwene, P. M., Dietrich, F. S., & McCusker, J. H. (2015). The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Research*, *25*(5), 762–774. https://doi.org/10.1101/gr.185538.114

Suiter, C. C., Moriyama, T., Matreyek, K. A., Yang, W., Scaletti, E. R., Nishii, R., Yang, W., Hoshitsuki, K., Singh, M., Trehan, A., Parish, C., Smith, C., Li, L., Bhojwani, D., Yuen, L. Y. P., Li, C. kong, Li, C. ho, Yang, Y. li, Walker, G. J., … Yang, J. J. (2020). Massively parallel variant characterization identifies *NUDT15* alleles associated with thiopurine toxicity. *Proceedings of the National Academy of Sciences of the United States of America*,

117(10), 5394–5401. https://doi.org/10.1073/pnas.1915680117

Thompson, S., Zhang, Y., Ingle, C., Reynolds, K. A., & Kortemme, T. (2020). Altered expression of a quality control protease in *E. coli* reshapes the in vivo mutational landscape of a model enzyme. *ELife*, *9*, 1–47. https://doi.org/10.7554/eLife.53476

Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L., & Baker, D. (2013). Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*, *501*(7466), 212–216. https://doi.org/10.1038/nature12443

Torabi, N., & Kruglyak, L. (2011). Variants in SUP45 and TRM10 Underlie Natural Variation in Translation Termination Efficiency in *Saccharomyces cerevisiae*. *PLoS Genetics*, *7*(7), e1002211. https://doi.org/10.1371/journal.pgen.1002211

Torrents, D., Suyama, M., Zdobnov, E., & Bork, P. (2003). A genome-wide survey of human pseudogenes. *Genome Research*, *13*(12), 2559–2567. https://doi.org/10.1101/gr.1455503

Treusch, S., Albert, F. W., Bloom, J. S., Kotenko, I. E., & Kruglyak, L. (2015). Genetic Mapping of MAPK-Mediated Complex Traits Across *S. cerevisiae. PLoS Genetics*, *11*(1). https://doi.org/10.1371/journal.pgen.1004913

Vaishnav, E. D., de Boer, C. G., Molinet, J., Yassour, M., Fan, L., Adiconis, X., Thompson, D. A., Levin, J. Z., Cubillos, F. A., & Regev, A. (2022). The evolution, evolvability and engineering of gene regulatory DNA. *Nature*, *603*(7901), 455–463. https://doi.org/10.1038/s41586-022-04506-6

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. In *American Journal of Human Genetics* (Vol. 90, Issue 1, pp. 7–24). Cell Press. https://doi.org/10.1016/j.ajhg.2011.11.029

Voordeckers, K, Brown, C. A., Vanneste, K., Van Der Zande, E., & Voet, A. (2012). Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication. *PLoS Biol*, *10*(12), 1001446. https://doi.org/10.1371/journal.pbio.1001446

Voordeckers, Karin, De Maeyer, D., van der Zande, E., Vinces, M. D., Meert, W., Cloots, L., Ryan, O., Marchal, K., & Verstrepen, K. J. (2012). Identification of a complex genetic network underlying *Saccharomyces cerevisiae* colony morphology. *Molecular Microbiology*, *86*(1), 225–239. https://doi.org/10.1111/j.1365-2958.2012.08192.x

Wagih, O., Galardini, M., Busby, B. P., Memon, D., Typas, A., & Beltrao, P. (2018). A resource of variant effect predictions of single nucleotide variants in model organisms. *Molecular Systems Biology*, *14*(12). https://doi.org/10.15252/msb.20188430

Wagih, O., & Parts, L. (2014). Gitter: A robust and accurate method for quantification of colony sizes from plate images. *G3: Genes, Genomes, Genetics*, *4*(3), 547–552. https://doi.org/10.1534/g3.113.009431

Walker, M. E., Zhang, J., Sumby, K. M., Lee, A., Houlès, A., Li, S., & Jiranek, V. (2021). Sulfate transport mutants affect hydrogen sulfide and sulfite production during alcoholic fermentation. *Yeast*, *38*(6), 367–381. https://doi.org/10.1002/yea.3553

Wanat, J. J., Singh, N., & Alani, E. (2007). The effect of genetic background on the function of *Saccharomyces cerevisiae MLH1* alleles that correspond to HNPCC missense mutations. *Human Molecular Genetics*, *16*(4), 445–452. https://doi.org/10.1093/hmg/ddl479

Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., & Sabatini, D. M. (2015). Identification and characterization of essential genes in the human genome. *Science*, *350*(6264), 1096–1101. https://doi.org/10.1126/science.aac7041

Wang, Z., Qi, Q., Lin, Y., Guo, Y., Liu, Y., & Wang, Q. (2019). QTL analysis reveals genomic variants linked to high-temperature fermentation performance in the industrial yeast. *Biotechnol. Biofuels*, *12*, 59. https://doi.org/10.1186/s13068-019-1398-7

Weile, J., & Roth, F. P. (2018). Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. In *Human Genetics* (Vol. 137, Issue 9, pp. 665–678). Springer Verlag. https://doi.org/10.1007/s00439-018-1916-x

Weile, J., Sun, S., Cote, A. G., Knapp, J., Verby, M., Mellor, J. C., Wu, Y., Pons, C., Wong, C., van Lieshout, N., Yang, F., Tasan, M., Tan, G., Yang, S., Fowler, D. M., Nussbaum, R., Bloom, J. D., Vidal, M., Hill, D. E., … Roth, F. P. (2017). A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.*, *13*(12), 957. https://doi.org/10.15252/msb.20177908

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Wilkening, S., Lin, G., Fritsch, E. S., Tekkedil, M. M., Anders, S., Kuehn, R., Nguyen, M., Aiyar, R. S., Proctor, M., Sakhanenko, N. A., Galas, D. J., Gagneur, J., Deutschbauer, A., & Steinmetz, L. M. (2014). An evaluation of high-throughput approaches to QTL mapping in *Saccharomyces cerevisiae*. *Genetics*, *196*(3), 853–865. https://doi.org/10.1534/genetics.113.160291

Will, J. L., Kim, H. S., Clarke, J., Painter, J. C., Fay, J. C., & Gasch, A. P. (2010). Incipient

Balancing Selection through Adaptive Loss of Aquaporins in Natural *Saccharomyces cerevisiae* Populations. *PLoS Genetics*, *6*(4), e1000893. https://doi.org/10.1371/journal.pgen.1000893

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., & Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. In *Nature Reviews Genetics* (Vol. 14, Issue 7, pp. 507–515). Nature Publishing Group. https://doi.org/10.1038/nrg3457

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. https://doi.org/10.1038/ng.608

Yeh, C.-L. C., Tsouris, A., Schacherer, J., & Dunham, M. J. (2021). High-throughput functional analysis of natural variants in yeast. *BioRxiv*, 2021.02.26.433108. https://doi.org/10.1101/2021.02.26.433108

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Zhang, Jiajie, Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593

Zhang, Jianzhi. (2003). Evolution by gene duplication: an update. *Trends in Ecology and Evolution*, *18*(6), 292–298. https://doi.org/10.1016/S0169-5347(03)00033-8

Zhang, N., Wu, J., & Oliver, S. G. (2009). Gis1 is required for transcriptional reprogramming of

carbon metabolism and the stress response during Transition into stationary phase in yeast. *Microbiology*, *155*(5), 1690–1698. https://doi.org/10.1099/mic.0.026377-0

Zhang, X., Kuang, X., Cao, F., Chen, R., Fang, Z., Liu, W., Shi, P., Wang, H., Shen, Y., & Huang, Z. (2020). Effect of cadmium on mRNA mistranslation in *Saccharomyces cerevisiae*. *Journal of Basic Microbiology*. https://doi.org/10.1002/jobm.201900495

Zhou, X., & Cai, X. (2021). Joint eQTL mapping and inference of gene regulatory network improves power of detecting both *cis*- and *trans*-eQTLs. In *Bioinformatics* (Vol. 38, Issue 1, pp. 149–156). https://doi.org/10.1093/bioinformatics/btab609

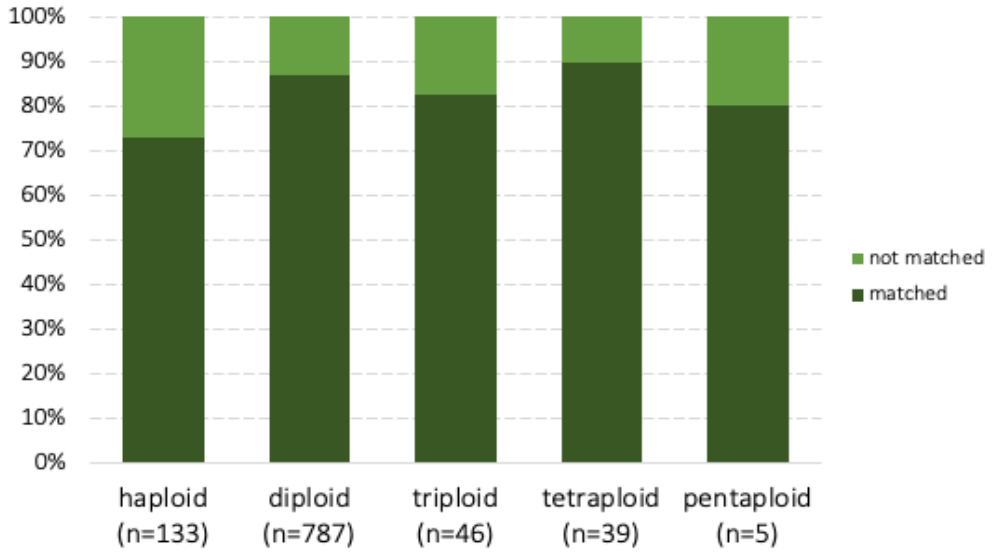
Zhu, Y. O., Sherlock, G., & Petrov, D. A. (2016). Whole genome analysis of 132 clinical *Saccharomyces cerevisiae* strains reveals extensive ploidy variation. *G3: Genes, Genomes, Genetics*, *6*(8), 2421–2434. https://doi.org/10.1534/g3.116.029397

Zimmer, C. T., Garrood, W. T., Singh, K. S., Randall, E., Lueke, B., Gutbrod, O., Matthiesen, S., Kohler, M., Nauen, R., Davies, T. G. E., & Bass, C. (2018). Neofunctionalization of Duplicated P450 Genes Drives the Evolution of Insecticide Resistance in the Brown Planthopper. *Current Biology*, *28*(2), 268-274.e5. https://doi.org/10.1016/j.cub.2017.11.060

