©Copyright 2019
Anne Elizabeth Clark

Causes and consequences of genetic variation in budding yeast

Anne Elizabeth Clark

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Joshua Akey, Chair

Aimée Dudley

Maitreya Dunham

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Causes and consequences of genetic variation in budding yeast

Anne Elizabeth Clark

Chair of the Supervisory Committee:
Professor Joshua Akey
Genome Sciences

The complex evolutionary forces that shape genomes result in sequences that are part utility and part history. The randomness involved in this process makes it difficult to imagine ever being able to read a genome and completely understand the effect of every base. But with large–scale mapping approaches, we can relate specific genetic variants present in natural populations to specific differences in traits of interest. The budding yeast *Saccharomyces cerevisiae* is one of the most useful model organisms in the study of genetics, and also has a long history of association with human activity. In this dissertation, I explore several methodological aspects related both to mapping traits in *S. cerevisiae* and to understanding how specific regions of budding yeast genomes have evolved over time. I begin by discussing the two main approaches to positionally mapping trait variation: association and linkage. Both of these approaches are aided by the recent sequencing of hundreds of *S. cerevisiae* isolates, along with the less extensive but complementary sequencing of other members of the *Saccharomyces* genus. Genome–wide association is a simple way of taking advantage of the extensive diversity of available sequences, but suffers from inflated false positive rates due to population structure. Linkage mapping has traditionally been more widely applied in yeast, since it is relatively simple to generate very large pools of individuals that are ideal for mapping. This approach remains powerful, but the great number of sequences now available makes it appealing to create larger crosses that incorporate more of this diversity.

However, simply mapping trait variation to specific loci does not directly inform us about the mechanisms by which genetic variants affect phenotypes, or about how this variation has arrived at its current state. I also discuss some specific evolutionary aspects of budding yeast genomes. In particular, the ability of these substantially diverged yeasts to hybridize creates interesting possibilities for adaptation. I describe the development and application of an approach for identifying introgressed regions that remain from past hybridization events, and that therefore may have played significant evolutionary roles. Overall, the extensive genetic and phenotypic variation uncovered in budding yeast opens up exciting opportunities to discover the genetic basis of complex traits, and to investigate the larger picture of how evolutionary forces have shaped and continue to shape these experimentally and economically important organisms.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor, Josh Akey, for his mentorship over the course of my time in graduate school, as well as the good company of all the members of the lab that he convinced to work with him (and, unfortunately, to move across the country with him). The Department of Genome Sciences has been an incredibly supportive place to grow as a scientist. In losing my home in one lab to the East Coast, I gain three new homes in its place. I'd like to thank the Dunham lab, the Harris lab, and particularly the Dudley lab for adopting me in various practical and intellectual ways. I'm also grateful to have been among my entire cohort in the department, which has been full of generally delightful people who have not only inspiring scientists, but also supportive triathlon companions, fun potluck buddies, and overall stellar friends. I'll miss them all as they disperse to new scientific opportunities. I would also like to thank my many other friends and family members who have contributed to keeping my sanity mostly intact over the past several years: Maria for reading recommendations and other healthy distractions, Kacyn for many adventures and laughs, Mark for an unwavering belief in my abilities, Rachel for a very long friendship, Lily for typing assistance, Rob for advice both useful and entertaining, Aaron for helping me dream, and my parents for everything.

Chapter 1: INTRODUCTION

1.1 What does it mean to understand a genome?

Although we sometimes talk about *reading* genomes, the term *sequencing* captures the limitations of the approach more accurately. Like a list of the 0s and 1s of a computer program's machine code, a genetic sequence is meaningless unless we know the rules for interpreting its components. If we know how a specific sequence of logical bits translates into corresponding programmatic commands, we can gain a complete understanding of the result of running a piece of code. But what does an equivalent kind of understanding look like for a genome? While computer programs are designed, genomes are evolved. There is chance involved, and the results are not necessarily optimal. Evolution has thus been aptly compared to the process of tinkering rather than engineering.[1,2] The way in which genomes develop over time makes them much more difficult to decode than blocks of machine code since they are the product of not only logical rules but also the randomness embedded in their history.[3] What would it mean to solve a genome—to fully comprehend the effects of every base it comprises?

Simulating every molecule in every cell of an organism, as well as the interactions between those molecules, could theoretically allow us to interpret any genetic variant. Although whole-cell simulations at the molecular level have been carried out and have successfully predicted novel effects of genetic perturbations in single-celled organisms,[4,5] these models require extensive prior knowledge of cellular networks and do not function at the resolution of individual nucleotide variants. There are limits to the complexity of interactions we can simulate, and we will probably never have a model detailed enough to tell us about everything happening inside of one cell or the trillions of cells in a single human.

But what we can more realistically strive for—and this goal lies at the heart of the study of genetics—is characterizing all of the observed genetic variants that underlie traits of interest.[6] Even if we do not have a complete model elegantly and completely describing how

sequence maps to function, we can create an extensive database of knowledge of all the genetic variants that are major contributors to various functions. Mapping phenotypic variance to genetic variance is incredibly important in medicine, agriculture, industrial applications, and beyond. With an understanding of how specific variants lead to trait differences, we can more accurately predict disease risk, breed better crops, and design more effective microbial tools.

1.2 Quantity and quality of genomic sequences

In the nearly thirty years since the initiation of the Human Genome Project, we have made astounding strides in technologies for sequencing genomes and cataloguing variants. We have gone from having a single bacterial genome sequence in 1995,[7] to a single eukaryotic genome in 1996,[8] a single invertebrate[9] and plant[10] genome a few years later, and a single complete human genome in 2001.[11,12] Today, so many distinct species and individuals within those species have been sequenced that there is no comprehensive list. We have not sequenced only *the* human genome, but hundreds of thousands of human genomes;[13] not just one individual yeast, but many yeast species and thousands of strains.[14,15] This great diversity of genomic sequences is crucial for inferring the evolutionary relationships between organisms, as well as for characterizing the effects of individual genetic variants.

But of equal importance to the diversity of available genomes is the quality of those genomes. Resequencing approaches that map short reads to a preexisting reference genome have trouble resolving larger structural variations, indels, copy number variants, and variants in regions that are AT– or GC–rich or highly diverged between individuals.[16] In addition, in the case of clinically important variation, the sequencing accuracy is paramount.[17] Our ability to sequence genomes to higher coverage, and to do so more quickly and cheaply, goes some way towards generating more complete and accurate genomes, but does not in itself guarantee high–quality genomes that will allow for the analysis of all types of variation of interest. *De novo* genome assembly is in general a much more challenging task, but algorithmic and technological improvements in recent years have continued to make the

approach more feasible.[16] Our ability to map phenotypic variation to specific genotypic variation will continue to advance with improvements in the quality of the genomic sequences these analyses rely upon.

1.3 How to decipher a genome

Even before the completion of the first human genome sequence, it was already realized that simply having this sequence would not be enough to tell us what it meant; new statistical and technological approaches would need to be developed to decipher its meaning.[18,19] While simply sequencing more genomes is not a panacea for all the difficulties of interpretation, aggregating many varying sequences in conjunction with varying phenotypes can reveal a great deal about the effects of genetic variation.

The difficulty of mapping observed traits to the ever–increasing pool of known genetic variants varies greatly depending on the specific phenotype of interest. In particular, monogenic traits that follow a Mendelian inheritance pattern are simpler to map because different affected individuals have causal variants that are identical or, at the least, located very close to each other in the genome. On the other hand, complex traits depend on variation at multiple locations in the genome that may differ between individuals. These traits are more difficult to map because the effect of any one genetic variant may be small and may depend on variants at other locations. Some disease phenotypes were early mapping successes following developments in sequencing technology—particularly common diseases with incidence mainly depending on only one or a few possible variants. In the first decade following the completion of sequencing the human genome, 2,850 genes involved in a variety of Mendelian diseases were identified, and more than 1,100 variants influencing more complex disorder were mapped.[20] A few years later, genes underlying approximately 50% of all known monogenic disorders had been discovered.[21] But more complex diseases and other traits have proved far more difficult to dissect, with vast numbers of variants contributing vanishingly small effects, or variants interacting in nonlinear ways. In addition, for traits that are rare, or that

depend on underlying variants that are rare, it is challenging to obtain sufficient sample sizes for mapping.[22]

The main approaches for connecting phenotypic variation to naturally occurring genetic variation are candidate gene screening, linkage mapping, and association studies.[22,23] In the simplest cases, we can employ candidate gene screening, in which we make educated guesses about genes that might be involved in a trait based on genetic functions that have already been characterized, sometimes in other organisms. But given the complex, polygenic nature of many traits—in addition to the pleiotropic nature of many genes—this approach often fails. For example, as researchers first began to investigate the genetic factors underlying psychiatric disorders, the only apparent candidates were the neurotransmitters that existing drugs targeted, but no association was found between variation in these genes and those disorders.[24] In cases like these where the current literature does not point to useful gene candidates, positional approaches to pinpointing relevant genetic variation have been pursued instead.

Linkage analysis, which utilizes variation within families, has historically been a powerful approach for understanding the genetic basic of Mendelian disorders and sometimes more complex conditions; its utility, however, is mostly confined to highly penetrant genetic variants, for which most individuals with the causal allele express the trait of interest.[25] Genome–wide association (GWA), on the other hand, utilizes the variation in large groups of unrelated individuals and gained popularity for its potential to provide insight into the causes of common diseases with more complex genetic architectures. I discuss the strengths and weaknesses of both of these mapping approaches in an experimentally and economically important organism—the budding yeast *Saccharomyces cerevisiae*—in more detail later in this introduction and in subsequent chapters.

1.4 Yeast: an old friend and a modern model

> *"Ever since man became sapient he has devised means of intoxicating himself, principally in order to create, albeit temporarily, a more pleasurable milieu. In all but a few cultures, the most common means of intoxication has resulted from the metabolic by-products of the anaerobic metabolism of certain species of yeast, a process that has historically been elicited in a variety of ways."* [26]

Yeast, and particularly the budding yeast *Saccharomyces cerevisiae*, has been an important part of human civilization for several millennia. Evidence of fermented beverages comes from China in 7000 BCE,[27] Iran in 6000 BCE,[28] and Egypt in 3000 BCE.[29] It is hypothesized that yeast–leavened bread originated in Egypt around 1500 BCE, either through the substitution of beer for water when making bread, or through pieces of dough being left to sit long enough that they collected wild yeast.[30] Yeast is also used in the fermentation of a variety of other food products, including cacao, in a practice that likely began in 1900 BCE or earlier.[31] The close association of yeast with human food and beverage production has shaped its evolution and current diversity in profound ways.[32–35]

The tradition of fermenting beverages, bread, and chocolate with the yeast *S. cerevisiae* and its close relatives has continued to the present day, but humanity has in recent history also found new uses for this organism. *S. cerevisiae* has been used for bioremediation in filtering heavy metal pollutants from industrial wastewater,[36] for generating biofuels,[37] and in the production of biopharmaceuticals, including insulin and vaccines for hepatitis and human papillomavirus.[38] In addition, *S. cerevisiae* has become one of the most important model organisms in genetics, and together with its relatives is one of the most useful systems in comparative evolutionary genomics.[39] *S. cerevisiae* was the first eukaryote to have its complete genome sequenced,[8] and currently has an extensive set of experimental tools available for its study: an extensively annotated reference sequence[40] housed in a well-curated genome database,[41] a large experimental toolkit,[42] and a wealth of other resources including the gene deletion collection[43,44] and detailed maps of genetic networks.[45]

Budding yeast has gained prominence as a model organism to due several attractive characteristics of its biology. The *S. cerevisiae* genome is approximately 12 million bp, about one two-hundredth the size of the human genome; in addition, it is much more compact, with far fewer introns and far less intergenic sequence. Budding yeast can reproduce both sexually and asexually depending on environmental conditions. This flexible lifestyle allows for the generation of large pools of (almost) genetically identical individuals, but also allows for the combination of different genotypes through mating and meiotic segregation. The generation time of yeast is also short, with cells dividing as quickly as every 90 minutes under ideal experimental conditions.[42]

Budding yeast are also, perhaps surprisingly, a good model for human cells in many ways. For example, it is estimated that nearly half of essential genes in *S. cerevisiae* that have a human ortholog can be functionally replaced with that ortholog; furthermore, these genes that can be successfully humanized also tend to cluster together in function, suggesting the possibility of humanizing entire metabolic pathways.[46] This complementation of yeast genes by human orthologs—or even paralogs—allows for testing the pathogenicity of variants in human disease genes in a much more tractable experimental system.[47]

1.5 Yeast: not always friendly

Despite providing an array of culinary and experimental benefits to humanity, *S. cerevisiae* is also an opportunistic pathogen. Although the incidence of *S. cerevisiae* infections is rare, it is thought to be increasing.[48,49] Some of its relatives, including *Candida albicans* and *Candida glabrata*, are currently much more clinically significant pathogens.[48] It is difficult to evaluate the mortality rates due fungal infections because they usually occur in the immunocompromised who have other medical conditions. But mortality for *C. albicans* infections is estimated at 30-40% and is even higher for some other types of fungal infections.[50,51] In addition, the number of fungal infections is growing, as is resistance to current antifungal treatments.[52] Yet there are currently only three structural classes of antifungal drugs, two

of which have been used since 1980.[50]

Drugs in the azole and polyene classes both target the sterol ergosterol, a cell membrane component unique to fungi. Amphotericin B is the only polyene used to treat systemic infections; it acts by binding ergosterol, but can also bind human cholesterol, and is thus highly toxic and used infrequently.[50] Azole antifungals inhibit ERG11, an important enzyme in the synthesis of ergosterol; although they are generally well tolerated, they can interfere with other drugs through cytochrome P450 inhibition. Drugs in the third class, echinocandins, inhibit the synthesis of an essential component of the fungal cell wall.[50] Resistance has been commonly acquired to both azoles and echinocandins. Developing antifungal drugs is a challenge because of the large degree of homology between humans and fungi. One approach to developing new antifungals is identifying molecules that synergize with existing antifungals.[50,53] And, in general, increasing our understanding of the genetic basis of resistance and susceptibility of pathogenic yeasts to existing antifungal drugs will help us explore new avenues of treatment.

1.6 The state of deciphering budding yeast genomes

Historically, both forward and reverse genetic approaches in yeast relied on random mutagenesis as a source of genetic variation.[54] In reverse genetics, we start with a particular genetic variant and attempt to determine any effects it has on phenotype, while in forward genetics, we start with trait variation of interest and attempt to locate the genetic variants that contribute to it.[54] In both cases, if we have just one or a few individuals of a species, we can mutate specific genes or randomly mutagenize the entire genome to create a pool of variation to study. But with the sequencing of a large number of genetically diverse *S. cerevisiae* strains, focus has shifted to mapping phenotypes resulting from naturally occurring variation, which is a useful way of limiting sequence space. As discussed earlier, the two main approaches to mapping phenotypic variance to specific locations in the genome are linkage analysis and genome–wide association, but the way these approaches are applied in

yeast is somewhat different than in humans.

Linkage analysis has typically been the more common approach for dissecting quantitative traits in yeast. In humans, this approach is limited by our ability to obtain genetic sequence data from a large number of families in which a trait is segregating, but in yeast it is possible to generate much larger, informative families by crossing two or more divergent parental strains—and to do so in a reasonable amount of time because the generation time in yeast is typically only a few hours. Linkage analysis is therefore primarily limited only by the genotypic and phenotypic divergence between the parental strains and the amount of recombination that occurs in the genetic cross cross. It requires more experimental work to generate a pool of segregants than to use a set of preexisting natural strains, but this approach has the advantage of being able to assign causation to variants rather than just correlation, due to the random nature of the segregant genomes. Isolating a variant from the genetic background in which it evolved makes statistical associations simpler, but it is also important to consider how it may not reflect the evolved genetic architecture of traits as accurately.

Early linkage analyses in yeast mapped the complex traits of high tolerance growth[55] and mRNA expression levels[56] to underlying genetic loci, and the success of these studies led to similar work across many other organisms.[57] In recent years, yeast crosses have been generated at a large enough scale to allow for the study of broader questions about the genetic architecture of traits and the ways in which many loci can collectively contribute to a quantitative trait.[58] Two–parent crosses have in general been the most common experimental design, but even with very large numbers of progeny, these crosses fail to capture much of existing natural genetic variation in the species and also limit the types of epistatic interactions that can be found.[59] Implementing new study designs that utilize linkage to map traits with greater precision in diverse yeast strains remains an important area of exploration.

Although genome–wide association was first pioneered in humans, the recent sequencing of hundreds of *S. cerevisiae* isolates from diverse geographical and ecological locations[14,60–62] has made the approach appealing in this species. Some characteristics of budding yeast can

make genome–wide association studies more powerful than in humans—including a high level of genetic diversity and a the fast breakdown of linkage disequilibrium[63]—but the potential for population structure to cause unacceptably elevated false positive rates is a known concern.[64] Statistical methods for reducing false positive rates exist, but the extent to which these methods succeed while maintaining power is of interest and is discussed in detail in the following chapter.

1.7 A brief tour of the *Saccharomyces* yeasts

Some of the characteristics that make budding yeast a useful model for molecular genetics also make it a somewhat challenging organism in which to study population genetics. The fact that yeast are single–celled means there they generally leave no fossil record, and also that it takes concerted sampling effort to determine where they live in the present day. In addition, their ability to reproduce both sexually and asexually, and to exist in both haploid and diploid states, makes the application of population genetic models less straightforward.

On the other hand, budding yeast have interesting patterns of diversity and evolutionary history. The *Saccharomyces* genus consists of eight known species, together comprising a large amount of both genetic and phenotypic diversity. These species, and strains within these species, exist in a wide variety of both human–associated and wild environments—from hospitals to vineyards, and primeval forests to suburban oak trees. Today, more than 2,000 yeast genomes have been sequenced, providing incredible opportunities for characterizing the variation among them. Pairs of these species are 15-30% diverged from each other at the nucleotide level. Despite this large divergence—and despite species often being defined by reproductive isolation—various pairs of distinct *Saccharomyces* species can mate and occasionally produce viable offspring in laboratory, industrial, and wild environments.[65]

1.8 Hybridization and introgression

Hybrid yeasts are of particular interest in part because they are frequently useful in commercial applications. In general, hybrids can have more extreme phenotypes than either of their parents, a phenomenon known as heterosis or hybrid vigor.[66] For example, the hybrid brewing yeast *S. pastorianus* (*S. cerevisiae* × *S. eubayanus*) is useful in the fermentation of lager ales because of its superior cold tolerance.[67] In addition to the existence of such hybrid diploid yeasts that can propagate clonally indefinitely, it is possible for hybrids to occasionally produce viable haploid spores. Prezygotic barriers in yeast are generally weak, implying that the spore viability of these hybrids must generally be low to maintain distinct species—and indeed, this viability is typically less than 1%.[68] However, the presence of introgressions—relatively short genomic segments originating from another species—suggests that these species barriers sometimes break down.

The mechanisms by which introgression occurs in budding yeast are not well understood, but a few possibilities have been proposed. In other eukaryotes, including plants and humans, introgression typically occurs through hybridization followed by repeated backcrossing with one of the original parental species, resulting in individuals that have genomes mainly matching one species but with interspersed introgressed segments from the other species.[65,68] This scenario is possible in yeast, but somewhat less plausible due to the large divergence between species, the low spore viability of hybrids, and the theorized infrequency of sexual reproduction.[68] There are, however, ways in which this process may be accelerated. For example, if aneuploidy occurs during the formation of ascospores, it could result in an individual with most chromosomes originating from one parent, and only a few from another. Alternatively, a single chromosome or portion of a chromosome could be directly transferred from the nucleus of one individual to another of a different species in incomplete karyogamy.[68] There is also the possibility that observed "introgressions" actually result from horizontal gene transfer through asexual means—a phenomenon that is common in bacteria but not well understood in *Saccharomyces*.[69]

Although many questions still remain about their origins, the introgressed segments that remain in modern yeast genomes are interesting for a few reasons. First, they may suggest interesting adaptive functions; that is, if specific introgressions remain from long-ago hybridization events, they may have persisted because they allowed yeast individuals adapt more effectively to new environments. Of course, these introgressions may also remain purely through random drift—and certainly some of them do—but it is possible that some of them have had more interesting evolutionary functions. In addition, introgressions can provide insight into how different species have interacted during their evolutionary history. Looking at overall patterns of introgression can inform us about how important of a role hybridization or other contact has played in the evolution of these the *Saccharomyces* yeasts, and perhaps how this role has varied among different species or populations.[65]

1.9 The state of identifying introgression in yeast

There are a variety of approaches for identifying introgression. One category of methods involves analyzing patterns of linkage disequilibrium (LD). These methods do not require a sequence for the species from which the introgression originated, but instead rely on the observation that admixture will leave signatures of LD for many generations in the population that was the recipient of introgression. For example, the S* statistic identifies long stretches of polymorphisms that are shared among a subset of individuals in a population that has experienced admixture, and that are absent from a non–admixed reference population.[70] This statistic has been used extensively to detect archaic hominin introgression in modern human genomes, and has been combined with archaic sequence data to detect introgressions from specific species.[71]

In yeast, it has been more common to rely on sequence identity from the beginning of the analysis—not only because making accurate demographic model assumptions in yeast is more difficult, but also because there are generally still surviving descendants of both species involved in the genetic exchange. In addition, incomplete lineage sorting (ILS) is not

as significant of a concern as a source of false positives in analyzing yeast genomes, since the divergence times between yeast species are so much larger; the potential role of ILS as a confounding factor in identifying introgression in yeast genomes is discussed in more detail in Chapter 4.

Researchers using sequence identity to infer introgression in yeast have typically focused on entire genes or ORFs, determining whether they more closely match those found in individuals of the same species or those found in a different species, according to preset identity thresholds.[14,62] This approach is often partially successful, but is unable to identify smaller pieces of genes that are introgressed, intergenic introgressions, and larger introgressed regions that span many genes. It also provides no information about the breakpoints or length distributions of introgressed regions. Methods that instead scan across the genome site–by–site or in windows—without regard to gene boundaries—can potentially address these shortcomings. Hidden Markov models (HMMs), which infer an unobserved state underlying each observation in a sequence, are one possible way to consider the possibility of introgression at every site in a statistically principled way. PhyloNet–HMM is a computational tool designed for detecting introgression in eukaryotes that simultaneously infers the gene tree at each site[72] and another, simpler HMM–based approach is discussed in detail in Chapters 4 and 5.

1.10 Not just what is, but what could be

Returning to our original discussion of what it means to fully understand a genome, there are a few new points to add. While having a complete catalogue describing the way that existing genetic variants map to phenotypes of interest would be a powerful resource that would help us figure out how to treat many types of disease and solve a variety of other problems, it would still not provide a complete picture. Such associations between genotype and phenotype tell us about the current state of genomes, but do not tell us how genomes came to be the way they are or, conversely, how they might change in the future. Understanding the

evolutionary forces that have shaped and continue to shape genomes is crucial both in helping us understand the variation we uncover, and in predicting outcomes such as how pathogens are likely to develop resistance to current treatments, how microbial communities might respond to environmental perturbations, or how a whole ecosystem might adapt to a changing climate. Developing tools and techniques for understanding the past, present, and future of genomes is crucial to answering many practical questions in genetics.

1.11 Outline of this dissertation

In the chapters that follow, I describe my work in several areas related to characterizing genetic variation in the *Saccharomyces* yeasts. In Chapter 2, I summarize the extent to which population structure contributes to false positives in GWA studies in yeast. I evaluate the performance of existing methods of correcting for this phenomenon, as well as some novel extensions to those methods. In Chapter 3, I outline several aspects of the design of a large and diverse yeast cross for mapping quantitative traits with unprecedented precision. This cross has recently been completed by members of the Dudley lab, and work is now underway to map a variety of phenotypes in the final pool of segregants. In Chapter 4, I discuss the development and application of a hidden Markov model–based method for identifying introgression in yeast genomes. In Chapter 5, I extend this method to incorporate more of the diversity within *Saccharomyces*. Finally, I conclude with some thoughts on the future of characterizing genetic variation in budding yeast and beyond.

Chapter 2: A COMPARISON OF METHODS FOR CORRECTING FOR POPULATION STRUCTURE IN GENOME–WIDE ASSOCIATION STUDIES IN BUDDING YEAST

2.1 Abstract

In recent years, a large number of diverse budding yeast genomes have been sequenced, leading to an interest in conducting genome–wide association (GWA) studies in this model organism. In some ways, *S. cerevisiae* is ideal for GWA studies: strains are highly diverse, phenotypes are relatively easy to quantify, and linkage disequilibrium extends over a short distance. However, the extensive population structure among *S. cerevisiae* strains can lead to spurious associations and highly inflated false positive rates. In this chapter, we examine the population structure present in a set of 302 *S. cerevisiae* isolates. We simulate phenotype measurements for a variety of traits with different genetic architectures, and compare the performance of multiple methods of correcting for population structure when conducting GWA. We find that the mixed–model method GEMMA outperforms an uncorrected t–test, as well as the inverse–regression method GCAT. In addition, the original formulation of GEMMA outperforms several novel extensions that incorporate more localized information about population structure.

2.2 Introduction

How can we identify the genetic factors that underlie heritable trait variation? In the simplest cases, we can make educated guesses about genes that might be involved based on functions that have already been characterized. But given the complexity of most traits of interest, this approach often falls short, and statistical approaches to systematically test for a relationship between genotype and phenotype are necessary. Most broadly, we can locate relevant variation within the genome through either linkage analysis or association

mapping.[73]

In linkage studies, the genomes of related individuals with differing trait values are compared, and variants that are present in only the individuals with the trait of interest are candidates for being causative. With more families and/or more individuals within those families, we can narrow down the region in which a causal variant lies. Linkage studies have been used extensively to map traits in yeast, in part because it is relatively simple to generate very large pools of segregants that are ideal for mapping. But with the great diversity of yeast genomes now being sequenced, there has in recent years been interest in conducting genome-wide association (GWA) studies in yeast, as well.

The GWA study design was pioneered in humans, for which obtaining the large, informative families necessary for successful linkage analysis is challenging. In GWA studies, individuals in a population are typically divided into affected and control groups; every marker or SNP in the population is then tested for a statistically different frequency in the two groups, with a correction for the large number of SNPs being tested.[24] GWA studies can also be used to analyze quantitative—rather than binary or case/control—traits by regressing the measured trait values on the genotype for every individual at each SNP.[74]

Early GWA studies were undertaken with the hope that they would help fulfill the promise of the Human Genome Project "to provide tools for identifying genetic factors that contribute to common, complex diseases such as cancer and diabetes," and ultimately suggest possible treatments.[23] Although we are still far from solving all the mysteries of heritable disease and health traits, GWA studies have proved to be a powerful investigative tool in many cases. Early successes of GWA studies included mapping risk of myocardial infarction to two SNPs in a single gene,[75] mapping risk of age-related macular degeneration to a single SNP that increases risk approximately fourfold,[76] and identifying the gene in which variation can lead to Crohn's disease.[77]

The success of GWA studies depend on several factors of the study design as well as the genetic architecture of the trait, including the heritability and the number of variants underlying it, as well as the frequencies and effect sizes of those variants[78] (Fig. 2.1). Most

early successes were in common diseases with a small number of underlying genetic variants of large effect, but subsequent work has identified variants underlying traits with increasingly complex underlying genetic architectures. By 2009, GWA studies had provided evidence of specific polymorphisms related to 40 different common diseases.[24] Hundreds of complex traits have now been linked to specific genetic variants, and genetic variants identified by GWA studies now total more than 10,000.[78]

Although linkage studies do not have the same downsides in yeast and other model organisms as they do in humans, GWA is still appealing because it takes advantage of naturally occurring variation rather than requiring the construction of new strains. The budding yeast *Saccharomyces cerevisiae* was the first eukaryote to have its whole genome sequenced,[8] but the sequencing of a greater diversity of strains within the species lagged by several years. In 2012, there were 36 *S. cerevisiae* genomes available from the *Saccharomyces* Genome Resequencing Project.[60,64] Today, there are more than one thousand.[14,61,62] In the past several years, GWA methodology has been explored in *S. cerevisiae*,[64,80] and has been applied broadly in the species to find variants associated with clinical versus nonclinical backgrounds[81] and variants associated with resistance to copper, lithium, and certain antifungal drugs,[62] among others.

S. cerevisiae is a good organism to use for this type of genetic mapping because strains are highly diverse and linkage disequilibrium extends over a relatively short distance.[14,33,64] A challenge for GWA studies in yeast, however, is the existence of extensive, complex population structure than can lead to high false positive rates. When subpopulations differ in their average trait values and in allele frequencies at certain sites, spurious associations can arise (Fig. 2.2). Even when using tools to correct for population structure, false positive rates can still be approximately double what they should be.[64]

In this chapter, we examine the population structure present in a set of 302 *S. cerevisiae* genome sequences (provided by Joseph Schacherer), which are now available as part of the published 1,011 genomes collection.[14] We characterize the power and false positive rate for mapping traits with a variety of genetic architectures in these strains, using simulated trait

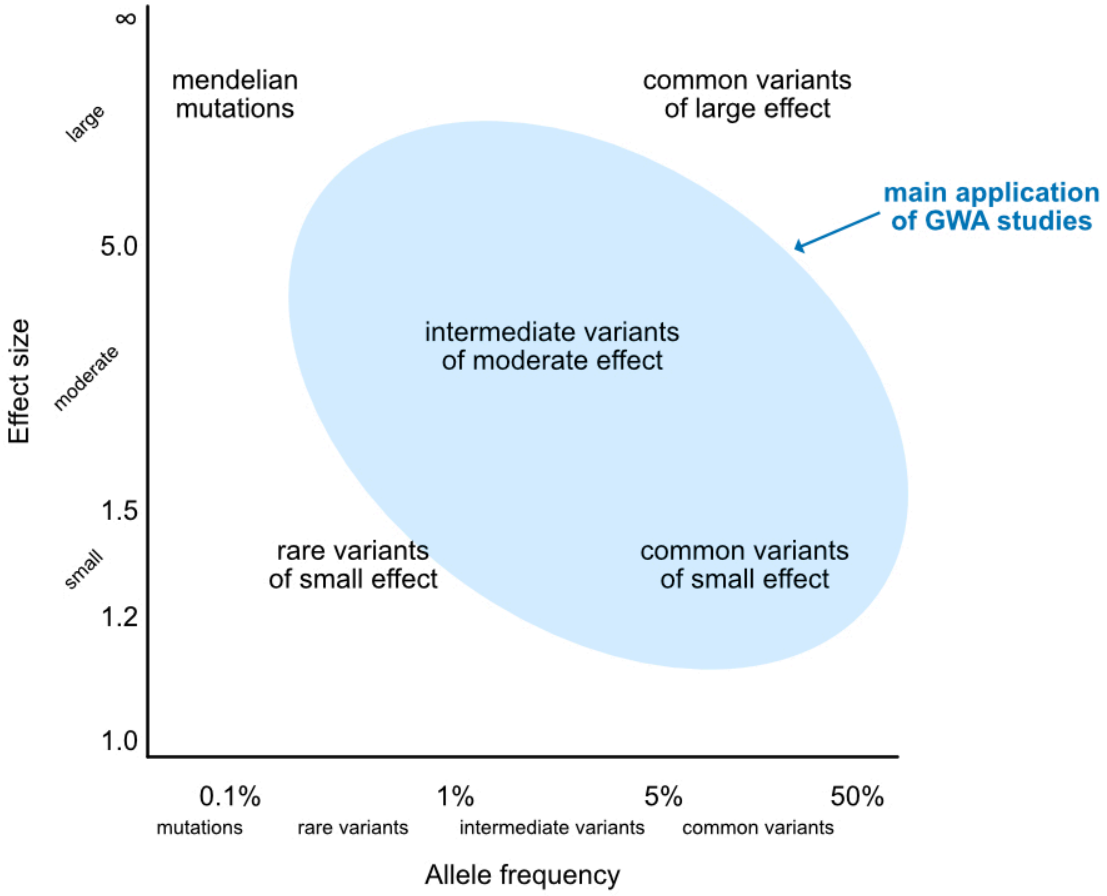

Figure 2.1: Applicability of GWA studies to identification of variants with varying effect sizes and frequencies in the population. GWA studies are primarily useful for identifying moderately common variants with small to intermediate effect sizes. For very rare variants (or "mutations"), other approaches, such as more targeted sequencing of affected families, may be necessary. Common variants of large effect are, of course, the easiest to detect but in general are not responsible for complex diseases. Adapted with modification from Bush and Moore 2012.[79]

data and a variety of methods for performing association tests with and without corrections for population structure.

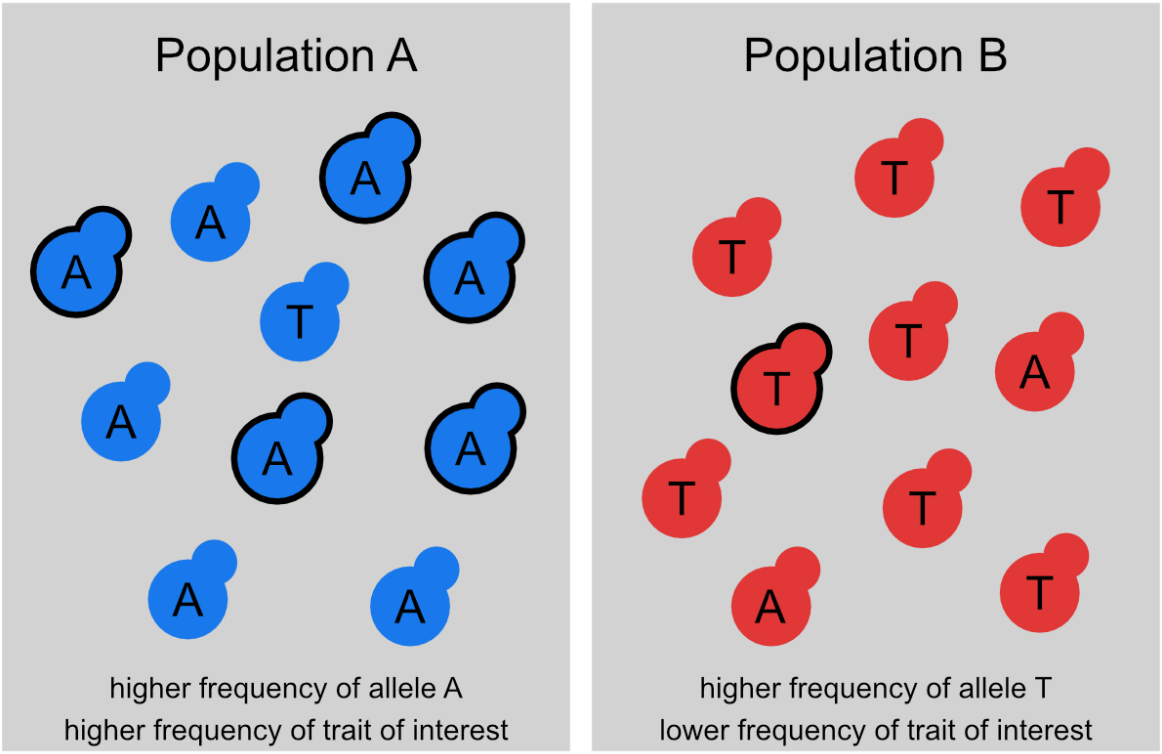

Figure 2.2: Population structure can result in spurious associations. If populations A and B express the phenotype of interest at different rates, then alleles that differ in frequency between the two populations can be spuriously associated with the phenotype. In this simple example, allele A is associated with a phenotype represented by a dark outline. The association is real but is due to population structure rather than a causal relationship.

2.3 Results

2.3.1 Population structure in 302 S. cerevisiae isolates

The genome sequences of 1,011 *S. cerevisiae* isolates from diverse geographical and ecological environments were recently published.[14] All analyses in this chapter are based on a subset of 302 of those sequences shared prior to the publication of the full data set. Using the maximum–likelihood program ADMIXTURE,[82] we estimate that there are eight distinct subpopulations represented in these strains (Fig. 2.3). This estimate falls between the five lineages recognized in the 39 *S. cerevisiae* strains of the Saccharomyces Genome Resequencing

Project[60] and the thirteen lineages proposed for all of *S. cerevisiae*.[83] The populations we identify correspond roughly to the location and/or type of environment from which the isolates were obtained, but approximately half of the strains show evidence of admixture. The minor allele frequency spectrum shows a small bump at approximately 0.35, possibly due to population structure or selective forces (Fig. 2.4).

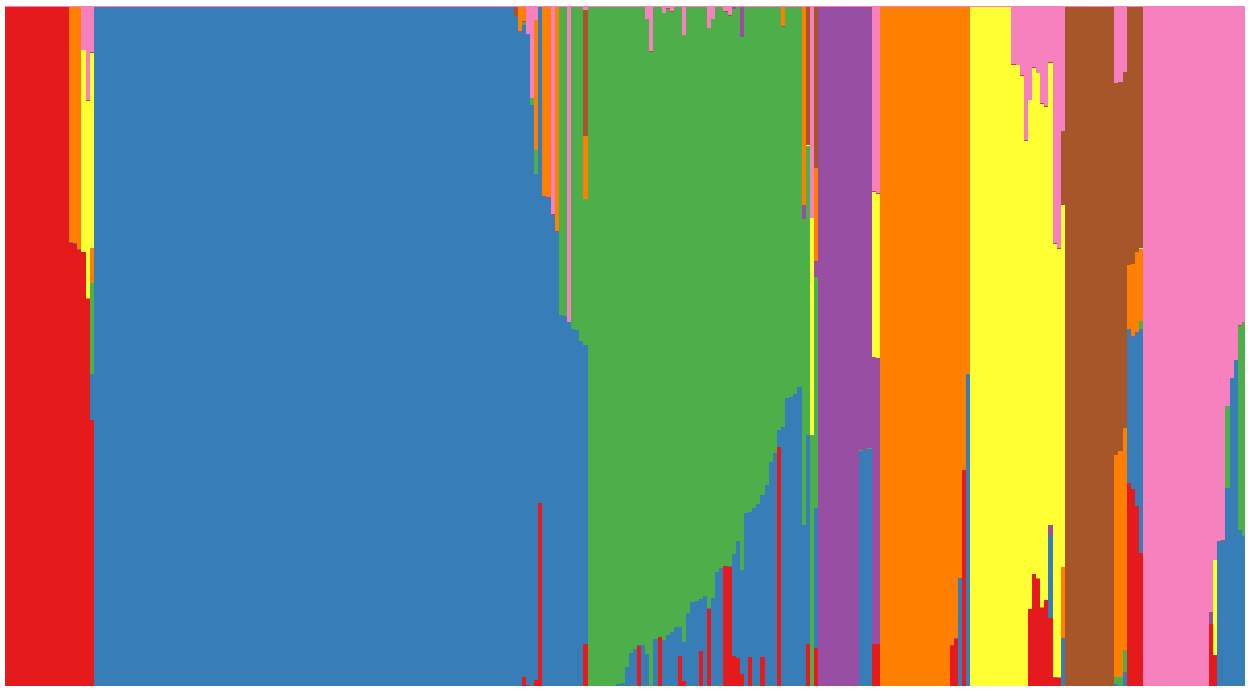

Figure 2.3: Populations determined by ADMIXTURE analysis. Each column shows the composition of one of the 302 strains, with the eight subpopulations represented by different colors. These subpopulations roughly correspond to assorted Asian strains (red), assorted European strains (blue), North and South American clinical strains (green), African wine strains (purple), European beer strains (orange), wild Asian strains (yellow), European wine strains (brown), and wild North American strains (pink).