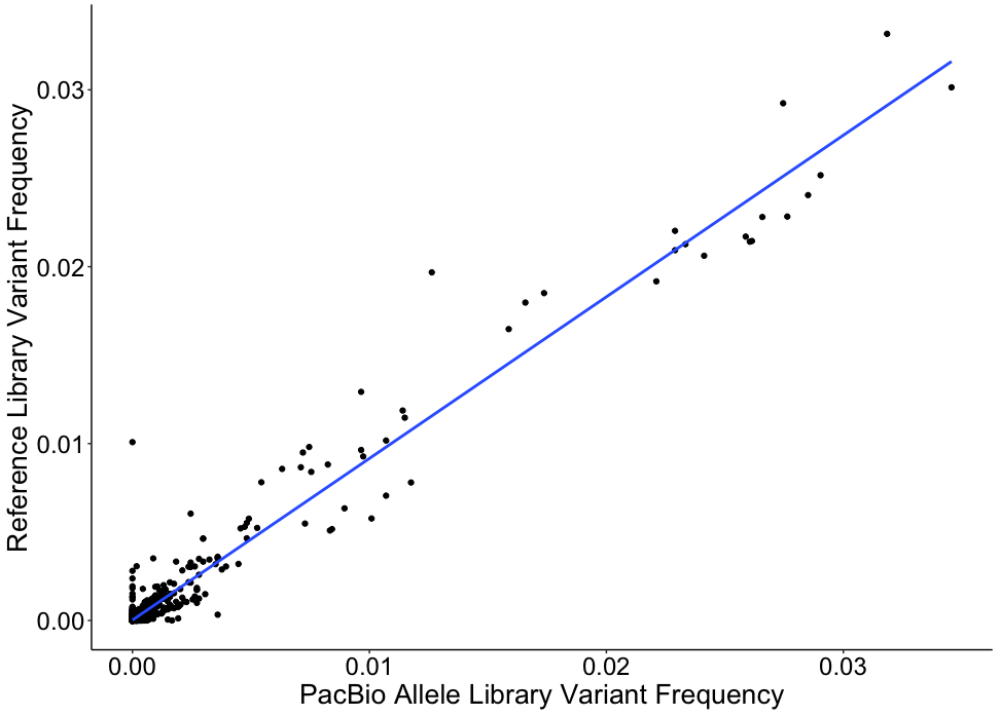
Zorgo, E., Gjuvsland, A., Cubillos, F. A., Louis, E. J., Liti, G., Blomberg, A., Omholt, S. W., & Warringer, J. (2012). Life History Shapes Trait Heredity by Accumulation of Loss-of-Function Alleles in Yeast. *Molecular Biology and Evolution*, *29*(7), 1781–1789. https://doi.org/10.1093/molbev/mss019

APPENDIX A: SUPPLEMENTARY FIGURES

Supplementary Figure 1. Percentage of strains for each genome-wide ploidy that matched to at least one PacBio read.

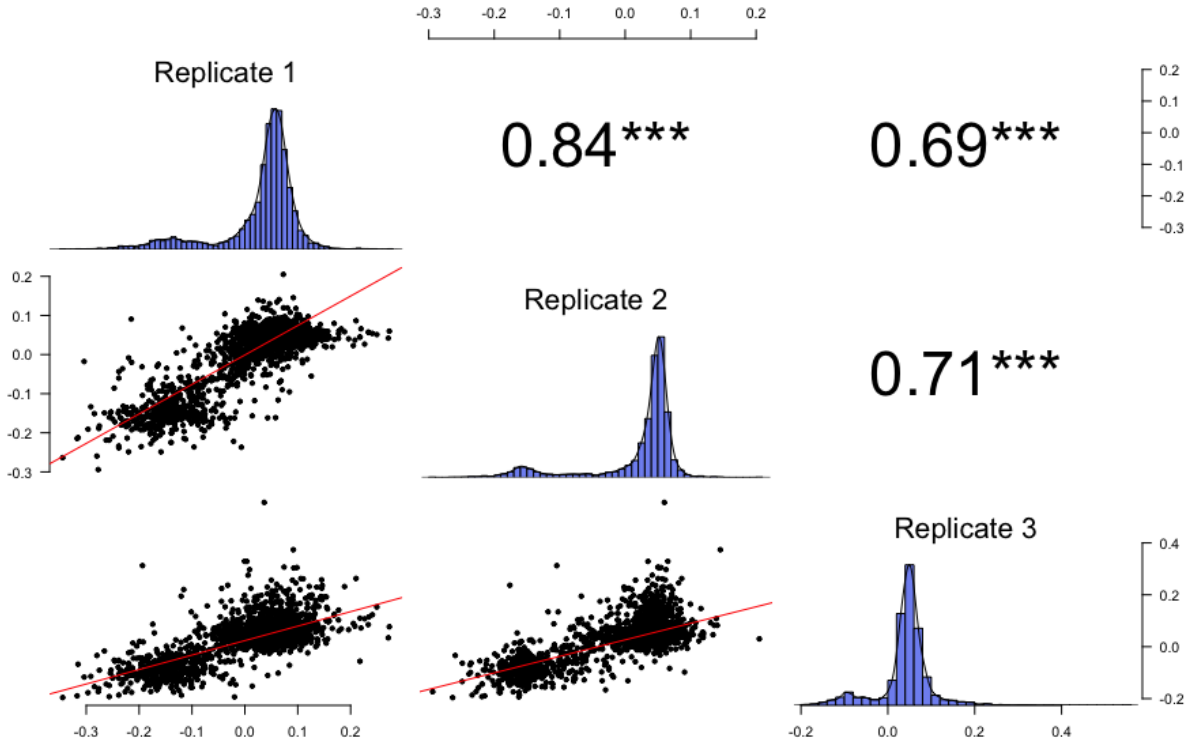


Supplementary Figure 2. Allele frequencies found in PacBio allele library reflect those found in the Illumina reference sequences (expected values).

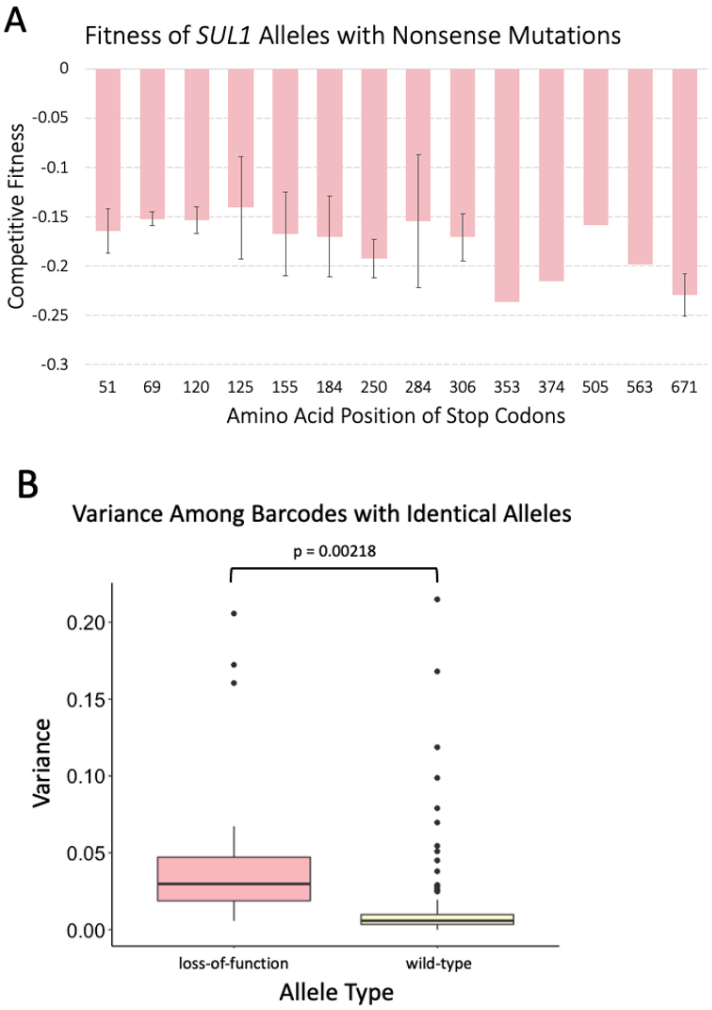


Supplementary Figure 3. Competitive fitness values calculated using FitSeq are well-correlated across replicates.

Pearson correlation coefficients *r* are listed on the top half. ***p<2.2e-16

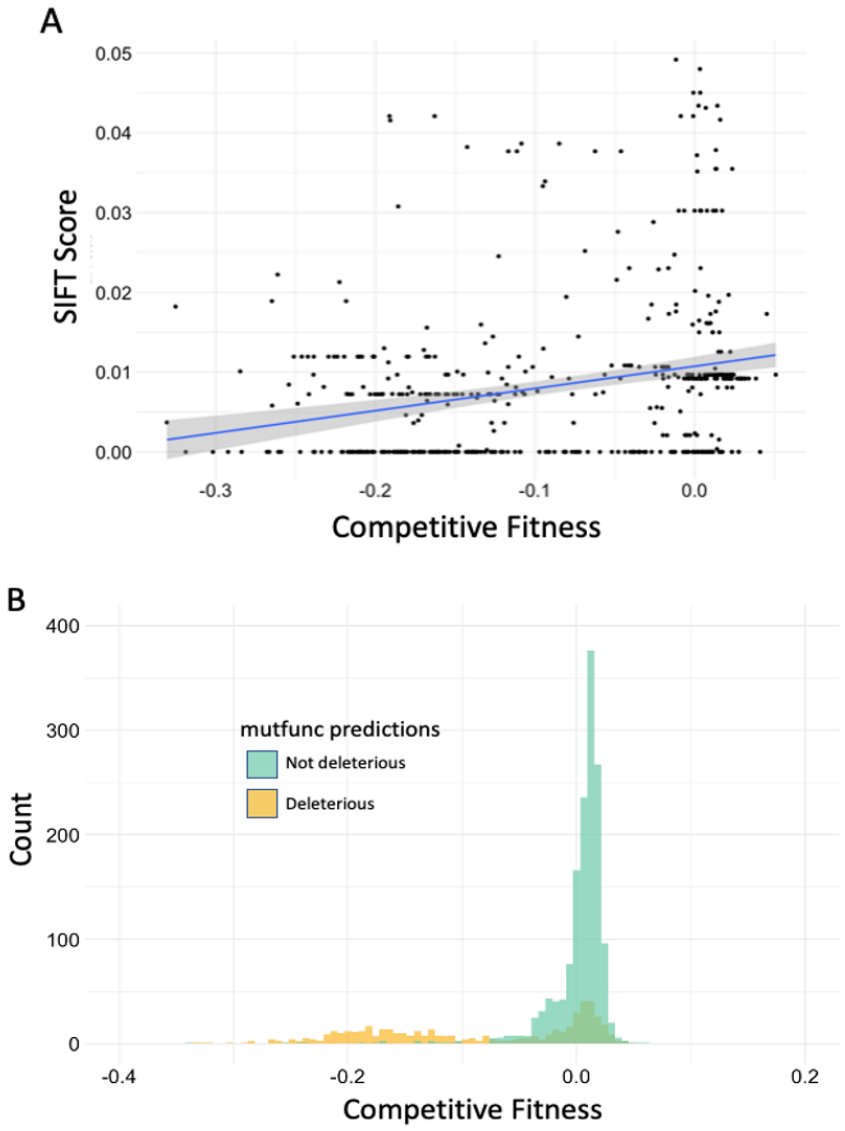


Supplementary Figure 4. Fitness and variance among loss-of-function alleles.
a) Barplot representing average fitness and standard deviation of barcodes categorized by location of premature stop codons (SUL1 contains 859 amino acids). Sites without error bars are represented by only one barcode. **b)** Barcodes associated with loss-of-function alleles tend to have greater variance compared to barcodes with wild-type fitness.



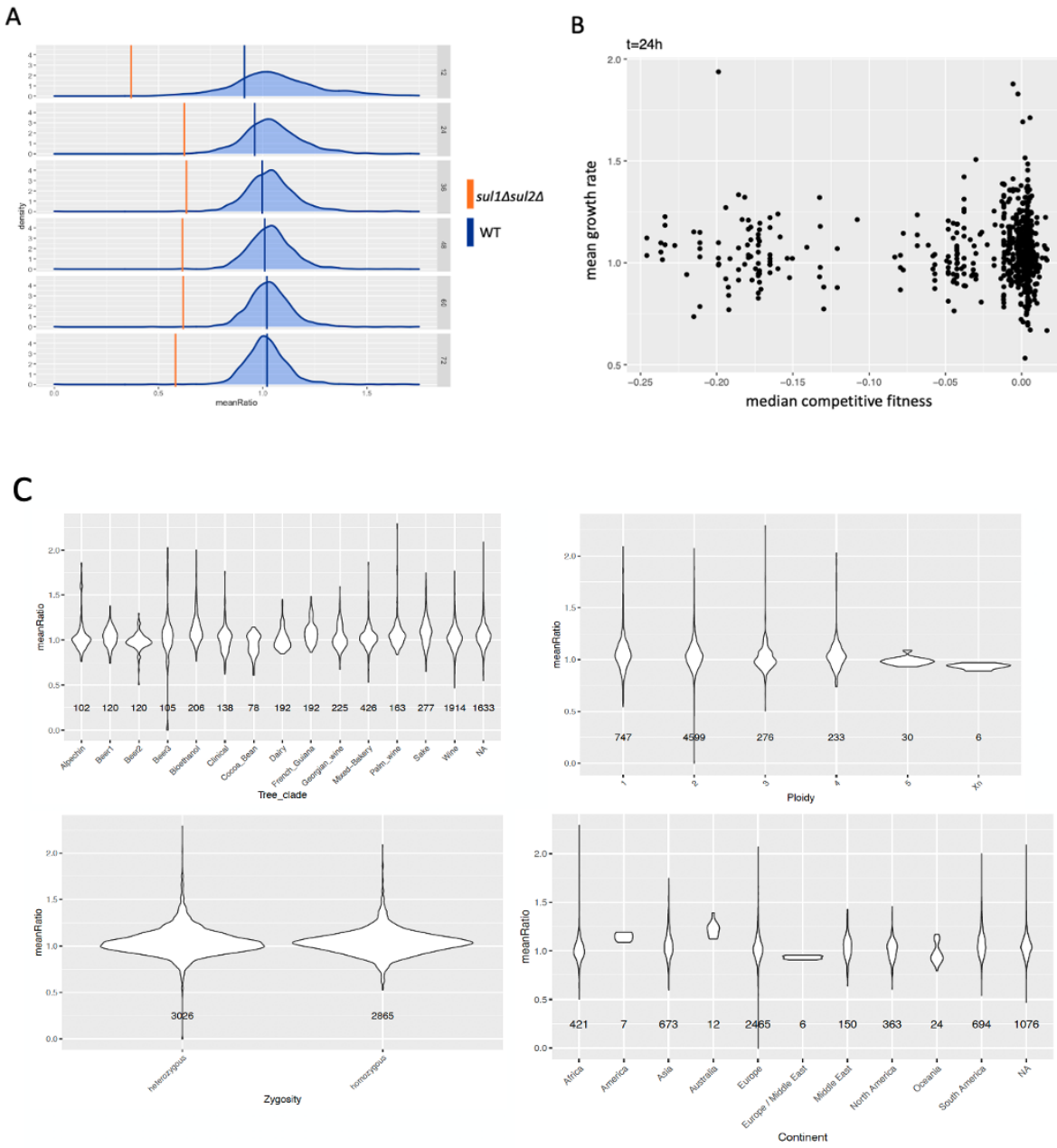
Supplementary Figure 5. mutfunc predicts which mutations are deleterious.

For our data, mutfunc returned SIFT scores for each mutation. We used the mutation with the most deleterious SIFT scores for each allele. **a)** Competitive fitness of allele from pooled natural variant library plotted against SIFT score of most deleterious mutation shows very little correlation (Pearson's correlation r=0.253). **b)** Distribution of experimentally assayed compared with mutfunc predictions of deleteriousness.



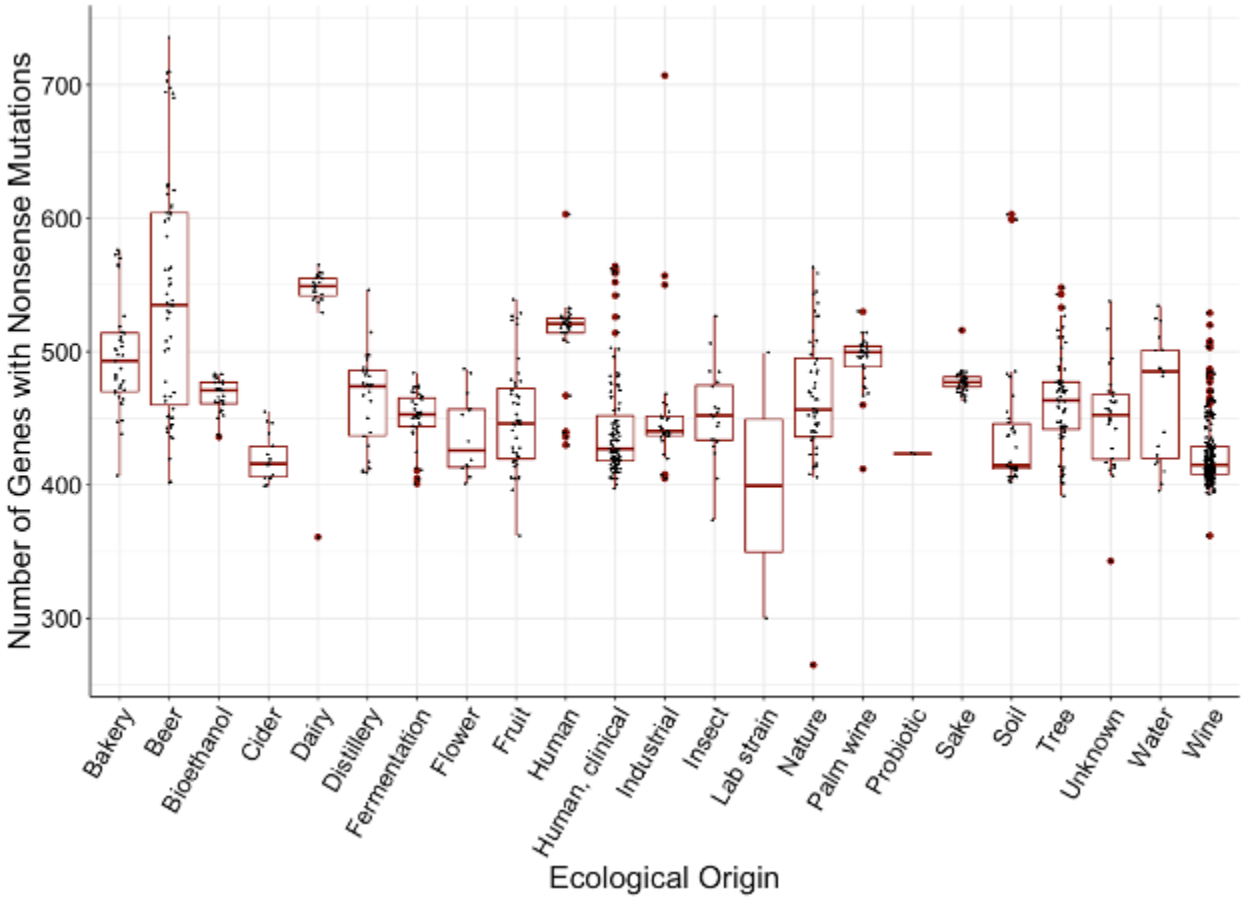
Supplementary Figure 6. Growth rates of unmodified isolates on sulfate limited solid plates.

For each plot, mean ratio indicates the growth rate on sulfate-limited media compared to growth on sulfate-rich media. **a)** Growth rate of *sul1*Δ*sul2*Δ strains (orange) and wild-type strain (blue) show differential growth on sulfate-limited media. **b)** Scatterplot comparing strain competitive fitness with growth rate on solid sulfate-limited media show no correlation. **c)** Grouped by clade, ploidy, zygosity, and continent, strains show no obvious pattern.



Supplementary Figure 7. Barplot of number of genes with premature stop codons per strain, grouped by ecological origins.

Variation within an ecological group may be due to geographical factors (for instance, yeast isolated from fruit in South America will have a different genetic makeup than yeast isolated from Southeast Asia).



APPENDIX B: SUPPLEMENTARY TABLES

Supplementary Table 1. Primers used in *SUL1* natural allele study

Primer ID	Primer Name	Function	Oligo sequence
1	SUL1_1011_Forward	Forward	GAATTCCTGCAGCCCGGGGGATCTC Gccttaaagacaatctggcgaaacc
2	SUL1_1011_Reverse	Reverse	GCGGCCGCTCTAGAACTAGTGGAT CTCGNNNNNNNNNaggacatttgtagaaaa taggctcaagc
25	pCY04_P25_amplifyForw ard	Forward	aatgatacggcgaccaccgagatctacacCCCAGT CACGACGTTGTAAAACGAC
26	PCY04_amplifyReverse_P 26	Reverse	caagcagaagacggcatacgagatTCTGTCTGG CCGTTTTCAGTAGTTTAAGGCGCTG
27	PCY04_amplifyReverse_P 27	Reverse	caagcagaagacggcatacgagatGGACAACCG CCGTTTTCAGTAGTTTAAGGCGCTG
28	PCY04_amplifyReverse_P 28	Reverse	caagcagaagacggcatacgagatGCTAGATGG CCGTTTTCAGTAGTTTAAGGCGCTG
29	PCY04_amplifyReverse_P 29	Reverse	caagcagaagacggcatacgagatCAATTTTCG CCGTTTTCAGTAGTTTAAGGCGCTG
30	PCY04_amplifyReverse_P 30	Reverse	caagcagaagacggcatacgagatGTGGGACAG CCGTTTTCAGTAGTTTAAGGCGCTG
31	PCY04_amplifyReverse_P 31	Reverse	caagcagaagacggcatacgagatCAGGACGCG CCGTTTTCAGTAGTTTAAGGCGCTG
32	PCY04_amplifyReverse_P 32	Reverse	caagcagaagacggcatacgagatCTAGAATAG CCGTTTTCAGTAGTTTAAGGCGCTG
33	PCY04_amplifyReverse_P 33	Reverse	caagcagaagacggcatacgagatTGGCTAGAG CCGTTTTCAGTAGTTTAAGGCGCTG
34	PCY04_amplifyReverse_P 34	Reverse	caagcagaagacggcatacgagatTCGCGGGCG CCGTTTTCAGTAGTTTAAGGCGCTG
35	PCY04_amplifyReverse_P 35	Reverse	caagcagaagacggcatacgagatCTGGGTGCG CCGTTTTCAGTAGTTTAAGGCGCTG
36	PCY04_amplifyReverse_P 36	Reverse	caagcagaagacggcatacgagatCTCCGGCTG CCGTTTTCAGTAGTTTAAGGCGCTG
37	PCY04_amplifyReverse_P 37	Reverse	caagcagaagacggcatacgagatGATAAGCTG CCGTTTTCAGTAGTTTAAGGCGCTG
38	PCY04_amplifyReverse_P 38	Reverse	caagcagaagacggcatacgagatAAACATGGG CCGTTTTCAGTAGTTTAAGGCGCTG
39	PCY04_amplifyReverse_P 39	Reverse	caagcagaagacggcatacgagatCCCGTGTAG CCGTTTTCAGTAGTTTAAGGCGCTG
40	PCY04_amplifyReverse_P 40	Reverse	caagcagaagacggcatacgagatATCCATTTG CCGTTTTCAGTAGTTTAAGGCGCTG
41	pCY04_SeqPrimer_P41_R ead1	Read1	CGGCCGCTCTAGAACTAGTGGATCT CG