2.3.2 False positive rates and power for different approaches to conducting GWA studies

We performed simulations to evaluate our ability to map quantitative traits with a variety of genetic architectures. For each simulation we (1) randomly chose between one and three

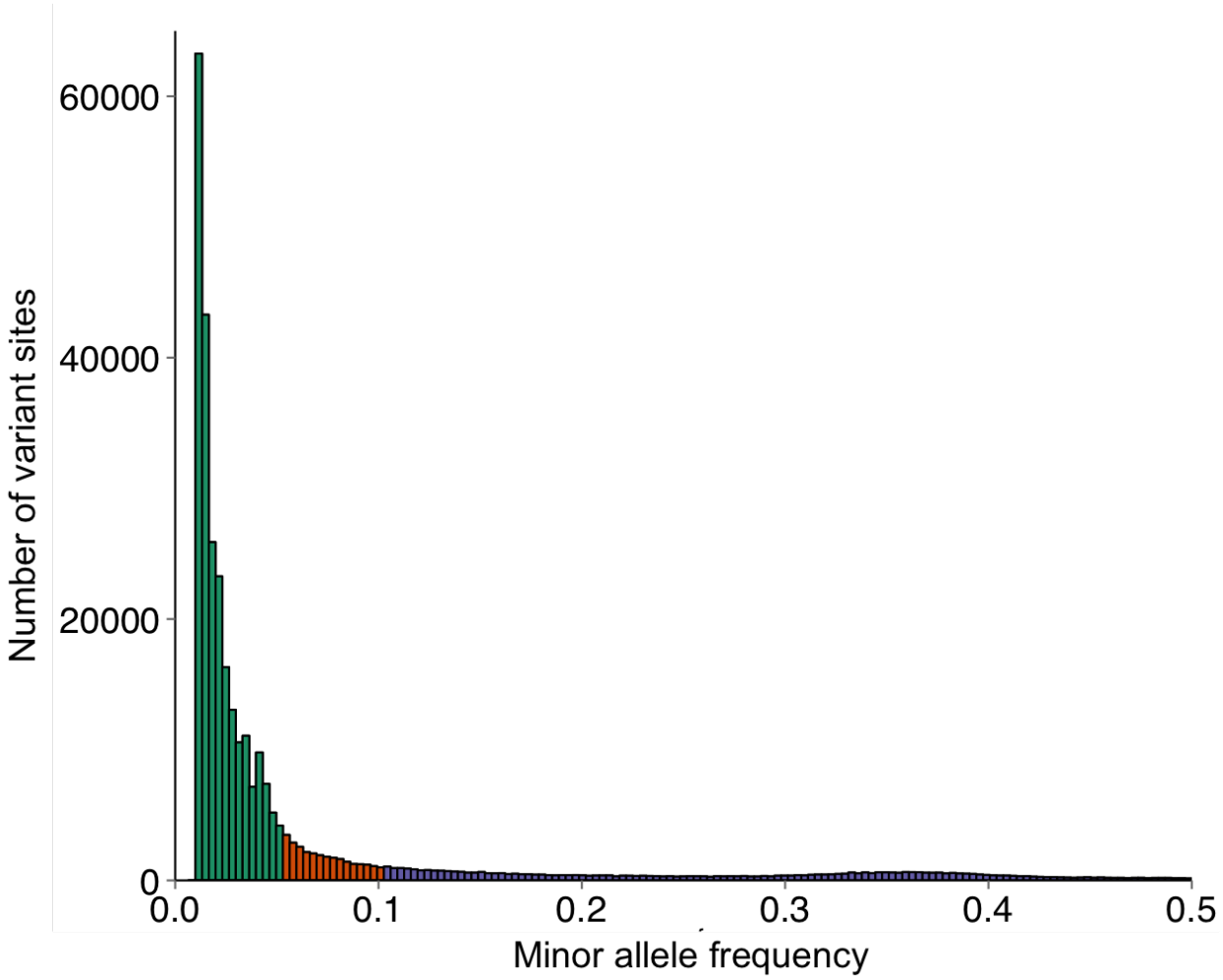

Figure 2.4: Minor allele frequency spectrum for 302 strains. Only variants present in at least three strains are shown. Variants with MAF less than 0.05 are considered rare (green), with MAF between 0.05 and 0.1 are considered intermediate (orange), and with MAF greater than 0.1 are considered common (purple).

quantitative trait loci (QTLs), (2) generated phenotypes for each strain based on the alleles at those loci, and (3) used multiple methods to calculate association between genotype and phenotype at every marker across the genome (Fig. 2.5). The methods we compared were a basic t-test, a linear mixed model method called GEMMA,[84] and an inverse regression method called GCAT.[85]

We set a threshold for significance such that we expect approximately seven false positives

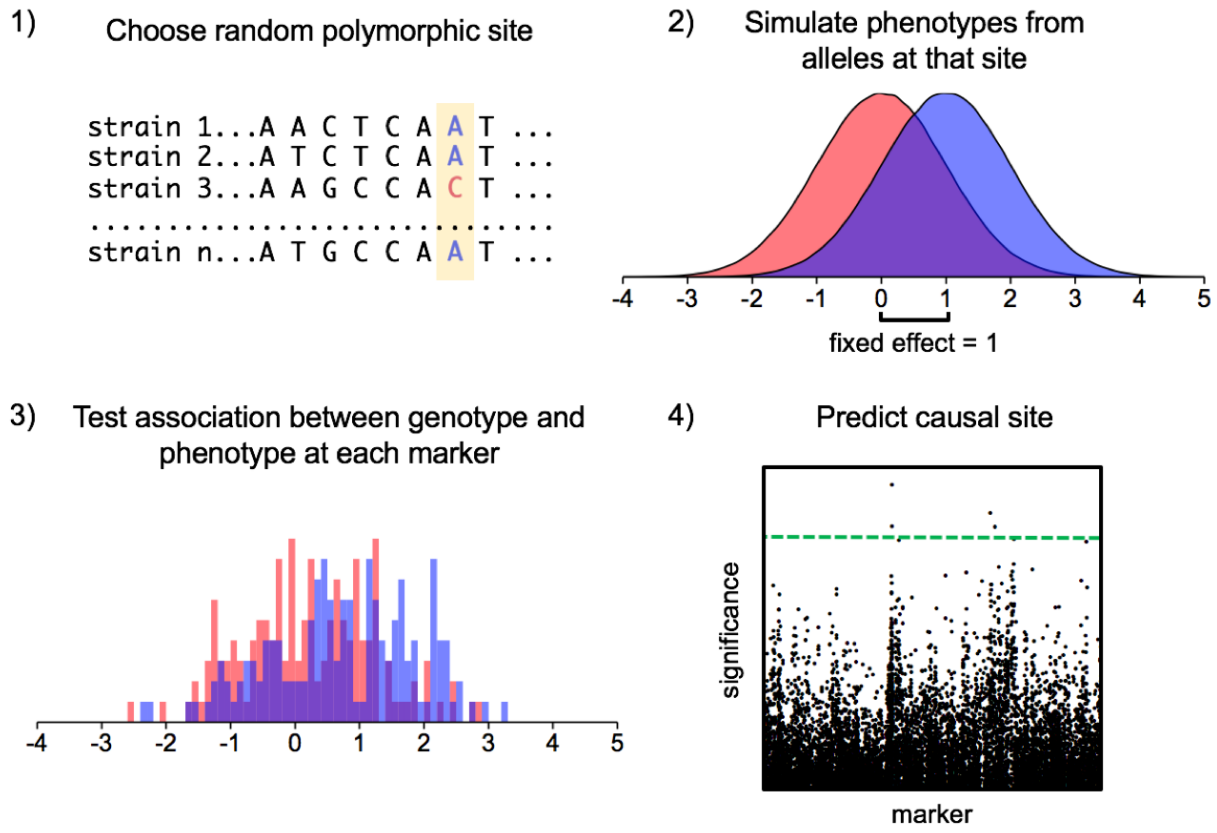

Figure 2.5: Simulation framework for comparing methods of associating phenotype and genotype. First, a random diallelic site is chosen as the causal site. Phenotypes are drawn for each individual based on their genotype at that site and the predetermined effect size. A variety of methods are used to test for association at every polymorphic site, and sites that exceed a specified threshold of significance are predicted to be causal.

per genome–wide scan (with 73,053 tested markers). We calculated the number of false positives observed in simulations with varying trait heritability, minor allele frequency, and number of QTLs (Fig. 2.6a). For the basic t–test, the observed false positive rate remains approximately the same when the total additive heritability of the trait is split between one, two, or three QTLs, but increases as the total heritability increases and as the minor allele frequency of each QTL increases.

When using GEMMA to correct for population structure, the false positive rate is much

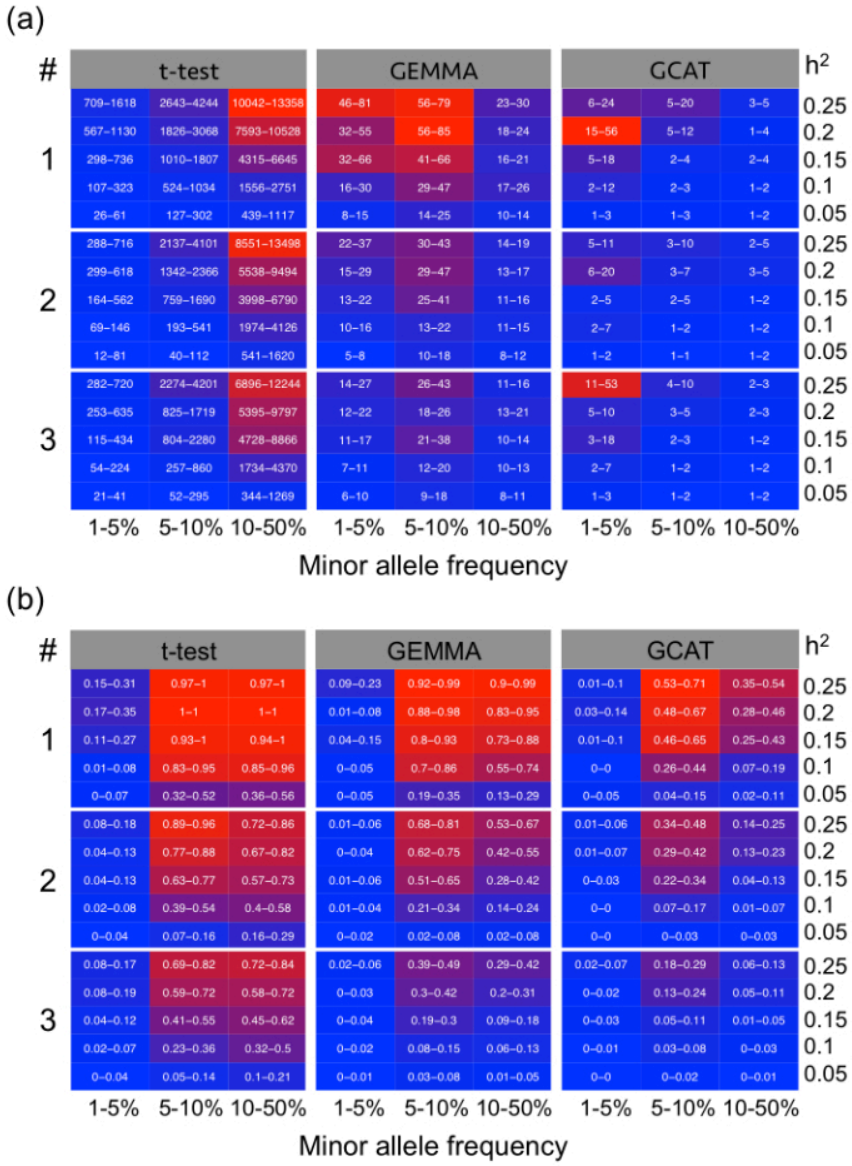

Figure 2.6: Comparison of the number of false positives and power for three different association methods. In both panels, the significance threshold is 10^{-4}, corresponding to an expected number of approximately 7 false positives per genome–wide scan. Blocks are separated vertically by the number of QTLs and horizontally by the method used. Each block is divided vertically by the total heritability of the trait and horizontally by the minor allele frequency of each QTL. (a) Colors correspond to the average number of false positives observed across approximately 100 simulations, ranging from blue for fewer to red for more in each block, with 95% bootstrapped confidence intervals shown for each combination of QTL number, heritability, minor allele frequency, and method. (b) Colors correspond to the average power observed across approximately 100 simulations, ranging from blue for low to red for high in each block, with 95% bootstrapped confidence intervals shown.

lower—approximately 2– to 700–fold depending on the specific simulation parameters. With this method, the false positive rate is highest for intermediate minor allele frequencies, and again increases with total heritability. The false positive rate decreases when the total heritability is split between more QTLs. When using GCAT, the false positive rate is in general lower than when using GEMMA; in fact, for most of the simulations, GCAT appears to overcorrect for population structure, frequently resulting in fewer than the expected seven false positives per scan. With GCAT, QTLs with lower minor allele frequency result in more false positives, as do traits with higher heritability.

For each set of simulations, we also calculated the power of each method to detect the causative QTL(s) (Fig. 2.6b). Across all simulations, the t-test has the highest power, followed by GEMMA and then GCAT. Combining the calculated power and false positive rates gives a better sense of the tradeoffs each method makes and which is overall most useful—reducing the false positive rate is beneficial, but not if it comes with too large of a reduction in power. For simulations with a total heritability of 15% divided evenly between 1-3 QTLs, we compared ROC curves for the three methods (Fig. 2.7a). GCAT is an improvement over the uncorrected t–test, but GEMMA outperforms both of these methods by a wide margin. A similar pattern holds for higher and lower heritabilities (data not shown).

In addition to evaluating the abilities of GEMMA and GCAT to correct for population structure in this data set, we also evaluated the performance of some modifications to GEMMA. We theorized that correcting for patterns of allele sharing in a way more specific to the marker currently being tested might result in a lower false positive rate. Accordingly, since population structure varies across the genome,[86] we replaced the global kinship matrix that GEMMA utilizes with a matrix computed from only sites within a variable distance from the marker being tested, ranging from 10 sites to the entire chromosome (Fig. 2.7b,d). We separately tried using a kinship matrix constructed from only sites with similar allele frequencies to the marker being tested, rather than all sites (Fig. 2.7c). Neither of these approaches resulted in improvements over GEMMA's performance when using a global kin-

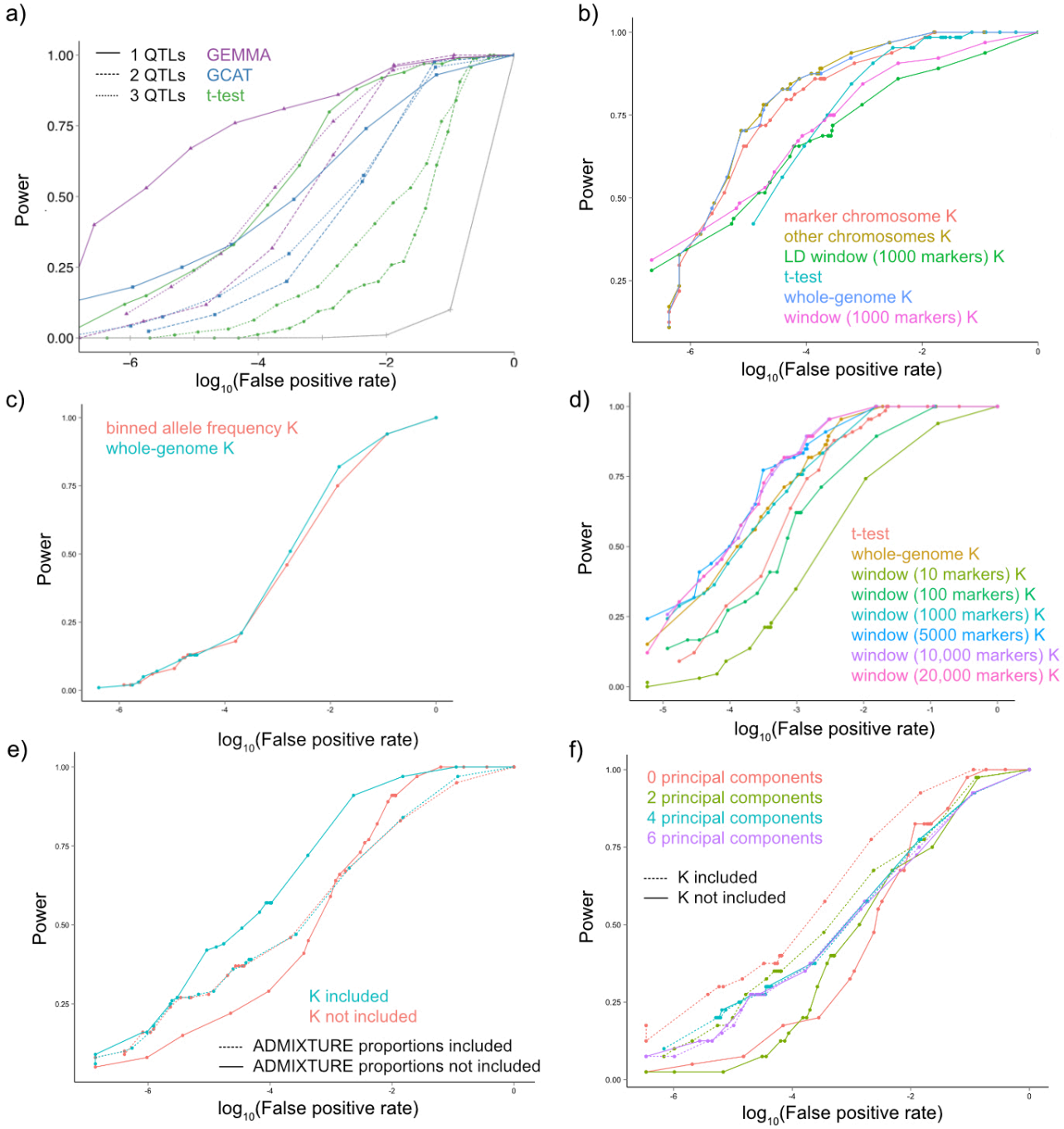

Figure 2.7: Figure legend continued on following page.

Figure 2.7: ROC curves comparing methods of correcting for population structure. (a) ROC curve for a total heritability of 15%. GEMMA outperforms both GCAT and a basic t-test. This result holds when the heritability is split equally between one, two, or three QTLs. The gray line shows the case where the power is equal to the false positive rate. (b) GEMMA performs similarly to mixed models with kinship matrices based on either all sites on the chromosome of the marker being tested ("marker chromosome K") or all sites not on that chromosome ("other chromosomes K"). Models where the kinship matrix is based on only sites within 1000 base pairs of the tested marker ("window (1000 markers) K"), or sites within 1000 markers of the closely linked markers surrounding the tested marker ("LD window (1000 markers) K"), perform similarly to the t–test. Trait values are determined by a single QTL and have a heritability of 10%. (c) A mixed model in which the kinship matrix is based only on sites with similar allele frequencies to the tested marker performs equivalently to the model with the whole genome kinship matrix. Trait values are determined by a single QTL and have a heritability of 5%. (d) Models with local kinship matrices based on numbers of sites ranging from 10 to 20,000 on either side of the tested marker show decreasing performance as they become more localized. Trait values are determined by a single QTL on chromosome I and have a heritability of 5%. (e) Adding the ADMIXTURE population fractions as a fixed effect in the model results in worse performance than using only the whole-genome kinship matrix. Trait values are determined by a single QTL and have a heritability of 10%. (f) Adding principle components from performing PCA on the genotypes of all individuals results in worse performance. Trait values are determined by a single QTL and have a heritability of 10%.

ship matrix. Finally, introducing principle components or admixture proportions as fixed effects in the model also failed to improve the performance of GEMMA's original model (Fig. 2.7e,f).

2.4 Methods

2.4.1 Sequence data, markers, population structure analysis

We obtained VCF files for 302 *S. cerevisiae* isolates prior to their publication as part of the 1,011 *S. cerevisiae* genomes collection.[14] In total, there are approximately 1,000,000 polymorphic sites in this set of isolates, and approximately 100,000 diallelic sites with minor allele frequency greater than 5%. Using `ldselect`,[87] we obtained a final set of 74,053 unlinked markers to test for association with simulated phenotypes.

2.4.2 Simulation framework

To compare the performance of different association mapping methods when applied to these strains, we simulated traits of varying heritabilities divided between different numbers of QTLs of differing allele frequencies. The simulation framework consists of the following steps.

1. The heritability is set to a value between 5 and 25%, and a minor allele frequency range is chosen (1-5%, 5-10%, or 10-50%). Between one and three random diallelic variant sites with allele frequencies in that range are chosen to be the causal loci.

2. A phenotype is generated for each isolate based on the allele it has at each causal locus chosen in the previous step. These phenotypes are drawn from a normal distribution with mean zero for the major allele and mean a for the minor allele, with a variance of 1 for each allele. The fixed effect a is calculated based on the heritability, h^2, using the formula $a = \sqrt{\frac{h^2(n-1)}{nq(q-1)(h^2-1)}}$, where n is the number of strains and q is the minor allele frequency.[88]

3. Each marker is tested for association with the simulated phenotypes using a t–test or one of the multiple association mapping methods described below.

4. The power and false positive rate for the different mapping methods are calculated, given a defined significance threshold (or expected false positive rate).

2.4.3 Approaches to correct for population structure

Genome-wide efficient mixed model analysis (GEMMA)

GEMMA[84] is an efficient implementation of a linear mixed model with a random effect that depends on the relatedness between strains: $y = \alpha + x\beta + u + \epsilon$, where y is the vector of quantitative trait measurements, α is the intercept, x is the vector of genotypes at the site being tested for association, β is the effect of the genotype, u is the random effect, and ϵ is

the residual error. The random effect, u, is distributed normally with mean zero and variance $\sigma_g^2 K$, which is the variance in phenotype due to genotype multiplied by the kinship matrix. The kinship matrix represents the relatedness between all pairs of strains. Thus, when the random effect based on the kinship can explain most of the phenotypic variance, our estimate of the genotypic effect, β, will be reduced. In this way, we can reduce the number of false positives due solely to population structure. We will, however, also reduce our power since in some cases the population structure may be correlated with the causal variant. In addition to testing the efficacy of GEMMA in its standard implementation, we also tested versions with modified estimates of the kinship matrix, described later.

Genotype-conditional association test (GCAT)

GCAT[85] is a recently developed method that uses an inverse regression approach. According to its model, non–genetic factors—such as population structure, lifestyle, and environment—influence the trait both directly and through genotype. The test for association thus becomes: $\Pr(x_i|y, \pi_i(z)) = \Pr(x_i|\pi_i(z))$, where x_i represents the alleles at the ith site being tested for association, y represents the trait measurements, z represents the genetic factors, and π_i is the function describing how population structure and other non–genetic factors affect the distribution of observed alleles. If the above equality holds, then the population structure is sufficient to explain the observed genotype. If it does not hold, then there is a significant relationship between the genotype and phenotype that is not solely explained by z. This method was designed to handle more arbitrarily complex situations in which population structure and other non–genetic factors may be correlated.

Modifications to GEMMA kinship matrix and model

Since the patterns of relatedness between strains vary substantially across the genome, it is possible that estimating the kinship matrix K from the entire genome is not ideal. We therefore implemented mixed models that use more localized versions of K for each marker,

computed based on a variety of window sizes ranging from the nearest 10–20,000 sites on either side of the marker. In addition, we computed kinship matrices based on only the chromosome containing the marker, and based on every chromosome except the one containing the marker. Finally, we implemented a mixed model that contained fixed effects instead of, or in addition to, the random effect of the kinship matrix; these fixed effects were either the population fractions obtained from ADMIXTURE or one of a range of numbers of principal components (the first 2, 4, or 6) obtained from conducting PCA on the genotype data.

2.5 Discussion

As more *S. cerevisiae* genomes have been sequenced from isolates collected from diverse environments, the complex patterns of population structure within the species have become more evident.[14,83] When performing association studies in this species, existing methods of controlling for population structure do, based on simulated phenotype data, substantially reduce inflated false positive rates. We find that the mixed-model method GEMMA[84] outperforms an uncorrected t-test, as well the inverse-regression method GCAT.[85] In addition, GEMMA outperforms several novel extensions to its model incorporating more localized information about population structure.

However, for quantitative traits influenced by QTLs of large effect sizes, false positives remain prevalent even when using these more sophisticated models of testing for association. Although GWA studies in yeast are appealing due to the ease with which they can be conducted on the large number of sequenced isolates now available, results should be carefully verified to avoid attributing trait variation to spuriously related genetic variation. As always, positional mapping methods are only a starting point for understanding the molecular causes of phenotypic differences between individuals.

Chapter 3: A LARGE AND DIVERSE CROSS FOR MAPPING QUANTITATIVE TRAITS IN *S. CEREVISIAE*

3.1 Abstract

The loci we identify underlying complex traits generally only account for a small fraction of the total expected heritability. In addition, when we identify causal loci, the resolution is generally not fine enough to suggest possible molecular mechanisms. Designing a cross that generates a large number of very diverse segregants can help us achieve both the power to identify loci with small effect sizes and the resolution to locate them within individual genes or regulatory regions. The recent sequencing of large numbers of diverse *S. cerevisiae* strains, combined with advances in experimental techniques for mating and sequencing large numbers of these strains, makes it possible to generate a pool of segregants ideal for quantitative trait mapping. In this chapter, we outline several aspects of the design of a funnel cross between eight diverse parental strains. We select the parental strains to capture a maximal fraction of all observed polymorphic sites. We then simulate a cross between these eight strains to estimate the mapping power and resolution for traits with varying heritabilities influenced additively by multiple loci. We find that with a pool of 10,000 segregants, we should be able to map loci with small effect sizes, often with single–gene resolution. By using a hidden Markov model to impute missing sequence data in the segregants, low–cost sequencing approaches will be sufficient, making this experiment an efficient approach to dissecting quantitative traits in *S. cerevisiae*.

3.2 Introduction

One of the primary goals of the study of genetics is to associate observed trait variation with the underlying genetic variation that contributes to it. As discussed in the previous chapter, genome–wide association studies can allow us to link naturally occurring variation

in a population to differences in trait values. However, these types of studies result only in correlation, rather than causation. In addition, they can suffer from inflated false positive rates due to population structure, and they can also lack power in cases where it is difficult to establish a well–matched control group.[24] Instead of relying on natural variation that occurs in existing populations, we can use linkage to analyze families of related individuals that have randomized and potentially more informative combinations of variants.

In humans, linkage analysis is limited by the number of families that are informative for a given trait that can be found and sequenced. For example, to detect an allele at moderate frequency that confers a twofold greater risk for a trait, we would need to analyze thousands of families.[18] In model organisms such as budding yeast, however, the possibility of generating very large, diverse pools of segregants from genetic crosses makes linkage mapping a very powerful approach. *Saccharomyces cerevisiae* is an ideal organism for mapping quantitative traits for several other reasons, as well. Linkage disequilibrium extends over a relatively short distance, and the species contains a large amount of diversity in both genotype and phenotype.[14] These characteristics allow for the fine–scale mapping of a wide variety traits.

For these reasons, linkage mapping has been a successful technique for dissecting complex traits in yeast for decades, but there still in exists a significant amount of "missing heritability" for most traits that have been studied.[58,89] That is, the fraction of the total estimated heritability explained by all of the causal variants identified is generally less than 1, and often substantially less.[89,90] There are several hypotheses about the source of this missing heritability, and it likely involves a combination of rare variants that we have not yet had the power to detect, variants with small effect sizes, and higher–order interactions between variants at different locations in the genome.[58,91]

Another challenge of linkage mapping, even in a model organism such as *S. cerevisiae*, is moving from large regions that are identified as causal to specific, mechanistic variants in genes or regulatory sequences.[92] Understanding the mechanisms by which genetic variation leads to phenotype will always rely on more extensive knowledge of gene annotations and careful functional experiments, but performing genetic crosses that result in greater mapping

resolution can help us narrow in on specific genes or regulatory mechanisms more easily. Mapping resolution is determined by both the density of markers, or polymorphisms, that exist among the strains used in a cross and the amount of recombination that has occurred to randomize the combinations of alleles at those sites.

In addition to the resolution with which we can locate relevant genetic variants, the power of a cross to identify QTLs that influence variation in a trait of interest is also a crucial aspect of experimental design. The power of a cross depends on the total number of individuals we examine, as well as the genetic architecture of the trait, which we do not have any influence over. Traits with higher heritability—for which more of their variance is explained by genetics—are easier to detect, as are traits that are influenced by QTLs with individually larger effect sizes. Conversely, traits that depend heavily on higher–order interactions between loci, or that depend on many QTLs of very small effect, are more difficult to map.

QTL mapping in yeast has typically been carried by analyzing the progeny from a cross of two divergent parental strains. Such crosses have successfully identified QTLs underlying a wide range of traits, but there is a limit to the amount of diversity that can be captured—and relatedly, the mapping resolution that can be achieved—when using only two strains. To address this shortcoming, a study was carried out with a cross of four parental strains from diverse lineages.[93] Dowever, since only a few hundred segregants were generated, the resulting power was low and mapping intervals were large. With the extreme QTL mapping (X–QTL) approach, millions of segregants from this same four–parent cross were subjected to a variety of selective pressures, and those with the most extreme growth traits were examined.[89] Pooled analysis methods such as X–QTL can result in high power for detecting additive loci with small effect sizes, but they are not effective for detecting epistasis and can only be applied to traits that can be reasonably selected for.[94] In addition, they can be confounded by aneuploidy and by different genotypes causing similar phenotypes.[95] Implementing new study designs that utilize linkage to map traits with greater precision in diverse yeast strains remains an important area of exploration.

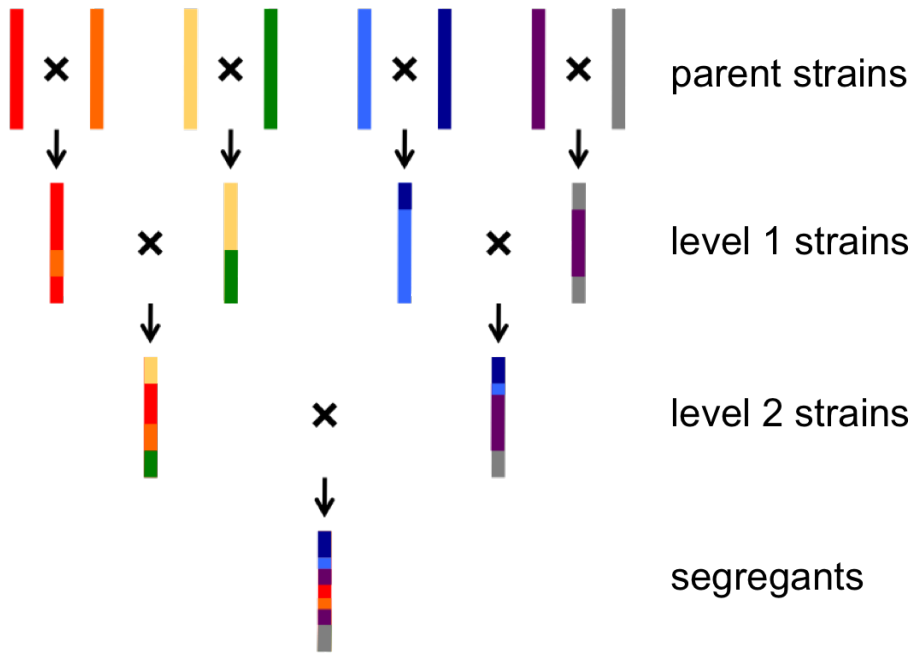

Figure 3.1: Eight parent funnel cross.

In general, crosses that utilize more diverse parents, that involve more recombination, and that produce more segregants will give us the greatest mapping power and resolution for a given trait. We have therefore designed a cross to make full use of our experimental capabilities that we anticipate will have unprecedented power and resolution for mapping QTLs. Our funnel cross design (Fig. 3.1) is inspired by the Mouse Collaborative Cross,[96] but while this approach in mice required a decade and an entire consortium to develop,[97] a similar cross in yeast can be carried out in yeast in a few years by a single laboratory.

Before carrying out this cross in the laboratory, however, we investigated several aspects of its design using simulations. In particular, we wanted to determine how successfully we could expect to map traits with a variety of underlying genetic architectures. We first selected specific parental strains to use. To do so, we calculated the fraction of the total known genetic diversity in *S. cerevisiae* that could be captured using varying numbers of parental strains, and ultimately selected a highly diverse set of eight strains. We used simulations to estimate the power and resolution a cross of these parental strains should provide. Finally,

we investigated approaches to sequencing a large pool of segregants that the cross could generate. While whole genome sequencing has become inexpensive enough to apply to a large number of strains, it is more cost–effective to perform low coverage sequencing and impute the missing information, a process that should be straightforward since we have high quality sequences of the parental strains. We implemented a hidden Markov model to perform this imputation and used simulations to assess its accuracy.

3.3 Results

3.3.1 Choosing parental strains for funnel cross

We chose parental strains for the cross to capture a large fraction of the diversity present in total set of 302 strains, using a simple greedy algorithm. We chose sets of strains of size two through sixteen, then calculated the fraction of polymorphic sites in the set of all strains that were captured in each subset of a given size (Fig. 3.2). With eight strains, it is possible to more than half of all variants present in the whole set of strains, and approximately 89% of common variants (defined as those being present in at least 5% of strains). Although there is not a large difference in the fraction of variants captured with slightly more or fewer than eight strains, we chose this number as a compromise between total diversity and experimental difficulty. With eight strains, we can use a simple three–level funnel cross design (Fig. 3.1). In the course of performing the crosses between the parental strains we chose for their diversity, some pairs had low spore viability and thus were replaced with others chosen to maintain as high of a level of diversity as possible.

3.3.2 Mapping power and resolution of eight–parent funnel cross

We performed simulations of this eight–parent funnel cross to analyze both the power (the ability to detect a QTL when present) and the resolution (the size of the region containing the correctly–detected QTL). In all simulations, we assumed that QTLs influenced traits additively, but we varied the heritability of the trait. We also varied the size of the final

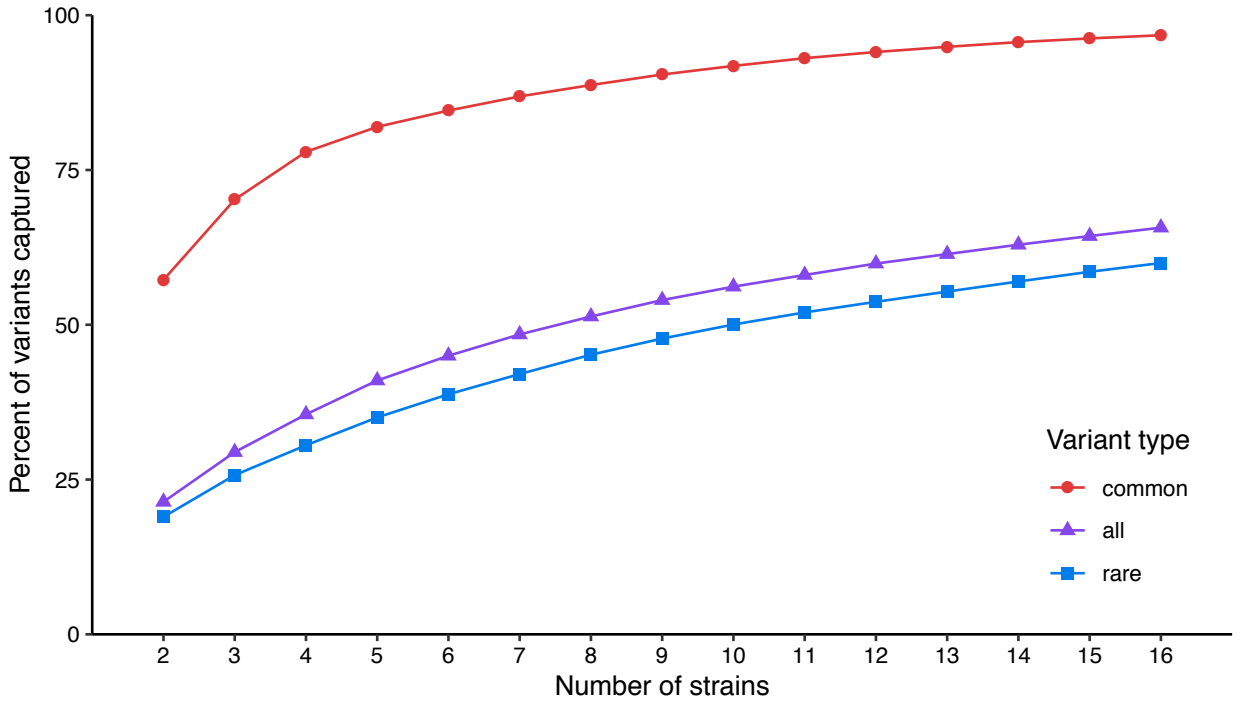

Figure 3.2: Fraction of total diversity captured by different–sized subsets of strains. It takes fewer strains to capture the same amount of common variants (present in >5% of strains) than rare variants (present in <5% of strains).

segregant pool between 1,000 and 10,000. For a pool of 5,000-10,000 segregants and total heritability greater than 5%, the mapping resolution is less than the average gene size in *S. cerevisiae* (Fig. 3.3a; Table 3.1). For 5,000 segregants, power is very high for loci contributing a heritability of 1%; for 10,000 segregants, the power is very high for those contributing a heritability of 0.5% (Fig. 3.3b).

Although these simulations are useful for comparing mapping performance on traits of differing heritabilities and pools of differing sizes, it is important to acknowledge their limitations. We have only simulated traits with simple additive genetic architectures, while mapping more complex traits is generally more difficult. In addition, we have made simplistic assumptions about the way that recombination will occur in the cross: we have not modeled the heterogeneity of recombination across the genome, instead assuming crossovers

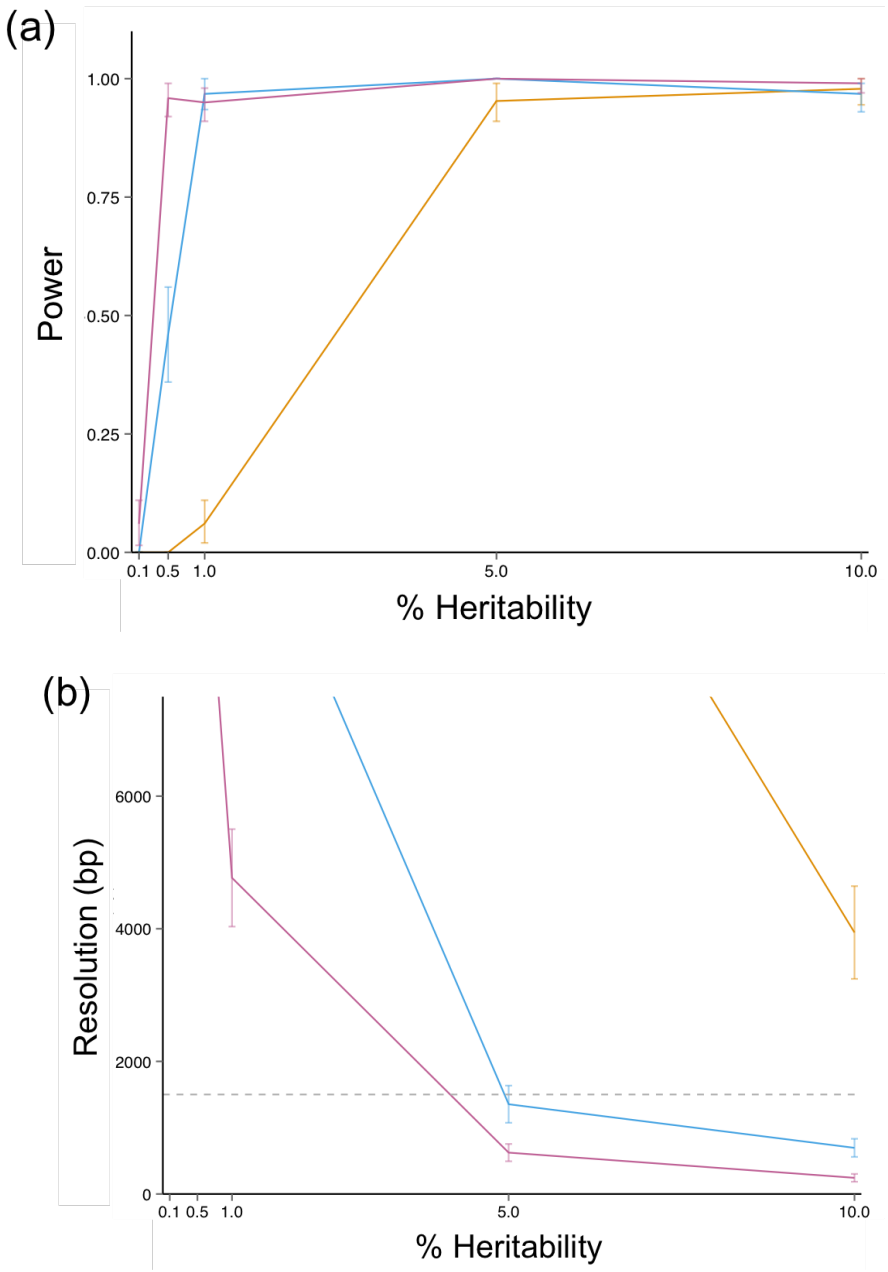

Figure 3.3: Power and resolution of eight-parent funnel cross. (a) The average power from simulations mapping a single QTL with varying heritability (at a LOD score threshold of 5.64) for 1,000 (orange), 5,000 (blue), and 10,000 (purple) segregants. Error bars are 95% bootstrapped confidence intervals. (b) The mean size of the Bayesian credible interval containing the detected locus (with LOD score above 5.64) for 1,000 (orange), 5,000 (blue), and 10,000 (purple) segregants. Error bars are the standard error of the mean.

Table 3.1: Resolution for mapping simulated QTLs of varying heritabilities.

Heritability explained	Avg. Interval Size (bp)	95% CI (bp)
1%	4768	4033–5502
5%	624	494–754
10%	244	185–302

occur according to a Poisson process with a uniform rate across the genome.

3.3.3 Imputation of segregant haplotypes from incomplete sequencing data

In order to reduce the cost of performing this cross, we investigated the possibility of using RAD–seq rather than whole–genome sequencing on the segregant genomes, followed by imputation of the missing sites based on the high–quality parental genomic sequences. We implemented a hidden Markov model to impute the missing sequence data and tested its performance on simulated segregant genomes. For the case of a restriction cut site every 900 bp on average (which is what we would expect with the proposed restriction enzymes), imputation accuracy is approximately 99.5%, or 32 out of 6703 markers incorrect on chromosome I (Fig. 3.4). Further tuning of the model parameters would likely result in even greater accuracy. In addition, it may be possible to develop a more sophisticated model that considers the sequences of all segregants at once to maximize the information about where recombination events occurred. In this case, sequencing some of the strains at intermediate levels of the cross could provide more information about the locations of recombination events if necessary. Finally, this model could also be used for imputation with low–coverage whole–genome sequencing data if that approach was used instead.