44	pCY04_Read2	Read2	GTCTAGAAGTTAACTTGAGCTTGAG CCTATTTTCTACAAATGTCCT
77	P77_YJM320_SUL1_F	Forward	tctagaactagtggatctcgAAGTTTCCCCCAC TTTTC
78	P78_YJM320_SUL1_R	Forward	ctgcagcccgggggatctcgTTGTAAAGGGTA AGCTCATTC
79	P79_YJM320_SUL1_A	Forward	accgttcatattttcttgaacgctgtc
80	P80_YJM320_SUL1_B	Forward	agacgtcttgtacgattctattccagc
81	P81_YJM320_SUL1_C	Forward	ctgctcttatgggatacaactcattgg
82	P82_YJM320_SUL1_D	Forward	aacgactataaagtcgtccctgacc
83	P83_YJM320_SUL1_E	Forward	gatgctgtgaaatcatctagtgaattaccc
84	P84_YJM320_SUL1_F	Forward	gcccaagttgattccaccgc
85	P85_YJM320_SUL1_G	Forward	gggaaaagcgaaaataaacttttgtacgc
100	CY_pCY03_seq_Index_Primer_P100	Index	AGGACATTTGTAGAAAATAGGCTC AAGC
120	PCY04_amplifyReverse_P120	Reverse	caagcagaagacggcatacgagatATCTTGGGG CCGTTTTCAGTAGTTTAAGGCGCTG
121	PCY04_amplifyReverse_P121	Reverse	caagcagaagacggcatacgagatAGGGCATCG CCGTTTTCAGTAGTTTAAGGCGCTG
122	PCY04_amplifyReverse_P122	Reverse	caagcagaagacggcatacgagatCTGGACTTG CCGTTTTCAGTAGTTTAAGGCGCTG
123	PCY04_amplifyReverse_P123	Reverse	caagcagaagacggcatacgagatTCAAGAAGG CCGTTTTCAGTAGTTTAAGGCGCTG
124	PCY04_amplifyReverse_P124	Reverse	caagcagaagacggcatacgagatTGGAGTATG CCGTTTTCAGTAGTTTAAGGCGCTG
125	PCY04_amplifyReverse_P125	Reverse	caagcagaagacggcatacgagatCGTTATGGG CCGTTTTCAGTAGTTTAAGGCGCTG
126	PCY04_amplifyReverse_P126	Reverse	caagcagaagacggcatacgagatAGAACTATG CCGTTTTCAGTAGTTTAAGGCGCTG
127	PCY04_amplifyReverse_P127	Reverse	caagcagaagacggcatacgagatTATGTGAGG CCGTTTTCAGTAGTTTAAGGCGCTG
128	PCY04_amplifyReverse_P128	Reverse	caagcagaagacggcatacgagatAGCGTTTGG CCGTTTTCAGTAGTTTAAGGCGCTG
147	YJM450_pIL37_Gibson_F	Forward	gcggccgctctagaactagtggatctcgcgaTGGAT ATTTGTAAAGGGTAAGCTCA
148	YJM450_pIL37_Gibson_R	Reverse	gaattcctgcagcccgggggatctcgtgcatACTTA AAGAAAATTTGGCGAAACCT
149	SUL1_seq_A	Forward	CGTCTCAAAACCAGGTCAATA
150	SUL1_seq_B	Forward	CCTATTATAAAATGGTTTCCTC
151	SUL1_seq_C	Forward	TCCGCATTTAACATCATCTG
152	SUL1_seq_D	Forward	TGAGGTGGGAGTTATGAAAAT
153	SUL1_seq_E	Forward	CCACCTGGACCTTCTGGAAG
154	SUL1_seq_F	Forward	AGTCTATCGTTTGGGTGATAG
155	SUL1_seq_G	Forward	TCTCCATGGATCAAAAGAAGT

Supplementary Table 2. Plasmids and strains used in *SUL1* allele study

Strain	Alt ID	Collection	Genotype	Mating type	Species	Notes
DBY7284	FY3	DBY	*ura3-52 Gal+*	MATa	*S. cerevisiae*	
YMD1214	BDY127a	YMD	*hoΔ::GFP-KANMX*	MATa	*S. cerevisiae*	
DBY11069	FY4	DBY	prototroph	MATa	*S. cerevisiae*	
YMD4321	ILY53	YMD	*ura3-52 sul1Δ::URA3-KANMX*	MATa	*S. cerevisiae*	
YMD4322	ILY56	YMD	*ura3-52 sul2Δ::URA3-KANMX*	MATa	*S. cerevisiae*	
YMD4323	ILY67	YMD	*ura3-52 sul1Δ::URA3-KanMX sul2Δ::URA3-KanMX*	MATa	*S. cerevisiae*	
YMD2307	pIL37	YMD			*E. coli*	pRS316 vector with an NruI site inserted in the BamHI site. Used for cloning in *SUL1*. Plasmid map here: https://benchling.com/s/seq-GlLbFAfgfnaU9P8UMeP3
YMD2890	pMS10	YMD			*E. coli*	pIL37 with *S. paradoxus* (CBS432) *SUL1* cloned in NruI site
YMD4324	YCY69	YMD	*ura3-52 Gal+*	MATa	*S. cerevisiae*	DBY7284 with AAV *SUL1* inserted in pIL37 NruI site
YMD4325	YCY104	YMD	*ura3-52 Gal+*	MATa	*S. cerevisiae*	DBY7284 SACE_YCC *SUL1* inserted in pIL37 NruI site
YMD4326	YCY105	YMD	*ura3-52 Gal+*	MATa	*S. cerevisiae*	DBY7284 with AQM *SUL1* inserted in pIL37 NruI site

YMD 4327	YCY110	YMD	*ura3-52 Gal+*	MATa	*S. cerevisiae*	DBY7284 with BGM *SUL1* inserted in pIL37 NruI site
AAV	YMD1405, YJM320	YMD, 1,011			*S. cerevisiae*	Competed *SUL1* allele found to be introgressed from *S. paradoxus* (CBS432) vs YMD1214
SACE_YCC	YJM1355	1,011			*S. cerevisiae*	Competed *SUL1* allele found to be introgressed from *S. paradoxus* (CBS432) vs YMD1214
AKN	YPS128, YMD1749	YMD, 1,011			*S. cerevisiae*	Competed *SUL1* allele vs YMD1214 (previously done)
AQM	CBS420	1,011			*S. cerevisiae*	Competed *SUL1* allele vs YMD1214 (previously done)
BGM	CLIB560	1,011			*S. cerevisiae*	Competed *SUL1* allele vs YMD1214
BII	DBVPG1106, YMD1750 (previously done)	YMD, 1,011			*S. cerevisiae*	Competed *SUL1* allele vs YMD1214
ADA	Y55_1b, YMD1736	YMD, 1,011			*S. cerevisiae*	Competed *SUL1* allele vs YMD1214 (previously done)
AGK	CBS2361	1,011			*S. cerevisiae*	Tested growth in liquid sulfate-limited media
ACN	YJM269	1,011			*S. cerevisiae*	Tested growth in liquid sulfate-limited media
ABM	CBS3093	1,011			*S. cerevisiae*	Tested growth in liquid sulfate-limited media
BLP	DBVPG1661	1,011			*S. cerevisiae*	Tested growth in liquid sulfate-limited media
AQR	CBS1190	1,011			*S. cerevisiae*	Tested growth in liquid sulfate-limited media
BML	NCYC2780	1,011			*S. cerevisiae*	Tested growth in liquid sulfate-limited media
BEK	CLIB651	1,011			*S. cerevisiae*	Tested growth in liquid sulfate-limited media

CBB	1-9	1,011			*S. cerevisiae*	Tested growth in liquid sulfate-limited media
CLH	NCYC 2743	1,011			*S. cerevisiae*	Tested growth in liquid sulfate-limited media
ATV	CECT1462	1,011			*S. cerevisiae*	Tested growth in liquid sulfate-limited media
BGR	CLIB564	1,011			*S. cerevisiae*	Sanger sequenced *SUL1*
BDQ	CEY616	1,011			*S. cerevisiae*	Sanger sequenced *SUL1*
BBP	1663	1,011			*S. cerevisiae*	Sanger sequenced *SUL1*
BMQ	2720	1,011			*S. cerevisiae*	Sanger sequenced *SUL1*
BSR	L-958	1,011			*S. cerevisiae*	Sanger sequenced *SUL1*
AKT	CLQCA_10-003	1,011			*S. cerevisiae*	Sanger sequenced *SUL1*

Supplementary Table 3. Allele metadata for *SUL1*

Due to its size, this table was split into two parts. Use the Allele ID column for this section and the second table to pair strains with associated alleles and fitness values. One strain can be matched with multiple alleles since some strains in the 1,011 *S. cerevisiae* collection are polyploid and are heterozygous for *SUL1*. The full table can be found at https://github.com/cindyyeh/SUL1_natural_variants/blob/main/Supplementary%20Tables.xlsx

Table 3, part 1

Allele ID	Strains from 1,011 *S. cerevisiae* collection matched to allele
1	BKD,BNP
2	CFT,BSK,ABD,ATC,AGT,YDN
3	AGL,BPT,BQM,BQS,BQE,BQR,BLT,BPF,AGF,AQQ,ASG,BLG,BLI,ALL,AVH,AHR,YAD,AGH,BFQ,CIA,BFR,CIB,CBG,ACF,YCD,ARC,BER,ARK,GAP,BHC,BHF,BLB,BPP,CRH,AHT,ATN,YBO,BNK,BNR,BEQ,AIK,CAH,CBA,CAI,CAK,CAL,BSL,YBX,BSC,CFQ,CFR,YCQ,BNF,BQI,AGA,AGS,BLH,CRG,AIL,AMS,BQB,AMR,ABT,BRA,BRF,BSM,BSN,CPE,BPV,CRC,BHN,BBB,BAV,CAE,CAF,CRL,BBA,ACR,BIH,BKC,BLS,CBH,CRB,CHD,BIF,BKG,BKF,BLL,CKI,CKL,CKH,CKD,CKE,CIS,BMN,CIT,CKP,CKQ,CKB,CKC,CIQ,CIM,CIN,CIR,CKF,CKG,CKM,CIP,BCA,AMT,CIV,CKT,ABS,BLV,BQG,CRD,BVM,CQV,CRK,BLF,CEA,CEB,BLD,BNH,BLC,CEE,ANQ,AVK,BIR,BNG,AIM,CFB,CHF,AHP,AEK,AQP,CAB,CAP,BPR,CAM,CAQ,CBB,CBE,CAS,CAR,CAT,CBC,CAV,CBD,AIC,YAF,BSR,BSS,BTR,BTP,ASR,YDJ,AQA,AHL,BRK,YBR,AGN,BHD,BHM,BHL,BHK,BTV,BVA,CNK,CNP,CNH,BPI,BPN,YDB,AQH,YCF,BHG,ATS,BBP,BMG,ADP,AMK,AMI,AMG,AMQ,APK,APN,CIE,ATR,APM,API,ASL,BTF,ASP,BCR,BCS,BCT,ALV,CFP,BBF,AAP,ALN,ALP,BBD,APL,CFI,CFN,CHK,ADM,CHG,CHR,BML,BTE,BMM,CBF,CKK,CKS,CLA,YBW,BDQ,YAX,YCG,CBR,CFG,CFH,CNC,BMQ,ACN,CBK,CBL,CBM,ACA,ACB,ADT,GAV,CQC,CQD,CQG,CQH,CQL,ANN,CQM,CQE,CQK,CQF,CQN,CQP,AKS,ALG,YAP,YAQ,ABK,ABA,YAV,AAS,CEN,AAG,AHA,CHQ,CIH
6	ASD,ASE,BQA,BFB,ALS,BBH,BPQ,BEP,BTI
7	YCT,BHH,BHI,MAB,CKA,CHS,CHT
9	AAD,AQN,CLM,CMN
10	CEP,CGN,CGP,ACM,AKL,ABB,ACK,ADH
11	CLC,CLB,CLD
12	BTT,BVH,CNN,CNF,CNB,BVG,CNE,CNL,BVD,BVF,CNS,CNT
13	BDK
14	CHH,ADL,BDL,BFM,BFE,BFG,BTD,ADM,AKT,CHR,BTE,BMM,CBF,CKK,CLA,CBR,CFG,CFH
15	BNT
16	ASS,BSD,CQT