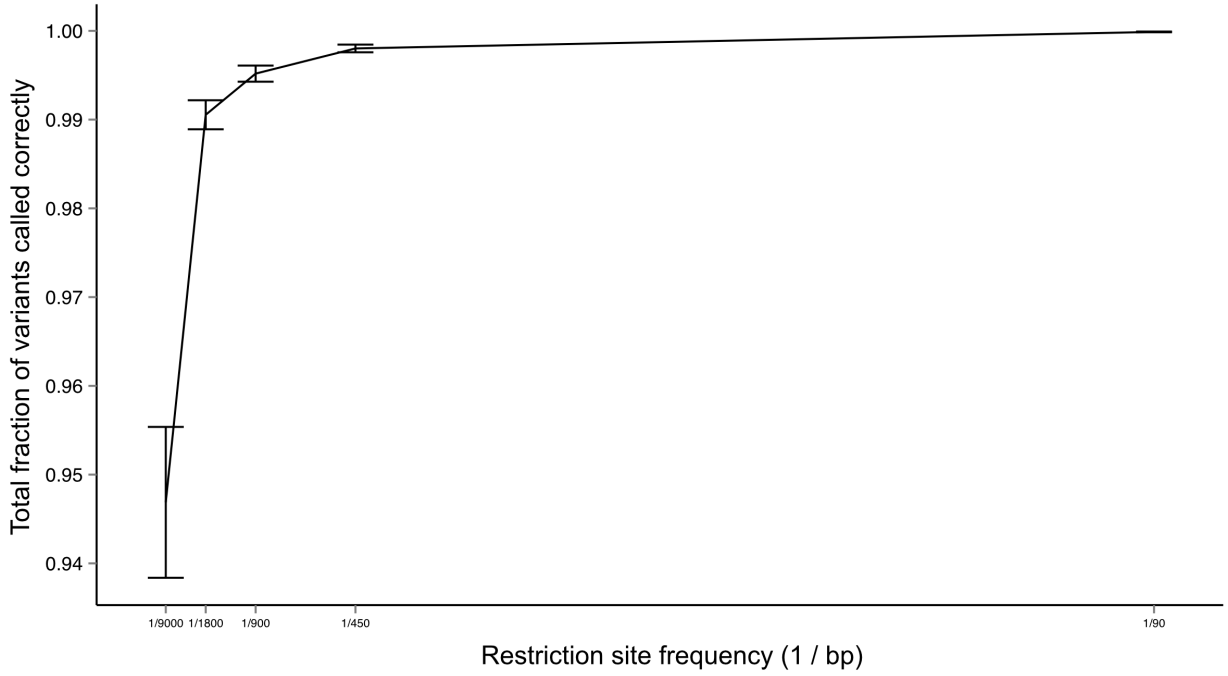

Figure 3.4: Imputation accuracy as a function of cut site frequency. Error bars are 95% bootstrapped confidence intervals.

3.4 Methods

3.4.1 Parental strain choice

The total set of strains we considered consisted of 302 sequenced *S. cerevisiae* isolates provided by Joseph Schacherer, which are a subset of the recently published 1,011 genomes.[14] To choose parental strains that captured a large amount of the total diversity in this set of strains, we implemented the following greedy algorithm:

1. Choose the two most highly diverged strains.

2. To choose k strains, use the algorithm to choose $k-1$ strains. Then add the strain that results in the greatest increase in diversity over those $k-1$ strains.

This algorithm is not guaranteed to maximize the diversity, but considering all strain sets of size k is generally not feasible (for example, with $k = 8$ strains, there are $\binom{1000}{8} \approx 2.4 \times 10^{19}$ possibilities).

Only diallelic sites present in at least three different strain sequences were considered. Rare variants were defined as those with the minor allele present in fewer than 5% of strains, while common variants had the minor allele present in at least 5% of strains. Sites with missing sequence calls in some strains were retained, but any sites with missing data in the current subset of strains being considered were ignored in diversity calculations.

3.4.2 Simulations to inform cross design

We simulated the generation of segregant sequences from an eight–parent funnel cross, based on the chromosome I sequences from the parental strains previously selected for their diversity. In each mating, one crossover was required on the chromosome, with additional crossovers occurring according to a Poisson distribution with an expected value equal to the chromosome size multiplied by the empirically determined 6.1×10^{-6}.[98] All crossover locations were chosen randomly, which is a notable simplification since there exist highly conserved recombination hotspots in yeast;[99] assuming that recombination events are instead uniformly distributed across the genome likely results in estimates of mapping resolution that have too low of variance.

We next randomly selected one of the four gametes from a given simulated mating to use in the next level of the cross, and repeated this process to populate the whole level of the cross. We simulated random mating in the same way to produce the next level of the cross. We generated a specified number of segregants (1,000, 5,000, or 10,000) in the final pool, and used these sequences for mapping simulated trait values.

To map traits and determine mapping resolution, we performed a t–test at each marker, corrected for multiple testing. From the residuals, we calculated a logarithm of odds (LOD)

score with the formula:

$$\mathrm{LOD} = \frac{N}{2}\log_{10}\left(\frac{RSS_{\mathrm{total}}}{RSS_{\mathrm{major}} + RSS_{\mathrm{minor}}}\right),$$

where N is the number of segregants from the cross, RSS_{total} is the residual sum of squares for the phenotype across all segregants, and RSS_{major} and RSS_{minor} are the residual sum of squares of the phenotypes for individuals with the major and minor alleles at the marker being tested. If RSS_{major} and RSS_{minor} are small, then the marker successfully partitions individuals by phenotype, and the LOD score is larger. Using the LOD scores, we then determined the Bayesian credible interval by finding the region that contained 95% of the area under 10^{LOD} for the maximum LOD score, as implemented in R/qtl.[100]

3.4.3 Sequence imputation

One strategy for efficiently sequencing the genomes of the 10,000 segregants is RAD-seq. With this approach, the genome is cut with two restriction enzymes, MboI and ApoI, resulting in one cut every 900 base pairs on average. A barcode is then ligated onto the cut sites, and the fragments are sequenced with 150 base pair reads, resulting in approximately 140 bp of usable sequence. In the case of our eight–parent funnel cross, there would be one variant site roughly every 40 base pairs, and so we would sequence only about 15% of the variant sites in each segregant. Given the whole genome sequences of the parental strains, however, it should be possible to impute many of the missing variant sites. We have implemented a Hidden Markov Model to perform this imputation (Fig. 3.5).

In this model, there are eight states for each variant site, each representing one of the eight parental genomes. Transitions between states can only occur at adjacent variant sites. There is a probability of transitioning between states for different parental genomes at adjacent sites that is proportional to the genetic distance between those sites, but there is a larger probability of remaining in the state for the same parent. The highest probability path through the states is determined using Viterbi decoding, resulting in a predicted allele at each variant site in a given segregant.

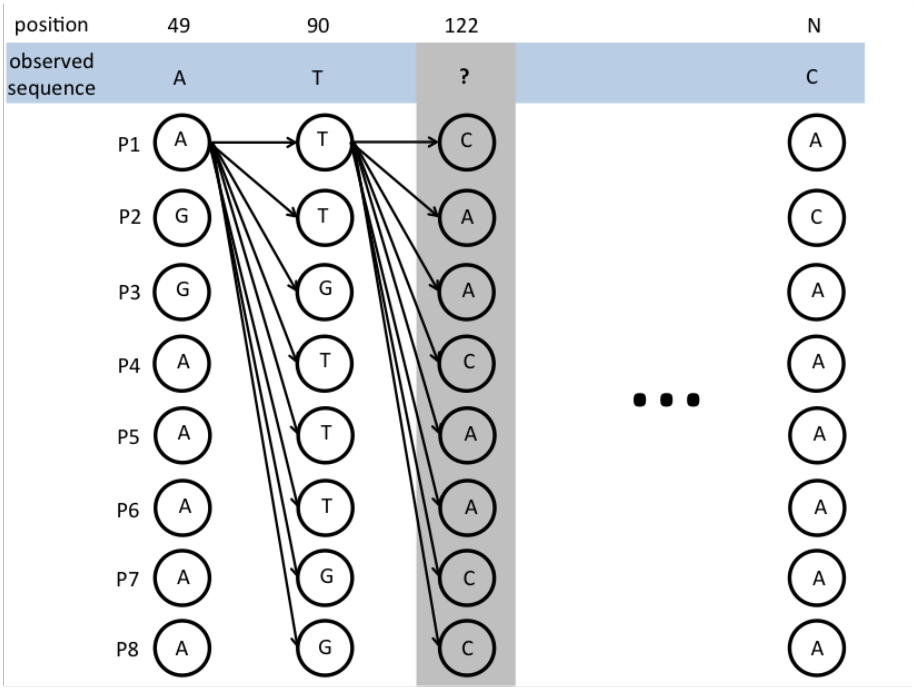

Figure 3.5: Hidden Markov model for imputing missing segregant sequence data. There are eight possible states at each site, representing each parental strain. Each state at site i can transition to any state at site $i+1$. For simplicity, only transitions from the P1 states are shown with arrows in this diagram.

To assess the accuracy of this model, we sampled 140 bp segments of simulated segregant genomes, geometrically distributed with an average spacing of 900 bp. We then imputed alleles at the missing variant sites using the described hidden Markov model and determined the overall accuracy of the imputed bases.

3.5 Discussion

In this chapter, we have outlined several aspects of the design of an eight–parent funnel cross to be used for QTL mapping in *S. cerevisiae*. Members of the Dudley lab have carried out this cross in the laboratory, with 12,000 segregants now generated. These segregants will be

phenotyped under a variety of conditions, including various sugar sources and small molecules studied in previous crosses,[58,101] as well as different categories of antifungal drugs. After the segregants are phenotyped and sequenced, traits will be mapped using the LOD score approach discussed in the Methods. A 2-dimensional scan for pairwise interactions between loci will also be conducted. We expect to be able map QTLs contributing additively to traits with low heritability with single–gene precision in some cases, though our simulations likely underestimate the variation in mapping resolution we will obtain for different traits.

Near the completion of this cross, two preprints were recently published describing multiparent crosses in *S. cerevisiae* on a similar scale.[92,102] In one,[92] sixteen diverse individuals were selected from the 1,011 *S. cerevisiae* strain collection; this subset of strains captures 82% of common diallelic polymorphisms present in the whole set of strains. 13,950 segregants were generated using a round–robin cross design and were phenotyped for 38 fitness traits. Across all traits, the authors mapped 4,552 QTLs (at an FDR of 5%) and were able to resolve approximately 9% of these QTLs to individual genes (at an FDR of 20%). In the other recent study,[102] 55 strains from the 1,011 strains collection were crossed in pairs to form 2,970 different heterozygous hybrids and 55 homozygous diploids, which were phenotyped for 49 growth traits. Through a GWA study on this panel, the authors were able to locate 1,723 SNPs significantly associated with the measured traits. Both of these studies reported a disproportionate impact of rare variants on phenotypic variance.

It will be interesting to see how the mapping ability of our cross compares to that of these other large–scale crosses. In particular, the sixteen–parent round–robin cross incorporates more of the variation within *S. cerevisiae* in total than our eight–parent funnel cross does, but less within each recombinant segregant. Overall, the availability of such diverse strains, combined with progress in experimental techniques for crossing and sequencing strains, has ushered in an era in which QTL mapping in *S. cerevisiae* may be more fruitful than ever. It remains to be seen which specific experimental designs will allow us to successfully dissect complex traits with a variety of genetic architectures.

Chapter 4: THE GENOMIC LANDSCAPE OF *S. PARADOXUS* INTROGRESSION IN GEOGRAPHICALLY DIVERSE *S. CEREVISIAE* STRAINS

4.1 Abstract

Natural hybrids between many pairs of the *Saccharomyces* species have been observed. Hybridization can allow for rapid adaptation to new environments, and when followed by repeated backcrossing can lead to small pieces of introgressed sequence from an individual of one species remaining in the genome of an individual of another species. Introgressed sequences that persist over time are interesting both because they provide evidence of past hybridization events and because they may contribute to functional and phenotypic diversity. Previous studies have identified some examples of introgressed sequences in *S. cerevisiae*, but we sought to gain a more comprehensive understanding of the landscape of introgression in this species. We developed a simple, flexible approach based on a hidden Markov model to identify introgression in yeast genomes. We used our method to look for introgression in 93 diverse *S. cerevisiae* genomes from its closest relative, *S. paradoxus*. We found evidence of introgression in all strains we considered, but the amount and location of introgression varied widely. We show that introgression contributes substantially to the total genetic diversity within these strains, and has the potential to confound inferences of their evolutionary relationships. Further characterizing introgression across diverse *Saccharomyces* species may help us better understand their evolution and the role that hybridization has played in adaptation to new environments.

4.2 Introduction

Since the original formulation of the theory of evolution, it has been apparent that the evolution of different species cannot be understood in isolation. Charles Darwin wrote of

the "complex relations of all animals and plants to each other in the struggle for existence," drawing on many examples of the evolution of one species being influenced by another.[103] As species compete or cooperate in shared environments, their interactions influence the composition of the genomes that evolve. But species can also affect each other's genomes more directly via hybridization or horizontal gene transfer. The ongoing improvement of sequencing technology has enabled more detailed studies of the movement of genetic material between species.[104]

The acquisition of genetic material from another species or strain can allow for more rapid adaptation to new environments by taking advantage of existing evolutionary innovation. For example, horizontally transferred genes have allowed for the evolution of multiple-drug-resistant *Staphylococcus aureus*.[105] In sexual organisms, hybridization can also facilitate adaptation, and hybrids are frequently useful in agricultural and industrial applications.[104,106,107] When hybridization is followed by repeated backcrossing, it can result in genomes that are mainly one species with interspersed introgressed sequences remaining from the other species. Introgressed sequences that originated from relatively old hybridizations but are still present in modern genomes may have remained because they are adaptive.[104,108]

The *Saccharomyces* yeast species were originally defined by the relatively low spore viability between them. However, this viability is not zero,[107] and the diversity of naturally-occurring hybrids that has been observed suggests that they hybridize relatively frequently.[68] One of the first such hybrids was found in the common brewing strain *S. carlsbergensis* (now *S. pastorianus*), and more extensive sequencing has since revealed additional hybrids within the genus, as well as a number of shorter introgressed sequences.[68] In particular, the sequencing of a great variety of strains of multiple *Saccharomyces* species has made it possible to consider global patterns of introgression among these organisms, and to examine how this introgression may have influenced their evolution.[109,110]

In addition to including the most well-characterized laboratory strain, *S. cerevisiae* has had the most strains sequenced out of all the *Saccharomyces* species. In particular, the 100-genomes strain collection provided 93 new *S. cerevisiae* genomes assembled de novo,

with quality approaching that of the S288c reference genome.[62] Hundreds of other wild, fermentation, and clinical strains have also been sequenced in the past few years,[14,61,111,112] though variants have most often been called using S288c as a reference.

S. paradoxus is on average ∼14% diverged from *S. cerevisiae* at the nucleotide level. Examples of *S. paradoxus* introgressions in *S. cerevisiae* have been identified that consist of very large segments[68] or individual genes.[62] These analyses suggest that introgression is widespread in *S. cerevisiae*. In addition, some introgressions have been found in *S. paradoxus* from *S. cerevisiae*, most notably a large subtelomeric segment on the left arm of chromosome XIV.[113] However, the methodological approaches used to identify introgression in yeast have relied on sequence identity—broadly, searching for regions of the genome that match a *S. paradoxus* reference better than a *S. cerevisiae* one. Although this pattern is a reasonable expectation for introgressed sequence, it may also be a consequence of incomplete lineage sorting (ILS),[104,114] which occurs when the divergence time between species is recent relative to the divergence time between individuals within the same species. In this scenario, lineages of different species can share a more recent common ancestor than lineages of the same species, resulting in patterns of sequence similarity that can look like introgression. To determine whether ILS was likely to confound our predictions of introgression, we first used coalescent theory to establish that relying on sequence identity is appropriate given the evolutionary history of these organisms.

We then developed a hidden Markov model (HMM)–based method to predict the species of origin of each site in a yeast genome. This approach allowed us to infer the boundaries of introgressed regions more precisely, instead of focusing only on individual genes. We evaluated the performance of our method on simulated sequence data, and found it to be faster, more accurate, and simpler to parameterize than an existing phylo–HMM–based method for identifying introgression.

We used our HMM–based approach to search for *S. paradoxus* introgression in the set of 93 diverse *S. cerevisiae* strains sequenced for the 100-genomes collection.[62] We found some level of introgression on nearly every chromosome of every strain considered, though strains

vary substantially in the amount of introgression that contain. Overall, the introgressions we predict represent a substantial contribution to genetic diversity, accounting for approximately 7% of the total nucleotide diversity among the strains. In some cases, introgressed sequences can influence our inference of the phylogenetic relationship of the strains. We identify genes previously suggested to be introgressed but also many novel introgressions. In general, this method has allowed us to characterize the global distribution of *S. paradoxus* introgression across a set of diverse *S. cerevisiae* genomes, and may allow us to further expand our understanding of introgression within the *Saccharomyces* yeasts.

4.3 Results

4.3.1 Establishing the validity of sequence identity–based approaches for identifying introgression

Incomplete lineage sorting (ILS) can occur when lineages within a species coalesce farther back in time than when that species diverged from another species.[115] When such ancient coalescences exist, individuals from different species may look more similar to each other than to other members of their own species (Fig. 4.1a), potentially resulting in false positives when using sequence identity to infer the presence of introgression. But ILS is only likely to occur when the time of divergence is recent relative to the effective population sizes of the species.[116] Looking backward in time, when the species are more diverged there is more time for within-species coalescences to finish before the species join, so ILS is less likely (Fig. 4.1b); conversely, when population sizes are larger the coalescences finish less quickly, so ILS is more likely. Given the divergence time and effective population sizes of two species, we can analytically calculate the probability of ILS for these species using equations based on coalescent theory.[117]

Although the divergence time and effective population sizes for *S. cerevisiae* and *S. paradoxus* are not known with certainty, we find that the probability of ILS is essentially very small for much of the parameter space that is plausible based on empirical data (Fig. 4.1c).

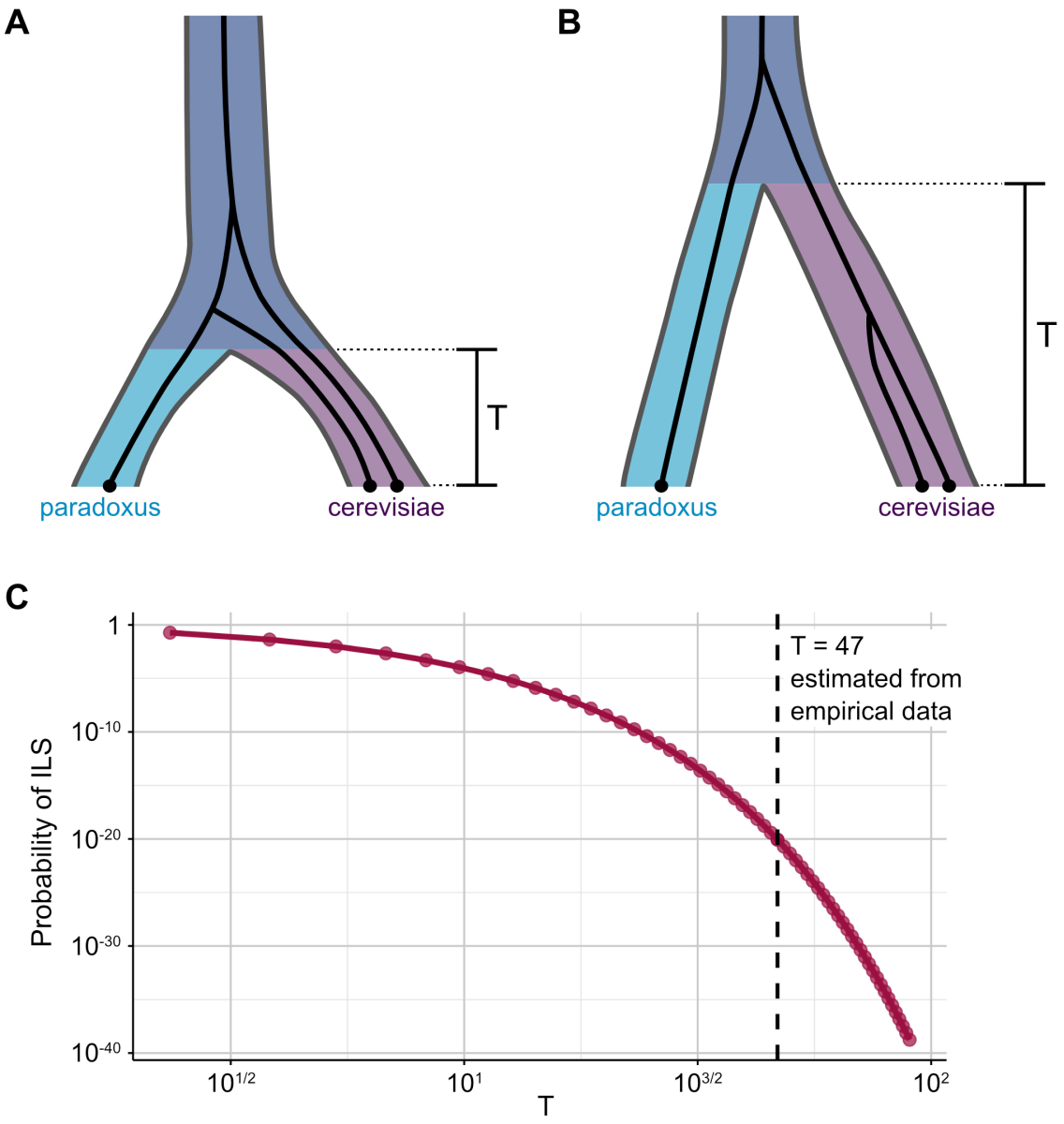

Figure 4.1: The probability of ILS in *S. cerevisiae* and *S. paradoxus*. (A) ILS can occur when the divergence time between species is short relative to their effective population sizes. (B) A longer divergence time (or a smaller population size) makes it more likely that coalescences occur within species than between them. (C) We calculate the probability of ILS for a range of values of the scaled divergence time—which is directly proportional to the actual divergence time, but inversely proportional to the effective population size. For our best estimate of $T = 47$, the probability of ILS is very small ($\sim 10^{-20}$), but for a value of T an order of magnitude smaller, the probability of ILS is substantial ($\sim 10^{-2}$).

For example, assuming an effective population size of 8×10^6 for each species[118] and a divergence time of 3.75×10^8 generations (see Methods for the rationale behind these values), the probability of ILS is $\sim 10^{-20}$. But if we allow for uncertainty in our parameter estimates by shortening the divergence time or increasing the effective population size, the probability of ILS increases. In general, however, the probability of ILS is small over the most likely range of the parameter space. Thus, these data show that ILS is unlikely to significantly confound detection of introgressed *S. paradoxus* sequences in *S. cerevisiae* and justify the use of sequence based approaches.

4.3.2 Developing and evaluating a hidden Markov model to detect introgressed sequences

To detect introgression, we implemented a hidden Markov model (HMM) that has one state for each reference species (*S. cerevisiae* and *S. paradoxus*) in addition to one unknown state (Fig. 4.2a). The model is used to assign a state to each alignment column in a given three-way alignment. Because the genomes of *S. cerevisiae* and *S. paradoxus* are almost entirely collinear,[119] it is possible to align and analyze entire chromosomes. Only alignment columns that are polymorphic and contain no gaps are considered.

In order to rigorously evaluate our method, we tested its performance on sequences with introgression generated using the coalescent simulator `ms`[120] with a variety of migration rates. We evaluated the true and false positives rates of our method (Fig. 4.2b), as well as those of an existing method for detecting introgression, PhyloNet–HMM.[72] Instead of simply inferring the species of origin of individual alignment columns, PhyloNet–HMM infers the underlying gene tree and species tree, which results in six possible states at each site compared to our three. As a result of this reduced complexity, our method requires fewer parameter estimates and also runs ∼40x faster on the simulated sequence data.

Both methods were run on sets of sequences generated using a range of migration rates. The performance of both is highly dependent on the specific parameters used to simulate the sequences; in particular, for higher migration rates, the level of introgression in the reference sequence makes it difficult for either method to detect introgression. For all sets of sequences

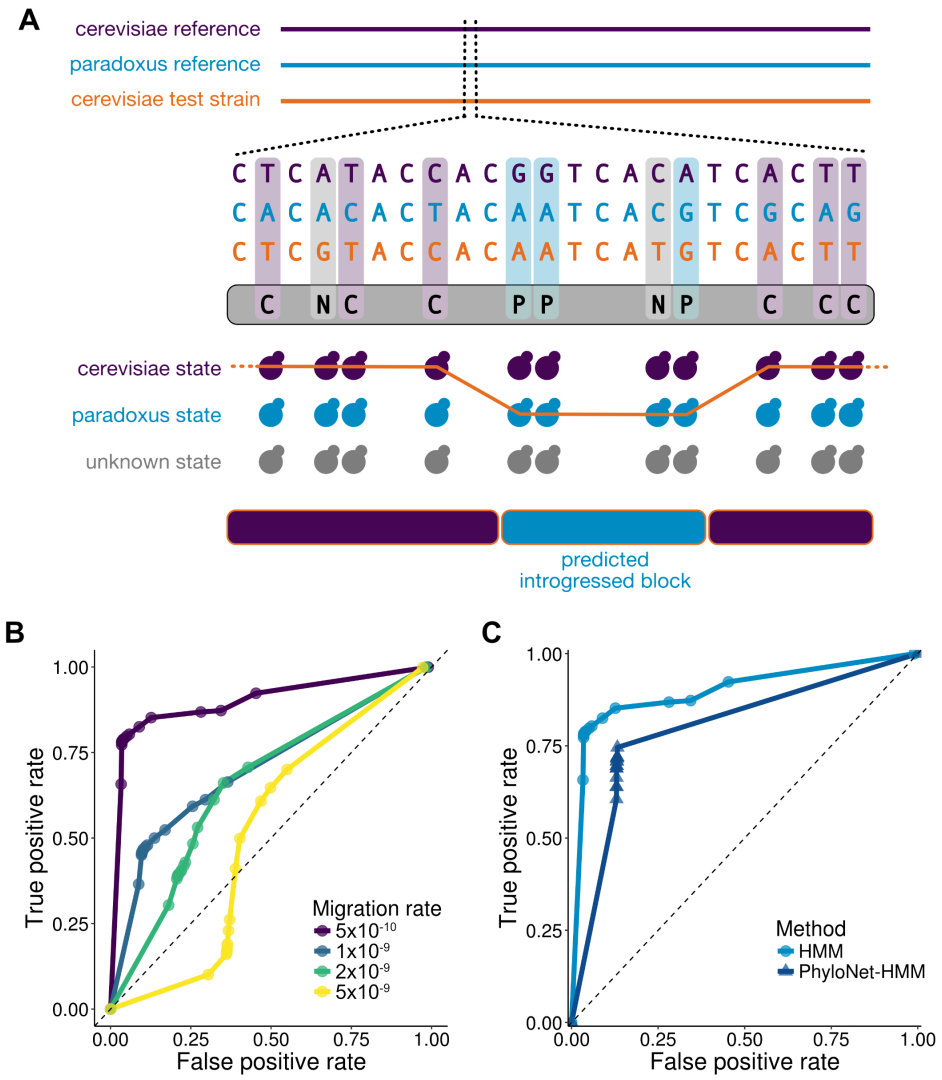

Figure 4.2: Structure and performance of our HMM for identifying introgression. (A) The chromosome of the current test strain is aligned to the corresponding chromosomes of the reference strains, and polymorphic alignment columns are coded by whether the test strain matches each of the references. The HMM is then used to assign a state to each of these alignment columns, with the paradoxus state indicating introgression. Consecutive occurrences of each state are grouped into regions. (B) True positive versus false positive rate for our method for a variety of migration rates, where the units are the fraction of the *S. cerevisiae* population made up of *S. paradoxus* individuals in each generation. Our method performs worse for higher migration rates, primarily because the high level of introgression in the reference cerevisiae sequence makes it difficult to identify introgression in the test sequence. (C) Comparison of performance of our HMM method and PhyloNet–HMM for a migration rate of 5×10^{-10}. Our method outperformed PhyloNet–HMM for all migration rates tested, as shown in Fig. A.1.

analyzed, however, our method outperformed PhyloNet–HMM (Fig. 4.2c).

Landscape of S. paradoxus–like introgressed sequences in S. cerevisiae strains and comparison to previously–identified introgressions

We analyzed the genomes of the 93 strains that were newly sequenced for the 100–genomes collection.[62] Each chromosome of each strain was aligned to *S. cerevisiae* S288c and *S. paradoxus* CBS432 using MAFFT.[121] Using our hidden Markov model–based approach, we predict introgressions on every chromosome (Fig. 4.3a). In total, we identify 3,147 introgressed regions across the 93 strains, ranging in size from 27 to 35,072 bp (median = 149 bp, mean = 1,071 bp; Table A.1). These introgressions collectively cover approximately 7.7% of the genome.

As expected, the introgressions we identify generally match the *S. paradoxus* reference more closely than the *S. cerevisiae* one, with the greatest concentration falling near 100% identity with *S. paradoxus* and 85-90% identity with *S. cerevisiae* (Fig. A.2). However, some regions we identify do not match the *S. paradoxus* reference well. For example, approximately 30% of regions have a sequence identity of less than 90% to the *S. paradoxus* reference, but because these tend to be shorter regions, they only account for approximately 6% of introgressed bases we identify. These regions may be a better match to another *S. paradoxus* strain or another species—or may just be in poorly–aligned regions.

Previously, 287 genes were identified as putatively introgressed from S. *paradoxus* in these 93 strains;[62] these genes were found by individually comparing the sequence identity with *S. cerevisiae* and *S. paradoxus* references across the entire gene to predefined thresholds. We identify nearly all of these genes, as well as an additional 201 genes (Fig. 4.3b). The five genes we fail to identify are due to translocations that we do not account for in our alignments, while the additional 201 genes we identify are mainly due to our ability to identify smaller parts of genes that appear to be introgressed (Fig. 4.3c).

We were concerned that some of the genes we identified might have been subject to paralogous gene conversion rather than introgression. To evaluate whether this was the case,

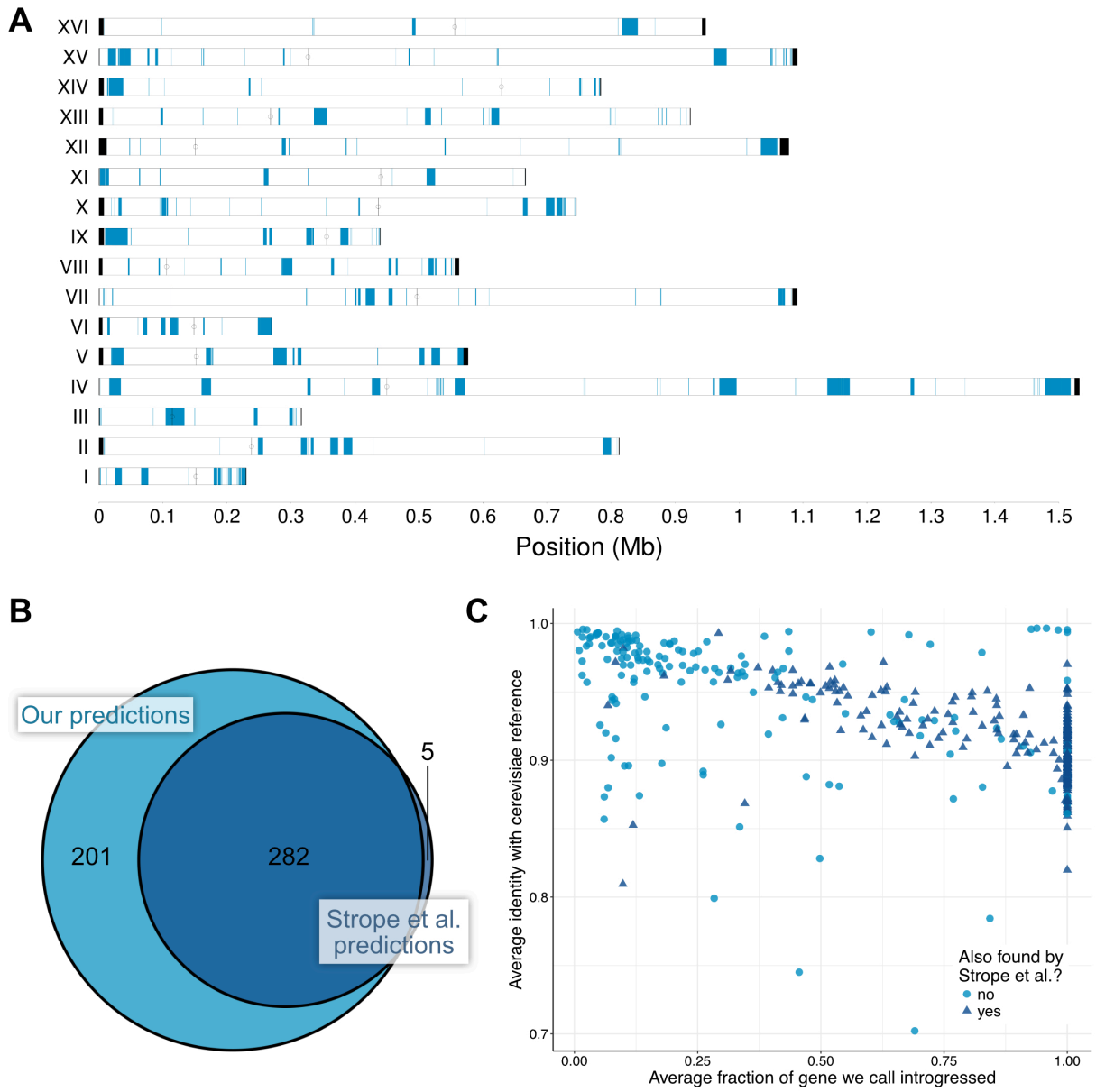

Figure 4.3: The introgressed regions and genes we identify, compared to previously identified genes. (A) The distribution of introgression across the entire genome. Regions that we identify as introgressed in one or more strains are colored blue. Telomeres are colored black, and centromeres are marked with a black line and circle. (B) The introgressed regions we identify overlap a total of 483 genes, 282 of which were previously identified as introgressed.[62] We fail to find five previously-identified genes because of translocations. (C) The genes we newly identified as introgressed tend to have only a fraction of their sequence introgressed, resulting in greater identity with the *S. cerevisiae* reference across the whole gene sequence.

we examined the 104 genes we identified that had a paralog. We used BLAST[122] to compare the introgressed portion of each gene to (1) the gene from the *S. cerevisiae* reference, (2) the gene from the *S. paradoxus* reference, (3) the paralog from the *S. cerevisiae* reference, and (4) the paralog from the *S. paradoxus* reference. We then examined the genes for which the best hit was in the *S. cerevisiae* paralog (rather than the *S. paradoxus* gene, as we would expect for introgression). In the case of four genes (RPL8B, SSB2, SSF1, and SSF2), paralogous gene conversion appeared to be a more likely explanation than introgression. Thus, this phenomenon does appear to result in some false positives in our predictions, but is not likely to be responsible for a substantial fraction of the paralogs we identify as introgressed.

4.3.3 Introgression in specific strains

The amount of the genome we predict to be introgressed ranges from 0.03% to 4.4% across all the strains (Fig. 4.4a). As noted previously,[62] strains YJM1252, YJM248, and YJM1078 appear to have far more introgression than the other strains. Much of this introgression is shared (Fig. 4.4b) and is comprised of longer regions (Fig. 4.4c), suggesting that an ancestor of these three strains hybridized with *S. paradoxus* relatively recently.

Most of the introgressed genes we identify are introgressed in only a few strains, but some are more widespread (Fig. 4.5). The most frequent number of strains for an introgressed gene is three because of the large amount of introgression shared between the three strains mentioned above; however, there are also many examples of gene introgressions shared among other sets of strains. For example, the same small part of the gene SIR4—which is involved in the assembly of silent chromatin domains—is introgressed in YJM1199, YJM1202, and YJM1304. A 603 bp introgressed region of this gene has introduced 142 polymorphisms and one 3 bp insertion into these three strains, resulting in 82 amino acid changes and one insertion.

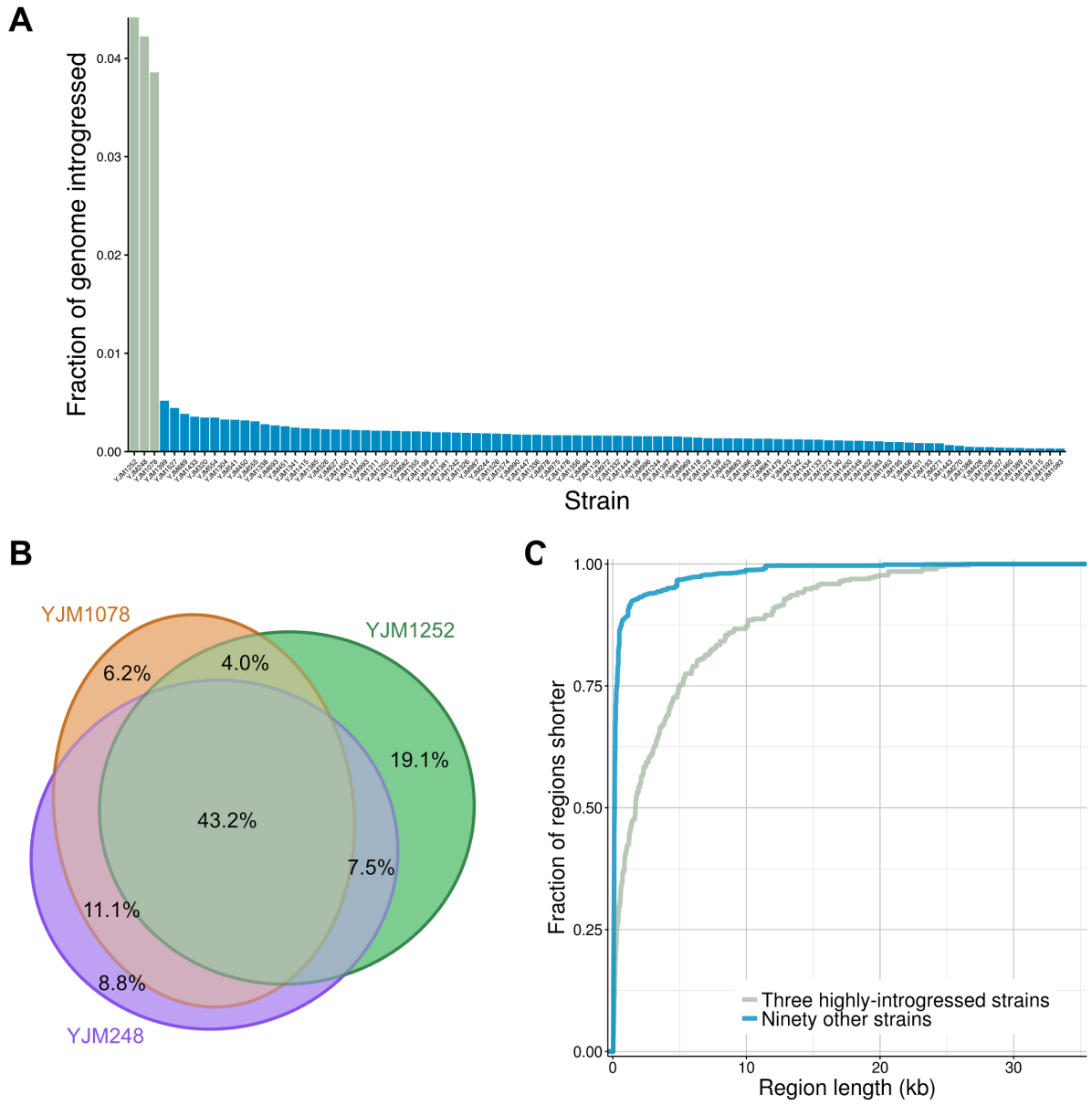

Figure 4.4: Introgression across all strains, and in three highly–introgressed strains. (A) We predict less than half a percent of most genomes to be introgressed from *S. paradoxus*, but three strains have a much higher level of introgression. (B) Much of this introgression is shared among all three strains, and about two-thirds is shared between at least two of the strains. (C) Most of the introgressed regions we find are short (less than 150 bp), but these three highly-introgressed strains have many longer introgressed regions, including a quarter of the regions that are greater than 5 kb in length.

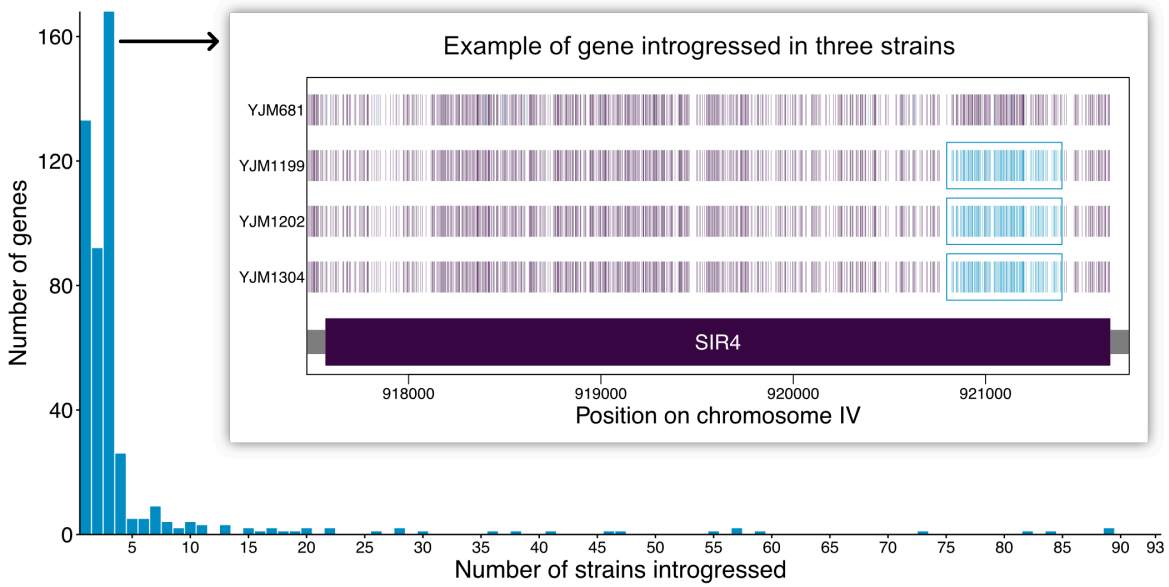

Figure 4.5: The number of strains each introgressed gene is found to be introgressed in. Inset shows an example of one gene, SIR4, that is partially introgressed in three strains. A strain without introgression in this gene is shown for comparison. For each strain, vertical lines indicate alignment columns at which the strain matches the *S. cerevisiae* reference but not the S. paradoxus reference (purple), the *S. paradoxus* reference but not the *S. cerevisiae* reference (blue), or neither reference (gray). Alignment columns containing gaps or ambiguous sequence calls are not shown.

4.3.4 Impact of introgression on global polymorphism and phylogenetic inference

Although we only predict a small fraction of most *S. cerevisiae* genomes to be introgressed from *S. paradoxus*, the level of divergence between the two species suggests that this small amount of sequence could account for a relatively large amount of the nucleotide diversity present in *S. cerevisiae*. We calculated nucleotide diversity among the 93 strains across the entire genome to be 0.81%. We separately calculated nucleotide diversity excluding sites predicted to be introgressed to be 0.76%, indicating that 6.6% of the nucleotide diversity is contributed by these putatively introgressed regions (Figure 6a). Repeating the same analysis only in coding regions, we find that slightly less (5.5%) of the nucleotide diversity is contributed by the introgressed regions, consistent with the hypothesis of purifying selection acting more strongly on protein versus non-coding regions. Furthermore, there is marked heterogeneity in the contribution of introgressed sequences to nucleotide diversity across chromosomes (Fig. 4.6a).

Given the relatively large effect of these introgressed sequences on nucleotide diversity, we next tested whether failing to account for introgression would influence inferences of the evolutionary history of these strains. We found that excluding introgressed regions resulted in a substantially different phylogeny than one constructed using all sites (Fig. 4.6b). Unsurprisingly, removing introgressed sites moves the three highly–introgressed strains YJM1252, YJM248, and YJM1078 to be more closely related to other S. cerevisiae strains, but several other strains also move to different clades in the phylogeny for less obvious reasons.

We were also interested in whether introgressions cluster together on a phylogeny of these species, which would indicate they may have come from common hybridization events. We looked at all genes that overlapped an introgressed region in at least one strain, and clustered the genes based on their phylogenetic distribution, using a tree constructed from sites without introgression (Fig. A.3). If most introgressions are due to just a few hybridization events, we would expect to see distinct clusters falling within clades. There are some examples of such patterns, and further analysis of specific shared introgressions may allow for the inference of

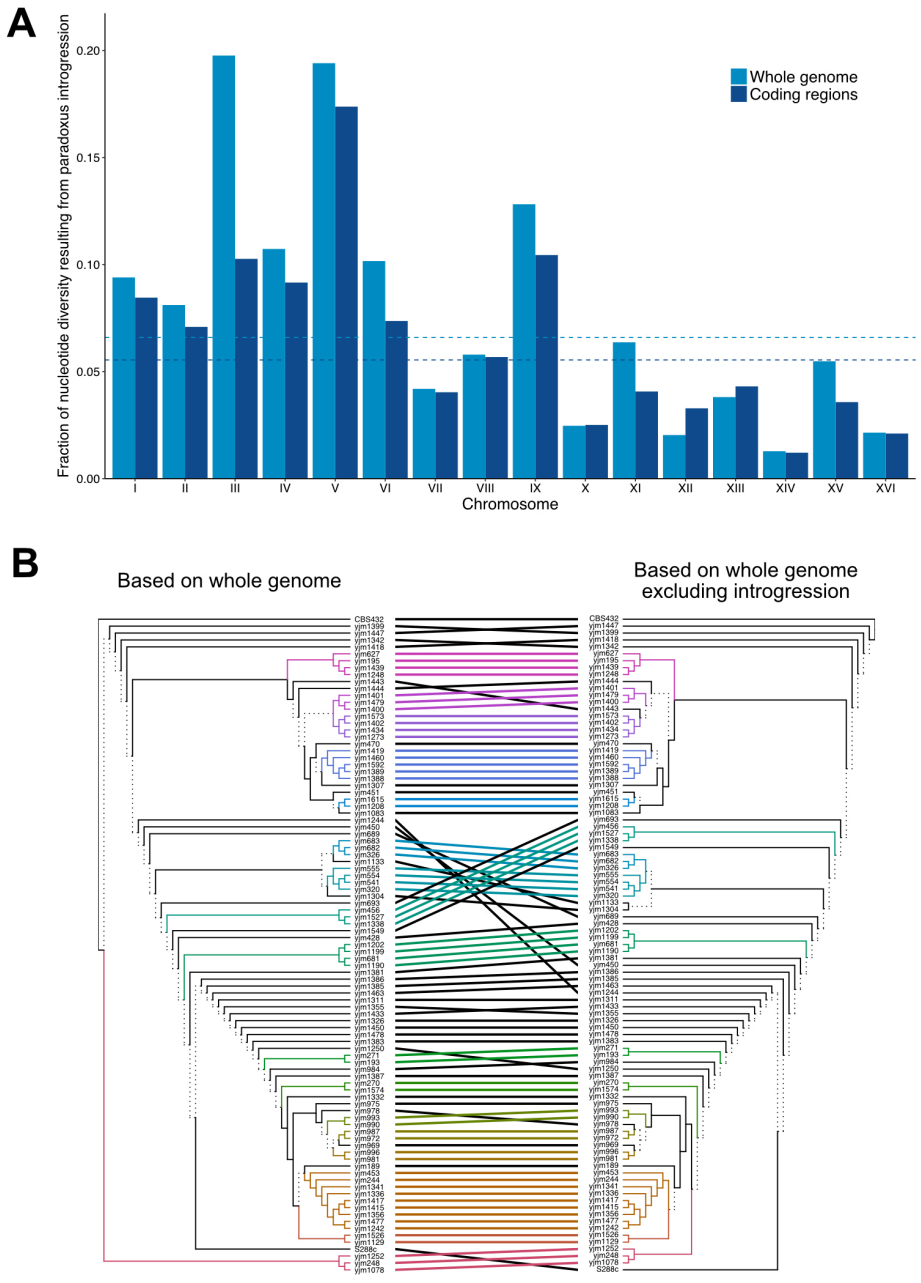

Figure 4.6: The impact of introgression on overall diversity and phylogenetic relationships within *S. cerevisiae*. (A) The fraction of nucleotide diversity contributed by introgressed regions on each chromosome, for all sites and all coding sites. Genome-wide averages are shown with dotted lines. (B) Tanglegram of phylogenies constructed from sites across the entire genome and from only non-introgressed regions. The whole-genome phylogeny is a consensus tree of 50 jackknife replicates generated by removing a number of random sites equal to the number of introgressed sites removed for the non-introgressed tree.

when hybridizations are likely to have occurred.

4.4 Methods

4.4.1 Analytical theory to estimate probability of incomplete lineage sorting

The probability of incomplete lineage sorting (ILS) is given by one minus the probability that a given gene tree and species tree are concordant. The formula for this probability can be derived using coalescent theory, and depends on the number of samples for each species and the scaled divergence time between the species.[117] The scaled divergence time is given by $T = t/N$, where t is the divergence time in generations, and N is the effective population size of each species. We are specifically interested in monophyletic concordance, the scenario in which both the gene tree and species tree are monophyletic. The probability of monophyletic concordance can be calculated as

$$P(r,s,t) = \sum_{m=1}^{r}\sum_{n=1}^{s} g_{rm}(T)g_{sn}(T) \times [1 - F_2^{A,B}(m,n)],$$

where $g_{ij}(T)$ is the probability that i lineages derive from j lineages that existed T coalescent time units in the past, and is given by

$$g_{ij}(T) = \sum_{k=j}^{i} e^{-k(k-1)T/2} \frac{(2k-1)(-1)^{k-j} j_{(k-1)} i_{[k]}}{j!(k-j)! i_{(k}}.$$

The quantity $F_k^{A,B}(a,b)$ is the probability that an interspecific coalescence occurs in coalescing a and b lineages from species A and B respectively to k total lineages:

$$F_k^{A,B}(a,b) = \frac{ab}{\binom{a+b}{2}} + F_k^{A,B}(a-1,b)\frac{\binom{a}{2}}{\binom{a+b}{2}} + F_k^{A,B}(a,b-1)\frac{\binom{b}{2}}{\binom{a+b}{2}}$$

To estimate the probability of ILS when looking at *S. paradoxus* and *S. cerevisiae* sequences, we assume one individual from the former species and 94 individuals from the latter, to correspond to the numbers of genomes we examine later. We estimate the time, T, from coalescent simulations (described in more detail in Simulated Sequences) that produce the

expected approximate sequence identity between the species of 86%. These simulations require as parameters the effective population size and mutation rate. The effective population size of *S. paradoxus* has been estimated as 8.6×10^6 for the European clade and 7.2×10^6 for the Far Eastern clade, and these estimates are in turn based on estimates of the mutation rate and inbreeding coefficient.[118] We set the effective population sizes for *S. cerevisiae* and *S. paradoxus* as 8×10^6 individuals each. Using a mutation rate of 1.84×10^{-10},[33] we estimated the time of divergence as 3.75×10^8 generations.

4.4.2 Hidden Markov model

The hidden Markov model has one state for each reference species as well as one unknown state. The model was used to assign a state to individual alignment columns in each three-way whole chromosome alignment. Only alignment columns that were polymorphic and contained no gaps were considered. The initial, emission, and transition probabilities were set by a combination of (1) initial estimates of the number and sizes of introgressed regions, and (2) the number of alignment columns in which the test strain matched each of the reference strains. These probabilities were updated by Baum-Welch training until the log likelihood decreased by less than 0.1% between iterations.

In analyzing simulated sequences, alignment columns with a high posterior probability (above a given threshold) of being in the *S. paradoxus* (introgressed) or unknown state were assigned those respective states, and were assigned the *S. cerevisiae* (non-introgressed) state otherwise. In analyzing actual chromosomes, the Viterbi algorithm was used to find the most likely state sequence. In both cases, introgressed regions were defined as continuous blocks of alignment columns assigned to the introgressed state, including any intervening alignment columns that were monomorphic or contained gaps. All regions are indexed relative to the *S. cerevisiae* reference genome.

4.4.3 Simulated Sequences

Sequences were simulated using `ms`.[120] Two samples were simulated for *S. cerevisiae*, one for *S. paradoxus*, and one for an outgroup. The effective population size was set to 8×10^6 for all species. This value was previously estimated for the European clade of *S. paradoxus* based on calculating variation at neutrally evolving regions in the genome;[118] for simplicity and for lack of a reliable estimate of the effective population size of *S. cerevisiae*, we set all effective population sizes to this same value.

Migration rate, expressed as the fraction of the *S. cerevisiae* population made up of *S. paradoxus* individuals in each generation, was set to either zero or a value ranging from 5×10^{-10} to 5×10^{-8}, and was set to only occur in the most recent 5% of time since the species diverged. The recombination rate was set to 7.425×10^{-6}, which was calculated from the formula $1 + 6.1$ crossovers/Mb,[98] using the average *S. cerevisiae* chromosome size. The mutation parameter, $\Theta = 4N_0\mu$, was set using a mutation rate of $\mu = 1.84 \times 10^{-10}$.[33] Simulated sequences were 100,000 bp in length. An example `ms` command with a migration rate of 5×10^{-10} is:

```
ms 4 100 -t 5.888 -r 0.09494496 100000 -p 8 -I 3 2 1 1 -m 1 2 0.32 -m 1 3 0.0
-em 0.5859375 1 2 0 -em 11.71875 1 3 0 -ej 11.71875 2 1 -ej 35.15625 3 1 -T
```

4.4.4 PhyloNet–HMM

PhyloNet–HMM[72] was run with trees of three species (representing *S. cerevisiae*, *S. paradoxus*, and an outgroup), for a total of six possible states for each alignment column. Sequences were generated from the `ms` simulation output using `seq-gen` (https://github.com/rambaut/Seq-Gen/releases/tag/1.3.4).

4.4.5 Strain genomes, annotations, and alignments

The reference *S. cerevisiae* S288c and *S. paradoxus* CBS432 genomes were downloaded from SGD (https://downloads.yeastgenome.org/sequence/). The 93 non-reference strains were downloaded from GenBank (https://www.ncbi.nlm.nih.gov/genbank/) based on the table of accession numbers provided by Strope *et al.* 2015.[62] The lists of all verified ORFs and all paralogs in *S. cerevisiae* S288c were downloaded from SGD using YeastMine.

Three–way alignments between the *S. cerevisiae* and *S. paradoxus* reference genomes and each *S. cerevisiae* test strain were performed on each chromosome separately using `MAFFT`[121] with default settings. The two reference genomes were also aligned separately. Changing the `ep` parameter from 0.0 to 0.321 did not result in substantially shorter alignments.

4.4.6 Filtering

Low-complexity regions of individual genomes were masked using dustmasker from BLAST version 2.7.1 (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.7.1/). Introgressed regions predicted by the HMM were included in final analysis if and only if they met all of the following criteria:

1. the fraction of alignment columns containing at least one gap or at least one masked base did not exceed 0.5,

2. the number of sites at which the test strain matched the *S. paradoxus* reference but not the *S. cerevisiae* reference was at least seven, and

3. the divergence between the test strain and the *S. cerevisiae* reference (calculated on all alignment columns without gaps or masked bases) was less than 0.3.

4.4.7 Phylogenies

A FASTA file was generated with a position for every site in the reference *S. cerevisiae* S288c genome. The corresponding nucleotides for the other 93 *S. cerevisiae* strains were extracted

from the three-way alignments, and for the *S. paradoxus* CBS432 reference from a two-way alignment. All alignment columns with gaps were removed, leaving a total of 6,033,510 sites. Then, all 376,536 of these sites that overlapped an introgressed region in any strain were removed, leaving a total of 5,656,975 sites. A phylogeny was constructed from these non-introgressed, non-gapped sites.

Phylogenies were constructed using `PHYLIP` (http://evolution.genetics.washington.edu/phylip/) dnadist and neighbor, using the UPGMA algorithm. In addition, a phylogeny was constructed in the same way from a dataset generated by removing a random set of sites equal in size to the set of introgressed sites; this sampling was repeated 50 times using `PHYLIP seqboot`, and a consensus tree was generated using `PHYLIP consense`. The tanglegram was plotted using the R package `dendextend`.[123]

4.5 Discussion

We have implemented a simple, flexible, and powerful method for identifying introgression among the *Saccharomyces* yeasts. By applying this method to look for *S. paradoxus* introgression in a set of 93 geographically and ecologically diverse *S. cerevisiae* strains, we have found introgression to be pervasive but highly variable across strains.

Although we have identified some amount of introgression in all of these strains, it is important to recognize that our inferences are dependent on our choice of reference genomes. In particular, we are unlikely to identify introgressions in *S. cerevisiae* strains that are also present in the *S. cerevisiae* reference we are using. Using a different *S. cerevisiae* reference genome may allow us to identify additional introgressions; however, sequence identity–based approaches like ours will always struggle to identify introgressions that are relatively old or that fall within highly conserved regions of the genome. Conversely, our set of introgressed regions likely also contains some false positives. We have filtered out many of these, and the ones that remain are on the whole more likely to be due to poorly aligned regions than to incomplete lineage sorting or paralogous gene conversion.

Our analysis of introgression has so far only concerned the presence of *S. paradoxus* sequence in *S. cerevisiae* genomes, but we would expect to find evidence of introgression in the opposite direction, as well. Furthermore, there is a substantial amount of population structure within *S. paradoxus*, with divergences between its different populations ranging from 1-3%.[124] Our model can be easily extended to include additional *S. paradoxus* reference genomes or even references for other *Saccharomyces* species; such an expanded analysis could allow us to identify more introgression, and perhaps to gain a clearer picture of when and where hybridization occurs.

Although we have mainly focused in this study on general patterns of introgression among the strains we analyzed, our method may also allow for the identification of specific examples of adaptive introgression that could be further characterized experimentally. Overall, we hope this approach may yield further insights into the evolution of *Saccharomyces* yeasts, and specifically the role that hybridization has played and continues to play in their adaptation to new environments.

4.6 Data access

The implementation of our hidden Markov model and downstream analyses are available for download on GitHub (https://github.com/hyperboliccake/introgression).

Chapter 5: DIVERSITY WITHIN *S. CEREVISIAE* AND *S. PARADOXUS* REVEALS A MORE COMPLETE PICTURE OF INTROGRESSION

5.1 Abstract

The *Saccharomyces* yeasts are among the most genetically well–characterized model eukaryotes, but we understand relatively little about how their genomes have changed over time. Despite the substantial sequence divergence within the genus, these species hybridize in natural and industrial environments. Previous work has identified introgressions in some of these species remaining from past hybridization events with other species. We previously developed a simple, flexible approach based on a hidden Markov model to identify introgression in yeast genomes. Here we extend this approach to incorporate more of the diversity within *S. cerevisiae* and *S. paradoxus*, which allows us to partially account for introgression in the reference strains we use, and to make more specific inferences about the origins of introgression. Overall, taking advantage of the full range of known diversity in *S. cerevisiae* and *S. paradoxus* results in a clearer picture of the pervasive exchange of genetic material through hybridization between these species.

5.2 Introduction

With the recent sequencing of hundreds of budding yeast genomes from geographically and ecologically diverse environments,[14,61,62,111,112] there has been a growing interest in disentangling the complex evolutionary history of these organisms. The *Saccharomyces* yeasts are among the most genetically well–characterized model eukaryotes, but we understand relatively little about how their genomes have changed over time. Their ability to live in both haploid and diploid states—and to reproduce both asexually and sexually depending on environmental conditions—creates unusual evolutionary opportunities. Although the species

within this genus are highly diverse, their flexible life cycle allows them to more easily form natural hybrids that in some cases can produce viable spores.[68,125]

Ecological sampling has revealed some locations where *S. cerevisiae* and *S. paradoxus* coexist,[126] but we have little in the way of historical or fossil evidence to determine where these species have lived and encountered each other in the past. With a large set of diverse genomes, however, we can make inferences about how different genomic regions have changed and been exchanged between different species over time. Identifying introgressed regions remaining from past hybridization events can tell us about the evolutionary history of these species, and can also potentially suggest specific adaptions that helped them colonize new environments.[127]

Some examples of large introgressions have been identified in *S. cerevisiae* and *S. paradoxus* strains,[113,128,129] but most research on introgression in yeast so far has focused on individual genes or ORFs.[14,62,111] This gene–centric approach is particularly useful when considering the pangenome of a species because it can explicitly handle differing gene content among strains, rather than making inferences about large insertions or deletions. We intend, however, to gain a more comprehensive view of introgression across the entire genome without regard to gene boundaries. By doing so, we can establish more precise boundaries of introgressed regions, and can also identify genes that are only partially introgressed. We previously developed a simple hidden Markov model–based approach to look for introgression in yeast genomes, and applied it to identify introgression from *S. paradoxus* in a set of 93 diverse *S. cerevisiae* genomes from the 100–genomes collection.[62] We now extend this analysis to explicitly consider that introgression may originate from detectably different *S. paradoxus* strains, and to account for the presence of introgression in the *S. cerevisiae* reference we use.

There is substantially more population structure within *S. paradoxus* than within *S. cerevisiae*.[60] The known *S. paradoxus* strains fall into four main clades, with European and Far Eastern strains more closely related, and American and Hawaiian strains more closely related (Fig. 5.1a). By looking at introgression from these highly diverged *S. paradoxus* strains, we thought it might be possible to more accurately pinpoint which introgressions

were due to different hybridization events, as well as when those events occurred relative to each other.

Within the 100–genomes *S. cerevisiae* strains, pairwise divergence is typically <1%, but the sequencing of strains from Chinese forests has greatly increased the amount of known diversity in *S. cerevisiae*.[111] The Chinese strain we use in this study is approximately as diverged from the 100–genomes strains as European *S. paradoxus* is from Far Eastern *S. paradoxus*, or American is from Hawaiian (Fig. 5.1b). We expected that using a more diverged strain as a reference for *S. cerevisiae* in our model would allow us to identify more introgression in the 100–genomes strains, since the Chinese forest strain should share less introgression with those other strains.

We have found that using these diverse references for *S. cerevisiae* and *S. paradoxus* instead of relying on a single representative for each species does allow us to identify more introgression in *S. cerevisiae*. Most of the introgression we identify in the 100–genomes *S. cerevisiae* strains appears to originate from European *S. paradoxus*, but we also find evidence of introgression from all four major clades of *S. paradoxus*. Strains vary widely in the amount of introgression they share with other strains. As expected, the Chinese strain shares little of its introgression with other strains, while the lab strain shares a much larger proportion—making it possible to identify more introgression when using the Chinese strain as the reference for *S. cerevisiae*. We identify multiple clusters of introgressed sites that are present in distinct sets of strains, suggesting that multiple hybridization events have led to the patterns we see. Overall, taking advantage of the full range of known diversity in *S. cerevisiae* and *S. paradoxus* results in a clearer picture of the pervasive exchange of genetic material through hybridization between these species.

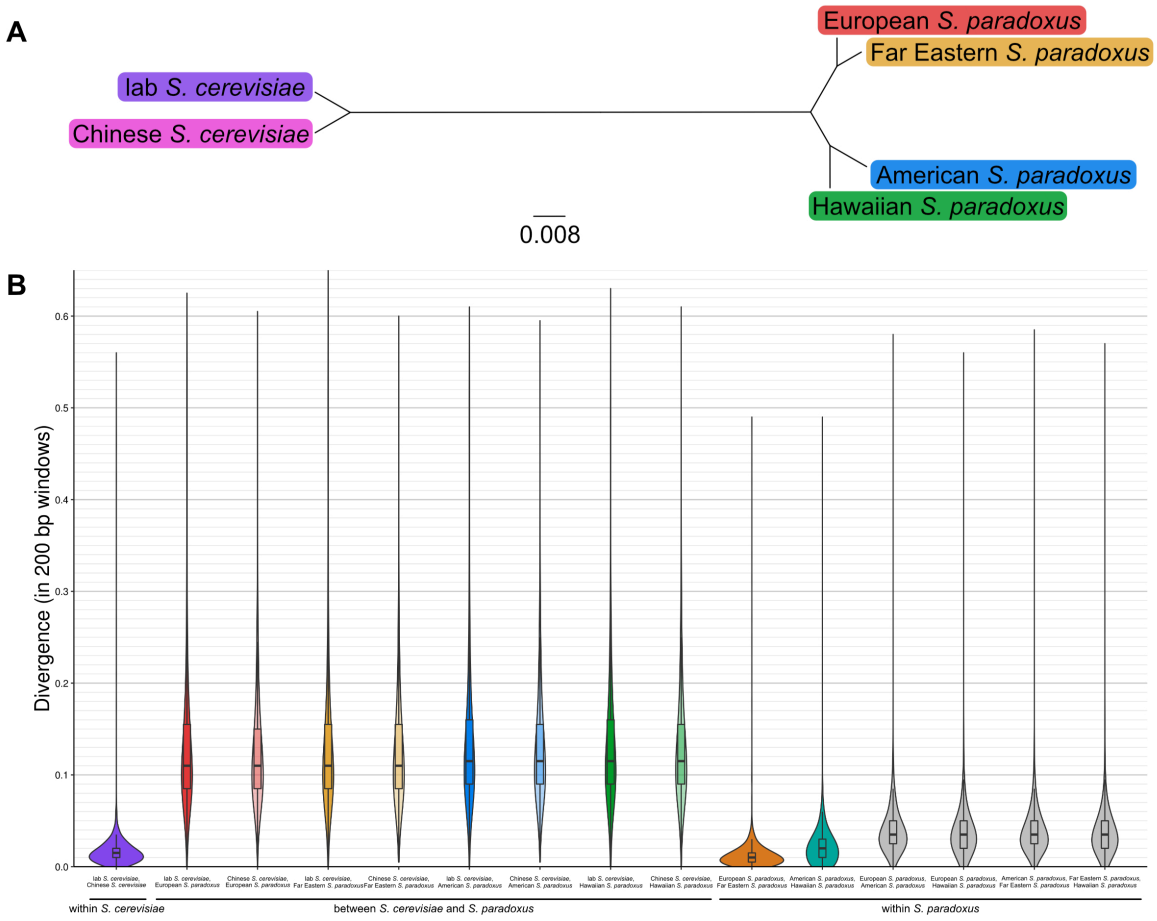

Figure 5.1: Divergence between *S. cerevisiae* and *S. paradoxus* reference strains used in this study. (A) Phylogeny of two *S. cerevisiae* reference strains (S288c, JXXY10.1) and four *S. paradoxus* reference strains (CBS432, N–45, DBVPG6304, UWOPS91–917.1). (B) Divergence between all pairs of reference strains in non–overlapping 200 bp windows across the genome.

5.3 Results

5.3.1 Divergence within S. paradoxus is great enough to distinguish between introgression originating from different populations

In order to distinguish between introgression in *S. cerevisiae* genomes that originates from different *S. paradoxus* populations, the *S. paradoxus* references we use for those populations need to be adequately diverged. Many of the introgressions we identified previously were relatively short, with a median length of 150 bp (see Chapter 4). If the divergence between *S. paradoxus* populations was relatively small, regions of this size might typically only vary by a few base pairs, which could make it impossible to confidently assign an introgressed region to a specific *S. paradoxus* origin.

We calculated divergence in 200 bp windows between all pairs of *S. cerevisiae* and *S. paradoxus* references used later in this study (Fig. 5.1b). The median window is 1% diverged between European and Far Eastern *S. paradoxus*, and about 2% between American and Hawaiian *S. paradoxus*. For shorter introgressed regions, this level of divergence would often not be large enough to confidently assign the region's origin to a specific population. But between the two larger clades of *S. paradoxus*, the median divergence is 3.5%, with approximately one–fifth of windows more than 5% diverged. Thus, it should be possible to more specifically classify some introgressions as originating from specific *S. paradoxus* populations, particularly if they are relatively recently introgressed.

Between *S. cerevisiae* and *S. paradoxus* the median window approximately 11% diverged, so in general it should at least be possible to detect the presence of introgression, if not to determine which specific population the introgression came from. The divergence between the two *S. cerevisiae* references used in this study is 1.5%, roughly equivalent to that between either the European and Far Eastern, or American and Hawaiian, *S. paradoxus* populations.

5.3.2 Extending hidden Markov model–based approach to incorporate multiple S. cerevisiae and S. paradoxus reference genomes

In order to incorporate multiple reference strains for *S. cerevisiae* and *S. paradoxus* in our analysis, we needed to extend the hidden Markov model (HMM)–based approach we previously developed (cite; Fig. 5.2 a). Our HMM can theoretically handle an arbitrary number of states, so it would be possible to simply include multiple *S. cerevisiae* and *S. paradoxus* references in the same model. For example, with two different *S. cerevisiae* references and four different *S. paradoxus* references, our model would have a total of seven states (including the unknown state), as opposed to the three states used in the previous analysis. However, this approach is problematic because by definition HMMs assign a single state to each alignment column, while logically there are some sites in the genome that should be assigned to multiple *S. cerevisiae* or *S. paradoxus* states because the test strain matches multiple references equally well. In practice, since the model must choose a single state, it assigns most of the introgressed sites to one of the *S. paradoxus* states, and assigns groups of sites to the other *S. paradoxus* states in an uninterpretable way. Incorporating multiple *S. cerevisiae* references results in similarly nonsensical predictions.

To resolve this issue, we could create one state for each possible combination of references, instead of only one state for each reference. However, with six references, this strategy would result in $2^6 = 64$ states. Such a model would take far longer to run and also be much more difficult to parameterize and train properly. In addition, comparing predictions obtained using different references would require running an entirely new, even larger model that incorporated all of the references at once. Thus, instead of introducing more states into our model, we ran the model separately for each pair of *S. cerevisiae* and *S. paradoxus* reference strains. We then combined the results from the eight different runs of the model in a principled way (Fig. 5.2b).

Specifically, after filtering out low–quality regions (as described in the Methods), we converted the predicted regions from all models into combined regions that were each predicted

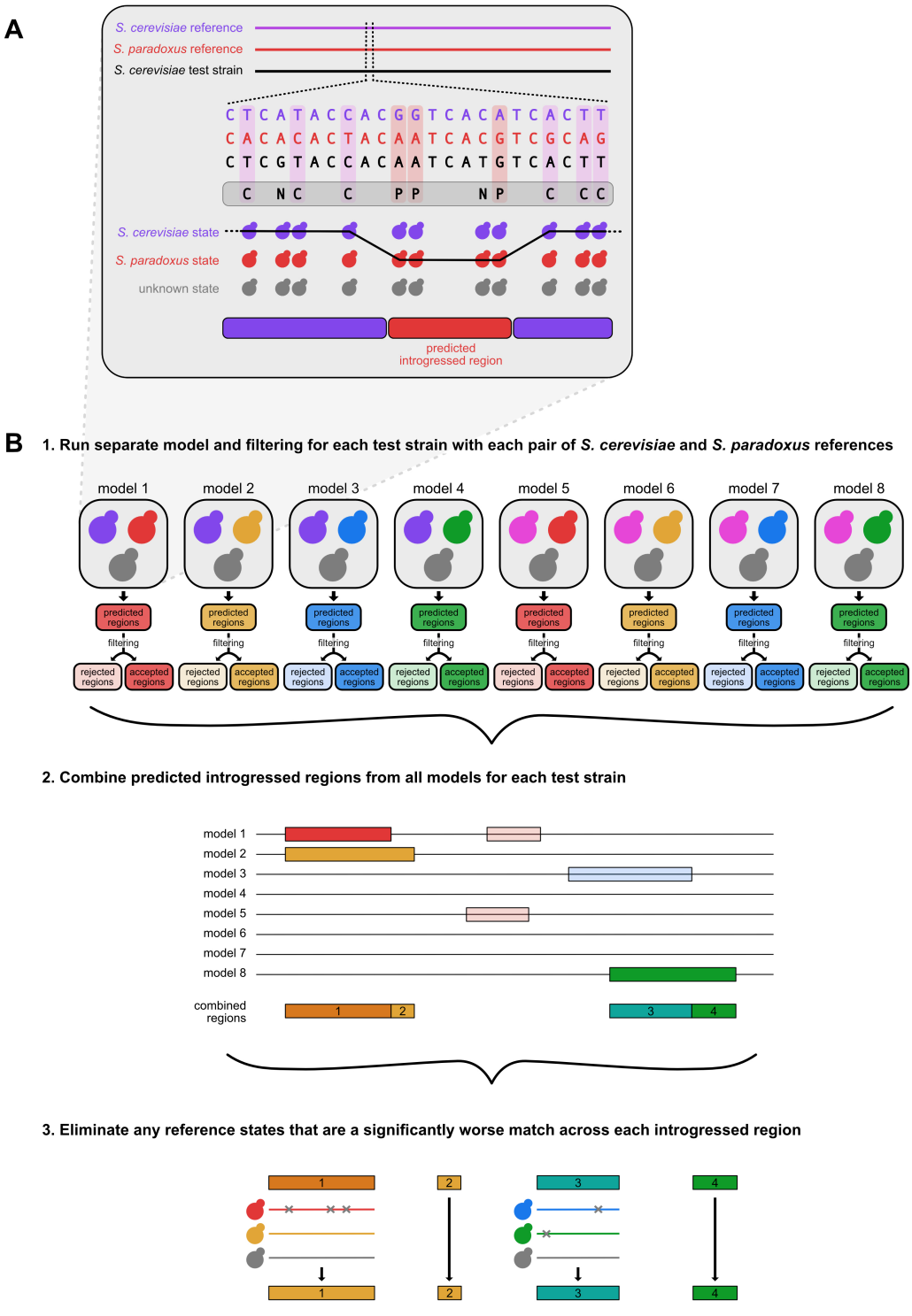

Figure 5.2: Figure legend continued on following page.

Figure 5.2: Extending HMM–based approach to utilize multiple references. (A) Outline of the method we described previously (see Chapter 4) for predicting introgression in a single test strain with one reference each for *S. cerevisiae* and *S. paradoxus*. (B) Outline of method for combining predicted introgressed regions from multiple runs of the model with different references. Regions predicted by each model are first filtered based on criteria including the identity of the test strain with both reference strains. The regions from different runs are then combined based on their overlap; in this step, regions rejected in filtering are added back into the analysis if they overlap a region from another model that passed filtering. After regions are combined, an attempt is made to increase the specificity of the reference(s) associated with each region. If one or more references are clearly better matches than the others across the length of a region, only those references are retained.

by the same set of models across the entire length. In combining these regions, we also considered the regions that failed the first filtering step, in order to avoid assigning a given region to one specific reference when it was on the border of being called by a model using another reference. We only added those filter–rejected regions back in if they overlapped at least one filter–accepted region from another model. After combining the regions, we attempted to assign them to *S. paradoxus* references with more specificity. To do so, we calculated the identity of the test strain with each predicted *S. paradoxus* reference strain across the region. If one or more references were significantly better matches than the others, we kept only those references. Thus, our final set of regions encompasses all sites collectively predicted by all of the contributing models, but with some predicted references eliminated as possibilities for some regions.

5.3.3 Chinese S. cerevisiae reference allows for the identification of more introgression than does lab reference

The reference–based approach to identifying introgression that we have taken cannot in general identify introgression that is shared with the reference strain. Therefore, in addition to using the well–annotated *S. cerevisiae* lab strain S288c as a reference, we also used the Chinese primeval forest strain JXXY10.1 from the CHN–IX population,[111] which comprises

the *S. cerevisiae* strains most diverged from the 100–genomes strains. We expected to find more introgression when using the Chinese reference than when using the lab reference, since the lab strain is more closely related to—and should thus share more introgression with—the 100–genomes strains. Overall, nearly half of all introgression we identified could be found using either of these references, but in almost all of strains, we found more introgression using the Chinese reference than using the lab reference (Fig. 5.3a).

We can also identify introgression in each of these *S. cerevisiae* reference strains by utilizing the other reference. The large majority of introgression we identified in the lab reference is shared with at least one other strain, whereas more than one–quarter of the introgression identified in the Chinese reference is unique among all strains we considered (Fig. 5.3b). Of course, even using the highly diverged Chinese reference when looking for introgression in other strains, we will still generally only be able to identify introgression that originated after the reference and test strains diverged.

5.3.4 S. cerevisiae strains differ in how much introgression they share with other strains

Because we have run our model separately with two different *S. cerevisiae* references, it is possible to identify introgression in each one of these strains while using the other as a reference. The Chinese *S. cerevisiae* reference strain has the greatest fraction of unique introgression, while the lab reference strain has a comparatively low fraction (Fig. 5.3b). The other test strains vary widely in the amount of introgression they share with other strains, as well as how many strains they share it with (Fig. 5.3c). Three strains (YJM248, YJM1078, YJM1252) uniquely share a large amount of introgression, as seen previously,[62] presumably due to a relatively recent hybridization event.

It is in general possible that a region we call introgressed in a set of strains is actually flagged due to the presence of introgression or some other type of sequence anomaly in the reference strain. This scenario is particularly likely to occur with introgressed regions that are shared among many strains. Although our filtering should remove many poorly aligned regions from the analysis, our current approach cannot always distinguish between

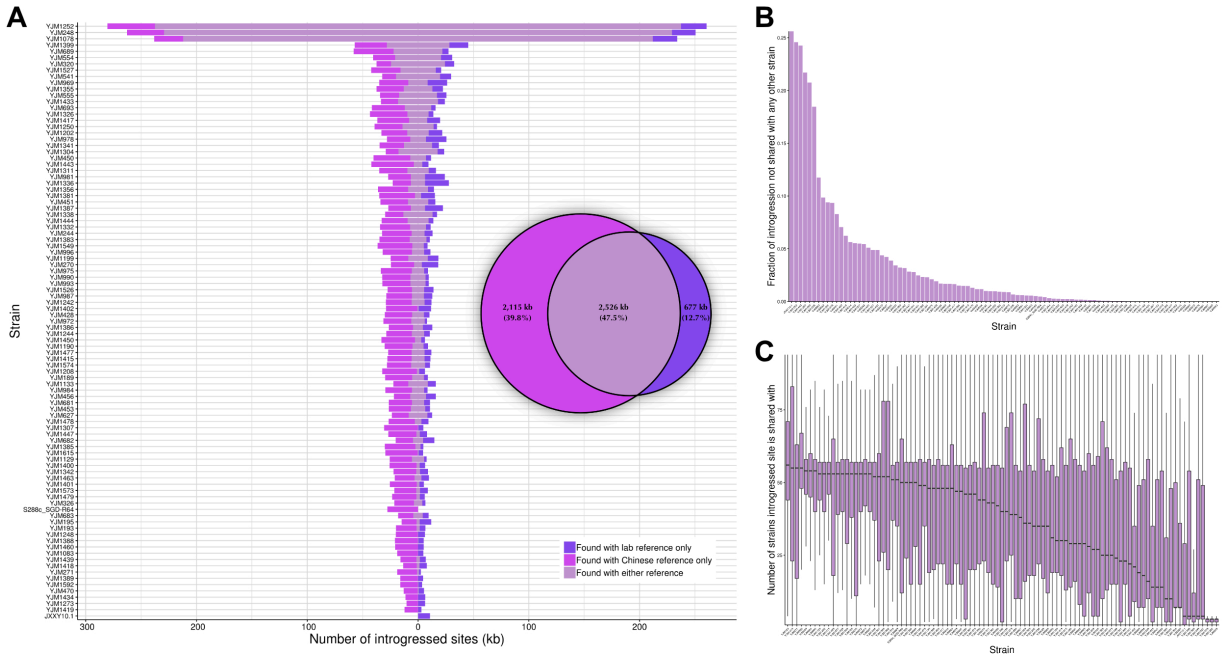

Figure 5.3: Amount of introgression identified using two different *S. cerevisiae* references, and differences in how much introgression is shared between strains. (A) Each bar represents a Venn diagram of the amount of introgression found when using each *S. cerevisiae* reference separately for each test strain. Strains are sorted by total amount of introgression predicted. Inset shows sites found with each reference across all strains. (B) Fraction of introgression in each strain that is uniquely predicted to be introgressed in that strain. The Chinese reference strain has the greatest fraction of unique introgression, but several other strains have nearly as much. (C) For each strain, a boxplot summarizes the distribution of the number of strains each introgressed site is shared with.

introgression in the test strain and introgression in the reference strain. The introduction of one additional state in the model to capture sites at which both reference strains match well, but differ from the test strain, could help separate these distinct phenomena.

5.3.5 The majority of introgression identified is from European S. paradoxus

In addition to running our model with two different *S. cerevisiae* strains as references, we also ran it with four different *S. paradoxus* references—one from each of the European, Far Eastern, American, and Hawaiian populations. Because our model does not handle non–mutually–exclusive states in a useful way (as explained above) we ran it separately with each *S. paradoxus* reference, combined the regions predicted by each model, then assigned each region to a specific *S. paradoxus* reference—or a combination of references—if possible. Strains differ widely in the amount of introgression originating from different *S. paradoxus* populations (Fig. 5.4a). But overall, most of the introgression we can assign to a single *S. paradoxus* reference is assigned to the European strain (Fig. 5.5a). This pattern is partially driven by the three strains mentioned above that share a large amount of introgression, which is mainly classified as European *S. paradoxus*, but most other strains also have more introgression assigned to the European reference than to any other reference (Fig. 5.4b).