17	BQC,ARL
18	BFP
20	AFN
22	APC,APD,AVC,CMD,CLP,CMC,APF,APE,AVD,AVM,CLT,CLS,CME,CLL,CLR,CML,CMM,CMB,AVL,CLV,CLK,CQQ,CQR
24	ANP,ATS,BBP,CDR,BMG,CHL,AMK,AML,AMI,AMG,AMQ,APK,APN,CIE,AEM,AQI,ATR,APM,API,ASL,BTF,BCR,BCS,BCT,AGK,AGE,BTC,ALV,CFP,BBF,AAP,ALN,ALP,BBD,AAN,APL,AEV,CFI,CFN,CHK,CHG,CHR,BML,BTE,BMM,CBF,CKK,CKS,CLA
26	BIT
27	BLG,BLI,AHN,BSG,YCA
28	ANP,ATS,BBP,BMG,ADP,AMK,AMI,AMG,AMQ,APK,APN,CIE,ATR,APM,API,ASL,BTF,ASP,BCR,BCS,BCT,ALV,CFP,BBF,AAP,ALN,ALP,BBD,APL,CFI,CFN,CHK,CHG,CHR,BML,BTE,BMM,CBF,CKK,CKS,CLA
29	BMA,AKM,BMB,BMC,MAL
30	AAM,AFE,ATR,ABV,ALV,BBC,AMN,CFP,BBF,AAP,ALN,ALP,BBD,BBE,CHI,APL,CFI,CFN,CHK,CHH,BFE,BFG,BTD,AEE,ARA
31	AQD,ATQ,AID,AIE,YAE,YDM
32	CAK,CAL,BSR,BSS,BTP
33	BNE,BLA,BKT,BIA,BKL,BIB,BKK,BLQ,YBY,ARQ,AAA,CFS,ATM,BAR,BAS,BAT,BRD,BPC,CPA
37	AQB,BRH
38	CGB,YAW,CMP,CCH,ACC,CCC,CCD,CCE,CCF,CCG,CCI,CMQ,GAT,YCX,BBR
41	BPK,YDF,YDO,CHC,BKS,BLP,BLN,CHF
42	BKI,ABR,ABQ,BNV,BFK,BID,ACP,ALT,BBK,BFB,ATH,YAA,BSP,BQP,AAL,AHV,AFT
43	AQG
44	ABE,YAM
45	ATS,BBP,BMG,ASP,ADL,BDL,BFM,BFE,BFG,BTD,AKT,CHG,BML,BTE,BMM,CBF,CKK,CKS,CLA,ARB
47	ABP,BIR,BHL,BHK,BTV,BVA,CNK,CNP,CNH,CNC
49	AAL
51	YAY,ADI,YBA,YAZ,BMI,YBE,BMK,YBF,BMH,ADD,YBB,YBG,YBH
52	AIM,AIP,CAN,CAB,CAP,AHK,BVT,CAA,CAC,BVV,CAD,CAM,CAQ,ART
53	YBM,ABC,YDL
56	BCG,BCB,BCM,BCF,BCL,BCD,BDP,BCC,BEC,BED,BCE,BCQ,BDV,BDR,BMV,BNA,BCN,BEA,BEE
57	AIN,BLR,BNL,BPG,BPH,BQT,AIQ,BKH
58	BHV
59	AEI,AFG,YBU,CKR,AFH,AGG,CEV,CGI
60	BFD,BQN,CBI,CCM,BBQ,BBS
61	AHE

62	ABL,YCB
63	ATS,BBP,BMG,ATR,ASP,ALV,CFP,BBF,AAP,ALN,ALP,BBD,APL,CFI,CFN,CHK,CHH,ADL,BDL,BFM,BFE,BFG,BTD,ADM,AKT,CHG,CHR,BML,BTE,BMM,CBF,CKK,CKS,CLA,AGB,ATV,AAC,CBR,CFG,CFH,AEC,AQT
65	BTV,BVA,CNK,CNP,CNH,BDQ,CNC
66	BLG,BLI,CAI,CAK,CAL,CHA,BSR,BSS,BTP
67	YCI,CMH,ARM,CLI,CMK
68	CHN,CHP
69	AFL,AEN,ASC
70	ATS,BBP,BMG,ADP,AMK,AMI,AMG,AMQ,APK,APN,CIE,ASP,CHG,BML,BTE,BMM,CBF,CKK,CKS,CLA,CQG
71	ACT,ACV,AKP,AKV,ALE,BHB
72	BKQ,BQB
73	AVI,ACD,ANF,ANH,ANC,ANE,ANG,AND,ANK,ANI,AKN,YBS,YCU
74	BFB,BRC,ACS
76	AFC
77	BCV,BDB,BDD,BDA,BDE,BDH,BDI
78	APT,AKI,ADA,ATE
79	BBV,BVK,CFE
83	AFM,APA,ANT,ANR,ANV,ADN,CMF
84	CBN,AFB,AFP,BTA
85	AEF,CMH,ARM,CLQ,CGP
86	BEV,BET,BST,ADS
88	AKQ,YBZ,BFA,BFL,AMB,AMC
90	BIR,BHL,BHK,CHV,AEG,AGM,CNQ,AFD,BTV,BVA,CMT,CNG,CNK,APG,CNM,CNP,CNR,BVB,CNI,CNH,BVE,CNV,YCE,ADK
94	CHN,CHP
95	YBK,YBL,ATD,CDQ,CMS,ADG,CDP,YDI
96	AQT
97	AVB,CPG
99	AIM,CHF,AAK,AHP,AQP,CAB,CAP,BPR,BVT,CAA,CAC,BVV,CAD,CAM,CAQ,CBB,CBE,ART
101	AVP,AVR,AVS,AVT
102	CPC
103	ALL,ALM
104	CGI,ALB,AKR,ALC,ALH,ALA,ALK,CGQ,BRG,CGK,CHM
105	AQC
106	CAH,CBA
108	BSE,BTG
109	BHE,ARE,CHN,CHP

110	YBN,APV,CFV,YDC
111	AHQ
113	ABM,AHM,AQA,CGD,CGE
114	AHG
119	ACI
120	BFC,BQM,BLT,BPF,AII,AQQ,BQB,CIH
122	BQM,BLT,BPF,AAI,ARV,BSA,ADR,BQB,BIL,BKS,BLP,BLN
124	ADQ
125	AGV,BKE,BIP,BIQ
126	BDT,BNB,BCH,BNC,BEB,BND
127	AMA,BRB,BFN,BGR,CPC
128	BIR,BHL,BHK,BTV,BVA,CNK,CNP,CNH
129	AQH,ATS,BBP,BMG,ASP,CQG,CHQ
130	ABF,YAL,BBT
131	BKP,CRA,BIR,BHL,BHK,ACG,ACH,BTV,BVA,CNK,CNP,CNH,YBV
132	CES,CER,CET
134	BIE,BPA
136	AQH,ATS,BBP,BMG,ASP,CQG,CHQ
137	BNN,BKS,BLP,BLN
138	BKM,CRI,CHF
142	BEK,BEM
144	ARI
146	BQM,AIF,AQR,BPF,BTH,BTL,BIG
148	BQQ
152	BGB,ARR,BFV,BGA,BGT,BGV,BGN,CKV
153	AEH,AHC,AHB,YBQ
154	APS
155	CID,YCS
156	BKR,CEC
157	ALR
159	BQM,BPF
160	BRP,BSI,CGC
161	CBR,CFG
163	CGI,BHP,BHQ,BHR,BHS
164	AQM
166	BEK,BEM
167	AGB,ATV
168	BGD,BGF,BGG,BGC,BGQ
169	BVC

171	BSB,ADC,YCP
174	BGP,ASN,BGK,BGS
175	CPD
178	CHQ
180	BRP,BSI,AEB,ASH,BBG,BRQ,CGC
181	YAB,AAR,AEA,BRM,AEQ,YAG,AGC
184	CHN,CHP
185	CPF
189	BRL
190	CDS
191	AQS
192	BII,BIS,BKS,BLP,BLN
193	YAU,YBT
194	BME
195	AFF
197	ACQ,CCN,YCR
198	CII,CIF
199	ARH
201	CGI
202	BKN,BNQ
208	BDQ
209	AFK
210	ATR,ALV,CFP,BBF,AAP,ALN,ALP,BBD,APL,CFI,CFN,CHK
211	CGR
212	BTB
215	BQM,BQE,BLT,BPF,BLG,BLI,BER,CAI,CAK,CAL,BQB,BKS,BLP,BLN,CHF,BSR, BSS,BTP
217	ATA
218	CGV
219	AFV,ART
220	BDC,BDF,BDG
221	YCN
224	BLK,AHI
225	ATR,ALV,CFP,BBF,AAP,ALN,ALP,BBD,APL,CFI,CFN,CHK
226	BEI,BEN,BHA
227	CGP,CGG
228	BLE
233	BTT
234	ADB

236	BAB,AVV,BAA,BAD,BAC,BAE
238	BNI
239	AQH
241	ALQ
242	BTN
243	AQP,CAB,CAP,AES,AFI,CBB,CBE
246	BIR,BHL,BHK,BTV,BVA,CNK,CNP,CNH
247	AFR,AFQ
249	BCK
250	CCT
251	BCM,BCF,BCI
257	BFT,BGE
259	BEF
260	CMI
268	CED,BKS,BLP,BLN,CDV
269	AGI,BQB
271	CAK,CAL,BSR,BSS,BTP
272	BLG,BLI,BQH,CAI,CAK,CAL,BSR,BSS,BTP
274	ATS,BBP,BMG,AMK,AMI,AMG,AMQ,APK,APN,CIE,ATR,APM,API,ASL,BTF,BCR,BCS,BCT,ALV,CFP,BBF,AAP,ALN,ALP,BBD,APL,CFI,CFN,CHK,CHG,CHR,BML,BTE,BMM,CBF,CKK,CKS,CLA
276	AIA,AAE,AIB
278	CCV
281	AKG,AKE
282	BPM
286	BCP,BDS,BMR
287	CHE
290	AHD
291	CPC
295	ANS
297	BIK,CLH,BII
298	ARP
304	AAB
305	CDB
307	BGR
309	BBM
310	CHN,CHP
312	CGT,CGS,CHD
320	BLM