Aside from the overall preponderance of European introgression, a few other trends stand out in the introgression assigned to other *S. paradoxus* populations. The amount of introgression from Hawaiian *S. paradoxus* varies less across the strains than does the amount from other *S. paradoxus* populations (Fig. 5.4b), which is consistent with the observation that a substantial fraction of the Hawaiian *S. paradoxus* introgression is shared among many strains (Fig. 5.5b). The amount of introgression from American *S. paradoxus*, on the other hand, is more highly variable between strains. The four strains (YJM1399, YJM451, YJM1355, YJM320) that have the largest quantity of introgression from American *S. paradoxus* (Fig. 5.4b) do not fall into any observed category: three are mosaic strains while one falls into the Wine/European population, two are clinical while one is from a cherry tree and one is from molasses, and two are from the United States while the other two are of

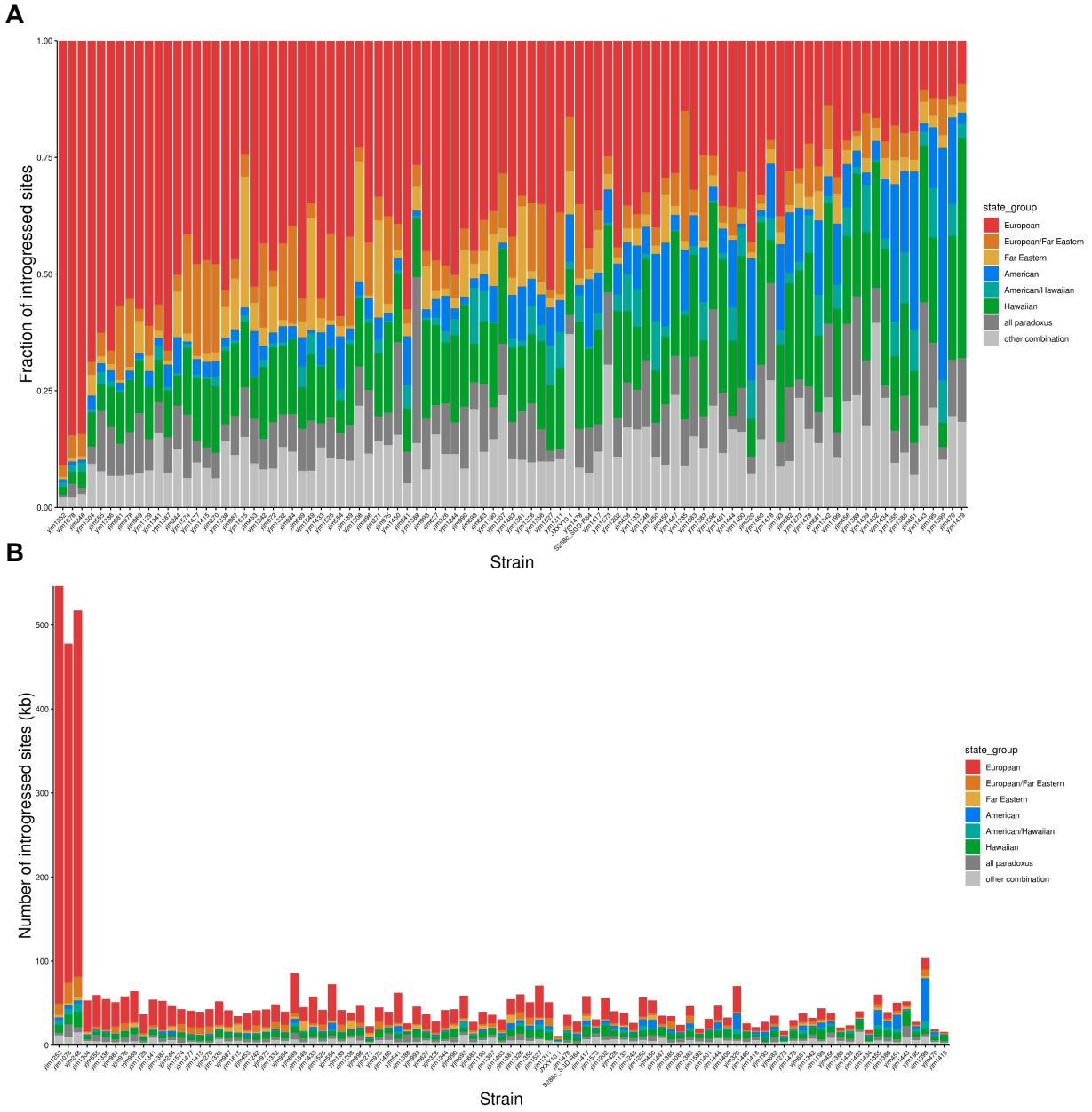

Figure 5.4: Amount of introgression from different *S. paradoxus* populations in individual test strains. (A) Fraction of introgression in each strain assigned to individual or combinations of *S. paradoxus* references. Strains are sorted by amount of introgression that is European and/or Far Eastern over amount of introgression that is American and/or Hawaiian. (B) Total amount of introgression in each strain assigned to each *S. paradoxus* reference.

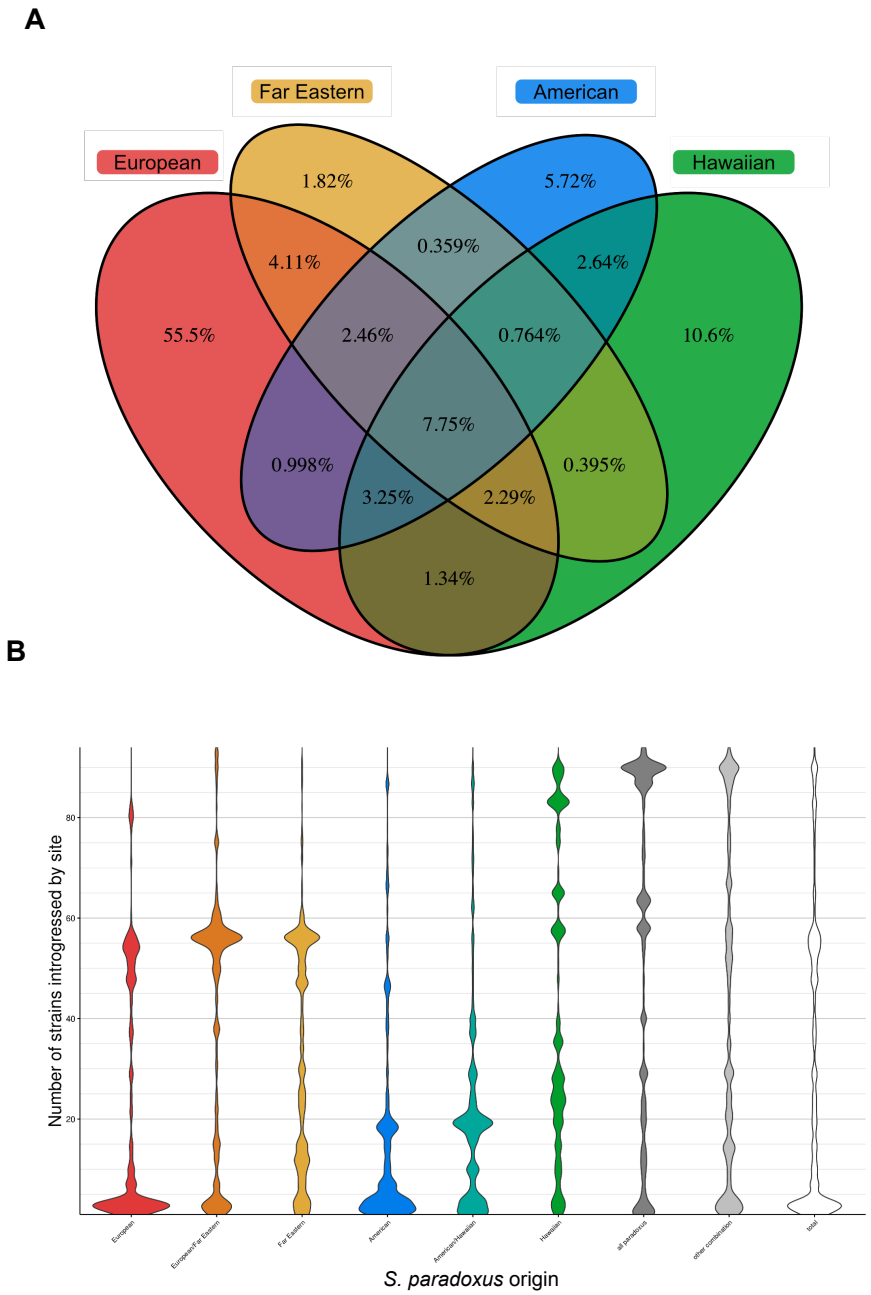

Figure 5.5: (A) Total amount of introgression assigned to every possible combination of *S. paradoxus* states across all strains. The majority of introgressed sites are assigned to the European reference alone. (B) Amount of introgression shared among different numbers of strains for each *S. paradoxus* reference or selected combination of references.

unknown geographical origin.

5.3.6 Introgressions cluster among distinct sets of strains, supporting the occurrence of multiple hybridization events

Since using the Chinese *S. cerevisiae* reference should allow us to identify some older introgressions that we were not able to previously, we wanted to see whether it was possible to group introgressed sites into larger, possibly discontinuous regions that are likely to have originated from the same hybridization event. We first grouped consecutive introgressed sites into regions introgressed in the same set of strains (regardless of their assigned *S. paradoxus* population(s)). We then hierarchically clustered these regions based on the pattern of presence/absence of introgression in all the strains. From the resulting dendrogram, seven clusters appeared to capture much of the variation. We then used k–means clustering to group the sites into seven clusters, which largely corresponded to the structure of the dendrogram. Although at least one of these clusters appears to be based on noise, several others identify groups of sites shared in similar sets of strains (Fig. 5.6). In addition, the sites in many of the clusters are mainly (though not entirely) predicted to be from the same *S. paradoxus* population(s), as we would expect if they were from the same hybridization event. While we are not able to pinpoint a precise number of distinct hybridization events, these data clearly support at least four distinct hybridization events occurring at different points in the evolutionary history of these strains.

5.3.7 Introgression in the S. cerevisiae lab strain S288c

As it is often appealing to think of the well–annotated lab strain *S. cerevisiae* S288c as a reference for all of *S. cerevisiae*, we were interested in the introgressions present in that strain—and particularly in any of those introgressions that are not common among the other strains we examined. We identified very few sites uniquely introgressed in *S. cerevisiae* S288c; most of this strain's introgressed sites are introgressed in more than half of the other strains

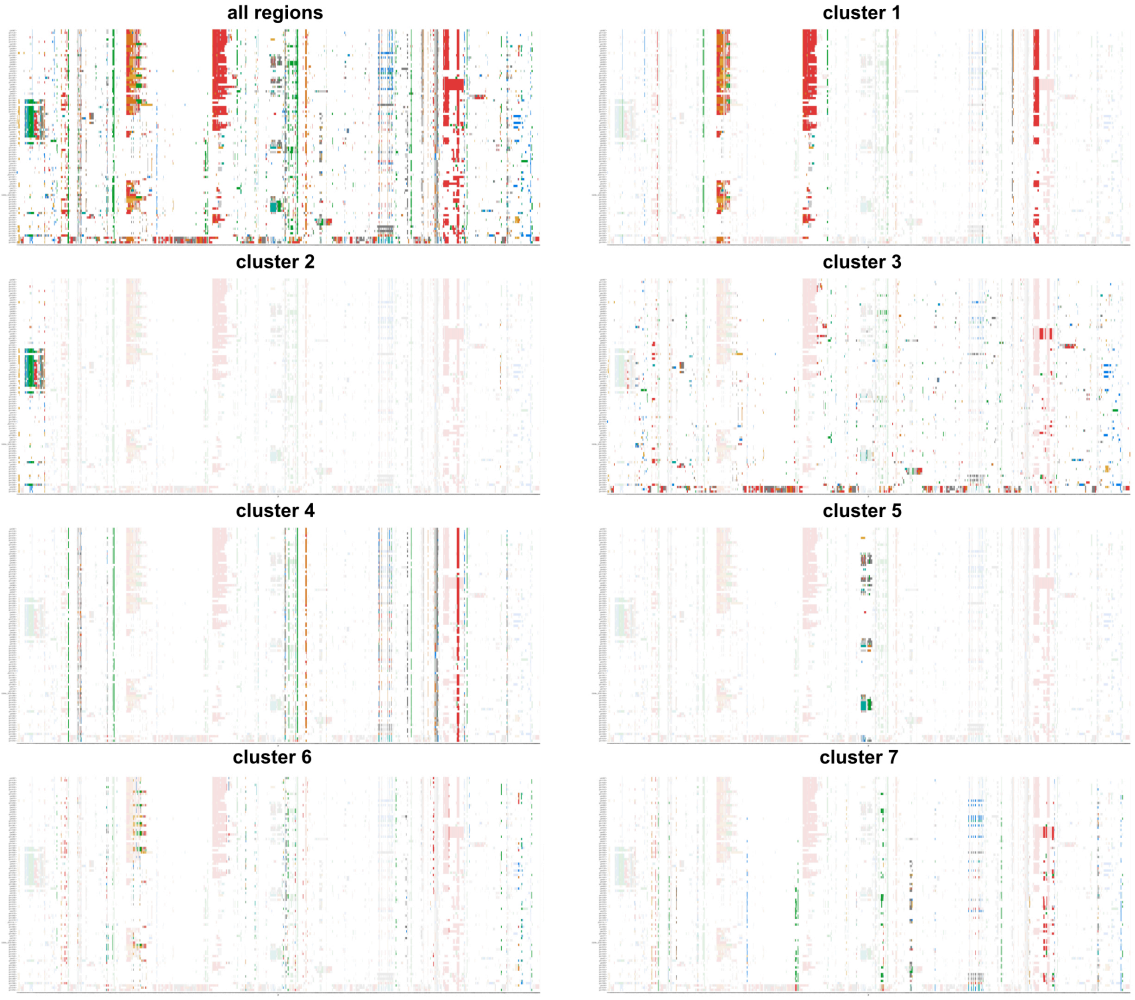

Figure 5.6: Introgressed regions clustered by the pattern of strains they are identified in. Top left shows matrix of strains by introgressed sites, with colors indicating the *S. paradoxus* reference(s) the region is assigned to. Regions are shown in sequential order, with all chromosomes concatenated. Strains are hierarchically clustered by their shared introgression across all sites. The width of each column is scaled by the square root of the corresponding region length in order to allow all regions to be visible. The regions in each of the seven clusters are individually highlighted in the remaining seven panels. Cluster 3 consists primarily of regions introgressed in the three strains YJM1078, YJM1252, and YJM248. Clusters 1, 2, and 5 appear to capture regions introgressed in distinct sets of strains. Clusters 6 and 7 are noisier, and cluster 4 captures regions introgressed in nearly all the strains.

(Fig. 5.3c). There are however, a few examples of sizeable introgressions in S288c that are uncommon among these strains.

One large introgression present in S288c and six other strains (YJM1381, YJM1385, YJM1386, YJM1399, YJM689, YJM1444) is a 3.9 kb region on the right arm of chromosome I, downstream of SWH1 and upstream of FLO1. Another region is near the centromere of chromosome VIII, containing part of the gene YHL008C, and is introgressed in S288c and a small number of other strains (YJM1307, YJM1401, YJM451, YJM1419, YJM1400, YJM1479). This predicted region is only 140 bp in length, but based on visually examination of the alignment, it clearly extends for another 1.8kb. It does not match any of the *S. paradoxus* references well, however, nor any of the genomes in the SGD fungal database (https://www.yeastgenome.org/blast-fungal).

Overall, however, we find few clear examples of introgressed regions that are introgressed in the lab strain S288c but not in many other strains. Thus, it is perhaps not a great concern that introgression has introduced dramatic changes that make the strain a poor model for other strains. On the other hand, it is possible that introgressions present in other strains cause those strains to be differ in important ways from S288c. In general, though, regions that are predicted to be introgressed in most strains, but excluding S288c, may be due to introgression in one of the reference strains we use, rather than in the test strains. Our current approach is not always able to distinguish between these possibilities.

5.4 Methods

5.4.1 Strain divergence, alignment, and phylogeny

The *S. cerevisiae* S288c genome was downloaded from the Saccharomyces Genome Database (https://downloads.yeastgenome.org/sequence/). The 93 *S. cerevisiae* strain genomes newly sequenced as part of the 100–genomes project,[62] as well as the *S. cerevisiae* JXXY10.1 genome,[111] were downloaded from GenBank (https://www.ncbi.nlm.nih.gov/genbank/). All *S. paradoxus* reference strains (European CBS432, Far Eastern N–45, American DBVPG6304,

Hawaiian UWOPS91–917.1) were downloaded from the Saccharomyces Genome Resequencing Project website (ftp://ftp.sanger.ac.uk/pub/users/dmc/yeast/latest). Alignments among reference strains and between reference and test strains were generated using `MAFFT`[121] with default settings. Divergence was calculated in 200 bp windows based on these alignments, excluding columns containing gaps or ambiguous characters. The phylogeny of reference strains was constructed from the 6–way alignment of the reference genomes using `PHYLIP dnadist` and `neighbor` (http://evolution.genetics.washington.edu/phylip/) with the UPGMA option.

5.4.2 Hidden Markov model for identifying introgression

We previously described a hidden Markov model–based approach to identifying introgression in budding yeast genomes (see Chapter 4 and Fig. 5.2). We first align a given *S. cerevisiae* test strain chromosome to the corresponding chromosomes of two reference strains—one for *S. cerevisiae* and one for *S. paradoxus*. We then run a hidden Markov model to assign the most likely of three states (*S. cerevisiae*/non–introgressed, *S. paradoxus*/introgressed, or unknown) to each polymorphic, non–gapped alignment column. The model is trained using the Baum–Welch algorithm until the likelihood between runs increases by less than 0.1%. Introgressed regions are defined as consecutive sets of sites in the *S. paradoxus* state, including any intervening non–polymorphic or gapped alignment columns.

5.4.3 Filtering predicted introgressed regions

Introgressed regions predicted by the HMM pass the first stage of filtering if they meet the following criteria, calculated across the entire region:

1. None of the three aligned sequences (*S. cerevisiae* reference, *S. paradoxus* reference, *S. cerevisiae* test strain) has greater than 50% gapped or ambiguous sites.

2. The identity of the test strain with the *S. cerevisiae* reference is at least 60%.

3. The identity of the test strain with the *S. paradoxus* reference is greater than the identity of the test strain with the *S. cerevisiae* reference.

4. The number of sites at which the test strain matches the *S. paradoxus* reference but not the *S. cerevisiae* reference is at least seven.

Regions that do not pass these filters are excluded from downstream analysis with one exception: when combining regions from multiple models (see below), if a filter–rejected region overlaps a filter–accepted region predicted by another model, it is added back in.

5.4.4 Combining predictions made using different references

To incorporate multiple reference genomes in our analysis, it would be possible to introduce additional states into the model. However, including multiple references for a single species results in uninformative predictions because those states are not mutually exclusive at every alignment column. To address this issue, we could create states for every possible combination of the references, but the model becomes slower and much more difficult to parameterize in that scenario. We therefore found it to be simplest and most extensible to run separate models for all references of interest, then combine the predictions in a principled way.

To combine the different sets of predictions, we first index all of the regions relative to the same reference genome, using alignments to convert between coordinates if necessary. We then iterate over the superset of regions, and generate a new set of combined regions, in which each region is comprised of a maximal set of consecutive sites called introgressed in the same set of models (Fig. 5.2b). Regions that failed to pass the first filtering step are added back in if they overlap a region that did pass filtering from a different model; this step helps prevent us from being overconfident in assigning a region to a specific model later if it was on the border of being called by a different model. Finally, each of these combined regions is assigned a final state or set of states depending on the sequence identity to introgressed references in different models. Specifically, if out of all the references the maximum identity was X_M bp over an aligned length of L_M bp, then we set the threshold

for retaining the other references at an identity of $1-(L_M - X_M + 2\sqrt{L_M - X_M})/L_M$. This threshold represents two standard deviations from the number of mismatches in the reference with the maximum identity, assuming mismatches are Poisson–distributed. In this way, some regions are assigned to a single reference genome, but others are assigned with less specificity.

5.4.5 Clustering introgressed sites

We converted the lists of regions for individual strains into one list of regions shared among strains by finding the largest consecutive blocks of sites called introgressed in the same set of strains. We then created a matrix of the presence/absence of introgression in each of these regions in each strain. Treating this as a binary matrix, we created a (hamming) distance matrix using the R function `rdist`.[130] We hierarchically clustered the regions based on these distances using the R function `hclust`.[131] We plotted the resulting dendrogram and visually determined that approximately seven clusters would capture much of the structure in the data. We then clustered the sites into seven clusters using the R function `kmeans`[131] with `nstart=20`. The resulting clusters corresponded reasonably well to the hierarchical clustering dendrogram. We also clustered the strains based on a (manhattan) distance matrix also generated with `rdist`, and clustered using `hclust`. We plotted the sites for the clustered strains using ggplot2.[132] We colored each region for each strain by the reference(s) it was finally assigned, as described above. We scaled the plotted lengths of the regions by the square root of the region length so that shorter regions would also be visible.

5.5 Discussion

We have extended our approach to identifying introgression in *S. cerevisiae* genomes from *S. paradoxus* by incorporating more diverse reference strains for both species. In doing so, we have identified more instances of introgression, and in some cases have been able to assign the origin of an introgression to a more specific population within *S. paradoxus*. In other cases, we are not able to determine a specific population of origin, due to the fact that too

little divergence is present in the region or the introgression originated too distantly in the past. It is also clear that we sometimes erroneously assign regions to a specific *S. paradoxus* population, particularly when nearby regions are assigned to different populations; this issue can arise because the sequence or alignment quality in the region is insufficient for some of the strains.

Despite the potential for further refinements to our approach, we can see general patterns in different levels of introgression among strains, and also in how often that introgression is shared with other strains. In the future, looking for introgression in the opposite direction—from *S. cerevisiae* in *S. paradoxus*—in combination with these findings may help clarify cases in which one or more of the reference genomes have introgression rather than the test strains. In general, introgression is quite prevalent in *S. cerevisiae* and likely within the *Saccharomyces* yeasts more broadly. We have shown that incorporating a larger range of within–species variation is useful for identifying these introgressions, but we will still always be limited in our ability to detect ancient introgressions.

Chapter 6: FINAL THOUGHTS

6.1 Lessons about mapping with association and linkage

In the preceding chapters, we have explored some of the benefits and challenges that come with different approaches for relating trait variation to naturally occurring genetic variation. We have seen how genome–wide association can be a simple approach to utilizing the large numbers of diverse *S. cerevisiae* genomes that have been sequenced in recent years, but have also seen how this approach can suffer due to the extensive population structure that exists within the species. We have discussed how we can instead use linkage to analyze the recombinant offspring from crossing a small but carefully chosen set of strains, and how this approach can provide very high power and fine mapping resolution for traits with amenable genetic architectures.

In general, our options for locating trait–related variation within the genome all involve comparing the genomes of individuals with a range of trait values, in order to find polymorphic sites at which specific alleles correspond to the observed trait differences. With more individual genomes that collectively contain more polymorphic sites, we are able to locate relevant variation with greater precision. But in some ways this type of positional mapping will always be inadequate. In particular, when variants at different locations in the genome interact with insignificant marginal effects, it can be difficult or even impossible to find them when only scanning for individual causal sites. In some cases it is reasonable to perform 2–dimensional scans to detect some of these interactions, but as the number of interacting sites grows, it quickly becomes infeasible to test all possible combinations of sites.

There are a variety of ways in which we might imagine dealing with this complexity. A candidate gene approach that relies on knowledge already existing in the literature may allow us to focus only on interactions between sites we have reason to expect might be related to the trait of interest, vastly constraining the total space of interactions that need to be

considered. But this strategy will of course be unable to help us uncover entirely novel effects of specific genetic variants. Other approaches such as X–QTL[89] or experimental evolution are able to screen many possible combinations of variants in bulk, ultimately retaining only the most successful ones. These experimental techniques are a powerful way of exploring a larger portion of sequence space and can also help us determine what types of genetic architectures we should be searching for. However, they are only useful for phenotypes that can be reasonably selected for, and they may still have difficulty finding interactions involving many sites. Other possibilities involve statistical heuristics, such as scanning for individual locus effects with a lenient cutoff, then examining interactions between only the smaller set of sites that remain.[133] The success of these various approaches will depend on the specific nature of interactions we are trying to discover, and continuing to develop an understanding of the frequency of different kinds of genetic architectures will therefore be important.

6.2 Leveraging diversity within *Saccharomyces*

In both mapping traits and identifying introgression, we have seen how the extensive variation within *S. cerevisiae* can be both a powerful experimental tool and a source of more questions. With a large pool of phenotypic variation, there are many possibilities of traits to explore and engineer, and understanding the workings of an individual genome becomes more tractable in the presence of many related genomes. Studying the effects of genetic loci in a single sequence in isolation is difficult because for every site we want to characterize, we need to engineer the sequence change and determine whether there is a phenotypic effect; or conversely, we can make random genetic changes, then try to identify those that underlie resulting phenotypic changes. But since sequence space is enormous, having a large pool of variation as it occurs in natural populations is a profoundly useful way to establish a smaller set of interesting genetic possibilities.

Moving from one well–sequenced and –annotated reference sequence for *S. cerevisiae* to a highly diverse collection of genomes within the species and genus is an immensely useful de-

velopment that will undoubtedly continue to contribute greatly to our understanding of the genomes of individual strains and entire species. The diversity within *Saccharomyces* more broadly provides opportunities for studying the unique evolutionary trajectories of closely related species. Despite the substantial sequence divergence between these species, hybridization appears to be a relatively frequent occurrence, providing interesting opportunities for rapid adaptation to new environments. We still have a lot to learn about the specific ways in which hybridization and introgression have shaped the evolution of domesticated, clinical, and wild yeasts, and how they may continue to do so in significant ways in the future.

6.3 Moving from location to mechanism

We have so far focused mainly on locating interesting variation within the genome—whether through association or linkage or by identifying regions that have a particular evolutionary history. But in doing so, what we have avoided is transitioning from identifying this variation to determining important details about *how* it leads to observed phenotypic differences. For example, in linkage mapping, we may identify a very specific causal site, but without knowledge of the mechanism by which it actually impacts a trait of interest, it will likely be difficult to utilize that information; we may find that a particular site causes increased resistance to an antifungal drug, and while this is an important first step, it does not directly provide us with a strategy for preventing or combating that resistance.

Thus, positional mapping approaches are only one step in understanding the complex ways in which genomes produce phenotypic effects, and they in general need to be complemented by carefully designed experiments. The same is true for the identification of introgressed regions; certainly some of the regions we identify have simply persisted over time by chance, while others may have had—or continue to have—an adaptive benefit. Distinguishing between these different scenarios requires us to test the effects of the introgressed regions in different genetic backgrounds, and to observe any effects on phenotypes that we think they might be involved in. Although there a few known examples of introgressions linked

to specific functional consequences, we still have work to do in understanding the overall impact that hybridization and introgression have had in the evolution of the *Saccharomyces* yeasts.

6.4 A more comprehensive reference

While the longstanding *S. cerevisiae* reference genome continues to be an invaluable resource, it also seems increasingly inadequate given the vast quantities of genomic sequences and related data currently being generated. How can we move from curating a single—and in many ways arbitrary—reference genome, to maintaining a resource that is more representative of the total range of variation within the species while still being tractable to work with? The Saccharomyces Genome Database has taken steps towards generating a panel of several reference genomes, rather than providing only the single S288c reference,[134] but this panel still lacks much of the diversity present within the species. This challenge of dealing with significant variation between populations—not only in terms of the single nucleotide polymorphisms we have focused on here, but also in larger differences in gene content and structure—is not specific to yeast genetics. Similar concerns arise whenever species are sampled more thoroughly; for example, in the study of human genomes, it has been recently observed that a pan–genome constructed from many individual African genomes differs substantially from the standard (mostly–European) reference genome.[135]

With the advent of next–generation sequencing technologies, it often seems that genomes can be sequenced much faster than the data we generate can be understood. There is immense potential for discovery in data sets that we have already collected if we can find ways to more effectively connect disparate findings. Developing methods of integrating diverse data sets in a way that can be easily maintained, updated, and searched continues to be an important problem in the field of genomics. Ultimately, the utility of data from large–scale experiments is only useful to the extent that they can be linked with past and future findings to generate more a more comprehensive understanding of genetic principles. In

addition to integrating many diverse genomes into more comprehensive reference panels, there also remains the challenge of integrating data from many types of assays related to genetic function. For example, having an easily accessible way to see current knowledge about the evolution of a region could be informative when we are trying to determine the function of an uncharacterized gene. Or as another example, having a comprehensive picture of the variation in a given gene among all strains from a similar environment could help us put the gene sequence from a particular strain in context. Moving forward, the effective management of data will be a crucial aspect of large–scale genomics experiments, as will the task of moving from one specific genome to a large cloud of genomes that are related in complex ways. The future of genomics lies in appreciating the extraordinary diversity of life.

BIBLIOGRAPHY

[1] Jacob, F. Evolution and tinkering. *Science* **196**, 1161–1166 (1977). URL https://science.sciencemag.org/content/196/4295/1161.

[2] Duboule, D. & Wilkins, A. S. The evolution of 'bricolage'. *Trends in Genetics* **14**, 54–59 (1998). URL http://www.sciencedirect.com/science/article/pii/S0168952597013589.

[3] Rose, M. R. & Oakley, T. H. The new biology: beyond the Modern Synthesis. *Biology Direct* **2**, 30 (2007). URL http://biologydirect.biomedcentral.com/articles/10.1186/1745-6150-2-30.

[4] Karr, J. *et al.* A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell* **150**, 389–401 (2012). URL http://www.sciencedirect.com/science/article/pii/S0092867412007763.

[5] OâĂŹBrien, E., Monk, J. & Palsson, B. Using Genome-scale Models to Predict Biological Capabilities. *Cell* **161**, 971–987 (2015). URL http://www.sciencedirect.com/science/article/pii/S0092867415005681.

[6] Hawkins, R. D., Hon, G. C. & Ren, B. Next-generation genomics: an integrative approach. *Nature Reviews Genetics* **11**, 476–486 (2010). URL https://www.nature.com/articles/nrg2795.

[7] Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)* **269**, 496–512 (1995).

[8] Goffeau, A. *et al.* Life with 6000 genes. *Science (New York, N.Y.)* **274**, 546, 563–567 (1996).

[9] Consortium*, T. C. e. S. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science* **282**, 2012–2018 (1998). URL https://science.sciencemag.org/content/282/5396/2012.

[10] Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796–815 (2000).

[11] Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001). URL https://science.sciencemag.org/content/291/5507/1304.

[12] Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001). URL https://doi.org/10.1038/35057062.

[13] Lappalainen, T., Scott, A. J., Brandt, M. & Hall, I. M. Genomic Analysis in the Age of Human Genome Sequencing. *Cell* **177**, 70–84 (2019). URL https://linkinghub.elsevier.com/retrieve/pii/S0092867419302156.

[14] Peter, J. *et al.* Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**, 339 (2018). URL https://www.nature.com/articles/s41586-018-0030-5.

[15] Naseeb, S. *et al.* Whole Genome Sequencing, de Novo Assembly and Phenotypic Profiling for the New Budding Yeast Species Saccharomyces jurei. *G3: Genes|Genomes|Genetics* **8**, 2967–2977 (2018). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6118302/.

[16] Chaisson, M. J. P., Wilson, R. K. & Eichler, E. E. Genetic variation and the de novo assembly of human genomes. *Nature Reviews. Genetics* **16**, 627–640 (2015).

[17] Levy, S. E. & Myers, R. M. Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics* **17**, 95–115 (2016). URL https://doi.org/10.1146/annurev-genom-083115-022413.

[18] Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science (New York, N.Y.)* **273**, 1516–1517 (1996).

[19] Lander, E. S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).

[20] Lander, E. S. Initial impact of the sequencing of the human genome. *Nature; London* **470**, 187–97 (2011). URL http://search.proquest.com/docview/852753306/abstract/A9C6D0E0CD774417PQ/1.

[21] Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics* **14**, 681–691 (2013). URL https://www.nature.com/articles/nrg3555.

[22] Witte, J. S. Genome-Wide Association Studies and Beyond. *Annual Review of Public Health* **31**, 9–20 (2010). URL https://doi.org/10.1146/annurev.publhealth.012809.103723.

[23] Daiger, S. P. Was the Human Genome Project Worth the Effort? *Science (New York, N.Y.)* **308**, 362–364 (2005). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2582021/.

[24] Genomewide association studies: History, rationale and prospects for psychiatric disorders. *The American journal of psychiatry* **166**, 540–556 (2009). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3894622/.

[25] Ott, J., Wang, J. & Leal, S. M. Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics* **16**, 275–284 (2015). URL https://www.nature.com/articles/nrg3908.

[26] Hornsey, I. S. *A History of Beer and Brewing* (Royal Society of Chemistry, 2003). Google-Books-ID: QqnvNsgas20C.

[27] McGovern, P. E. *et al.* Fermented beverages of pre- and proto-historic China. *Proceedings of the National Academy of Sciences* **101**, 17593–17598 (2004). URL https://www.pnas.org/content/101/51/17593.

[28] McGovern, P. E., Hartung, U., Badler, V. R., Glusker, D. L. & Exner, L. J. The beginnings of winemaking and viniculture in the ancient Near East and Egypt. *Expedition* **39**, 3–21 (1997). URL https://www.bcin.ca/bcin/detail.app?id=179115.

[29] Cavalieri, D., McGovern, P. E., Hartl, D. L., Mortimer, R. & Polsinelli, M. Evidence for S. cerevisiae Fermentation in Ancient Wine. *Journal of Molecular Evolution* **57**, S226–S232 (2003). URL https://doi.org/10.1007/s00239-003-0031-2.

[30] Suas, M. *Advanced Bread and Pastry* (Cengage Learning, Detroit, 2008), 1 edition edn.

[31] Powis, T. G. *et al.* Oldest chocolate in the New World. *Antiquity* **81**, 302–305 (2007).

[32] Fay, J. C. *et al.* A polyploid admixed origin of beer yeasts derived from European and Asian wine populations. *PLOS Biology* **17**, e3000147 (2019). URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000147.

[33] Fay, J. C. & Benavides, J. A. Evidence for Domesticated and Wild Populations of Saccharomyces cerevisiae. *PLOS Genetics* **1**, e5 (2005). URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0010005.

[34] Legras, J.-L., Merdinoglu, D., Cornuet, J.-M. & Karst, F. Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history. *Molecular Ecology* **16**, 2091–2102 (2007). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-294X.2007.03266.x.

[35] Ludlow, C. L. *et al.* Independent origins of yeast associated with coffee and cacao fermentation. *Current biology : CB* **26**, 965–971 (2016). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4821677/.

[36] Soares, E. V. & Soares, H. M. V. M. Bioremediation of industrial effluents containing heavy metals using brewing cells of Saccharomyces cerevisiae as a green technology: a review. *Environmental Science and Pollution Research* **19**, 1066–1083 (2012). URL https://doi.org/10.1007/s11356-011-0671-5.

[37] Buijs, N. A., Siewers, V. & Nielsen, J. Advanced biofuel production by the yeast Saccharomyces cerevisiae. *Current Opinion in Chemical Biology* **17**, 480–488 (2013). URL http://www.sciencedirect.com/science/article/pii/S1367593113000598.

[38] Nielsen, J. Production of biopharmaceutical proteins by yeast. *Bioengineered* **4**, 207–211 (2013). URL https://doi.org/10.4161/bioe.22856.

[39] Scannell, D. R. *et al.* The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. *G3: Genes, Genomes, Genetics* **1**, 11–25 (2011). URL https://www.g3journal.org/content/1/1/11.

[40] Engel, S. R. *et al.* The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. *G3: Genes|Genomes|Genetics* **4**, 389–398 (2013). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3962479/.

[41] Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research* **40**, D700–705 (2012).

[42] Duina, A. A., Miller, M. E. & Keeney, J. B. Budding Yeast for Budding Geneticists: A Primer on the Saccharomyces cerevisiae Model System. *Genetics* **197**, 33–48 (2014). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4012490/.

[43] Giaever, G. *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387–391 (2002).

[44] Giaever, G. & Nislow, C. The Yeast Deletion Collection: A Decade of Functional Genomics. *Genetics* **197**, 451–465 (2014). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4063906/.

[45] Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science (New York, N.Y.)* **353** (2016).

[46] Kachroo, A. H. *et al.* Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science (New York, N.Y.)* **348**, 921–925 (2015). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4718922/.

[47] Yang, F. *et al.* Identifying pathogenicity of human variants via paralog-based yeast complementation. *PLoS Genetics* **13** (2017). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5466341/.

[48] Mccusker, J. H. Saccharomyces cerevisiae: an Emerging and Model Pathogenic Fungus. *Molecular Principles of Fungal Pathogenesis* 245–259 (2006). URL http://www.asmscience.org/content/book/10.1128/9781555815776.ch18.

[49] PÃľrez-Torrado, R. & Querol, A. Opportunistic Strains of Saccharomyces cerevisiae: A Potential Risk Sold in Food Products. *Frontiers in Microbiology* **6** (2016). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4705302/.

[50] Roemer, T. & Krysan, D. J. Antifungal Drug Development: Challenges, Unmet Clinical Needs, and New Approaches. *Cold Spring Harbor Perspectives in Medicine* **4** (2014). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3996373/.

[51] Butts, A. & Krysan, D. J. Antifungal Drug Discovery: Something Old and Something New. *PLOS Pathogens* **8**, e1002870 (2012). URL https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1002870.

[52] Pfaller, M. A. Antifungal drug resistance: mechanisms, epidemiology, and consequences for treatment. *The American Journal of Medicine* **125**, S3–13 (2012).

[53] Mukherjee, P. K., Sheehan, D. J., Hitchcock, C. A. & Ghannoum, M. A. Combination Treatment of Invasive Fungal Infections. *Clinical Microbiology Reviews* **18**, 163–194 (2005). URL https://cmr.asm.org/content/18/1/163.

[54] Nieduszynski, C. A. & Liti, G. From sequence to function: Insights from natural variation in budding yeasts. *Biochimica et Biophysica Acta* **1810**, 959–966 (2011). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3271348/.

[55] Steinmetz, L. M. *et al.* Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**, 326 (2002). URL https://www.nature.com/articles/416326a.

[56] Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science (New York, N.Y.)* **296**, 752–755 (2002).

[57] Ehrenreich, I., Gerke, J. & Kruglyak, L. Genetic Dissection of Complex Traits in Yeast: Insights from Studies of Gene Expression and Other Phenotypes in the BYÃůRM Cross. *Cold Spring Harbor symposia on quantitative biology* **74**, 145–153 (2009). URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2888688/`.

[58] Bloom, J. S., Ehrenreich, I. M., Loo, W., Lite, T.-L. V. & Kruglyak, L. Finding the sources of missing heritability in a yeast cross. *Nature* **494**, 234–237 (2013). URL `http://arxiv.org/abs/1208.2865`. ArXiv: 1208.2865.

[59] Cubillos, F. A. *et al.* High-resolution mapping of complex traits with a four-parent advanced intercross yeast population. *Genetics* **195**, 1141–1155 (2013).

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

[61] Gallone, B. *et al.* Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. *Cell* **166**, 1397–1410.e16 (2016). URL `https://www.cell.com/cell/abstract/S0092-8674(16)31071-6`.

[62] Strope, P. K. *et al.* 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**, 762–774 (2015). URL `http://genome.cshlp.org/content/25/5/762`.

[63] Maclean, C. J. *et al.* Deciphering the Genic Basis of Yeast Fitness Variation by Simultaneous Forward and Reverse Genetics. *Molecular Biology and Evolution* **34**, 2486–2502 (2017). URL `https://academic.oup.com/mbe/article/34/10/2486/3797322`.

[64] Connelly, C. F. & Akey, J. M. On the Prospects of Whole-Genome Association Mapping in Saccharomyces cerevisiae. *Genetics* **191**, 1345–1353 (2012). URL `https://www.genetics.org/content/191/4/1345`.

[65] Dujon, B. A. & Louis, E. J. Genome Diversity and Evolution in the Budding Yeasts (Saccharomycotina). *Genetics* **206**, 717–750 (2017). URL `https://www.genetics.org/content/206/2/717`.

[66] Shapira, R., Levy, T., Shaked, S., Fridman, E. & David, L. Extensive heterosis in growth of yeast hybrids is explained by a combination of genetic models. *Heredity* **113**, 316–326 (2014).

[67] Dunn, B. & Sherlock, G. Reconstruction of the genome origins and evolution of the hybrid lager yeast Saccharomyces pastorianus. *Genome Research* **18**, 1610–1623 (2008). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2556262/.

[68] Morales, L. & Dujon, B. Evolutionary Role of Interspecies Hybridization and Genetic Exchanges in Yeasts. *Microbiology and Molecular Biology Reviews* **76**, 721–739 (2012). URL http://mmbr.asm.org/content/76/4/721.

[69] Fitzpatrick, D. A. Horizontal gene transfer in fungi. *FEMS Microbiology Letters* **329**, 1–8 (2012). URL https://academic.oup.com/femsle/article/329/1/1/627174.

[70] Plagnol, V. & Wall, J. D. Possible Ancestral Structure in Human Populations. *PLoS Genetics* **2** (2006). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1523253/.

[71] Vernot, B. & Akey, J. M. Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. *Science* **343**, 1017–1021 (2014). URL https://science.sciencemag.org/content/343/6174/1017.

[72] Liu, K. J. *et al.* An HMM-Based Comparative Genomic Framework for Detecting Introgression in Eukaryotes. *PLoS Computational Biology* **10**, e1003649 (2014). URL http://dx.plos.org/10.1371/journal.pcbi.1003649.

[73] Ott, J., Kamatani, Y. & Lathrop, M. Family-based designs for genome-wide association studies. *Nature Reviews Genetics* **12**, 465–474 (2011). URL https://www.nature.com/articles/nrg2989.

[74] Yang, J., Wray, N. R. & Visscher, P. M. Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. *Genetic Epidemiology* **34**, 254–257 (2010).

[75] Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature Genetics* **32**, 650–654 (2002).

[76] Klein, R. J. *et al.* Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science (New York, N.Y.)* **308**, 385–389 (2005). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1512523/.

[77] Yamazaki, K. *et al.* Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. *Human Molecular Genetics* **14**, 3499–3506 (2005).

[78] Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics* **101**, 5–22 (2017). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5501872/.

[79] Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology* **8** (2012). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531285/.

[80] Mehmood, T., Martens, H., SÃębÃÿ, S., Warringer, J. & Snipen, L. Mining for genotype-phenotype relations in Saccharomyces using partial least squares. *BMC Bioinformatics* **12**, 318 (2011). URL https://doi.org/10.1186/1471-2105-12-318.

[81] Muller, L. a. H., Lucas, J. E., Georgianna, D. R. & McCusker, J. H. Genome-wide association analysis of clinical vs. nonclinical origin provides insights into Saccharomyces cerevisiae pathogenesis. *Molecular Ecology* **20**, 4085–4097 (2011).

[82] Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655–1664 (2009).

[83] Wang, Q.-M., Liu, W.-Q., Liti, G., Wang, S.-A. & Bai, F.-Y. Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity. *Molecular Ecology* **21**, 5404–5417 (2012). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-294X.2012.05732.x.

[84] Zhou, X. & Stephens, M. Genome-wide Efficient Mixed Model Analysis for Association Studies. *Nature genetics* **44**, 821–824 (2012). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3386377/.

[85] Song, M., Hao, W. & Storey, J. D. Testing for genetic associations in arbitrarily structured populations. *Nature Genetics* **47**, 550–554 (2015).

[86] Cromie, G. A. *et al.* Genomic sequence diversity and population structure of Saccharomyces cerevisiae assessed by RAD-seq. *G3 (Bethesda, Md.)* **3**, 2163–2171 (2013).

[87] Carlson, C. S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics* **74**, 106–120 (2004).

[88] Long, A. D. & Langley, C. H. The Power of Association Studies to Detect the Contribution of Candidate Genetic Loci to Variation in Complex Traits. *Genome Research* **9**, 720–731 (1999). URL http://genome.cshlp.org/content/9/8/720.

[89] Ehrenreich, I. M. *et al.* Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* **464**, 1039–1042 (2010).

[90] Koning, D.-J. d. & McIntyre, L. M. Back to the Future: Multiparent Populations Provide the Key to Unlocking the Genetic Basis of Complex Traits. *Genetics* **206**, 527–529 (2017). URL https://www.genetics.org/content/206/2/527.

[91] Deutschbauer, A. M. & Davis, R. W. Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nature Genetics* **37**, 1333 (2005). URL https://www.nature.com/articles/ng1674.

[92] Bloom, J. S. *et al.* Rare variants contribute disproportionately to quantitative trait variation in yeast. *bioRxiv* (2019). URL http://biorxiv.org/lookup/doi/10.1101/607291.

[93] Cubillos, F. A. *et al.* Assessing the complex architecture of polygenic traits in diverged yeast populations. *Molecular Ecology* **20**, 1401–1413 (2011).

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

[95] Sirr, A. *et al.* Allelic variation, aneuploidy, and nongenetic mechanisms suppress a monogenic trait in yeast. *Genetics* **199**, 247–262 (2015).

[96] Churchill, G. A. *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genetics* **36**, 1133–1137 (2004).

[97] Threadgill, D. W. & Churchill, G. A. Ten Years of the Collaborative Cross. *Genetics* **190**, 291–294 (2012). URL https://www.genetics.org/content/190/2/291.

[98] Mancera, E., Bourgon, R., Brozzi, A., Huber, W. & Steinmetz, L. M. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* **454**, 479–485 (2008). URL https://www.nature.com/articles/nature07135.

[99] Tsai, I. J., Burt, A. & Koufopanou, V. Conservation of recombination hotspots in yeast. *Proceedings of the National Academy of Sciences* **107**, 7847–7852 (2010). URL https://www.pnas.org/content/107/17/7847.

[100] Broman, K. W., Wu, H., Sen, Å. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890 (2003). URL https://academic.oup.com/bioinformatics/article/19/7/889/197785.

[101] Dudley, A. M., Janse, D. M., Tanay, A., Shamir, R. & Church, G. M. A global view of pleiotropy and phenotypically derived gene function in yeast. *Molecular Systems Biology* **1**, 2005.0001 (2005).

[102] Fournier, T. *et al.* Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. *bioRxiv* 609917 (2019). URL `https://www.biorxiv.org/content/10.1101/609917v1`.

[103] Darwin, C. *On the origin of species by means of natural selection or the preservation of favoured races in the struggle for life* (Project Gutenberg, 1859). URL `https://www.gutenberg.org/files/1228/1228-h/1228-h.htm`.

[104] Goulet, B. E., Roda, F. & Hopkins, R. Hybridization in Plants: Old Ideas, New Techniques[OPEN]. *Plant Physiology* **173**, 65–78 (2017). URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210733/`.

[105] Price, L. B. *et al.* Staphylococcus aureus CC398: Host Adaptation and Emergence of Methicillin Resistance in Livestock. *mBio* **3**, e00305–11 (2012). URL `http://mbio.asm.org/content/3/1/e00305-11`.

[106] Krogerus, K., MagalhÃčes, F., Vidgren, V. & Gibson, B. Novel brewing yeast hybrids: creation and application. *Applied Microbiology and Biotechnology* **101**, 65–78 (2017). URL `https://doi.org/10.1007/s00253-016-8007-5`.

[107] Peris, D., PÃĭrezâĂŘTorrado, R., Hittinger, C. T., Barrio, E. & Querol, A. On the origins and industrial applications of Saccharomyces cerevisiae ÃŮ Saccharomyces kudriavzevii hybrids. *Yeast* **35**, 51–69 (2018). URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/yea.3283`.

[108] Anderson, E. & Stebbins, G. L. Hybridization as an Evolutionary Stimulus. *Evolution* **8**, 378–388 (1954). URL `https://www.jstor.org/stable/2405784`.

[109] Almeida, P. *et al.* A Gondwanan imprint on global diversity and domestication of wine and cider yeast Saccharomyces uvarum. *Nature Communications* **5**, 4044 (2014).

[110] Barbosa, R. *et al.* Evidence of Natural Hybridization in Brazilian Wild Lineages of Saccharomyces cerevisiae. *Genome Biology and Evolution* **8**, 317–329 (2016). URL `https://academic.oup.com/gbe/article/8/2/317/2574019`.

[111] Duan, S.-F. *et al.* The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. *Nature Communications* **9** (2018).

[112] Zhu, Y. O., Sherlock, G. & Petrov, D. A. Whole Genome Analysis of 132 Clinical Saccharomyces cerevisiae Strains Reveals Extensive Ploidy Variation. *G3: Genes|Genomes|Genetics* **6**, 2421–2434 (2016). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4978896/.

[113] Liti, G., Barton, D. B. H. & Louis, E. J. Sequence Diversity, Reproductive Isolation and Species Concepts in Saccharomyces. *Genetics* **174**, 839–850 (2006). URL https://www.genetics.org/content/174/2/839.

[114] Maddison, W. P., Knowles, L. L. & Collins, T. Inferring Phylogeny Despite Incomplete Lineage Sorting. *Systematic Biology* **55**, 21–30 (2006). URL https://academic.oup.com/sysbio/article/55/1/21/2842934.

[115] Maddison, W. P. & Wiens, J. J. Gene Trees in Species Trees. *Systematic Biology* **46**, 523–536 (1997). URL https://academic.oup.com/sysbio/article/46/3/523/1651369.

[116] Wakeley, J. *Coalescent Theory: An Introduction* (2009).

[117] Rosenberg, N. A. The Probability of Topological Concordance of Gene Trees and Species Trees. *Theoretical Population Biology* **61**, 225–247 (2002). URL http://linkinghub.elsevier.com/retrieve/pii/S0040580901915680.

[118] Tsai, I. J., Bensasson, D., Burt, A. & Koufopanou, V. Population genomics of the wild yeast Saccharomyces paradoxus: Quantifying the life cycle. *Proceedings of the National Academy of Sciences* **105**, 4957–4962 (2008). URL http://www.pnas.org/content/105/12/4957.

[119] Byrne, K. P. & Wolfe, K. H. The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research* **15**, 1456–1461 (2005). URL http://genome.cshlp.org/content/15/10/1456.

[120] Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002). URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/18.2.337.

[121] Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).

[122] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

[123] Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718–3720 (2015). URL https://academic.oup.com/bioinformatics/article/31/22/3718/240978.

[124] Leducq, J.-B. *et al.* Local climatic adaptation in a widespread microorganism. *Proceedings of the Royal Society B: Biological Sciences* **281** (2014). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3896012/.

[125] Sipiczki, M. Interspecies hybridization and recombination in Saccharomyces wine yeasts. *FEMS Yeast Research* **8**, 996–1007 (2008). URL https://academic.oup.com/femsyr/article/8/7/996/495125.

[126] Sniegowski, P. D., Dombrowski, P. G. & Fingerman, E. Saccharomyces cerevisiae and Saccharomyces paradoxus coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS yeast research* **1**, 299–306 (2002).

[127] Marsit, S. & Dequin, S. Diversity and adaptive evolution of Saccharomyces wine yeast: a review. *FEMS Yeast Research* **15** (2015). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4629790/.

[128] Wei, W. *et al.* Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 12825–12830 (2007). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933262/.

[129] Doniger, S. W. *et al.* A Catalog of Neutral and Deleterious Polymorphism in Yeast. *PLOS Genetics* **4**, e1000183 (2008). URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000183.

[130] Blaser, N. *rdist: Calculate Pairwise Distances* (2018). URL https://CRAN.R-project.org/package=rdist.

[131] R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2019). URL https://www.R-project.org/.

[132] Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016). URL https://ggplot2.tidyverse.org.

[133] Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**, 413–

(2005). URL http://link.galegroup.com/apps/doc/A183410965/AONE?u=wash_main&sid=AONE&xid=13ec0620. 413.

[134] Engel, S. R. *et al.* From one to many: expanding the Saccharomyces cerevisiae reference genome panel. *Database* **2016** (2016). URL https://academic.oup.com/database/article/doi/10.1093/database/baw020/2630216.

[135] Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics* **51**, 30 (2019). URL https://www.nature.com/articles/s41588-018-0273-y.

Appendix A: CHAPTER 4 SUPPLEMENT

A.1 Supplementary Figures

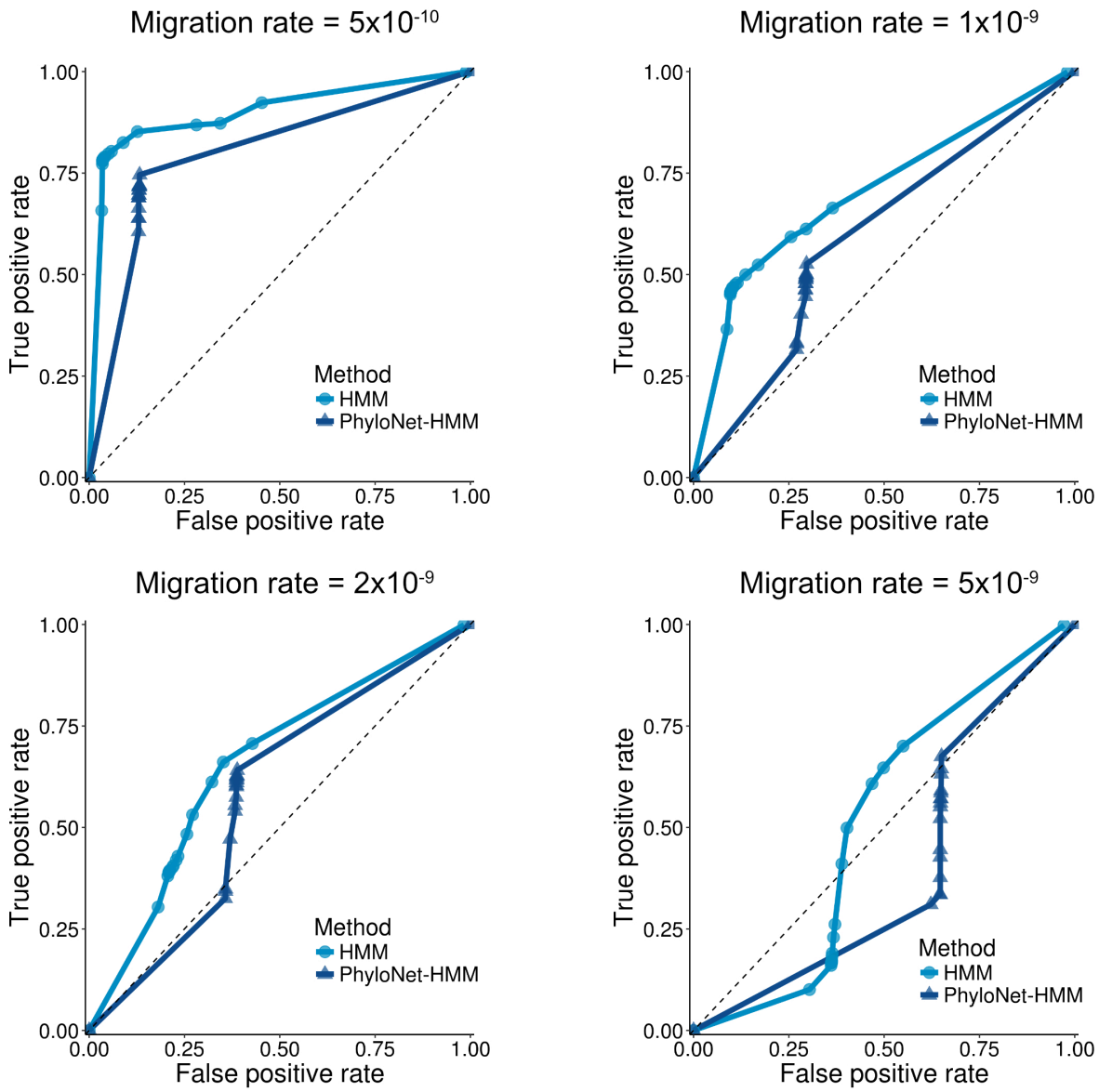

Figure A.1: Comparison of performance to PhyloNet–HMM for a variety of migration rates. Our method outperforms PhyloNet–HMM in all cases. Both methods perform worse for higher migration rates because more introgression is shared with the non-introgressed reference genome, making it difficult to detect.

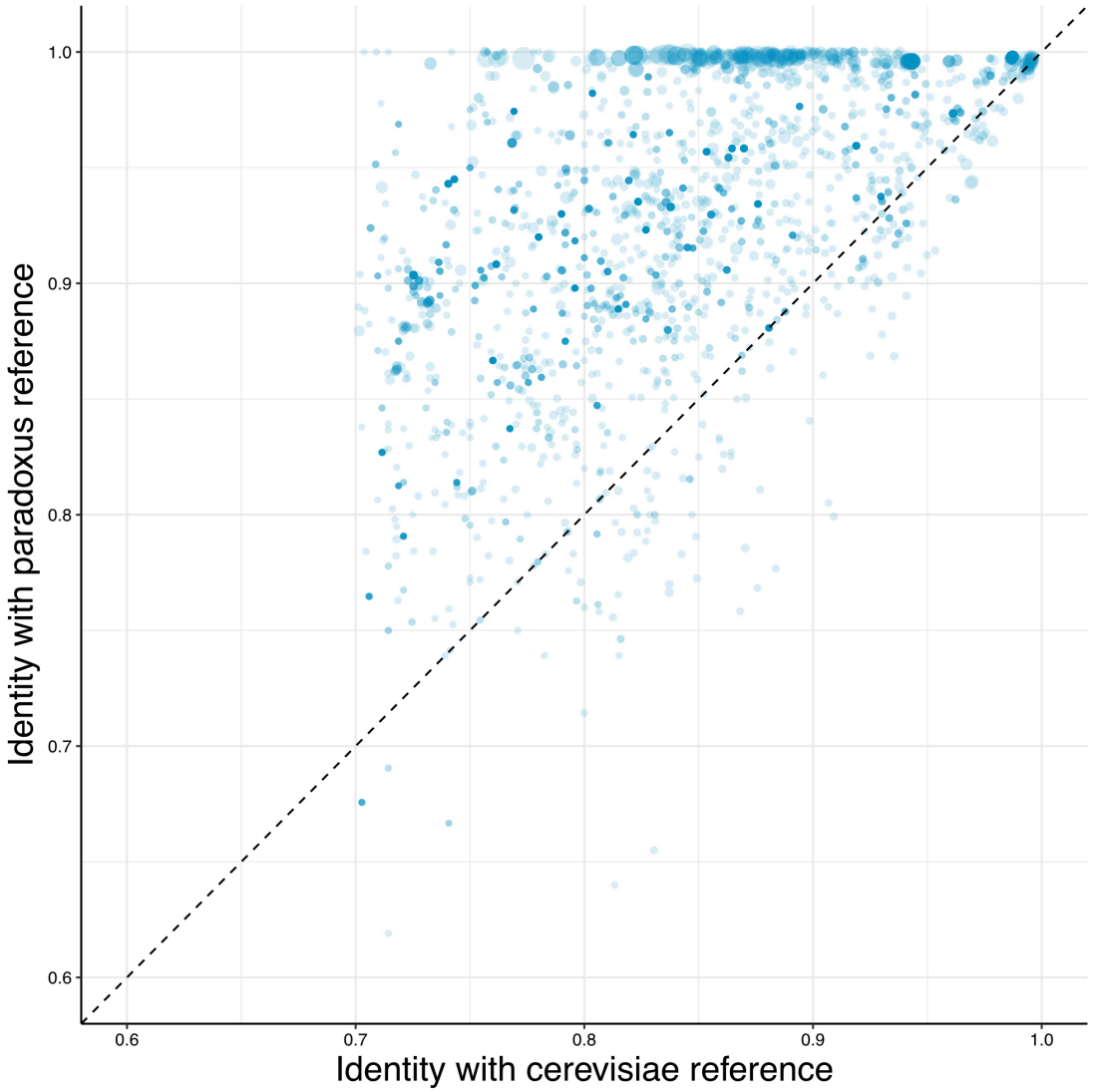

Figure A.2: Identity of introgressed regions with introgressed and non–introgressed reference strains. Points are scaled by the region length, with most longer regions falling near 100% identity with the introgressed reference and 80-90% identity with the non–introgressed reference.

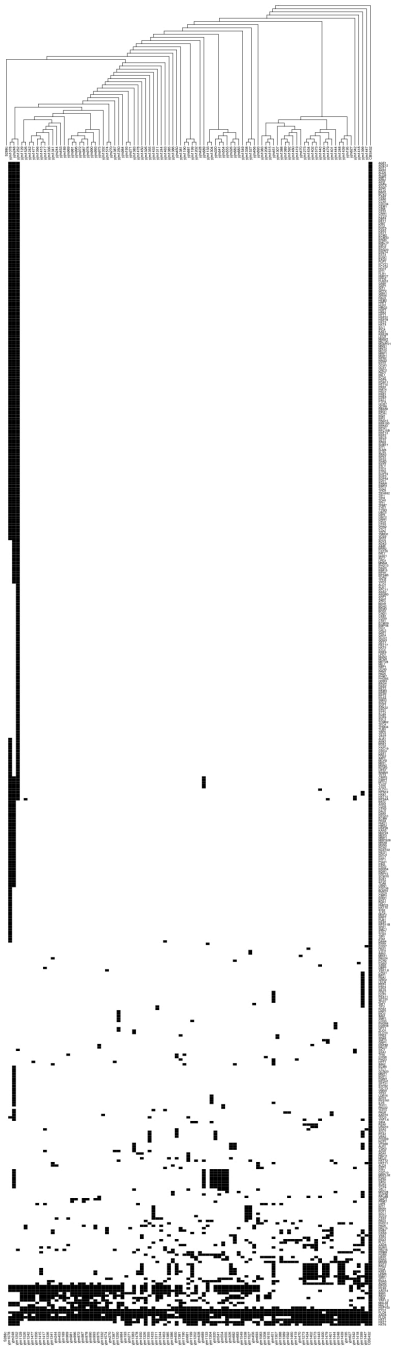

Figure A.3: Clusters of introgressed genes shared among closely related strains. Columns are strains, ordered by phylogenetic relationship shown at top. Rows are genes, and cells are colored black if any portion of the given gene is introgressed in the given strain. Gene rows are hierarchically clustered.

A.2 Supplementary Tables

Table A.1: Introgressed regions across all strains.

region id	strain	chromosome	start	end
r1	yjm470	I	527	664
r2	yjm1401	I	1780	2067
r3	yjm248	I	1845	2067
r4	yjm1389	I	2112	2261
r5	yjm1592	I	2112	2261
r6	yjm1338	I	2476	2542
r7	yjm1574	I	2500	2610
r8	yjm1450	I	2511	2644
r9	yjm1419	I	12189	12362
r10	yjm1389	I	12266	12395
r11	yjm1527	I	25400	25802
r12	yjm1418	I	25439	25508
r13	yjm1447	I	25448	25535
r14	yjm326	I	25451	26150
r15	yjm450	I	25451	25514
r16	yjm555	I	25451	25514
r17	yjm682	I	25451	26150
r18	yjm1342	I	25505	25628
r19	yjm1573	I	25518	25607
r20	yjm1252	I	25574	25618
r21	yjm1434	I	25712	25928
r22	yjm1615	I	25712	25943
r23	yjm554	I	25757	25815
r24	yjm1388	I	25826	26078
r25	yjm450	I	25871	26047
r26	yjm1273	I	25892	26048
r27	yjm978	I	25925	26063
r28	yjm1342	I	25970	26047
r29	yjm1573	I	25994	26225
r30	yjm1389	I	26022	26085
r31	yjm1460	I	26022	26085
r32	yjm1592	I	26022	26085
r33	yjm451	I	26027	26085
r34	yjm1083	I	26033	26078
r35	yjm1479	I	26033	26120
r36	yjm1433	I	26061	26258
r37	yjm1389	I	26157	26213
r38	yjm1460	I	26157	26213
r39	yjm1592	I	26157	26213
r40	yjm1307	I	26170	26317
r41	yjm193	I	26170	26231
r42	yjm978	I	26170	26231
r43	yjm1434	I	26292	26375
r44	yjm1342	I	26303	26414
r45	yjm1450	I	26304	26333
r46	yjm450	I	26304	26387
r47	yjm682	I	26304	27209
r48	yjm1273	I	26307	26507
r49	yjm1527	I	26307	26519
r50	yjm1389	I	26317	26386
r51	yjm1307	I	26411	26452
r52	yjm451	I	26427	26507
r53	yjm1083	I	26439	26528
r54	yjm450	I	26439	26693
r55	yjm1479	I	26441	26558
r56	yjm1434	I	26442	26495
r57	yjm1083	I	26564	26657
r58	yjm1273	I	26564	26693
r59	yjm1433	I	26564	26642
r60	yjm1434	I	26564	26693
r61	yjm1527	I	26564	26648
r62	yjm195	I	26564	26625
r63	yjm1479	I	26574	26693
r64	yjm1133	I	26699	26747
r65	yjm1527	I	26699	26744
r66	yjm470	I	26699	26933
r67	yjm1401	I	26709	26828
r68	yjm450	I	26709	27065
r69	yjm1615	I	26711	26792
r70	yjm1573	I	26712	26747
r71	yjm555	I	26712	26939
r72	yjm1252	I	26777	33128
r73	yjm1273	I	26834	26939
r74	yjm1434	I	26834	26939
r75	yjm195	I	26834	26882
r76	yjm451	I	26834	26939
r77	yjm1401	I	26877	26963
r78	yjm1527	I	26877	27209
r79	yjm1273	I	26969	27098
r80	yjm1389	I	26969	27098
r81	yjm1433	I	26969	27098
r82	yjm1434	I	26969	27098
r83	yjm1460	I	26969	27098
r84	yjm1477	I	26969	27098
r85	yjm195	I	26969	27098
r86	yjm987	I	26969	27098

region id	strain	chromosome	start	end
r87	yjm1133	I	27289	27404
r88	yjm1342	I	27289	27422
r89	yjm1388	I	27289	27452
r90	yjm1389	I	27289	27452
r91	yjm1443	I	27289	27422
r92	yjm1447	I	27289	27563
r93	yjm1460	I	27289	27452
r94	yjm1592	I	27289	27452
r95	yjm320	I	27289	27404
r96	yjm541	I	27289	27404
r97	yjm554	I	27289	27404
r98	yjm428	I	27323	27404
r99	yjm1129	I	27355	27422
r100	yjm470	I	27458	27822
r101	yjm1418	I	27469	27822
r102	yjm1549	I	27469	27824
r103	yjm1133	I	27475	27557
r104	yjm1208	I	27475	27824
r105	yjm1248	I	27475	27824
r106	yjm1273	I	27475	27824
r107	yjm1307	I	27475	27824
r108	yjm1342	I	27475	27822
r109	yjm1381	I	27475	27505
r110	yjm1388	I	27475	27822
r111	yjm1389	I	27475	27557
r112	yjm1399	I	27475	27822
r113	yjm1401	I	27475	27557
r114	yjm1419	I	27475	27824
r115	yjm1433	I	27475	27824
r116	yjm1439	I	27475	27824
r117	yjm1443	I	27475	27557
r118	yjm1460	I	27475	27822
r119	yjm1478	I	27475	27621
r120	yjm1527	I	27475	27824
r121	yjm1574	I	27475	27824
r122	yjm1592	I	27475	27557
r123	yjm193	I	27475	27824
r124	yjm428	I	27475	27557
r125	yjm451	I	27475	27824
r126	yjm541	I	27475	27557
r127	yjm554	I	27475	27505
r128	yjm627	I	27475	27824
r129	yjm693	I	27475	27824
r130	yjm984	I	27475	27621
r131	yjm1129	I	27485	27618
r132	yjm1190	I	27485	27824
r133	yjm1199	I	27485	27621
r134	yjm1202	I	27485	27621
r135	yjm1242	I	27485	27621
r136	yjm1244	I	27485	27621
r137	yjm1250	I	27485	27824
r138	yjm1304	I	27485	27824
r139	yjm1326	I	27485	27621
r140	yjm1332	I	27485	27824
r141	yjm1336	I	27485	27824
r142	yjm1338	I	27485	27621
r143	yjm1341	I	27485	27824
r144	yjm1356	I	27485	27621
r145	yjm1383	I	27485	27621
r146	yjm1387	I	27485	27618
r147	yjm1402	I	27485	27824
r148	yjm1415	I	27485	27621
r149	yjm1434	I	27485	27824
r150	yjm1450	I	27485	27824
r151	yjm1463	I	27485	27842
r152	yjm1477	I	27485	27847
r153	yjm1526	I	27485	27824
r154	yjm1573	I	27485	27621
r155	yjm189	I	27485	27824
r156	yjm270	I	27485	27824
r157	yjm271	I	27485	27618
r158	yjm453	I	27485	27824
r159	yjm681	I	27485	27621
r160	yjm683	I	27485	27621
r161	yjm689	I	27485	27621
r162	yjm969	I	27485	27621
r163	yjm972	I	27485	27621
r164	yjm975	I	27485	27621
r165	yjm978	I	27485	27847
r166	yjm981	I	27485	27621
r167	yjm987	I	27485	27847
r168	yjm990	I	27485	27621
r169	yjm993	I	27485	27621
r170	yjm1078	I	27590	27621
r171	yjm1311	I	27590	27621
r172	yjm1444	I	27590	27621
r173	yjm248	I	27590	27621
r174	yjm456	I	27590	27621
r175	yjm1415	I	27662	27824
r176	yjm1478	I	27662	27824
r177	yjm1129	I	27667	27824
r178	yjm1401	I	27683	27824
r179	yjm1389	I	27699	27822
r180	yjm1592	I	27699	27822
r181	yjm541	I	27700	27822

region id	strain	chromosome	start	end
r182	yjm1133	I	27702	27822
r183	yjm1202	I	27702	27824
r184	yjm1242	I	27702	27809
r185	yjm1244	I	27702	27824
r186	yjm1338	I	27702	27824
r187	yjm1387	I	27702	27824
r188	yjm1443	I	27702	27822
r189	yjm1447	I	27702	27822
r190	yjm1573	I	27702	27824
r191	yjm244	I	27702	27824
r192	yjm271	I	27702	27824
r193	yjm428	I	27702	27822
r194	yjm681	I	27702	27824
r195	yjm969	I	27702	27809
r196	yjm972	I	27702	27847
r197	yjm975	I	27702	27809
r198	yjm981	I	27702	27809
r199	yjm984	I	27702	27847
r200	yjm990	I	27702	27847
r201	yjm993	I	27702	27847
r202	yjm1078	I	27715	27760
r203	yjm1199	I	27715	27824
r204	yjm1311	I	27715	27824
r205	yjm1326	I	27715	27824
r206	yjm1355	I	27715	27822
r207	yjm1356	I	27715	27824
r208	yjm1383	I	27715	27824
r209	yjm1417	I	27715	27824
r210	yjm1444	I	27715	27824
r211	yjm248	I	27715	27760
r212	yjm320	I	27715	27822
r213	yjm456	I	27715	27824
r214	yjm554	I	27715	27822
r215	yjm683	I	27715	27824
r216	yjm689	I	27715	27824
r217	yjm470	I	30264	30480
r218	yjm1252	I	33164	35606
r219	yjm1078	I	66116	76608
r220	yjm248	I	66116	69416
r221	yjm1252	I	68204	69416
r222	yjm1252	I	72621	77043
r223	yjm248	I	74682	76608
r224	yjm1342	I	139967	140156
r225	yjm1356	I	179600	179728
r226	yjm1399	I	179600	180926
r227	yjm320	I	179600	179728
r228	yjm456	I	179600	179728
r229	yjm541	I	179600	179728
r230	yjm1199	I	179874	180126
r231	yjm1383	I	179874	180126
r232	yjm681	I	179874	180126
r233	yjm689	I	179874	180092
r234	yjm1311	I	179940	180048
r235	yjm1311	I	180400	180885
r236	yjm1444	I	180400	180926
r237	yjm1386	I	181196	182604
r238	yjm1399	I	181196	182604
r239	yjm541	I	181196	182604
r240	yjm689	I	181196	182604
r241	yjm1133	I	181263	182502
r242	yjm1199	I	181263	182502
r243	yjm1202	I	181263	182502
r244	yjm1311	I	181263	182502
r245	yjm1326	I	181263	182502
r246	yjm1355	I	181263	182502
r247	yjm1356	I	181263	182502
r248	yjm1383	I	181263	182502
r249	yjm1417	I	181263	182502
r250	yjm1444	I	181263	181839
r251	yjm320	I	181263	182502
r252	yjm428	I	181263	182502
r253	yjm554	I	181263	182502
r254	yjm681	I	181263	182502
r255	yjm456	I	181268	182502
r256	yjm1444	I	182091	182502
r257	yjm1356	I	183406	184749
r258	yjm1417	I	183406	184749
r259	yjm450	I	183674	183726
r260	yjm1208	I	183814	183932
r261	yjm1388	I	183834	183921
r262	yjm1083	I	184077	184182
r263	yjm1478	I	184107	184247
r264	yjm451	I	184137	184321
r265	yjm1208	I	184272	184519
r266	yjm683	I	186167	190146
r267	yjm555	I	186301	186354
r268	yjm1386	I	186886	190146
r269	yjm1399	I	186886	189084
r270	yjm1199	I	187081	187176
r271	yjm1383	I	187081	187176
r272	yjm428	I	187081	187176
r273	yjm456	I	187081	187176
r274	yjm554	I	187081	187176
r275	yjm681	I	187081	187176
r276	yjm1439	I	187374	187651

region id	strain	chromosome	start	end
r277	yjm195	I	187374	187461
r278	yjm1415	I	187419	187458
r279	yjm1248	I	187435	187531
r280	yjm627	I	187435	187531
r281	yjm1385	I	187507	187556
r282	yjm555	I	188105	188150
r283	yjm1444	I	188441	189803
r284	yjm1199	I	189771	190146
r285	yjm1202	I	189771	190146
r286	yjm1311	I	189771	190146
r287	yjm1326	I	189771	190146
r288	yjm428	I	189771	190146
r289	yjm456	I	189771	190177
r290	yjm554	I	189771	190146
r291	yjm681	I	189771	190146
r292	yjm689	I	189771	190146
r293	yjm1444	I	190061	190146
r294	yjm456	I	190929	191181
r295	yjm1399	I	192195	192281
r296	yjm470	I	198556	198675
r297	yjm682	I	198559	198712
r298	yjm1248	I	198581	198712
r299	yjm1439	I	198581	198712
r300	yjm1083	I	198600	198733
r301	yjm1190	I	198600	198733
r302	yjm1208	I	198600	198733
r303	yjm1615	I	198600	198733
r304	yjm450	I	198600	198733
r305	yjm1129	I	198604	198712
r306	yjm1133	I	198604	198712
r307	yjm1199	I	198604	198712
r308	yjm1202	I	198604	198712
r309	yjm1242	I	198604	198712
r310	yjm1244	I	198604	198712
r311	yjm1250	I	198604	198712
r312	yjm1273	I	198604	198712
r313	yjm1304	I	198604	198712
r314	yjm1307	I	198604	198712
r315	yjm1311	I	198604	198712
r316	yjm1326	I	198604	198712
r317	yjm1332	I	198604	198712
r318	yjm1336	I	198604	198712
r319	yjm1338	I	198604	198712
r320	yjm1341	I	198604	198712
r321	yjm1342	I	198604	198744
r322	yjm1356	I	198604	198712
r323	yjm1381	I	198604	198712
r324	yjm1385	I	198604	198712
r325	yjm1387	I	198604	198712
r326	yjm1388	I	198604	198712
r327	yjm1389	I	198604	198712
r328	yjm1401	I	198604	198712
r329	yjm1415	I	198604	198712
r330	yjm1434	I	198604	198712
r331	yjm1443	I	198604	198712
r332	yjm1450	I	198604	198712
r333	yjm1460	I	198604	198712
r334	yjm1463	I	198604	198712
r335	yjm1477	I	198604	198712
r336	yjm1478	I	198604	198675
r337	yjm1526	I	198604	198712
r338	yjm1527	I	198604	198712
r339	yjm1549	I	198604	198675
r340	yjm1574	I	198604	198712
r341	yjm1592	I	198604	198712
r342	yjm189	I	198604	198712
r343	yjm193	I	198604	198712
r344	yjm244	I	198604	198712
r345	yjm270	I	198604	198712
r346	yjm271	I	198604	198675
r347	yjm326	I	198604	198712
r348	yjm451	I	198604	198712
r349	yjm453	I	198604	198712
r350	yjm554	I	198604	198712
r351	yjm555	I	198604	198712
r352	yjm627	I	198604	198712
r353	yjm681	I	198604	198712
r354	yjm683	I	198604	198712
r355	yjm693	I	198604	198712
r356	yjm969	I	198604	198712
r357	yjm972	I	198604	198712
r358	yjm975	I	198604	198712
r359	yjm978	I	198604	198712
r360	yjm981	I	198604	198712
r361	yjm984	I	198604	198712
r362	yjm987	I	198604	198712
r363	yjm990	I	198604	198712
r364	yjm993	I	198604	198712
r365	yjm996	I	198604	198712
r366	yjm1355	I	198618	198712
r367	yjm1383	I	198618	198712
r368	yjm1386	I	198618	198712
r369	yjm1417	I	198618	198712
r370	yjm1419	I	198618	198712
r371	yjm195	I	198618	198712

region id	strain	chromosome	start	end
r372	yjm248	I	198618	198712
r373	yjm320	I	198618	198712
r374	yjm456	I	198618	198712
r375	yjm541	I	198618	198712
r376	yjm689	I	198618	198712
r377	yjm1400	I	198641	198712
r378	yjm428	I	198647	198712
r379	yjm1399	I	198725	198927
r380	yjm1208	I	200904	200982
r381	yjm1083	I	202625	202739
r382	yjm1133	I	202625	202709
r383	yjm1190	I	202625	202739
r384	yjm1199	I	202625	202709
r385	yjm1202	I	202625	202709
r386	yjm1208	I	202625	202762
r387	yjm1307	I	202625	202762
r388	yjm1326	I	202625	202709
r389	yjm1388	I	202625	202762
r390	yjm1389	I	202625	202762
r391	yjm1400	I	202625	202739
r392	yjm1402	I	202625	202762
r393	yjm1433	I	202625	202762
r394	yjm1450	I	202625	202739
r395	yjm1460	I	202625	202762
r396	yjm1479	I	202625	202739
r397	yjm1573	I	202625	202762
r398	yjm1574	I	202625	202731
r399	yjm1592	I	202625	202762
r400	yjm1615	I	202625	202762
r401	yjm326	I	202625	202762
r402	yjm428	I	202625	202709
r403	yjm450	I	202625	202762
r404	yjm683	I	202625	202709
r405	yjm693	I	202625	202762
r406	yjm1399	I	202630	203132
r407	yjm1399	I	203247	204228
r408	yjm1573	I	205172	205219
r409	yjm1419	I	205197	205279
r410	yjm682	I	205322	205615
r411	yjm554	I	205442	205549
r412	yjm1573	I	205585	205657
r413	yjm1399	I	205759	205927
r414	yjm1433	I	206008	206083
r415	yjm1389	I	206044	206392
r416	yjm1388	I	206359	206551
r417	yjm1400	I	206509	206585
r418	yjm1527	I	206935	207034
r419	yjm1447	I	215923	215961
r420	yjm1447	I	216018	216104
r421	yjm1447	I	216923	217003
r422	yjm1208	I	218637	219078
r423	yjm1615	I	218637	218913
r424	yjm1419	I	218646	218913
r425	yjm1083	I	218757	218820
r426	yjm1273	I	218757	218820
r427	yjm1386	I	218757	218820
r428	yjm1418	I	218757	218913
r429	yjm1434	I	218757	218820
r430	yjm1447	I	218757	218820
r431	yjm1573	I	218757	218820
r432	yjm451	I	218757	218820
r433	yjm1399	I	219686	221335
r434	yjm1385	I	220208	220381
r435	yjm1383	I	220214	220381
r436	yjm689	I	222514	222958
r437	yjm689	I	223018	223368
r438	yjm1355	I	223305	226026
r439	yjm554	I	223877	223917
r440	yjm627	I	223877	224003
r441	yjm683	I	223877	223917
r442	yjm554	I	224219	224458
r443	yjm683	I	224219	224458
r444	yjm554	I	224734	224808
r445	yjm683	I	224734	224808
r446	yjm1248	I	225359	225548
r447	yjm1307	I	225359	225464
r448	yjm1439	I	225359	225548
r449	yjm1460	I	225359	225464
r450	yjm627	I	225359	225548
r451	yjm693	I	225359	225635
r452	yjm1385	I	225400	226142
r453	yjm1083	I	225959	226094
r454	yjm1355	I	226220	226556
r455	yjm1304	I	227300	227422
r456	yjm1304	I	227482	227548
r457	yjm1355	I	227680	228136
r458	yjm1399	I	227680	228136
r459	yjm1399	I	228350	229091
r460	yjm1273	I	228377	228701
r461	yjm1355	I	228377	228505
r462	yjm1434	I	228377	228672
r463	yjm1250	I	228553	228672
r464	yjm1385	I	228553	228672
r465	yjm244	I	228553	228755
r466	yjm450	I	228553	228652

region id	strain	chromosome	start	end
r467	yjm451	I	228553	228790
r468	yjm981	I	228553	228672
r469	yjm993	I	228553	228672
r470	yjm1338	I	228577	228672
r471	yjm1402	I	228577	228790
r472	yjm1444	I	228577	228652
r473	yjm1573	I	228577	228790
r474	yjm320	I	228577	228672
r475	yjm554	I	228577	228672
r476	yjm1355	I	228611	228881
r477	yjm1383	I	228611	228672
r478	yjm1417	I	228611	228672
r479	yjm1355	I	228989	229091
r480	yjm271	II	642	1777
r481	yjm1463	II	2194	2272
r482	yjm271	II	2353	2467
r483	yjm1479	II	6525	6618
r484	yjm1615	II	6537	6594
r485	yjm1248	II	6543	6591
r486	yjm1381	II	6543	6591
r487	yjm1386	II	6543	6592
r488	yjm627	II	6543	6592
r489	yjm1311	II	7760	8046
r490	yjm1338	II	7823	8083
r491	yjm189	II	7823	8083
r492	yjm1381	II	7913	8016
r493	yjm682	II	7913	8018
r494	yjm1479	II	7917	8032
r495	yjm1385	II	7918	8016
r496	yjm1463	II	7918	8016
r497	yjm1242	II	7932	8046
r498	yjm1477	II	7932	8046
r499	yjm627	II	7932	8046
r500	yjm990	II	7932	8046
r501	yjm993	II	7932	8046
r502	yjm1450	II	7944	8033
r503	yjm975	II	7969	8046
r504	yjm1433	II	8045	8083
r505	yjm451	II	8220	8330
r506	yjm1450	II	9167	9315
r507	yjm1399	II	188646	188787
r508	yjm1252	II	248609	252607
r509	yjm248	II	248609	252607
r510	yjm1078	II	250480	252607
r511	yjm1078	II	252787	256305
r512	yjm1252	II	252787	256305
r513	yjm248	II	252787	256305
r514	yjm1252	II	315708	324670
r515	yjm248	II	319785	321057
r516	yjm248	II	321165	321210
r517	yjm1338	II	327077	327196
r518	yjm456	II	327077	327196
r519	yjm1078	II	331268	335120
r520	yjm1078	II	335210	335504
r521	yjm1078	II	361644	373569
r522	yjm1252	II	361644	373569
r523	yjm248	II	361644	373569
r524	yjm1078	II	382334	391428
r525	yjm248	II	382334	391428
r526	yjm1078	II	391449	395664
r527	yjm248	II	391449	395664
r528	yjm1252	II	391834	395664
r529	yjm1381	II	427997	428063
r530	yjm1419	II	427997	428063
r531	yjm451	II	427997	428063
r532	yjm682	II	602076	602254
r533	yjm1355	II	787152	793900
r534	yjm1399	II	787152	793824
r535	yjm320	II	787152	792487
r536	yjm450	II	787152	793282
r537	yjm320	II	793135	793612
r538	yjm320	II	793795	793896
r539	yjm1355	II	793966	799581
r540	yjm320	II	793975	794341
r541	yjm320	II	794447	795773
r542	yjm320	II	796685	800336
r543	yjm1078	II	801660	801692
r544	yjm1133	II	801660	801813
r545	yjm1190	II	801660	801698
r546	yjm1250	II	801660	801698
r547	yjm1252	II	801660	801692
r548	yjm1326	II	801660	801698
r549	yjm1336	II	801660	801698
r550	yjm1356	II	801660	801698
r551	yjm1381	II	801660	801698
r552	yjm1383	II	801660	801698
r553	yjm1549	II	801660	801698
r554	yjm1574	II	801660	801698
r555	yjm248	II	801660	801692
r556	yjm270	II	801660	801698
r557	yjm326	II	801660	801698
r558	yjm541	II	801660	801698
r559	yjm554	II	801660	801698
r560	yjm683	II	801660	801698
r561	yjm969	II	801660	801698