321	CEF
324	BTT,APH
325	CAK,CAL,BSR,BSS,BTP
327	AAT
328	GAL,APQ,APR,ATP,YBP,YCV
330	BRP,BSI,CGC
331	BTV,BVA,CNK,CNP,CNH,BDQ
332	AGR
333	ASV,CLG
334	AIR,AIT
342	ADF,AKB,AKC,AKH
345	CLN
346	ABI,YAT
348	CIC
349	APP
350	BDQ,AGP,YDH
351	CEK
353	BFH
359	CHB
363	AST
365	BKS,BLP,BLN
367	YCM,YDE
370	BPI,BPN
373	YAC
374	BTT,BVH,CNN,CNF,CNB,BVG,CNE,CNL,BVF,CNS,CNT
382	CFD,CFM
383	AIG,YDD,ALR,CHN,CHP
384	MAA,YCZ
389	ATS,BBP,BMG,AMK,AMI,AMG,AMQ,APK,APN,CIE,ATR,APM,API,ASL,BTF,BCR,BCS,BCT,ALV,CFP,BBF,AAP,ALN,ALP,BBD,APL,CFI,CFN,CHK,CHG,CHR,BML,BTE,BMM,CBF,CKK,CKS,CLA
390	CHQ
393	AEC
400	AAQ,ASB,AAH
402	BRE
403	CDL
405	CGL,BEH

Table 3, part 2

Allele ID	Amino acid mutations	Median fitness	Standard deviation	Allele type	Promoter mutations
1	T7I,A99V,P323S,N588K,K856E	0.00260588	0.00051268	Nonsynonymous	ins-125A,ins-125A
2	V20I,I716V	0.00668779	0.01073587	Nonsynonymous	G-733C,ins-642T,A-121T,A-95G,G-52A,T-40A,ins-32T,T-28A
3	NA	0.0040422	0.06971981	Synonymous	NA
6	NA	0.00246911	0.1259488	Synonymous	NA
7	V20I,T221I,I716V	-0.0300245	0.01760643	Nonsynonymous	G-733C,ins-642T,A-609G,C-511G,A-408G,C-371T,A-121T,A-95G,G-52A,T-40A,ins-32T,T-28A
9	Y651*,I716V	NA	NA	Nonsense	C-793A,C-736T,ins-642T,G-456A,A-72G
10	A454P,D483N,I716V	-0.2110058	0.01702836	Nonsynonymous	C-793A,C-736T,ins-642T,G-456A,A-72G
11	NA	0.00390098	0.00583692	Synonymous	NA
12	NA	0.00378297	0.09868416	Synonymous	C-512T,ins-494A,A-482G,T-391C,G-381C,A-243G,T-170C
13	A99V,N588K,K856E	0.00401867	0.01174142	Nonsynonymous	ins-125A,ins-125A
14	Q184*,V749A	-0.1754401	0.04050917	Nonsense	NA
15	N588K,V598L	0.00369886	3.96E-05	Nonsynonymous	NA
16	NA	0.00476619	0.00763915	Synonymous	NA
17	NA	0.00380343	0.00702995	Synonymous	G-709C
18	NA	0.00847534	0	Synonymous	NA
20	R671*,I716V	-0.2275183	0.02937796	Nonsense	C-793A,C-736T,ins-642T,A-451G,T-230C,G-189A
22	I716V	-0.0116793	0.0498073	Nonsynonymous	C-793A,C-736T,ins-642T,A-451G,T-230C,A-72G
24	NA	0.00224413	0.04511842	Synonymous	A-571G,A-218G,G-51A

26	E45K,A190V,W306*,N588K	-0.1745094	0.0240452	Nonsense	ins-642T,T-120A
27	NA	0.0050192	0.00742821	Synonymous	NA
28	NA	0.00445578	0.01085219	Synonymous	A-218G,G-51A
29	F612Y,I716V	0.00729921	0.21100804	Nonsynonymous	ins-642T,C-556T,del-209A,del-208A,del-207A,del-206A,del-205G,del-204A,del-203A,del-202T,A-121T,C-97T,G-52A,ins-32T,T-28A
30	Q184*	-0.1727313	0.04107916	Nonsense	NA
31	NA	0.00381926	0.05098052	Synonymous	G-701A
32	N588K,K856E	0.00445204	0.00921734	Nonsynonymous	NA
33	A99V,N588K	0.00329738	0.01667405	Nonsynonymous	ins-125A,ins-125A
37	NA	-0.0069505	0.0278462	Synonymous	C-352T,C-255T
38	N595S,I716V	-0.0034444	0.08202838	Nonsynonymous	C-793A,C-785T,C-736T,ins-642T,C-593A,A-121T,A-95G
41	N588K	0.00066352	0.00988962	Nonsynonymous	G-64A
42	NA	0.00320577	0.02640487	Synonymous	NA
43	R29K,V533I	0.00221477	0.00612072	Nonsynonymous	ins-642T
44	S374*	-0.2150144	0	Nonsense	NA
45	Q184*	-0.1709992	0.04440701	Nonsense	NA
47	NA	-0.002565	0.00972076	Synonymous	NA
49	NA	0.0119962	0.00496793	Synonymous	A-571G,A-218G,G-51A
51	NA	0.0033728	0.01953584	Synonymous	NA
52	NA	0.00585929	0.00435239	Synonymous	A-527G,A-85C
53	E14D,V346I,I716V	0.00070826	0.00695283	Nonsynonymous	T-390A,A-121T,G-52A,T-40C,A-32T,T-28A
56	E45K	-0.0077705	0.00784661	Nonsynonymous	ins-723T,C-654T,A-370G,A-331G,T-41A
57	A99V,N588K,V598L	-0.000469	0.13698275	Nonsynonymous	ins-125A,ins-125A
58	N588K,V598L	-0.0014028	0.01143826	Nonsynonymous	NA
59	H41R,I716V	-0.0113396	0.00544813	Nonsynonymous	C-793A,C-736T,ins-642T,A-451G,T-

					230C,G-189A,A-72G
60	NA	0.00223935	0.03800177	Synonymous	C-670A
61	NA	0.00136885	0.00636551	Synonymous	NA
62	I716V	0.00396585	0.0067267	Nonsynonymous	C-650T,ins-642T,A-539G,A-532G,T-469C,C-338T,A-121T,A-95G,G-52A,ins-32T,T-28A
63	Q184*	-0.1913311	0.04904495	Nonsense	NA
65	NA	0.00202632	0.00174099	Synonymous	NA
66	A99V	0.0089908	0.00196686	Nonsynonymous	NA
67	R671*,I716V	-0.2338855	0.01184721	Nonsense	C-793A,C-736T,ins-642T,A-451G,T-230C,G-189A
68	NA	NA	NA	Synonymous	C-793A,C-736T,ins-642T
69	S367R	0.01660449	0.00741045	Nonsynonymous	NA
70	NA	0.0025552	0.00672036	Synonymous	NA
71	A99V,G380R,N588K,K856E	0.00538443	0.11865665	Nonsynonymous	ins-125A,ins-125A
72	N588K	-4.32E-05	0	Nonsynonymous	NA
73	I716V	0.00155558	0.00996298	Nonsynonymous	G-733C,ins-637A,G-52A,G-19A
74	A99V	0.00500974	0.00786516	Nonsynonymous	NA
76	NA	0.00803079	0.00719111	Synonymous	NA
77	A99V,N588K	0.00082677	0.00285151	Nonsynonymous	T-641A,ins-125A,ins-125A
78	NA	0.00756172	0.01324578	Synonymous	ins-125A,ins-125A
79	T7I,A99V,K856E	0.00927301	0.00424664	Nonsynonymous	ins-125A,ins-125A
83	I716V,R761G	-0.0481548	0.005039	Nonsynonymous	C-793A,C-736T,ins-642T,G-456A,A-72G
84	R29K	0.00254674	0.01144027	Nonsynonymous	ins-642T
85	I716V	-0.0311353	0.00471679	Nonsynonymous	C-793A,C-736T,ins-642T,G-456A,A-72G
86	A99V	0.00564602	0.01746402	Nonsynonymous	ins-125A,ins-125A
88	V20I,I716V	0.00489155	0.00378931	Nonsynonymous	G-733C,ins-642T,A-121T,A-95G,G-52A,T-40A,ins-32T,T-28A

90	NA	-0.0056512	0.00985183	Synonymous	C-352T,C-255T
94	NA	0.01285654	0	Synonymous	ins-642T
95	E14Q,I716V	-0.0003572	0.17220489	Nonsynonymous	C-759T,ins-642T,A-72G
96	Q184*	-0.1094067	0	Nonsense	NA
97	V496I	0.00290339	0.00593124	Nonsynonymous	C-383T,T-91C
99	NA	0.0033844	0.00402804	Synonymous	A-527G
101	I312V,V325A,P591L	0.01157028	0.00215187	Nonsynonymous	T-426C,ins-228A,A-107G
102	NA	-0.0005955	0.00034368	Synonymous	NA
103	NA	0.00062527	0.02893702	Synonymous	T-230C
104	H41R,I716V	-0.0377311	0.00626955	Nonsynonymous	C-793A,C-736T,ins-642T,G-456A,T-230C,G-189A,A-72G
105	NA	0.01054989	0.00770442	Synonymous	T-133C
106	NA	NA	NA	Synonymous	NA
108	NA	0.00693599	0.00521281	Synonymous	T-74C
109	NA	0.00622269	0.012805	Synonymous	ins-642T
110	NA	0.00087836	0.00457137	Synonymous	NA
111	H41R,A127T,I716V	-0.069237	0.02700295	Nonsynonymous	C-793A,C-736T,C-683T,ins-642T,G-456A,T-230C,G-189A,A-72G
113	NA	0.00548194	0.00586441	Synonymous	NA
114	S739F	-0.0135567	0.0009969	Nonsynonymous	NA
119	H41R,K516N,I716V	-0.0336921	0.00244471	Nonsynonymous	C-793A,C-736T,ins-642T,G-456A,T-230C,G-189A,A-72G
120	NA	0.00413953	0.00653931	Synonymous	NA
122	N588K	0.00306534	0.01651155	Nonsynonymous	NA
124	I716V	-0.0328567	0.00665592	Nonsynonymous	C-793A,C-736T,ins-642T,G-456A,G-419A,A-72G
125	A99V,N588K	0.00560279	0.00569371	Nonsynonymous	ins-125A,ins-125A
126	NA	0.0100521	0.00193563	Synonymous	C-512T,ins-494A,A-482G,T-391C,G-381C,A-243G,T-170C
127	NA	0.00529189	0.00653596	Synonymous	NA
128	NA	-0.0088571	0.00337017	Synonymous	C-352T