region id	strain	chromosome	start	end
r562	yjm972	II	801660	801698
r563	yjm975	II	801660	801698
r564	yjm978	II	801660	801698
r565	yjm981	II	801660	801698
r566	yjm984	II	801660	801698
r567	yjm987	II	801660	801698
r568	yjm990	II	801660	801698
r569	yjm993	II	801660	801698
r570	yjm996	II	801660	801698
r571	yjm1341	II	809397	809461
r572	yjm1401	III	1303	1404
r573	yjm1083	III	2041	2565
r574	yjm1326	III	4066	4185
r575	yjm1400	III	4076	4185
r576	yjm1478	III	4076	4183
r577	yjm1401	III	4083	4196
r578	yjm470	III	84492	84734
r579	yjm1304	III	104464	104936
r580	yjm248	III	104464	107957
r581	yjm451	III	104538	104813
r582	yjm1304	III	107459	133556
r583	yjm248	III	108047	108256
r584	yjm248	III	108397	133255
r585	yjm320	III	110480	130740
r586	yjm541	III	110480	130740
r587	yjm554	III	110480	130740
r588	yjm555	III	110480	130740
r589	yjm689	III	110480	130740
r590	yjm248	III	133298	133382
r591	yjm1399	III	149488	149781
r592	yjm1342	III	149745	149806
r593	yjm1252	III	242253	247693
r594	yjm1078	III	297618	302145
r595	yjm1252	III	297618	299930
r596	yjm248	III	297618	302145
r597	yjm1463	III	300960	301034
r598	yjm1326	III	301206	301667
r599	yjm1386	III	301206	301667
r600	yjm451	III	301206	301667
r601	yjm681	III	301206	301667
r602	yjm470	III	301726	301916
r603	yjm195	III	301789	301896
r604	yjm1326	III	302008	302079
r605	yjm1386	III	302008	302079
r606	yjm451	III	302008	302079
r607	yjm681	III	302008	302093
r608	yjm1326	III	302260	302585
r609	yjm1386	III	302260	302585
r610	yjm451	III	302260	302585
r611	yjm681	III	302260	302585
r612	yjm1326	III	304546	304681
r613	yjm1419	III	304546	304681
r614	yjm1439	III	304546	304607
r615	yjm1460	III	304546	304675
r616	yjm326	III	304546	304681
r617	yjm450	III	304546	304607
r618	yjm1389	III	307718	307854
r619	yjm1208	III	307734	307979
r620	yjm693	III	307734	307823
r621	yjm1133	III	307767	307979
r622	yjm1326	III	307767	307881
r623	yjm428	III	307832	307898
r624	yjm470	III	307832	307898
r625	yjm1304	III	307863	308042
r626	yjm1338	III	307863	307979
r627	yjm1478	IV	601	716
r628	yjm1447	IV	2225	2312
r629	yjm248	IV	16191	16616
r630	yjm1433	IV	16751	16880
r631	yjm248	IV	16770	34010
r632	yjm1463	IV	16976	17265
r633	yjm1311	IV	18194	18515
r634	yjm990	IV	18194	18388
r635	yjm1252	IV	18302	19601
r636	yjm1248	IV	18317	18498
r637	yjm1439	IV	18317	18498
r638	yjm195	IV	18317	18476
r639	yjm1083	IV	18324	18476
r640	yjm1129	IV	18324	18376
r641	yjm1133	IV	18324	18515
r642	yjm1190	IV	18324	18398
r643	yjm1199	IV	18324	18515
r644	yjm1202	IV	18324	18398
r645	yjm1208	IV	18324	18498
r646	yjm1242	IV	18324	18398
r647	yjm1244	IV	18324	18398
r648	yjm1250	IV	18324	18376
r649	yjm1273	IV	18324	18515
r650	yjm1307	IV	18324	18503
r651	yjm1326	IV	18324	18398
r652	yjm1336	IV	18324	18398
r653	yjm1338	IV	18324	18398
r654	yjm1355	IV	18324	18476
r655	yjm1356	IV	18324	18398
r656	yjm1381	IV	18324	18476

region id	strain	chromosome	start	end
r657	yjm1383	IV	18324	18476
r658	yjm1385	IV	18324	18476
r659	yjm1386	IV	18324	18398
r660	yjm1387	IV	18324	18376
r661	yjm1388	IV	18324	18515
r662	yjm1389	IV	18324	18503
r663	yjm1399	IV	18324	18515
r664	yjm1400	IV	18324	18515
r665	yjm1401	IV	18324	18515
r666	yjm1402	IV	18324	18515
r667	yjm1415	IV	18324	18388
r668	yjm1417	IV	18324	18398
r669	yjm1418	IV	18324	18515
r670	yjm1419	IV	18324	18515
r671	yjm1433	IV	18324	18398
r672	yjm1434	IV	18324	18515
r673	yjm1443	IV	18324	18515
r674	yjm1444	IV	18324	18515
r675	yjm1447	IV	18324	18515
r676	yjm1450	IV	18324	18476
r677	yjm1460	IV	18324	18503
r678	yjm1463	IV	18324	18515
r679	yjm1478	IV	18324	18398
r680	yjm1479	IV	18324	18515
r681	yjm1527	IV	18324	18398
r682	yjm1549	IV	18324	18515
r683	yjm1573	IV	18324	18515
r684	yjm1574	IV	18324	18376
r685	yjm1592	IV	18324	18476
r686	yjm1615	IV	18324	18498
r687	yjm193	IV	18324	18515
r688	yjm244	IV	18324	18376
r689	yjm270	IV	18324	18398
r690	yjm271	IV	18324	18476
r691	yjm320	IV	18324	18398
r692	yjm326	IV	18324	18376
r693	yjm428	IV	18324	18398
r694	yjm451	IV	18324	18515
r695	yjm456	IV	18324	18512
r696	yjm470	IV	18324	18398
r697	yjm541	IV	18324	18398
r698	yjm554	IV	18324	18398
r699	yjm555	IV	18324	18515
r700	yjm627	IV	18324	18398
r701	yjm681	IV	18324	18498
r702	yjm682	IV	18324	18398
r703	yjm683	IV	18324	18398
r704	yjm689	IV	18324	18503
r705	yjm693	IV	18324	18398
r706	yjm969	IV	18324	18376
r707	yjm972	IV	18324	18376
r708	yjm975	IV	18324	18398
r709	yjm978	IV	18324	18398
r710	yjm981	IV	18324	18398
r711	yjm984	IV	18324	18398
r712	yjm987	IV	18324	18398
r713	yjm993	IV	18324	18388
r714	yjm996	IV	18324	18398
r715	yjm1342	IV	18341	18515
r716	yjm1252	IV	19627	34010
r717	yjm1078	IV	160478	167372
r718	yjm1078	IV	167416	175063
r719	yjm1078	IV	325705	330618
r720	yjm1252	IV	325705	330618
r721	yjm1342	IV	383492	383539
r722	yjm320	IV	383937	383988
r723	yjm981	IV	383937	383988
r724	yjm987	IV	383937	383988
r725	yjm1419	IV	384082	384124
r726	yjm456	IV	384188	384224
r727	yjm248	IV	426510	428271
r728	yjm1078	IV	428547	432400
r729	yjm1252	IV	429197	439146
r730	yjm248	IV	429197	432400
r731	yjm1478	IV	513300	513396
r732	yjm1208	IV	513314	513396
r733	yjm1615	IV	513314	513396
r734	yjm1133	IV	527332	527523
r735	yjm1273	IV	527332	527523
r736	yjm1418	IV	527332	527523
r737	yjm320	IV	527332	527523
r738	yjm456	IV	527332	527523
r739	yjm683	IV	527332	527523
r740	yjm1463	IV	527333	528012
r741	yjm1341	IV	527481	527600
r742	yjm1133	IV	527904	528012
r743	yjm1199	IV	527904	528012
r744	yjm1248	IV	527904	528012
r745	yjm1273	IV	527904	528012
r746	yjm1304	IV	527904	528012
r747	yjm1342	IV	527904	528012
r748	yjm1418	IV	527904	528012
r749	yjm1433	IV	527904	528012
r750	yjm1439	IV	527904	528012
r751	yjm1447	IV	527904	528012

region id	strain	chromosome	start	end
r752	yjm320	IV	527904	528012
r753	yjm456	IV	527904	528012
r754	yjm541	IV	527904	528012
r755	yjm683	IV	527904	528012
r756	yjm1133	IV	529743	529797
r757	yjm1273	IV	529743	529797
r758	yjm1418	IV	529743	529797
r759	yjm320	IV	529743	529797
r760	yjm456	IV	529743	529797
r761	yjm683	IV	529743	529797
r762	yjm1433	IV	529760	529797
r763	yjm1199	IV	530349	530465
r764	yjm1248	IV	530349	530465
r765	yjm1273	IV	530349	530514
r766	yjm1304	IV	530349	530465
r767	yjm1307	IV	530349	530449
r768	yjm1336	IV	530349	530449
r769	yjm1342	IV	530349	530465
r770	yjm1400	IV	530349	530449
r771	yjm1418	IV	530349	530514
r772	yjm1433	IV	530349	530465
r773	yjm1439	IV	530349	530465
r774	yjm1447	IV	530349	530465
r775	yjm1463	IV	530349	530465
r776	yjm1549	IV	530349	530514
r777	yjm320	IV	530349	530514
r778	yjm451	IV	530349	530514
r779	yjm456	IV	530349	530465
r780	yjm541	IV	530349	530514
r781	yjm683	IV	530349	530514
r782	yjm689	IV	530349	530514
r783	yjm1133	IV	530379	530514
r784	yjm1202	IV	531789	531897
r785	yjm1386	IV	531789	531897
r786	yjm1388	IV	531789	531897
r787	yjm1402	IV	531789	531897
r788	yjm1443	IV	531789	531897
r789	yjm1444	IV	531789	531897
r790	yjm1527	IV	531789	531897
r791	yjm195	IV	531789	531897
r792	yjm554	IV	531789	531897
r793	yjm555	IV	531789	531897
r794	yjm627	IV	531789	531897
r795	yjm1443	IV	533628	533682
r796	yjm555	IV	533628	533682
r797	yjm1388	IV	533645	533682
r798	yjm1402	IV	533645	533682
r799	yjm1202	IV	534234	534350
r800	yjm1386	IV	534234	534350
r801	yjm1388	IV	534234	534350
r802	yjm1402	IV	534234	534399
r803	yjm1443	IV	534234	534350
r804	yjm1527	IV	534234	534350
r805	yjm555	IV	534234	534399
r806	yjm627	IV	534234	534350
r807	yjm1444	IV	534264	534363
r808	yjm195	IV	534264	534350
r809	yjm554	IV	534264	534350
r810	yjm1338	IV	535674	535782
r811	yjm1389	IV	535674	535782
r812	yjm1399	IV	535674	535782
r813	yjm1401	IV	535674	535782
r814	yjm1419	IV	535674	535782
r815	yjm1434	IV	535674	535782
r816	yjm1460	IV	535674	535782
r817	yjm1573	IV	535674	535782
r818	yjm1592	IV	535674	535782
r819	yjm470	IV	535674	535782
r820	yjm682	IV	535674	535782
r821	yjm1419	IV	537513	537567
r822	yjm1434	IV	537513	537567
r823	yjm682	IV	537513	537567
r824	yjm1460	IV	537530	537567
r825	yjm1573	IV	537530	537567
r826	yjm1389	IV	538119	538220
r827	yjm1399	IV	538119	538220
r828	yjm1401	IV	538119	538220
r829	yjm1419	IV	538119	538220
r830	yjm1434	IV	538119	538220
r831	yjm1460	IV	538119	538220
r832	yjm1573	IV	538119	538220
r833	yjm450	IV	538119	538182
r834	yjm470	IV	538119	538220
r835	yjm682	IV	538119	538220
r836	yjm1338	IV	538149	538220
r837	yjm1592	IV	538149	538220
r838	yjm1244	IV	538242	538383
r839	yjm1078	IV	556184	560535
r840	yjm1252	IV	556184	571463
r841	yjm248	IV	556184	560535
r842	yjm1078	IV	561356	571463
r843	yjm248	IV	561356	571463
r844	yjm1402	IV	757579	757634
r845	yjm1415	IV	757579	757634
r846	yjm1573	IV	757579	757634

region id	strain	chromosome	start	end
r847	yjm456	IV	757579	757634
r848	yjm1463	IV	759541	759607
r849	yjm1336	IV	759562	759607
r850	yjm1355	IV	759562	759607
r851	yjm1401	IV	759562	759607
r852	yjm451	IV	759562	759607
r853	yjm1273	IV	759565	759607
r854	yjm1386	IV	871836	872129
r855	yjm1415	IV	871838	871909
r856	yjm1399	IV	871932	872059
r857	yjm1402	IV	877655	877757
r858	yjm1199	IV	920798	921397
r859	yjm1202	IV	920798	921397
r860	yjm1304	IV	920798	921397
r861	yjm1252	IV	959196	962306
r862	yjm1252	IV	969551	996226
r863	yjm1078	IV	973071	996226
r864	yjm248	IV	973071	996226
r865	yjm1355	IV	986847	987030
r866	yjm1460	IV	986990	987036
r867	yjm1399	IV	1088292	1088482
r868	yjm1273	IV	1088326	1088482
r869	yjm1450	IV	1088326	1088482
r870	yjm451	IV	1088326	1088482
r871	yjm682	IV	1088326	1088482
r872	yjm1400	IV	1088410	1088482
r873	yjm1434	IV	1088410	1088482
r874	yjm1479	IV	1088410	1088482
r875	yjm1078	IV	1138207	1151026
r876	yjm1252	IV	1138207	1140929
r877	yjm248	IV	1138207	1151026
r878	yjm1252	IV	1149220	1150965
r879	yjm1078	IV	1151226	1153896
r880	yjm248	IV	1151226	1153896
r881	yjm1078	IV	1153916	1154204
r882	yjm1252	IV	1153916	1154204
r883	yjm248	IV	1153916	1154204
r884	yjm1133	IV	1154467	1154582
r885	yjm1190	IV	1154467	1154582
r886	yjm1199	IV	1154467	1154582
r887	yjm1202	IV	1154467	1154582
r888	yjm1208	IV	1154467	1154582
r889	yjm1248	IV	1154467	1154582
r890	yjm1273	IV	1154467	1154582
r891	yjm1304	IV	1154467	1154582
r892	yjm1326	IV	1154467	1154582
r893	yjm1342	IV	1154467	1154582
r894	yjm1383	IV	1154467	1154582
r895	yjm1389	IV	1154467	1154582
r896	yjm1399	IV	1154467	1154582
r897	yjm1400	IV	1154467	1154582
r898	yjm1402	IV	1154467	1154582
r899	yjm1418	IV	1154467	1154593
r900	yjm1419	IV	1154467	1154582
r901	yjm1433	IV	1154467	1154582
r902	yjm1434	IV	1154467	1154582
r903	yjm1439	IV	1154467	1154582
r904	yjm1478	IV	1154467	1154582
r905	yjm1479	IV	1154467	1154582
r906	yjm1549	IV	1154467	1154582
r907	yjm1573	IV	1154467	1154582
r908	yjm1592	IV	1154467	1154582
r909	yjm1615	IV	1154467	1154582
r910	yjm195	IV	1154467	1154582
r911	yjm320	IV	1154467	1154582
r912	yjm450	IV	1154467	1154582
r913	yjm451	IV	1154467	1154582
r914	yjm470	IV	1154467	1154582
r915	yjm541	IV	1154467	1154582
r916	yjm554	IV	1154467	1154582
r917	yjm555	IV	1154467	1154582
r918	yjm627	IV	1154467	1154582
r919	yjm682	IV	1154467	1154582
r920	yjm683	IV	1154467	1154582
r921	yjm689	IV	1154467	1154582
r922	yjm693	IV	1154467	1154582
r923	yjm1307	IV	1154506	1154582
r924	yjm1385	IV	1154506	1154582
r925	yjm1460	IV	1154506	1154582
r926	yjm456	IV	1154506	1154582
r927	yjm1387	IV	1154508	1154582
r928	yjm1078	IV	1154599	1154760
r929	yjm248	IV	1154599	1154760
r930	yjm1078	IV	1155010	1155166
r931	yjm248	IV	1155010	1155166
r932	yjm1133	IV	1155151	1155257
r933	yjm554	IV	1155151	1155257
r934	yjm193	IV	1155196	1155262
r935	yjm1190	IV	1155204	1155340
r936	yjm1199	IV	1155204	1155340
r937	yjm1202	IV	1155204	1155340
r938	yjm1208	IV	1155204	1155340
r939	yjm1248	IV	1155204	1155340
r940	yjm1273	IV	1155204	1155340
r941	yjm1304	IV	1155204	1155340

region id	strain	chromosome	start	end
r942	yjm1307	IV	1155204	1155325
r943	yjm1311	IV	1155204	1155340
r944	yjm1326	IV	1155204	1155340
r945	yjm1399	IV	1155204	1155257
r946	yjm1400	IV	1155204	1155340
r947	yjm1401	IV	1155204	1155340
r948	yjm1402	IV	1155204	1155340
r949	yjm1418	IV	1155204	1155340
r950	yjm1419	IV	1155204	1155340
r951	yjm1433	IV	1155204	1155340
r952	yjm1434	IV	1155204	1155340
r953	yjm1439	IV	1155204	1155340
r954	yjm1443	IV	1155204	1155340
r955	yjm1447	IV	1155204	1155340
r956	yjm1460	IV	1155204	1155262
r957	yjm1479	IV	1155204	1155340
r958	yjm1549	IV	1155204	1155340
r959	yjm1573	IV	1155204	1155340
r960	yjm1615	IV	1155204	1155340
r961	yjm195	IV	1155204	1155340
r962	yjm320	IV	1155204	1155340
r963	yjm326	IV	1155204	1155340
r964	yjm450	IV	1155204	1155340
r965	yjm451	IV	1155204	1155340
r966	yjm541	IV	1155204	1155340
r967	yjm555	IV	1155204	1155340
r968	yjm627	IV	1155204	1155340
r969	yjm681	IV	1155204	1155340
r970	yjm682	IV	1155204	1155340
r971	yjm683	IV	1155204	1155340
r972	yjm689	IV	1155204	1155340
r973	yjm1078	IV	1155369	1155475
r974	yjm1252	IV	1155369	1155475
r975	yjm248	IV	1155369	1155475
r976	yjm1199	IV	1155439	1155660
r977	yjm1202	IV	1155439	1155660
r978	yjm1208	IV	1155439	1155660
r979	yjm1304	IV	1155439	1155660
r980	yjm1311	IV	1155439	1155712
r981	yjm1326	IV	1155439	1155660
r982	yjm1400	IV	1155439	1155660
r983	yjm1401	IV	1155439	1155660
r984	yjm1402	IV	1155439	1155660
r985	yjm1419	IV	1155439	1155660
r986	yjm1433	IV	1155439	1155660
r987	yjm1434	IV	1155439	1155660
r988	yjm1443	IV	1155439	1155660
r989	yjm1479	IV	1155439	1155660
r990	yjm1549	IV	1155439	1155660
r991	yjm1573	IV	1155439	1155660
r992	yjm1615	IV	1155439	1155660
r993	yjm271	IV	1155439	1155660
r994	yjm320	IV	1155439	1155697
r995	yjm326	IV	1155439	1155660
r996	yjm450	IV	1155439	1155697
r997	yjm451	IV	1155439	1155660
r998	yjm541	IV	1155439	1155697
r999	yjm555	IV	1155439	1155660
r1000	yjm682	IV	1155439	1155697
r1001	yjm683	IV	1155439	1155660
r1002	yjm689	IV	1155439	1155660
r1003	yjm1273	IV	1155442	1155660
r1004	yjm1190	IV	1155475	1155660
r1005	yjm1248	IV	1155475	1155660
r1006	yjm1439	IV	1155475	1155660
r1007	yjm195	IV	1155475	1155660
r1008	yjm693	IV	1155475	1155660
r1009	yjm1307	IV	1155498	1155660
r1010	yjm1342	IV	1155498	1155660
r1011	yjm1418	IV	1155498	1155660
r1012	yjm193	IV	1155498	1155660
r1013	yjm456	IV	1155498	1155660
r1014	yjm1389	IV	1155528	1155660
r1015	yjm1592	IV	1155531	1155660
r1016	yjm1387	IV	1155606	1155754
r1017	yjm1078	IV	1155706	1158952
r1018	yjm1252	IV	1155706	1158952
r1019	yjm248	IV	1155706	1158952
r1020	yjm1078	IV	1158964	1159658
r1021	yjm248	IV	1158964	1159796
r1022	yjm1388	IV	1159859	1159974
r1023	yjm1078	IV	1159991	1160152
r1024	yjm1078	IV	1160309	1160558
r1025	yjm320	IV	1160543	1160649
r1026	yjm450	IV	1160543	1160649
r1027	yjm682	IV	1160543	1160649
r1028	yjm1199	IV	1160596	1160649
r1029	yjm1202	IV	1160596	1160649
r1030	yjm1208	IV	1160596	1160649
r1031	yjm1273	IV	1160596	1160717
r1032	yjm1304	IV	1160596	1160649
r1033	yjm1326	IV	1160596	1160649
r1034	yjm1401	IV	1160596	1160717
r1035	yjm1402	IV	1160596	1160649
r1036	yjm1418	IV	1160596	1160649

region id	strain	chromosome	start	end
r1037	yjm1419	IV	1160596	1160649
r1038	yjm1433	IV	1160596	1160649
r1039	yjm1434	IV	1160596	1160717
r1040	yjm1443	IV	1160596	1160717
r1041	yjm1444	IV	1160596	1160649
r1042	yjm1447	IV	1160596	1160717
r1043	yjm1463	IV	1160596	1160717
r1044	yjm1549	IV	1160596	1160649
r1045	yjm1573	IV	1160596	1160649
r1046	yjm1615	IV	1160596	1160649
r1047	yjm193	IV	1160596	1160732
r1048	yjm271	IV	1160596	1160732
r1049	yjm451	IV	1160596	1160649
r1050	yjm555	IV	1160596	1160649
r1051	yjm681	IV	1160596	1160649
r1052	yjm683	IV	1160596	1160649
r1053	yjm689	IV	1160596	1160649
r1054	yjm693	IV	1160596	1160649
r1055	yjm1078	IV	1160761	1160969
r1056	yjm1252	IV	1160761	1160969
r1057	yjm1078	IV	1161011	1161167
r1058	yjm1252	IV	1161011	1162649
r1059	yjm248	IV	1161011	1161167
r1060	yjm1199	IV	1161205	1161363
r1061	yjm1202	IV	1161205	1161363
r1062	yjm1208	IV	1161205	1161363
r1063	yjm1273	IV	1161205	1161363
r1064	yjm1304	IV	1161205	1161363
r1065	yjm1326	IV	1161205	1161363
r1066	yjm1342	IV	1161205	1161363
r1067	yjm1388	IV	1161205	1161363
r1068	yjm1389	IV	1161205	1161363
r1069	yjm1401	IV	1161205	1161363
r1070	yjm1402	IV	1161205	1161363
r1071	yjm1418	IV	1161205	1161363
r1072	yjm1419	IV	1161205	1161363
r1073	yjm1433	IV	1161205	1161363
r1074	yjm1434	IV	1161205	1161363
r1075	yjm1443	IV	1161205	1161363
r1076	yjm1444	IV	1161205	1161363
r1077	yjm1549	IV	1161205	1161363
r1078	yjm1573	IV	1161205	1161363
r1079	yjm1592	IV	1161205	1161363
r1080	yjm1615	IV	1161205	1161363
r1081	yjm193	IV	1161205	1161363
r1082	yjm271	IV	1161205	1161363
r1083	yjm451	IV	1161205	1161363
r1084	yjm555	IV	1161205	1161363
r1085	yjm627	IV	1161205	1161363
r1086	yjm681	IV	1161205	1161363
r1087	yjm683	IV	1161205	1161363
r1088	yjm689	IV	1161205	1161363
r1089	yjm693	IV	1161205	1161363
r1090	yjm1248	IV	1161212	1161363
r1091	yjm1439	IV	1161212	1161363
r1092	yjm195	IV	1161212	1161363
r1093	yjm1463	IV	1161241	1161363
r1094	yjm1078	IV	1161406	1162649
r1095	yjm248	IV	1161406	1162649
r1096	yjm1078	IV	1162676	1162965
r1097	yjm1252	IV	1162676	1162965
r1098	yjm248	IV	1162750	1162965
r1099	yjm1078	IV	1163313	1164186
r1100	yjm1252	IV	1163313	1164186
r1101	yjm248	IV	1163313	1164186
r1102	yjm1078	IV	1164453	1171656
r1103	yjm1252	IV	1164453	1173132
r1104	yjm248	IV	1164453	1171656
r1105	yjm1078	IV	1268257	1273523
r1106	yjm1252	IV	1268257	1273523
r1107	yjm248	IV	1268257	1273523
r1108	yjm1083	IV	1307666	1307716
r1109	yjm1199	IV	1307666	1307716
r1110	yjm1202	IV	1307666	1307716
r1111	yjm1208	IV	1307666	1307716
r1112	yjm1388	IV	1307666	1307716
r1113	yjm1433	IV	1307666	1307721
r1114	yjm1444	IV	1307666	1307716
r1115	yjm1460	IV	1307666	1307716
r1116	yjm1549	IV	1307666	1307721
r1117	yjm1615	IV	1307666	1307716
r1118	yjm320	IV	1307666	1307716
r1119	yjm451	IV	1307666	1307721
r1120	yjm541	IV	1307666	1307716
r1121	yjm555	IV	1307666	1307716
r1122	yjm1573	IV	1307669	1307714
r1123	yjm1401	IV	1352901	1352967
r1124	yjm975	IV	1460797	1460931
r1125	yjm978	IV	1460797	1460931
r1126	yjm981	IV	1460797	1460931
r1127	yjm987	IV	1460797	1460931
r1128	yjm993	IV	1460797	1460931
r1129	yjm996	IV	1460797	1460931
r1130	yjm1244	IV	1461967	1462015
r1131	yjm1479	IV	1467034	1467064

region id	strain	chromosome	start	end
r1132	yjm1400	IV	1469396	1469554
r1133	yjm1479	IV	1469398	1469554
r1134	yjm1078	IV	1477834	1490900
r1135	yjm1252	IV	1477834	1485568
r1136	yjm248	IV	1477834	1485568
r1137	yjm248	IV	1486399	1490900
r1138	yjm1252	IV	1487964	1490900
r1139	yjm1252	IV	1490922	1503752
r1140	yjm1078	IV	1490964	1503752
r1141	yjm248	IV	1490964	1503752
r1142	yjm1252	IV	1504000	1518287
r1143	yjm248	IV	1504053	1504381
r1144	yjm248	IV	1504480	1518287
r1145	yjm1078	IV	1504481	1518287
r1146	yjm248	IV	1525053	1525097
r1147	yjm470	V	18890	19548
r1148	yjm1078	V	20046	20367
r1149	yjm1252	V	20097	20157
r1150	yjm1252	V	20292	22340
r1151	yjm1078	V	21221	38199
r1152	yjm248	V	21221	38199
r1153	yjm1342	V	21289	21557
r1154	yjm1252	V	22664	38199
r1155	yjm1078	V	167344	169057
r1156	yjm1252	V	167344	169057
r1157	yjm1078	V	169082	174008
r1158	yjm1252	V	169094	174008
r1159	yjm248	V	171779	172813
r1160	yjm1078	V	174034	174935
r1161	yjm1252	V	174034	174935
r1162	yjm248	V	174034	174935
r1163	yjm1078	V	175387	175477
r1164	yjm1078	V	176094	176222
r1165	yjm1078	V	177171	178175
r1166	yjm1252	V	177171	178175
r1167	yjm248	V	177171	178112
r1168	yjm1078	V	272724	281150
r1169	yjm1252	V	272724	281150
r1170	yjm248	V	272724	281150
r1171	yjm1078	V	281160	293195
r1172	yjm1252	V	281160	293195
r1173	yjm248	V	281160	293195
r1174	yjm1078	V	303116	305228
r1175	yjm1252	V	303116	305228
r1176	yjm248	V	303116	305228
r1177	yjm1078	V	310719	311270
r1178	yjm1252	V	310719	311270
r1179	yjm248	V	310719	311270
r1180	yjm1078	V	311355	311456
r1181	yjm1252	V	311355	311456
r1182	yjm248	V	311355	311456
r1183	yjm470	V	311613	312495
r1184	yjm1078	V	311696	313618
r1185	yjm1252	V	311696	313618
r1186	yjm248	V	311696	313618
r1187	yjm1401	V	311762	312150
r1188	yjm428	V	311785	312495
r1189	yjm195	V	311788	312495
r1190	yjm1307	V	311836	312495
r1191	yjm1342	V	311836	312495
r1192	yjm1388	V	311836	312495
r1193	yjm1389	V	311836	312150
r1194	yjm1450	V	311836	312495
r1195	yjm1460	V	311836	312495
r1196	yjm1592	V	311836	312150
r1197	yjm682	V	311836	312495
r1198	yjm683	V	311836	312495
r1199	yjm470	V	313587	316221
r1200	yjm1248	V	435139	435225
r1201	yjm1439	V	435139	435225
r1202	yjm1248	V	435305	435344
r1203	yjm1439	V	435305	435344
r1204	yjm1388	V	435308	435344
r1205	yjm451	V	435308	435344
r1206	yjm682	V	435885	435998
r1207	yjm683	V	435885	435998
r1208	yjm1252	V	501091	508420
r1209	yjm1078	V	519698	522360
r1210	yjm1252	V	519698	532988
r1211	yjm248	V	519698	532988
r1212	yjm1342	V	560640	566138
r1213	yjm470	V	560640	566138
r1214	yjm1129	V	561144	572598
r1215	yjm1133	V	561144	565207
r1216	yjm1190	V	561144	565207
r1217	yjm1199	V	561144	572057
r1218	yjm1202	V	561144	572598
r1219	yjm1242	V	561144	572598
r1220	yjm1244	V	561144	571067
r1221	yjm1250	V	561144	565207
r1222	yjm1304	V	561144	565207
r1223	yjm1311	V	561144	572598
r1224	yjm1326	V	561144	572447
r1225	yjm1332	V	561144	571721
r1226	yjm1336	V	561144	572433

region id	strain	chromosome	start	end
r1227	yjm1338	V	561144	561485
r1228	yjm1341	V	561144	568970
r1229	yjm1356	V	561144	571067
r1230	yjm1387	V	561144	571067
r1231	yjm1415	V	561144	572598
r1232	yjm1417	V	561144	572598
r1233	yjm1433	V	561144	571067
r1234	yjm1444	V	561144	569833
r1235	yjm1477	V	561144	572598
r1236	yjm1478	V	561144	572598
r1237	yjm1526	V	561144	572598
r1238	yjm1527	V	561144	571067
r1239	yjm1549	V	561144	565207
r1240	yjm1574	V	561144	572150
r1241	yjm189	V	561144	572447
r1242	yjm244	V	561144	572598
r1243	yjm248	V	561144	571067
r1244	yjm450	V	561144	572433
r1245	yjm453	V	561144	570075
r1246	yjm541	V	561144	561460
r1247	yjm554	V	561144	561460
r1248	yjm555	V	561144	561460
r1249	yjm627	V	561144	572598
r1250	yjm693	V	561144	572433
r1251	yjm969	V	561144	568460
r1252	yjm972	V	561144	572447
r1253	yjm975	V	561144	571067
r1254	yjm978	V	561144	572447
r1255	yjm981	V	561144	571067
r1256	yjm984	V	561144	570982
r1257	yjm987	V	561144	572598
r1258	yjm990	V	561144	571067
r1259	yjm993	V	561144	572598
r1260	yjm996	V	561144	572447
r1261	yjm541	V	561544	565207
r1262	yjm554	V	561544	565207
r1263	yjm555	V	561544	565207
r1264	yjm1386	V	561701	566138
r1265	yjm195	V	561701	565006
r1266	yjm689	V	562517	572598
r1267	yjm1399	V	562880	563512
r1268	yjm1381	V	562915	572433
r1269	yjm682	V	562915	563969
r1270	yjm683	V	562915	563969
r1271	yjm1383	V	563248	563619
r1272	yjm1385	V	564808	565207
r1273	yjm193	V	565207	571067
r1274	yjm271	V	565207	571067
r1275	yjm1304	V	565650	569283
r1276	yjm1385	V	565650	569640
r1277	yjm541	V	565650	569789
r1278	yjm554	V	565650	568896
r1279	yjm195	V	565736	566138
r1280	yjm1386	V	567019	568460
r1281	yjm195	V	567019	567742
r1282	yjm470	V	567019	568848
r1283	yjm195	V	567854	568460
r1284	yjm555	V	571896	572598
r1285	yjm554	V	571931	572598
r1286	yjm1129	V	572823	573248
r1287	yjm1202	V	572823	573020
r1288	yjm1242	V	572823	575057
r1289	yjm1311	V	572823	573020
r1290	yjm1415	V	572823	574993
r1291	yjm1417	V	572823	574244
r1292	yjm1477	V	572823	575057
r1293	yjm1478	V	572823	575496
r1294	yjm1526	V	572823	575219
r1295	yjm244	V	572823	575496
r1296	yjm554	V	572823	575101
r1297	yjm555	V	572823	575496
r1298	yjm627	V	572823	575034
r1299	yjm689	V	572823	573020
r1300	yjm987	V	572823	574993
r1301	yjm993	V	572823	573113
r1302	yjm1202	V	573206	575496
r1303	yjm1311	V	573206	575057
r1304	yjm1417	V	574450	575101
r1305	yjm1415	V	575120	575496
r1306	yjm1242	V	575132	575496
r1307	yjm1417	V	575165	575264
r1308	yjm627	V	575165	575219
r1309	yjm1477	V	575168	575496
r1310	yjm682	VI	4974	5154
r1311	yjm682	VI	5239	5454
r1312	yjm450	VI	13251	13455
r1313	yjm969	VI	13341	15032
r1314	yjm1190	VI	13746	15032
r1315	yjm1381	VI	13954	15032
r1316	yjm1399	VI	14403	16829
r1317	yjm1273	VI	14427	14511
r1318	yjm1342	VI	14427	14511
r1319	yjm1402	VI	14427	14511
r1320	yjm1418	VI	14427	14511
r1321	yjm1434	VI	14427	14511

region id	strain	chromosome	start	end
r1322	yjm1573	VI	14427	14511
r1323	yjm320	VI	14427	14511
r1324	yjm456	VI	14427	14511
r1325	yjm554	VI	14427	14511
r1326	yjm1190	VI	15110	15317
r1327	yjm1381	VI	15110	16829
r1328	yjm969	VI	15110	15317
r1329	yjm1190	VI	15392	16829
r1330	yjm969	VI	15392	16829
r1331	yjm248	VI	60883	61060
r1332	yjm1252	VI	68317	75107
r1333	yjm248	VI	72120	75107
r1334	yjm1252	VI	97299	103868
r1335	yjm1078	VI	111036	118535
r1336	yjm1252	VI	115087	118069
r1337	yjm1078	VI	118562	121848
r1338	yjm1078	VI	122229	123762
r1339	yjm248	VI	162798	164936
r1340	yjm681	VI	192044	192171
r1341	yjm1078	VI	248548	249960
r1342	yjm1252	VI	248548	252650
r1343	yjm248	VI	248548	249960
r1344	yjm1399	VI	252773	259527
r1345	yjm1078	VI	253076	253672
r1346	yjm1252	VI	253076	253672
r1347	yjm248	VI	253076	253672
r1348	yjm1252	VI	253864	254503
r1349	yjm1078	VI	253924	254503
r1350	yjm248	VI	253924	254503
r1351	yjm1252	VI	254848	270138
r1352	yjm248	VI	254848	269731
r1353	yjm1399	VI	259799	260058
r1354	yjm1399	VI	260494	260673
r1355	yjm681	VI	268665	268751
r1356	yjm1592	VI	269765	269791
r1357	yjm1307	VII	377	403
r1358	yjm1415	VII	377	403
r1359	yjm270	VII	377	403
r1360	yjm681	VII	6602	6731
r1361	yjm1338	VII	6962	7109
r1362	yjm1418	VII	6962	7088
r1363	yjm683	VII	6962	7088
r1364	yjm693	VII	8009	8406
r1365	yjm244	VII	8323	8406
r1366	yjm1326	VII	8332	8390
r1367	yjm1402	VII	8332	8390
r1368	yjm1419	VII	8332	8390
r1369	yjm1434	VII	8332	8390
r1370	yjm1450	VII	8332	8406
r1371	yjm1463	VII	8332	8406
r1372	yjm1477	VII	8332	8406
r1373	yjm1573	VII	8332	8390
r1374	yjm450	VII	8332	8379
r1375	yjm1307	VII	8357	8406
r1376	yjm271	VII	8357	8390
r1377	yjm456	VII	8357	8390
r1378	yjm541	VII	8357	8406
r1379	yjm1450	VII	10799	10855
r1380	yjm1447	VII	10802	10876
r1381	yjm1083	VII	10803	10876
r1382	yjm1190	VII	10803	10876
r1383	yjm1199	VII	10803	10855
r1384	yjm1202	VII	10803	10855
r1385	yjm1208	VII	10803	10876
r1386	yjm1250	VII	10803	10855
r1387	yjm1273	VII	10803	10876
r1388	yjm1304	VII	10803	10876
r1389	yjm1336	VII	10803	10855
r1390	yjm1338	VII	10803	10876
r1391	yjm1341	VII	10803	10855
r1392	yjm1342	VII	10803	10876
r1393	yjm1355	VII	10803	10855
r1394	yjm1381	VII	10803	10855
r1395	yjm1388	VII	10803	10876
r1396	yjm1401	VII	10803	10876
r1397	yjm1419	VII	10803	10876
r1398	yjm1434	VII	10803	10876
r1399	yjm1443	VII	10803	10876
r1400	yjm1444	VII	10803	10855
r1401	yjm1463	VII	10803	10892
r1402	yjm1477	VII	10803	10855
r1403	yjm1549	VII	10803	10855
r1404	yjm1573	VII	10803	10876
r1405	yjm1574	VII	10803	10855
r1406	yjm195	VII	10803	10876
r1407	yjm320	VII	10803	10876
r1408	yjm326	VII	10803	10876
r1409	yjm428	VII	10803	10855
r1410	yjm450	VII	10803	10876
r1411	yjm456	VII	10803	10876
r1412	yjm470	VII	10803	10876
r1413	yjm541	VII	10803	10855
r1414	yjm683	VII	10803	10876
r1415	yjm693	VII	10803	10892
r1416	yjm969	VII	10803	10855