129	NA	0.00909043	0.0065556	Synonymous	NA
130	N47K,W48L,K49E,V50G,N51*	-0.1816265	0	Nonsense	C-648T,ins-642T,T-230C,G-189A
131	NA	-0.0087293	0.05453093	Synonymous	C-352T,C-255T
132	S367R,N711S,Y830F	0.01569839	0.07900653	Nonsynonymous	C-736T,ins-642T,A-571G
134	H41R,M244I,P424S,I716V	-0.0442812	0	Nonsynonymous	C-793A,C-736T,ins-642T,A-451G,del-359A,T-230C,G-189A,A-72G
136	NA	0.01114801	0.00494171	Synonymous	NA
137	N588K	0.00577349	0.1679314	Nonsynonymous	G-64A
138	NA	0.00549552	0.0049324	Synonymous	G-64A
142	N42S,S367R,P591L,S827G,F846L	0.00424189	0	Nonsynonymous	A-807G,G-258A
144	S712N,I716V	0.00274794	0.00405283	Nonsynonymous	A-699G,ins-642T,C-329T,A-121T,T-103C,A-100T,G-52A,ins-32T,T-28A
146	I256L,N588K	-0.0022006	0.00682726	Nonsynonymous	NA
148	NA	0.01347668	0.00990649	Synonymous	C-821T
152	Y155*,G242C,K412E,S782C	-0.1650256	0.03684735	Nonsense	G-733C,ins-642T
153	NA	0.00185683	0.0104543	Synonymous	C-512T,ins-494A,A-482G,T-391C,G-381C,A-243G,T-170C
154	A99V	0.00841408	0.00849012	Nonsynonymous	NA
155	E14D,V346I,I716V	0.01011974	0.00235178	Nonsynonymous	A-777G,G-662C,ins-642T,T-390A,A-121T,G-52A,T-40C,A-32T,T-28A
156	NA	0.00641719	0.21477225	Synonymous	A-216G
157	R29K	0.00022905	0.00262626	Nonsynonymous	ins-642T
159	I256L,N588K	-0.0136778	0	Nonsynonymous	NA
160	T676I	0.00166653	0.00543886	Nonsynonymous	ins-642T,A-82G
161	Q184*,V749A	-0.1657337	0.00576903	Nonsense	NA
163	H41R,M244I,I716V	-0.0448887	0.0299149	Nonsynonymous	C-793A,C-736T,ins-642T,G-456A,A-72G
164	L125*	-0.1408701	0.05195171	Nonsense	A-619G,G-251T

166	L68R,S367R,I437L,P591L,S827G,F846L	0.01353867	0	Nonsynonymous	A-47G
167	Q184*,K187E,E427Q	-0.1753071	0.02795187	Nonsense	NA
168	G128E,Y155*	-0.1719376	0.03022883	Nonsense	G-811A,ins-723T,ins-642T
169	NA	-0.0125837	0.00466284	Synonymous	C-352T,C-255T
171	S586P	0.00337818	0.00304191	Nonsynonymous	NA
174	T7I,G128E,Y155*,G242C,D797N	-0.1856799	0.06318132	Nonsense	G-733C,ins-642T
175	I716V,V809I	0.00215588	0	Nonsynonymous	G-733C,ins-642T,T-230C,G-189A
178	NA	0.00187864	0	Synonymous	C-793A,C-736T
180	Q184*	-0.1988257	0.03800455	Nonsense	ins-642T,A-463G,A-82G
181	R29K,P694L,I716V	-0.0014903	0.00502028	Nonsynonymous	ins-642T
184	NA	-0.0345731	0.00679012	Synonymous	C-793A,C-736T,ins-642T,G-456A
185	NA	NA	NA	Synonymous	NA
189	NA	0.00806218	0.00148593	Synonymous	NA
190	H282Q,L283S,P284*	-0.1716636	0.06731263	Nonsense	NA
191	A454P,D483N,I716V	-0.2196272	0.04661055	Nonsynonymous	C-793A,C-736T,ins-642T,G-456A,A-72G
192	N588K,V598L	0.00089936	0.02535026	Nonsynonymous	NA
193	H41R,I716V	-0.0082844	0.00270806	Nonsynonymous	C-793A,C-736T,ins-642T,A-451G,T-230C,G-189A,A-72G
194	Y155*,G242C,K281I,A758S	-0.1924096	0	Nonsense	G-811A,G-733C,G-658A,A-619G,C-116T
195	Q184*,I716V,D815G	-0.1920711	0.02342695	Nonsense	NA
197	I716V	0.00638286	0.00866287	Nonsynonymous	del-723T,ins-642T,G-52A,ins-32T,T-28A
198	N47K,W48L,K49E,V50G,N51*	-0.1598178	0.02261423	Nonsense	C-648T,ins-642T,T-230C,G-189A
199	S827I	-0.0174063	0.00400494	Nonsynonymous	C-670G,ins-642T,G-589A,C-567A,C-

					319G,G-162A,C-153T,A-140G,A-121T,G-96A
201	I716V	-0.0119713	0.00341702	Nonsynonymous	C-793A,C-736T,ins-642T,A-451G,T-230C,A-72G
202	V343I,G803A	-0.0007192	0.00640144	Nonsynonymous	C-352T,C-255T
208	NA	0.00285336	0.00623146	Synonymous	NA
209	K516N,I716V	-0.0015078	0.00562345	Nonsynonymous	ins-642T,T-592C,A-543G,C-492G,A-167G,A-121T,A-95G
210	NA	0.00139633	0.00291057	Synonymous	NA
211	H41R,A454P,D483N,I716V	-0.2355735	0.01125774	Nonsynonymous	C-793A,C-736T,ins-642T,C-412T,T-230C,G-189A,A-72G
212	W120*,E270Q,P694L,I716V	-0.1537223	0.01362576	Nonsense	C-793A,C-736T,ins-642T,A-451G,T-230C,G-189A
215	N588K	-0.0020989	0.00557056	Nonsynonymous	NA
217	L40F	0.00631758	0.00549977	Nonsynonymous	NA
218	A99V,P296L,N588K	NA	NA	Nonsynonymous	ins-125A,ins-125A
219	N238H	0.00159339	5.03E-05	Nonsynonymous	A-527G,A-85C
220	N595S,I716V	0.00524321	0.00164693	Nonsynonymous	G-733C,ins-642T,A-121T,A-95G,G-52A,ins-32T,T-28A
221	H41R,I716V	-0.0272671	0.01592473	Nonsynonymous	C-793A,C-736T,ins-642T,C-412T,T-230C,G-189A,A-72G
224	T7I,A99V,N588K,K856E	0.00841317	0.0032877	Nonsynonymous	ins-125A,ins-125A
225	A99V,P296L,N588K	NA	NA	Nonsynonymous	ins-125A,ins-125A
226	S367R,P591L,R705C,Y830F,I851T	-0.006818	0.00471906	Nonsynonymous	T-426C,ins-228A,A-47G
227	V325I,I716V	-0.0073903	0.00474914	Nonsynonymous	C-793A,C-736T,ins-642T,A-451G,T-230C,G-189A
228	A99V,I437V,L731V	-0.0022476	0	Nonsynonymous	ins-125A,ins-125A

233	NA	0.00485861	0.00381621	Synonymous	C-512T,ins-494A,A-482G,T-391C,G-381C,A-243G,T-170C
234	R29K,I716V	0.00513469	0.00096979	Nonsynonymous	ins-642T
236	A123D,K516N,I716V	-0.0020995	0.00549764	Nonsynonymous	ins-642T,T-592C,A-543G,C-492G,A-121T,A-95G
238	N588K	-0.0013084	0.0139719	Nonsynonymous	G-709C
239	NA	-0.0120355	0.01266316	Synonymous	NA
241	A99V,L388V,N588K	-0.0084083	0.00281812	Nonsynonymous	ins-125A,ins-125A
242	NA	-0.0069621	0.20556877	Synonymous	T-262C
243	T669K	-0.2458508	0	Nonsynonymous	NA
246	NA	0.00968582	0	Synonymous	C-255T
247	I826M	0.00814294	0.00025581	Nonsynonymous	C-512T,G-483A,A-482G,G-381C,A-243G
249	A810V	-0.0010431	0	Nonsynonymous	C-512T,ins-494A,A-482G,T-391C,G-381C,A-243G,T-170C
250	I716V,A839V	0.00117139	0.00963949	Nonsynonymous	ins-642T,T-592C,A-543G,C-492G,A-121T,A-95G
251	E45K	0.00402381	0.0046702	Nonsynonymous	ins-723T,C-654T,A-370G,A-331G,T-41A
257	Y155*,G242C,R364C,K412E,S782C	-0.1940504	0	Nonsense	G-733C,ins-642T
259	H41R,I716V	-0.0077734	0.00687615	Nonsynonymous	C-793A,C-736T,ins-642T,A-451G,T-230C,G-189A,A-72G
260	A99V,P296L,N588K	NA	NA	Nonsynonymous	ins-125A,ins-125A
268	N588K,V598L,F738L,I775M	-0.0019513	0.00628316	Nonsynonymous	G-64A
269	N588K,I675T	0.00423699	0.00203207	Nonsynonymous	NA
271	A99V,P296L,N588K	NA	NA	Nonsynonymous	ins-125A,ins-125A
272	A99V,P296L,N588K	NA	NA	Nonsynonymous	ins-125A,ins-125A

274	NA	0.00445163	0.00311307	Synonymous	A-571G
276	I384T	-0.0180566	0.00345725	Nonsynonymous	NA
278	I716V	0.00302495	0.00698085	Nonsynonymous	ins-642T,T-592C,A-543G,C-492G,A-121T,A-95G
281	V20I,V527I,P608L,F701I,I716V	0.00574547	0.00099715	Nonsynonymous	G-733C,ins-637A,A-121T,A-95G,G-52A,A-39G,ins-32T,T-28A
282	NA	0.00165479	0.00247244	Synonymous	NA
286	A810V	-0.0012986	0.00155016	Nonsynonymous	C-512T,ins-494A,A-482G,T-391C,G-381C,A-243G
287	A99V,N588K	-0.0395358	0	Nonsynonymous	T-641A,T-450C,ins-125A,ins-125A
290	I256L	0.00179127	0.0047329	Nonsynonymous	NA
291	NA	-0.0008971	0.00335467	Synonymous	NA
295	I716V	-0.0260844	0	Nonsynonymous	C-793A,C-736T,ins-642T,G-456A,A-72G
297	A99V,P296L,N588K	-0.1260505	NA	Nonsynonymous	ins-125A,ins-125A
298	P694S,I716V	-0.0612323	0.0203894	Nonsynonymous	C-793A,C-736T,ins-642T,G-456A,A-72G
304	R29K,E63D	0.00416011	0.00341728	Nonsynonymous	ins-642T
305	I716V	0.00594383	0.0051731	Nonsynonymous	G-733C,ins-642T,A-121T,A-95G,G-52A,ins-32T,T-28A
307	S782C	-0.0182512	0.0039585	Nonsynonymous	NA
309	L223V,I225Y,L226T,R227A,L228F,G229R,F230L,L231F,V232S,E233G,L234T,I235Y,S236F,L237S,N238K,A239C,V240C,A241C,G242W,F243L,M244H,T245D,G246R,S247F,A248R,F249I,N250*	-0.1921588	0.0193739	Nonsense	A-742G,C-512T,G-381C,A-243G

310	NA	-0.0488871	0	Synonymous	C-793A,C-736T,ins-642T,G-456A,A-72G
312	S699L	-0.0776178	0.16039301	Nonsynonymous	NA
320	N588K	-0.0041501	0.02487121	Nonsynonymous	G-129T
321	N588K	-0.0080535	0.00964949	Nonsynonymous	ins-642T
324	NA	0.00703846	0.00357527	Synonymous	C-512T,ins-494A,A-482G,T-391C,G-381C,A-243G,T-170C
325	A99V,G380R,N588K,K856E	0.00355811	0	Nonsynonymous	NA
327	H41R	-0.0067704	0.00578463	Nonsynonymous	C-793A,C-736T,ins-642T,A-451G,C-362A,T-230C,G-189A,A-72G
328	V20I,V527I,P608L,F701I,I716V	-0.0020599	0.00209727	Nonsynonymous	A-790G,G-733C,ins-637A,A-121T,A-95G,G-52A,A-39G,ins-32T,T-28A
330	NA	-0.0081508	0.00190708	Synonymous	ins-642T,A-82G
331	NA	-0.00934	0	Synonymous	T-391C,G-381C,A-243G,T-170C
332	N47D,S712N,I716V	0.00755639	0	Nonsynonymous	C-670G,ins-642T,C-567A,C-319G,G-162A,C-153T,A-140G,A-121T,G-96A
333	NA	-0.0002132	0.02890556	Synonymous	NA
334	NA	-0.0036786	0.01189646	Synonymous	G-328A
342	V20I,V527I,P608L,F701I,I716V	-0.0089829	0.01299654	Nonsynonymous	G-733C,ins-637A,A-121T,A-95G,G-52A,A-39G,ins-32T,T-28A
345	D316E,I716V	-0.0225235	0.0139829	Nonsynonymous	C-793A,C-736T,ins-642T,A-451G,T-230C,G-189A,A-72G
346	H41R	-0.0118931	0	Nonsynonymous	C-793A,C-736T,ins-642T,A-451G,T-230C,G-189A,A-72G
348	V20I,T221I,P373S,N588K	NA	NA	Nonsynonymous	G-733C,ins-642T,C-511G,A-408G,C-

					371T,A-121T,A-95G,G-52A,T-40A,ins-32T,T-28A
349	Y69*,D70N,I716V	-0.1522268	0.00672011	Nonsense	G-733C,ins-642T,T-230C,G-189A
350	NA	0.00581325	0.00257939	Synonymous	NA
351	V20I,L505*	-0.1584013	0	Nonsense	C-793A,C-736T,ins-642T,C-508T,C-412T,A-121T,A-95G,G-52A,A-48G,T-40A,ins-32T,T-28A
353	Y155*,G242C,K412E,S782C	-0.107856	0.02329395	Nonsense	G-733C,ins-642T
359	NA	-0.00957	0	Synonymous	T-627A,G-51A
363	M260I,L553K,L554K,Q556K,A557T,F558G,P559F,A560P,G561C,K562W,F563*	-0.1986467	0	Nonsense	G-64A
365	N588K,V598L	-0.0009908	0	Nonsynonymous	G-64A
367	W353*,F558S,S827I	-0.2366081	0	Nonsense	C-670G,ins-642T,C-567A,C-319G,G-162A,C-153T,A-140G,A-121T,G-96A
370	E45K	-0.0042439	0.00459341	Nonsynonymous	T-120A,T-50G
373	V20I,V527I,P608L,F701I,I716V	0.00411649	0.00518258	Nonsynonymous	G-733C,ins-637A,A-121T,A-95G,G-52A,ins-32T,T-28A
374	NA	0.00429621	0	Synonymous	C-512T,ins-494A,A-482G,T-391C,G-381C,A-243G,T-170C
382	Q184*,G254S	-0.1297796	0	Nonsense	ins-642T,A-463G,A-82G
383	R29K	-0.000517	0.00981088	Nonsynonymous	ins-642T
384	A99V,N588K,K856E	0.00978734	9.27E-05	Nonsynonymous	ins-125A,ins-125A
389	NA	0.00743378	0	Synonymous	A-571G,G-51A
390	NA	-0.0057557	0	Synonymous	C-793A
393	Q184*,G254S	-0.1623546	0.00694162	Nonsense	A-82G
400	R29K	0.00206958	0.0074097	Nonsynonymous	ins-642T

402	N588K	0.01495159	0	Nonsynonymous	NA
403	I716V	0.00531234	0.00045975	Nonsynonymous	ins-642T,G-52A,ins-32T,T-28A
405	H41R,A99V,I716V	-0.0091594	0	Nonsynonymous	C-793A,C-736T,ins-642T,A-451G,T-230C,G-189A,A-72G

Supplementary Table 4. One-way ANOVA test of simulated libraries to determine if there is a significant difference when using PacRAT.

Each test was performed for simulations at all three coverages per gene. P-values for pairwise comparisons between methods were calculated using the Tukey test and reported here. Nearly all simulated libraries showed significant improvements ($p < 0.05$) with PacRAT except the library generated from a 759bp insert sequence across all three coverages (highlighted orange) and the 936 library at 10X coverage (highlighted purple). This indicates that PacRAT does not improve barcode-variant mapping for smaller inserts, but is more crucial as insert sizes (and error rates) increase.

Size of variable region (bp)	SimLoRD Parameter	ANOVA (df = 1)		Comparison between ABP and PacRAT (p-values)		
		F statistic	p-value	10X	5X	3X
759	adjusted	2666.134	7.7509E-18	0.99981365	0.69736541	0.78105659
936	adjusted	21798.43	2.6097E-23	0.62449709	2.0916E-05	2.3756E-05
1128	adjusted	10432.88	2.1694E-21	0.02317619	9.2129E-08	6.1294E-08
1710	adjusted	17714.41	9.0595E-23	4.563E-14	2.7311E-14	2.7089E-14
2166	adjusted	72159.58	1.9844E-26	2.7089E-14	2.7089E-14	2.7089E-14
2538	adjusted	145442.3	2.96E-28	2.7089E-14	2.7089E-14	2.7089E-14
2961	adjusted	156165.4	1.9317E-28	2.7089E-14	2.7089E-14	2.7089E-14
4086	adjusted	2234635	2.2503E-35	2.7089E-14	2.7089E-14	2.7089E-14
759	default	7414.062	1.68E-20	5.5423E-08	2.7255E-11	1.6157E-09
936	default	9752.343	3.25E-21	7.3386E-13	2.0513E-12	4.9183E-14
1128	default	32614.59	2.33E-24	2.7089E-14	2.7089E-14	2.7089E-14
1710	default	43628.94	4.06E-25	2.7089E-14	2.7089E-14	2.7089E-14
2166	default	473927.7	2.47E-31	2.7089E-14	2.7089E-14	2.7089E-14
2538	default	85208.39	7.32E-27	2.7089E-14	5.0848E-14	2.7089E-14
2961	default	34793.84	1.58E-24	2.7089E-14	1.7345E-10	2.7089E-14
4086	default	10728.29	1.83E-21	2.7089E-14	0.00568811	3.5973E-10

Supplementary Table 5. Primers used for amplifying and cloning *MALx3* alleles.

Primer ID	Primer Name	Primer sequence	Notes
P170	MAL33_F	GAATGGGACTGATTGGAACAGCGC	Amplifying *MAL33* only
P171	MAL33_R	AACTTCTGTGGTAGATGAAGGGCG	Amplifying *MAL33* only
P172	MAL33_F_Gibson	gcggccgctctagaactagtggatctcgGAATGGGACTGATTGGAACAGCGC	Amplifying *MAL33* with homology to plasmid YMD2307 NruI cut site. Used for mini library.
P173	MAL33_R_Gibson	gaattcctgcagcccgggggatctcgAACTTCTGTGGTAGATGAAGGGCG	Amplifying *MAL33* with homology to plasmid YMD2307 NruI cut site. Used for mini library.
P183	MAL13_2_F	ACGCAGTTAAGAGCTATTGTCCGG	Amplifying *MAL13* only
P184	MAL13_2_R	CCTCTGCCCTTCAACTAGTGCTCC	Amplifying *MAL13* only
P188	P188_MAL13_F_Gibson	gcggccgctctagaactagtggatctcgACGCAGTTAAGAGCTATTGTCCGG	Amplifying *MAL13* with homology to plasmid YMD2307 NruI cut site. Used for mini library.
P189	P189_MAL13_R_Gibson	gaattcctgcagcccgggggatctcgCCTCTGCCCTTCAACTAGTGCTCC	Amplifying *MAL13* with homology to plasmid YMD2307 NruI cut site. Used for mini library.
P190	P190_YPR196W_F	TTCTTTCCCAACTCTTTCGCCAGC	Amplifying *MAL73* only
P191	P190_YPR196W_R	TCGTAGTGTACATGGTCCTCTTCC	Amplifying *MAL73* only
P192	P192_YPR196W_F_Gibson	gcggccgctctagaactagtggatctcgTTCTTTCCCAACTCTTTCGCCAGC	Amplifying *MAL73* with homology to plasmid YMD2307 NruI cut site. Used for mini library.
P193	P193_YPR196W_R_Gibson	gaattcctgcagcccgggggatctcgTCGTAGTGTACATGGTCCTCTTCC	Amplifying *MAL73* with homology to plasmid YMD2307 NruI cut site. Used for mini library.
P194	MAL33_BC_Gibson_R	gaattcctgcagcccgggggatctcgNNNNNNNNNAACTTCTGTGGTAGATGAAGGGCG	Barcoded primer used with P172 to amplify and barcode unique *MAL33* alleles. Similar ones were designed for *MAL13* (P196) and *MAL73* (P195), but not used here.

P199	MAL33_barseq_F	aatgatacggcgaccaccgagatctacacAAACTTCAAAGCCGCAACAAAACC	Amplifying for barcode sequencing
P200	MAL33_barseq_Rev1	caagcagaagacggcatacgagatGTGCCTACtggccgattcattaatgcagctgg	Amplifying for barcode sequencing
P197	MAL33_barseq_Read1	CACAACAGATGAGGTGTTTCGCCCTTCATCTACCACAGAAGTT	Sequencing primer, Read 1
P198	MAL33_barseq_Read2	atcgaattcctgcagcccgggggatctcg	Sequencing primer, Read 2
P202	MAL33_barseq_Index	ggaaacctgtcgtgccagctgcattaatgaatcggcca	Index primer

Supplementary Table 6. Percent identity and similarity between *MALx3* sequences.

A) Percent identity between nucleotide sequences of *MALx3* genes.

	MAL13	MAL23	MAL33	MAL43	MAL63	MAL73
MAL13						
MAL23	70.24%					
MAL33	70.75%	77.14%				
MAL43	69.89%	94.27%	75.44%			
MAL63	71.78%	94.76%	74.17%	92.64%		
MAL73	67.44%	82.31%	74.45%	84.85%	78.20%	

B) Percent identity (top) and similarity (bottom) between amino acid sequences of MALx3 proteins.

	MAL13	MAL23	MAL33	MAL43	MAL63	MAL73
MAL13						
MAL23	69.68% 78.53%					
MAL33	68.08% 77.80%	74.26% 80.00%				
MAL43	69.05% 77.68%	94.89% 97.02%	72.13% 79.15%			
MAL63	70.53% 78.95%	94.47% 96.17%	71.06% 78.51%	92.98% 95.11%		
MAL73	64.63% 74.74%	85.32% 90.43%	70.64% 78.09%	87.45% 91.70%	81.91% 87.87%	

VITA

Chiann-Ling Cindy Yeh was born in Las Cruces, NM and lived between Taiwan and Las Cruces as a child. She attended New Mexico State University from 2012-2016 in Las Cruces and received a Bachelor of Sciences degree in Genetics and Biotechnology. As an undergraduate researcher, she worked in the laboratory of Dr. Graciela Unguez studying how the weakly electric fish, *Sternopygus macrurus*, dissociated its distal tail muscle fibers and transdifferentiated those muscle proteins into components of the electric organ. In the summer of 2015, Chiann-Ling worked with Drs. Jerald Radich and Amy Paguirigan in developing a multiplex PCR assay to detect mutations in cells that may result in minimal residual disease (MRD) following remission from acute myeloid leukemia. After her undergraduate tenure, Chiann-Ling joined the Genome Sciences PhD program at the University of Washington and worked with Dr. Maitreya Dunham to develop a high-throughput functional assay for measuring the effects of natural variants on a population level in yeast. Chiann-Ling served as president and on various board positions with the Women in Genome Sciences, hoping to bring a more intersectional approach to providing an inclusive community in Genome Sciences. During her tenure, she created a high school coding/genetics camp for young high schoolers in the Seattle area called Genome Hackers, mentoring over 60 girls during this time. In 2021, Chiann-Ling was recognized as a Husky 100 for her impact on her community at the University of Washington.