region id	strain	chromosome	start	end
r1417	yjm972	VII	10803	10855
r1418	yjm978	VII	10803	10855
r1419	yjm981	VII	10803	10855
r1420	yjm993	VII	10803	10855
r1421	yjm996	VII	10803	10855
r1422	yjm1383	VII	11163	11285
r1423	yjm541	VII	11163	11255
r1424	yjm554	VII	11163	11255
r1425	yjm555	VII	11163	11285
r1426	yjm681	VII	11163	11285
r1427	yjm693	VII	11207	11525
r1428	yjm428	VII	20913	21040
r1429	yjm693	VII	20913	21512
r1430	yjm1447	VII	20916	21040
r1431	yjm1083	VII	21217	21351
r1432	yjm428	VII	21217	21512
r1433	yjm470	VII	21217	21524
r1434	yjm1342	VII	21317	21512
r1435	yjm1273	VII	21327	21533
r1436	yjm1326	VII	21327	21512
r1437	yjm1385	VII	21327	21533
r1438	yjm1400	VII	21327	21512
r1439	yjm1402	VII	21327	21533
r1440	yjm1418	VII	21327	21524
r1441	yjm1434	VII	21327	21533
r1442	yjm1573	VII	21327	21533
r1443	yjm1574	VII	21327	21512
r1444	yjm195	VII	21327	21492
r1445	yjm270	VII	21327	21512
r1446	yjm326	VII	21327	21512
r1447	yjm541	VII	21327	21512
r1448	yjm554	VII	21327	21512
r1449	yjm683	VII	21327	21512
r1450	yjm555	VII	21349	21512
r1451	yjm681	VII	21349	21512
r1452	yjm1248	VII	21663	21721
r1453	yjm1326	VII	21663	21721
r1454	yjm1439	VII	21663	21721
r1455	yjm1574	VII	21663	21721
r1456	yjm195	VII	21663	21721
r1457	yjm270	VII	21663	21721
r1458	yjm326	VII	21663	21721
r1459	yjm627	VII	21663	21721
r1460	yjm683	VII	21663	21721
r1461	yjm1381	VII	110800	110847
r1462	yjm693	VII	323737	324061
r1463	yjm1419	VII	324534	324753
r1464	yjm1573	VII	324534	324753
r1465	yjm450	VII	324534	324753
r1466	yjm693	VII	324534	324804
r1467	yjm1083	VII	324600	324753
r1468	yjm1208	VII	324600	324753
r1469	yjm1401	VII	324600	324753
r1470	yjm1433	VII	324600	324753
r1471	yjm195	VII	324600	324753
r1472	yjm195	VII	327906	327994
r1473	yjm1083	VII	385518	385569
r1474	yjm1129	VII	385518	385569
r1475	yjm1133	VII	385518	385569
r1476	yjm1190	VII	385518	385569
r1477	yjm1199	VII	385518	385618
r1478	yjm1202	VII	385518	385618
r1479	yjm1208	VII	385518	385569
r1480	yjm1242	VII	385518	385569
r1481	yjm1244	VII	385518	385569
r1482	yjm1248	VII	385518	385569
r1483	yjm1250	VII	385518	385569
r1484	yjm1273	VII	385518	385569
r1485	yjm1304	VII	385518	385569
r1486	yjm1307	VII	385518	385569
r1487	yjm1311	VII	385518	385618
r1488	yjm1326	VII	385518	385569
r1489	yjm1332	VII	385518	385569
r1490	yjm1336	VII	385518	385569
r1491	yjm1338	VII	385518	385569
r1492	yjm1341	VII	385518	385569
r1493	yjm1342	VII	385518	385569
r1494	yjm1355	VII	385518	385569
r1495	yjm1356	VII	385518	385569
r1496	yjm1381	VII	385518	385618
r1497	yjm1383	VII	385518	385618
r1498	yjm1385	VII	385518	385618
r1499	yjm1386	VII	385518	385569
r1500	yjm1387	VII	385518	385569
r1501	yjm1388	VII	385518	385569
r1502	yjm1389	VII	385518	385569
r1503	yjm1399	VII	385518	385569
r1504	yjm1400	VII	385518	385569
r1505	yjm1401	VII	385518	385569
r1506	yjm1402	VII	385518	385618
r1507	yjm1415	VII	385518	385569
r1508	yjm1417	VII	385518	385569
r1509	yjm1418	VII	385518	385569
r1510	yjm1419	VII	385518	385618
r1511	yjm1433	VII	385518	385569

region id	strain	chromosome	start	end
r1512	yjm1434	VII	385518	385569
r1513	yjm1439	VII	385518	385569
r1514	yjm1443	VII	385518	385569
r1515	yjm1444	VII	385518	385569
r1516	yjm1450	VII	385518	385569
r1517	yjm1460	VII	385518	385569
r1518	yjm1463	VII	385518	385569
r1519	yjm1477	VII	385518	385569
r1520	yjm1478	VII	385518	385569
r1521	yjm1479	VII	385518	385569
r1522	yjm1526	VII	385518	385569
r1523	yjm1527	VII	385518	385569
r1524	yjm1549	VII	385518	385569
r1525	yjm1573	VII	385518	385618
r1526	yjm1574	VII	385518	385569
r1527	yjm1592	VII	385518	385569
r1528	yjm1615	VII	385518	385569
r1529	yjm189	VII	385518	385569
r1530	yjm193	VII	385518	385569
r1531	yjm195	VII	385518	385569
r1532	yjm244	VII	385518	385569
r1533	yjm270	VII	385518	385569
r1534	yjm271	VII	385518	385569
r1535	yjm320	VII	385518	385569
r1536	yjm326	VII	385518	385569
r1537	yjm428	VII	385518	385569
r1538	yjm450	VII	385518	385618
r1539	yjm451	VII	385518	385569
r1540	yjm453	VII	385518	385569
r1541	yjm456	VII	385518	385569
r1542	yjm470	VII	385518	385569
r1543	yjm541	VII	385518	385569
r1544	yjm554	VII	385518	385569
r1545	yjm555	VII	385518	385569
r1546	yjm627	VII	385518	385569
r1547	yjm681	VII	385518	385569
r1548	yjm682	VII	385518	385569
r1549	yjm683	VII	385518	385569
r1550	yjm689	VII	385518	385569
r1551	yjm693	VII	385518	385569
r1552	yjm969	VII	385518	385569
r1553	yjm972	VII	385518	385569
r1554	yjm975	VII	385518	385569
r1555	yjm978	VII	385518	385569
r1556	yjm981	VII	385518	385569
r1557	yjm984	VII	385518	385569
r1558	yjm987	VII	385518	385569
r1559	yjm990	VII	385518	385569
r1560	yjm993	VII	385518	385569
r1561	yjm996	VII	385518	385569
r1562	yjm1078	VII	399493	402273
r1563	yjm1252	VII	399493	401981
r1564	yjm248	VII	399493	401981
r1565	yjm1399	VII	401974	402281
r1566	yjm1252	VII	402034	402273
r1567	yjm248	VII	402046	402273
r1568	yjm1479	VII	402072	402128
r1569	yjm1252	VII	405237	408278
r1570	yjm248	VII	405237	408278
r1571	yjm1078	VII	405393	408278
r1572	yjm1078	VII	416839	416985
r1573	yjm248	VII	416839	416985
r1574	yjm1252	VII	416844	420785
r1575	yjm1078	VII	417402	420785
r1576	yjm248	VII	417402	420785
r1577	yjm1078	VII	420804	430918
r1578	yjm1252	VII	420804	430918
r1579	yjm248	VII	420804	430918
r1580	yjm1252	VII	452651	452717
r1581	yjm1252	VII	452882	458796
r1582	yjm1447	VII	480161	480626
r1583	yjm451	VII	480164	480383
r1584	yjm1199	VII	480191	480383
r1585	yjm1248	VII	480191	480443
r1586	yjm1273	VII	480191	480611
r1587	yjm1338	VII	480191	480611
r1588	yjm1400	VII	480191	480611
r1589	yjm1402	VII	480191	480611
r1590	yjm1418	VII	480191	480626
r1591	yjm1434	VII	480191	480611
r1592	yjm1439	VII	480191	480443
r1593	yjm1443	VII	480191	480383
r1594	yjm1444	VII	480191	480383
r1595	yjm1479	VII	480191	480611
r1596	yjm1573	VII	480191	480611
r1597	yjm193	VII	480191	480611
r1598	yjm195	VII	480191	480443
r1599	yjm693	VII	480191	480611
r1600	yjm1399	VII	480356	480611
r1601	yjm1208	VII	561966	562159
r1602	yjm1399	VII	561966	562159
r1603	yjm1615	VII	561966	562159
r1604	yjm1252	VII	588422	589536
r1605	yjm1450	VII	609643	609733
r1606	yjm1078	VII	838109	838561

region id	strain	chromosome	start	end
r1607	yjm248	VII	838109	838561
r1608	yjm1078	VII	877745	878464
r1609	yjm1252	VII	877745	878464
r1610	yjm248	VII	877745	878464
r1611	yjm1250	VII	1061962	1062358
r1612	yjm1383	VII	1061962	1062310
r1613	yjm1250	VII	1062700	1067314
r1614	yjm1250	VII	1067629	1068247
r1615	yjm1250	VII	1068343	1070343
r1616	yjm1242	VII	1070276	1071421
r1617	yjm1336	VII	1070276	1071638
r1618	yjm1341	VII	1070276	1071272
r1619	yjm1385	VII	1070276	1071638
r1620	yjm1477	VII	1070276	1071421
r1621	yjm1574	VII	1070276	1071602
r1622	yjm627	VII	1070276	1071638
r1623	yjm972	VII	1070276	1071638
r1624	yjm978	VII	1070276	1071602
r1625	yjm981	VII	1070276	1071638
r1626	yjm993	VII	1070276	1071638
r1627	yjm984	VII	1070354	1071602
r1628	yjm990	VII	1070354	1071602
r1629	yjm1381	VII	1070398	1071272
r1630	yjm682	VII	1070417	1071272
r1631	yjm1078	VII	1070850	1070953
r1632	yjm1252	VII	1070861	1070953
r1633	yjm1078	VII	1071076	1071270
r1634	yjm987	VII	1071076	1071270
r1635	yjm1252	VII	1071092	1071270
r1636	yjm248	VII	1071092	1071270
r1637	yjm1341	VII	1071387	1071602
r1638	yjm1381	VII	1071387	1071638
r1639	yjm1208	VII	1081847	1081915
r1640	yjm450	VII	1081847	1081915
r1641	yjm1083	VII	1084102	1084161
r1642	yjm1252	VIII	45505	47737
r1643	yjm1083	VIII	93416	93557
r1644	yjm1190	VIII	93416	93557
r1645	yjm1304	VIII	93416	93557
r1646	yjm1311	VIII	93416	93557
r1647	yjm1342	VIII	93416	93664
r1648	yjm1355	VIII	93416	93573
r1649	yjm1381	VIII	93416	93557
r1650	yjm1402	VIII	93416	93557
r1651	yjm1417	VIII	93416	93557
r1652	yjm1434	VIII	93416	93557
r1653	yjm1443	VIII	93416	93573
r1654	yjm1450	VIII	93416	93573
r1655	yjm1463	VIII	93416	93557
r1656	yjm1478	VIII	93416	93557
r1657	yjm1573	VIII	93416	93557
r1658	yjm193	VIII	93416	93557
r1659	yjm271	VIII	93416	93557
r1660	yjm450	VIII	93416	93557
r1661	yjm554	VIII	93416	93557
r1662	yjm681	VIII	93416	93557
r1663	yjm689	VIII	93416	93573
r1664	yjm693	VIII	93416	93573
r1665	yjm975	VIII	93416	93557
r1666	yjm984	VIII	93416	93573
r1667	yjm990	VIII	93416	93557
r1668	yjm993	VIII	93416	93557
r1669	yjm1342	VIII	93823	94980
r1670	yjm627	VIII	93823	95002
r1671	yjm1078	VIII	93838	94270
r1672	yjm1083	VIII	93838	94389
r1673	yjm1129	VIII	93838	94330
r1674	yjm1133	VIII	93838	94980
r1675	yjm1190	VIII	93838	94980
r1676	yjm1199	VIII	93838	94330
r1677	yjm1202	VIII	93838	94980
r1678	yjm1208	VIII	93838	94999
r1679	yjm1242	VIII	93838	94330
r1680	yjm1244	VIII	93838	94980
r1681	yjm1248	VIII	93838	95002
r1682	yjm1250	VIII	93838	94330
r1683	yjm1252	VIII	93838	94098
r1684	yjm1273	VIII	93838	94330
r1685	yjm1304	VIII	93838	94980
r1686	yjm1311	VIII	93838	94980
r1687	yjm1326	VIII	93838	94980
r1688	yjm1332	VIII	93838	94980
r1689	yjm1336	VIII	93838	94330
r1690	yjm1338	VIII	93838	94330
r1691	yjm1341	VIII	93838	94330
r1692	yjm1355	VIII	93838	94330
r1693	yjm1356	VIII	93838	94389
r1694	yjm1381	VIII	93838	94389
r1695	yjm1383	VIII	93838	94330
r1696	yjm1385	VIII	93838	94330
r1697	yjm1386	VIII	93838	94330
r1698	yjm1387	VIII	93838	94389
r1699	yjm1388	VIII	93838	94330
r1700	yjm1389	VIII	93838	95007
r1701	yjm1399	VIII	93838	94330

region id	strain	chromosome	start	end
r1702	yjm1402	VIII	93838	94980
r1703	yjm1415	VIII	93838	94976
r1704	yjm1417	VIII	93838	94980
r1705	yjm1418	VIII	93838	94330
r1706	yjm1433	VIII	93838	94330
r1707	yjm1434	VIII	93838	94980
r1708	yjm1439	VIII	93838	95002
r1709	yjm1443	VIII	93838	94330
r1710	yjm1444	VIII	93838	94330
r1711	yjm1447	VIII	93838	94980
r1712	yjm1450	VIII	93838	94999
r1713	yjm1460	VIII	93838	94999
r1714	yjm1463	VIII	93838	94980
r1715	yjm1477	VIII	93838	94270
r1716	yjm1478	VIII	93838	94389
r1717	yjm1526	VIII	93838	94330
r1718	yjm1527	VIII	93838	94330
r1719	yjm1549	VIII	93838	94330
r1720	yjm1573	VIII	93838	94980
r1721	yjm1574	VIII	93838	94330
r1722	yjm1592	VIII	93838	94330
r1723	yjm1615	VIII	93838	94999
r1724	yjm189	VIII	93838	94270
r1725	yjm193	VIII	93838	94980
r1726	yjm195	VIII	93838	94976
r1727	yjm244	VIII	93838	94330
r1728	yjm248	VIII	93838	94261
r1729	yjm270	VIII	93838	94330
r1730	yjm271	VIII	93838	94330
r1731	yjm320	VIII	93838	94980
r1732	yjm326	VIII	93838	94330
r1733	yjm428	VIII	93838	94330
r1734	yjm450	VIII	93838	94980
r1735	yjm453	VIII	93838	94330
r1736	yjm456	VIII	93838	94999
r1737	yjm470	VIII	93838	94330
r1738	yjm541	VIII	93838	94980
r1739	yjm554	VIII	93838	94980
r1740	yjm555	VIII	93838	94980
r1741	yjm681	VIII	93838	94980
r1742	yjm682	VIII	93838	94330
r1743	yjm683	VIII	93838	94980
r1744	yjm689	VIII	93838	94980
r1745	yjm693	VIII	93838	94389
r1746	yjm969	VIII	93838	94330
r1747	yjm972	VIII	93838	94330
r1748	yjm975	VIII	93838	94980
r1749	yjm978	VIII	93838	94330
r1750	yjm981	VIII	93838	94330
r1751	yjm984	VIII	93838	94980
r1752	yjm987	VIII	93838	94330
r1753	yjm990	VIII	93838	94980
r1754	yjm993	VIII	93838	94980
r1755	yjm996	VIII	93838	94330
r1756	yjm1252	VIII	94161	94261
r1757	yjm1342	VIII	133588	133652
r1758	yjm1355	VIII	190088	191619
r1759	yjm1381	VIII	229723	229927
r1760	yjm1083	VIII	229760	229927
r1761	yjm1342	VIII	229760	229927
r1762	yjm1463	VIII	229760	229927
r1763	yjm693	VIII	229760	229927
r1764	yjm1078	VIII	285603	287284
r1765	yjm1252	VIII	285603	287284
r1766	yjm248	VIII	285603	287284
r1767	yjm1078	VIII	287737	289490
r1768	yjm1252	VIII	287737	302083
r1769	yjm248	VIII	287737	289490
r1770	yjm1083	VIII	291155	291373
r1771	yjm1133	VIII	291155	291373
r1772	yjm1190	VIII	291155	291373
r1773	yjm1199	VIII	291155	291373
r1774	yjm1202	VIII	291155	291373
r1775	yjm1326	VIII	291155	291373
r1776	yjm1419	VIII	291155	291373
r1777	yjm1433	VIII	291155	291373
r1778	yjm1444	VIII	291155	291373
r1779	yjm1478	VIII	291155	291410
r1780	yjm1549	VIII	291155	291373
r1781	yjm428	VIII	291155	291373
r1782	yjm470	VIII	291155	291373
r1783	yjm555	VIII	291155	291373
r1784	yjm681	VIII	291155	291373
r1785	yjm689	VIII	291155	291373
r1786	yjm1078	VIII	291311	291370
r1787	yjm1129	VIII	291311	291410
r1788	yjm1242	VIII	291311	291410
r1789	yjm1244	VIII	291311	291410
r1790	yjm1248	VIII	291311	291410
r1791	yjm1273	VIII	291311	291410
r1792	yjm1307	VIII	291311	291410
r1793	yjm1311	VIII	291311	291410
r1794	yjm1332	VIII	291311	291410
r1795	yjm1336	VIII	291311	291443
r1796	yjm1338	VIII	291311	291410

region id	strain	chromosome	start	end
r1797	yjm1341	VIII	291311	291410
r1798	yjm1342	VIII	291311	291410
r1799	yjm1356	VIII	291311	291410
r1800	yjm1387	VIII	291311	291410
r1801	yjm1388	VIII	291311	291410
r1802	yjm1389	VIII	291311	291410
r1803	yjm1399	VIII	291311	291413
r1804	yjm1400	VIII	291311	291410
r1805	yjm1401	VIII	291311	291410
r1806	yjm1402	VIII	291311	291410
r1807	yjm1415	VIII	291311	291410
r1808	yjm1417	VIII	291311	291410
r1809	yjm1418	VIII	291311	291425
r1810	yjm1434	VIII	291311	291410
r1811	yjm1439	VIII	291311	291410
r1812	yjm1443	VIII	291311	291410
r1813	yjm1447	VIII	291311	291410
r1814	yjm1460	VIII	291311	291410
r1815	yjm1463	VIII	291311	291410
r1816	yjm1477	VIII	291311	291410
r1817	yjm1479	VIII	291311	291410
r1818	yjm1527	VIII	291311	291410
r1819	yjm1573	VIII	291311	291410
r1820	yjm1574	VIII	291311	291410
r1821	yjm1592	VIII	291311	291410
r1822	yjm189	VIII	291311	291410
r1823	yjm193	VIII	291311	291410
r1824	yjm195	VIII	291311	291410
r1825	yjm244	VIII	291311	291410
r1826	yjm248	VIII	291311	291370
r1827	yjm270	VIII	291311	291410
r1828	yjm271	VIII	291311	291410
r1829	yjm451	VIII	291311	291410
r1830	yjm453	VIII	291311	291410
r1831	yjm627	VIII	291311	291410
r1832	yjm969	VIII	291311	291410
r1833	yjm972	VIII	291311	291410
r1834	yjm975	VIII	291311	291410
r1835	yjm978	VIII	291311	291410
r1836	yjm981	VIII	291311	291410
r1837	yjm984	VIII	291311	291410
r1838	yjm987	VIII	291311	291410
r1839	yjm990	VIII	291311	291410
r1840	yjm993	VIII	291311	291410
r1841	yjm996	VIII	291311	291410
r1842	yjm1078	VIII	296754	302083
r1843	yjm248	VIII	296754	302083
r1844	yjm1252	VIII	362914	367255
r1845	yjm1381	VIII	389123	389169
r1846	yjm1078	VIII	452764	456995
r1847	yjm248	VIII	452764	456995
r1848	yjm1400	VIII	463887	465243
r1849	yjm1479	VIII	463887	465243
r1850	yjm627	VIII	464896	466484
r1851	yjm1332	VIII	465578	466498
r1852	yjm1418	VIII	465768	466142
r1853	yjm1190	VIII	465995	466484
r1854	yjm1242	VIII	465995	466484
r1855	yjm1244	VIII	465995	466498
r1856	yjm1304	VIII	465995	466484
r1857	yjm1341	VIII	465995	466484
r1858	yjm1356	VIII	465995	466484
r1859	yjm1386	VIII	465995	466484
r1860	yjm1387	VIII	465995	466484
r1861	yjm1402	VIII	465995	466484
r1862	yjm1415	VIII	465995	466484
r1863	yjm1417	VIII	465995	466484
r1864	yjm1433	VIII	465995	466484
r1865	yjm1463	VIII	465995	466484
r1866	yjm1478	VIII	465995	466484
r1867	yjm1527	VIII	465995	466484
r1868	yjm1573	VIII	465995	466484
r1869	yjm193	VIII	465995	466484
r1870	yjm195	VIII	465995	466484
r1871	yjm244	VIII	465995	466484
r1872	yjm271	VIII	465995	466484
r1873	yjm428	VIII	465995	466484
r1874	yjm451	VIII	465995	466484
r1875	yjm689	VIII	465995	466484
r1876	yjm693	VIII	465995	466484
r1877	yjm975	VIII	465995	466498
r1878	yjm984	VIII	465995	466484
r1879	yjm990	VIII	465995	466498
r1880	yjm993	VIII	465995	466484
r1881	yjm1443	VIII	466142	466497
r1882	yjm1248	VIII	466249	466484
r1883	yjm1439	VIII	466249	466484
r1884	yjm1199	VIII	466268	466484
r1885	yjm1336	VIII	466268	466498
r1886	yjm1355	VIII	466268	466484
r1887	yjm1400	VIII	466268	466484
r1888	yjm1477	VIII	466268	466484
r1889	yjm1479	VIII	466268	466484
r1890	yjm1549	VIII	466268	466484
r1891	yjm1574	VIII	466268	466484

region id	strain	chromosome	start	end
r1892	yjm270	VIII	466268	466484
r1893	yjm682	VIII	466268	466484
r1894	yjm987	VIII	466268	466484
r1895	yjm996	VIII	466268	466484
r1896	yjm1307	VIII	466323	466484
r1897	yjm1311	VIII	466323	466484
r1898	yjm1381	VIII	466323	466484
r1899	yjm1460	VIII	466323	466484
r1900	yjm450	VIII	466323	466484
r1901	yjm554	VIII	466323	466484
r1902	yjm1078	VIII	466333	466484
r1903	yjm1133	VIII	466333	466484
r1904	yjm1202	VIII	466333	466484
r1905	yjm1383	VIII	466333	466484
r1906	yjm1389	VIII	466333	466484
r1907	yjm1592	VIII	466333	466484
r1908	yjm681	VIII	466333	466484
r1909	yjm1252	VIII	504856	505120
r1910	yjm248	VIII	504856	505120
r1911	yjm1399	VIII	515129	520411
r1912	yjm1399	VIII	520639	522909
r1913	yjm1355	VIII	520945	522909
r1914	yjm1342	VIII	525100	525180
r1915	yjm1083	VIII	525440	525855
r1916	yjm1129	VIII	525440	525855
r1917	yjm1133	VIII	525440	525855
r1918	yjm1190	VIII	525440	525855
r1919	yjm1199	VIII	525440	525855
r1920	yjm1202	VIII	525440	525855
r1921	yjm1242	VIII	525440	525855
r1922	yjm1244	VIII	525440	525855
r1923	yjm1248	VIII	525440	525855
r1924	yjm1273	VIII	525440	525855
r1925	yjm1304	VIII	525440	525855
r1926	yjm1326	VIII	525440	525855
r1927	yjm1332	VIII	525440	525855
r1928	yjm1336	VIII	525440	525855
r1929	yjm1338	VIII	525440	525855
r1930	yjm1341	VIII	525440	525855
r1931	yjm1342	VIII	525440	525901
r1932	yjm1356	VIII	525440	525855
r1933	yjm1381	VIII	525440	525855
r1934	yjm1383	VIII	525440	525855
r1935	yjm1385	VIII	525440	525855
r1936	yjm1386	VIII	525440	525855
r1937	yjm1400	VIII	525440	525855
r1938	yjm1401	VIII	525440	525855
r1939	yjm1402	VIII	525440	525855
r1940	yjm1415	VIII	525440	525855
r1941	yjm1417	VIII	525440	525855
r1942	yjm1418	VIII	525440	525855
r1943	yjm1419	VIII	525440	525855
r1944	yjm1433	VIII	525440	525855
r1945	yjm1434	VIII	525440	525855
r1946	yjm1439	VIII	525440	525855
r1947	yjm1443	VIII	525440	525855
r1948	yjm1447	VIII	525440	525855
r1949	yjm1450	VIII	525440	525855
r1950	yjm1463	VIII	525440	525855
r1951	yjm1477	VIII	525440	525855
r1952	yjm1478	VIII	525440	525855
r1953	yjm1526	VIII	525440	525855
r1954	yjm1527	VIII	525440	525855
r1955	yjm1549	VIII	525440	525855
r1956	yjm1573	VIII	525440	525855
r1957	yjm1574	VIII	525440	525855
r1958	yjm189	VIII	525440	525855
r1959	yjm193	VIII	525440	525855
r1960	yjm195	VIII	525440	525855
r1961	yjm244	VIII	525440	525855
r1962	yjm270	VIII	525440	525855
r1963	yjm271	VIII	525440	525855
r1964	yjm320	VIII	525440	525855
r1965	yjm428	VIII	525440	525855
r1966	yjm450	VIII	525440	525855
r1967	yjm451	VIII	525440	525855
r1968	yjm453	VIII	525440	525855
r1969	yjm456	VIII	525440	525855
r1970	yjm470	VIII	525440	525855
r1971	yjm541	VIII	525440	525855
r1972	yjm555	VIII	525440	525855
r1973	yjm627	VIII	525440	525855
r1974	yjm681	VIII	525440	525855
r1975	yjm682	VIII	525440	525855
r1976	yjm689	VIII	525440	525855
r1977	yjm693	VIII	525440	525855
r1978	yjm969	VIII	525440	525855
r1979	yjm972	VIII	525440	525855
r1980	yjm975	VIII	525440	525855
r1981	yjm978	VIII	525440	525855
r1982	yjm981	VIII	525440	525855
r1983	yjm984	VIII	525440	525855
r1984	yjm987	VIII	525440	525855
r1985	yjm990	VIII	525440	525855
r1986	yjm993	VIII	525440	525855

region id	strain	chromosome	start	end
r1987	yjm996	VIII	525440	525855
r1988	yjm1479	VIII	525577	525855
r1989	yjm1078	VIII	525606	525855
r1990	yjm1252	VIII	525606	525855
r1991	yjm248	VIII	525606	525855
r1992	yjm1387	VIII	526026	526104
r1993	yjm1460	VIII	526026	526104
r1994	yjm1342	VIII	526230	526291
r1995	yjm1244	VIII	526263	526314
r1996	yjm1478	VIII	526263	526314
r1997	yjm975	VIII	526263	526314
r1998	yjm984	VIII	526263	526314
r1999	yjm990	VIII	526263	526314
r2000	yjm993	VIII	526263	526314
r2001	yjm326	VIII	526269	526461
r2002	yjm1463	VIII	526270	526410
r2003	yjm1342	VIII	526365	526467
r2004	yjm1443	VIII	526383	526602
r2005	yjm1399	VIII	526389	526704
r2006	yjm1273	VIII	526398	526602
r2007	yjm683	VIII	526404	526632
r2008	yjm271	VIII	526473	526551
r2009	yjm1402	VIII	526497	526560
r2010	yjm1381	VIII	526677	526794
r2011	yjm555	VIII	526728	526770
r2012	yjm1242	VIII	526731	526770
r2013	yjm1338	VIII	526731	526770
r2014	yjm1574	VIII	526731	526782
r2015	yjm270	VIII	526731	526770
r2016	yjm320	VIII	526731	526770
r2017	yjm681	VIII	526731	526770
r2018	yjm453	VIII	526857	526954
r2019	yjm1400	VIII	526938	527091
r2020	yjm1479	VIII	526938	527091
r2021	yjm1573	VIII	526989	527070
r2022	yjm450	VIII	526998	527091
r2023	yjm1463	VIII	527001	527273
r2024	yjm1450	VIII	527133	527232
r2025	yjm1355	VIII	540288	541986
r2026	yjm1399	VIII	540288	541922
r2027	yjm1383	VIII	540810	540972
r2028	yjm1385	VIII	540810	540972
r2029	yjm1388	VIII	550514	550615
r2030	yjm1447	VIII	550514	550615
r2031	yjm195	VIII	550514	550615
r2032	yjm1388	VIII	550856	551445
r2033	yjm1447	VIII	550856	551445
r2034	yjm195	VIII	550856	551445
r2035	yjm1418	VIII	554071	554279
r2036	yjm1527	IX	9728	44799
r2037	yjm1450	IX	18858	18964
r2038	yjm1615	IX	18858	18964
r2039	yjm554	IX	18891	18964
r2040	yjm1401	IX	20481	20888
r2041	yjm689	IX	20586	20888
r2042	yjm681	IX	20685	20888
r2043	yjm1250	IX	20888	21081
r2044	yjm1202	IX	21064	21195
r2045	yjm1401	IX	21064	21195
r2046	yjm681	IX	21064	21195
r2047	yjm1338	IX	21445	44799
r2048	yjm1433	IX	21445	44799
r2049	yjm693	IX	21445	30991
r2050	yjm1311	IX	24546	24795
r2051	yjm1244	IX	24744	24795
r2052	yjm244	IX	24744	24795
r2053	yjm969	IX	24744	24795
r2054	yjm1133	IX	27089	27131
r2055	yjm1199	IX	27089	27131
r2056	yjm1202	IX	27089	27131
r2057	yjm1208	IX	27089	27131
r2058	yjm1273	IX	27089	27131
r2059	yjm1307	IX	27089	27131
r2060	yjm1326	IX	27089	27131
r2061	yjm1342	IX	27089	27131
r2062	yjm1386	IX	27089	27131
r2063	yjm1399	IX	27089	27131
r2064	yjm1401	IX	27089	27131
r2065	yjm1402	IX	27089	27131
r2066	yjm1418	IX	27089	27131
r2067	yjm1434	IX	27089	27131
r2068	yjm1443	IX	27089	27131
r2069	yjm1444	IX	27089	27131
r2070	yjm1447	IX	27089	27131
r2071	yjm1450	IX	27089	27131
r2072	yjm1460	IX	27089	27131
r2073	yjm1478	IX	27089	27131
r2074	yjm1573	IX	27089	27131
r2075	yjm1615	IX	27089	27131
r2076	yjm320	IX	27089	27131
r2077	yjm456	IX	27089	27131
r2078	yjm554	IX	27089	27131
r2079	yjm555	IX	27089	27131
r2080	yjm681	IX	27089	27131
r2081	yjm682	IX	27089	27131

region id	strain	chromosome	start	end
r2082	yjm689	IX	27089	27131
r2083	yjm972	IX	27089	27131
r2084	yjm1202	IX	37428	37578
r2085	yjm1417	IX	49913	50618
r2086	yjm1202	IX	138638	138982
r2087	yjm1273	IX	138638	138982
r2088	yjm1434	IX	138638	138982
r2089	yjm1078	IX	257104	262104
r2090	yjm248	IX	257104	262104
r2091	yjm1078	IX	266113	270299
r2092	yjm248	IX	266113	270299
r2093	yjm1252	IX	269792	270299
r2094	yjm1078	IX	324075	325465
r2095	yjm1252	IX	324075	325465
r2096	yjm248	IX	324075	332458
r2097	yjm1419	IX	324825	324868
r2098	yjm1248	IX	325046	325129
r2099	yjm1381	IX	325046	325129
r2100	yjm1439	IX	325046	325129
r2101	yjm1252	IX	325499	328929
r2102	yjm1078	IX	325508	330924
r2103	yjm248	IX	333296	334005
r2104	yjm1078	IX	333387	334005
r2105	yjm1252	IX	333387	334005
r2106	yjm1078	IX	334621	335824
r2107	yjm1252	IX	334621	335824
r2108	yjm248	IX	334621	335824
r2109	yjm1252	IX	377249	389752
r2110	yjm1443	IX	388519	388585
r2111	yjm456	IX	388519	388585
r2112	yjm1443	IX	393048	393182
r2113	yjm1078	IX	393055	393105
r2114	yjm248	IX	393055	393105
r2115	yjm1479	IX	394649	394846
r2116	yjm1418	IX	426431	426461
r2117	yjm1381	IX	433911	434233
r2118	yjm1401	IX	433991	434245
r2119	yjm456	IX	433991	434101
r2120	yjm683	IX	433991	434259
r2121	yjm689	IX	433991	434050
r2122	yjm1419	IX	433996	434101
r2123	yjm1526	IX	433996	434155
r2124	yjm1574	IX	433996	434215
r2125	yjm271	IX	433996	434112
r2126	yjm993	IX	433996	434155
r2127	yjm1418	IX	437250	437344
r2128	yjm1418	IX	437630	437840
r2129	yjm1208	IX	438431	438481
r2130	yjm1443	IX	438431	438481
r2131	yjm195	IX	439032	439138
r2132	yjm1208	IX	439053	439131
r2133	yjm1444	IX	439080	439138
r2134	yjm456	IX	439083	439131
r2135	yjm1386	X	18841	18946
r2136	yjm195	X	18841	18946
r2137	yjm1439	X	18928	19047
r2138	yjm1311	X	24343	24441
r2139	yjm1338	X	24343	24441
r2140	yjm1381	X	24343	24441
r2141	yjm1387	X	24343	24487
r2142	yjm1450	X	24343	24487
r2143	yjm1527	X	24343	24441
r2144	yjm541	X	24343	24441
r2145	yjm1387	X	24709	25357
r2146	yjm541	X	24802	25357
r2147	yjm1381	X	24925	24985
r2148	yjm555	X	24925	24995
r2149	yjm1338	X	25132	25357
r2150	yjm1450	X	25132	25357
r2151	yjm1549	X	25132	25357
r2152	yjm1129	X	25150	25357
r2153	yjm1242	X	25150	25357
r2154	yjm1311	X	25150	25357
r2155	yjm1332	X	25150	25357
r2156	yjm1336	X	25150	25357
r2157	yjm1381	X	25150	25358
r2158	yjm1415	X	25150	25357
r2159	yjm1417	X	25150	25357
r2160	yjm1477	X	25150	25357
r2161	yjm1527	X	25150	25357
r2162	yjm990	X	25150	25357
r2163	yjm1433	X	25207	25357
r2164	yjm1463	X	25207	25357
r2165	yjm1447	X	25303	25384
r2166	yjm1190	X	25309	25357
r2167	yjm1199	X	25309	25357
r2168	yjm1202	X	25309	25357
r2169	yjm1244	X	25309	25357
r2170	yjm1341	X	25309	25357
r2171	yjm1356	X	25309	25357
r2172	yjm1383	X	25309	25357
r2173	yjm1385	X	25309	25384
r2174	yjm1526	X	25309	25357
r2175	yjm1615	X	25309	25357
r2176	yjm189	X	25309	25357

region id	strain	chromosome	start	end
r2177	yjm193	X	25309	25357
r2178	yjm244	X	25309	25357
r2179	yjm270	X	25309	25357
r2180	yjm271	X	25309	25357
r2181	yjm450	X	25309	25357
r2182	yjm554	X	25309	25384
r2183	yjm689	X	25309	25357
r2184	yjm972	X	25309	25357
r2185	yjm975	X	25309	25357
r2186	yjm978	X	25309	25357
r2187	yjm981	X	25309	25357
r2188	yjm984	X	25309	25357
r2189	yjm987	X	25309	25357
r2190	yjm993	X	25309	25357
r2191	yjm996	X	25309	25357
r2192	yjm1444	X	30723	33957
r2193	yjm681	X	30723	33831
r2194	yjm1250	X	32680	35262
r2195	yjm248	X	32689	32788
r2196	yjm1252	X	95080	95539
r2197	yjm1078	X	98270	104562
r2198	yjm1252	X	98270	104562
r2199	yjm248	X	98270	104562
r2200	yjm1078	X	105933	107623
r2201	yjm1252	X	105933	107623
r2202	yjm248	X	105933	107623
r2203	yjm1083	X	120867	120922
r2204	yjm1129	X	120867	120922
r2205	yjm1190	X	120867	120922
r2206	yjm1199	X	120867	120922
r2207	yjm1202	X	120867	120922
r2208	yjm1242	X	120867	120922
r2209	yjm1250	X	120867	121044
r2210	yjm1273	X	120867	120922
r2211	yjm1304	X	120867	120922
r2212	yjm1307	X	120867	120922
r2213	yjm1326	X	120867	120922
r2214	yjm1332	X	120867	120922
r2215	yjm1338	X	120867	121044
r2216	yjm1341	X	120867	120994
r2217	yjm1342	X	120867	120922
r2218	yjm1356	X	120867	120922
r2219	yjm1381	X	120867	120925
r2220	yjm1386	X	120867	120922
r2221	yjm1399	X	120867	120922
r2222	yjm1402	X	120867	120922
r2223	yjm1415	X	120867	120922
r2224	yjm1417	X	120867	120922
r2225	yjm1433	X	120867	120922
r2226	yjm1460	X	120867	120922
r2227	yjm1477	X	120867	120922
r2228	yjm1526	X	120867	120922
r2229	yjm1574	X	120867	120922
r2230	yjm189	X	120867	120994
r2231	yjm193	X	120867	120922
r2232	yjm428	X	120867	120925
r2233	yjm451	X	120867	120922
r2234	yjm453	X	120867	120922
r2235	yjm470	X	120867	120922
r2236	yjm554	X	120867	120922
r2237	yjm555	X	120867	120922
r2238	yjm681	X	120867	120972
r2239	yjm972	X	120867	120922
r2240	yjm975	X	120867	120922
r2241	yjm981	X	120867	120922
r2242	yjm987	X	120867	120922
r2243	yjm990	X	120867	121044
r2244	yjm993	X	120867	120922
r2245	yjm996	X	120867	120922
r2246	yjm1419	X	120873	120925
r2247	yjm450	X	120873	120972
r2248	yjm456	X	120873	120922
r2249	yjm1450	X	120948	121191
r2250	yjm1242	X	121020	121191
r2251	yjm1477	X	121020	121191
r2252	yjm1433	X	121056	121191
r2253	yjm1078	X	143501	143656
r2254	yjm1129	X	204278	204386
r2255	yjm1242	X	204278	204386
r2256	yjm1250	X	204278	204386
r2257	yjm1332	X	204278	204386
r2258	yjm1336	X	204278	204386
r2259	yjm1356	X	204278	204386
r2260	yjm1415	X	204278	204386
r2261	yjm1417	X	204278	204386
r2262	yjm1433	X	204278	204386
r2263	yjm1450	X	204278	204386
r2264	yjm1477	X	204278	204386
r2265	yjm972	X	204278	204386
r2266	yjm975	X	204278	204386
r2267	yjm978	X	204278	204386
r2268	yjm981	X	204278	204386
r2269	yjm984	X	204278	204386
r2270	yjm987	X	204278	204386
r2271	yjm990	X	204278	204386

region id	strain	chromosome	start	end
r2272	yjm993	X	204278	204386
r2273	yjm996	X	204278	204386
r2274	yjm1463	X	204308	204386
r2275	yjm1248	X	204338	204386
r2276	yjm1383	X	204338	204386
r2277	yjm1439	X	204338	204386
r2278	yjm193	X	204338	204386
r2279	yjm1478	X	204352	204386
r2280	yjm1574	X	204352	204386
r2281	yjm1311	X	253061	253449
r2282	yjm1242	X	355049	355136
r2283	yjm1387	X	355049	355136
r2284	yjm1417	X	355049	355136
r2285	yjm1477	X	355049	355136
r2286	yjm1244	X	355061	355100
r2287	yjm1332	X	355061	355100
r2288	yjm1341	X	355061	355100
r2289	yjm1356	X	355061	355100
r2290	yjm1526	X	355061	355100
r2291	yjm987	X	355061	355100
r2292	yjm993	X	355061	355100
r2293	yjm996	X	355061	355100
r2294	yjm1252	X	405492	407511
r2295	yjm248	X	405492	407511
r2296	yjm1418	X	606757	606791
r2297	yjm1443	X	606757	606791
r2298	yjm1447	X	606757	606791
r2299	yjm193	X	606757	606840
r2300	yjm1078	X	662573	667789
r2301	yjm248	X	662573	669487
r2302	yjm1252	X	662678	668929
r2303	yjm1078	X	698664	702323
r2304	yjm1252	X	698664	702128
r2305	yjm248	X	698664	707017
r2306	yjm1342	X	701943	702602
r2307	yjm1479	X	701943	702602
r2308	yjm1526	X	702128	702602
r2309	yjm1252	X	702407	702782
r2310	yjm1078	X	702596	708568
r2311	yjm248	X	707064	707797
r2312	yjm248	X	707938	708568
r2313	yjm1078	X	708596	711685
r2314	yjm248	X	708598	709437
r2315	yjm248	X	709649	709734
r2316	yjm248	X	709843	710135
r2317	yjm1463	X	714909	715022
r2318	yjm1078	X	715377	715984
r2319	yjm1078	X	716025	716530
r2320	yjm1078	X	716629	719113
r2321	yjm1078	X	719169	720804
r2322	yjm248	X	719895	720762
r2323	yjm1078	X	721156	721887
r2324	yjm248	X	721156	721303
r2325	yjm248	X	721805	721887
r2326	yjm1078	X	721952	723608
r2327	yjm248	X	721981	724087
r2328	yjm248	X	725564	726013
r2329	yjm248	X	726778	727062
r2330	yjm1401	X	727469	727556
r2331	yjm248	X	727610	727688
r2332	yjm1332	X	727736	727922
r2333	yjm248	X	728036	728142
r2334	yjm248	X	728533	728738
r2335	yjm554	X	729068	729185
r2336	yjm1199	X	742961	743123
r2337	yjm1250	X	743000	743357
r2338	yjm1250	X	743494	743574
r2339	yjm1386	XI	1759	2117
r2340	yjm1463	XI	1759	1922
r2341	yjm1478	XI	1783	2112
r2342	yjm1342	XI	1857	2117
r2343	yjm1573	XI	1857	2112
r2344	yjm1252	XI	1863	1920
r2345	yjm1388	XI	1863	2117
r2346	yjm1387	XI	1953	2112
r2347	yjm189	XI	1953	2117
r2348	yjm244	XI	1953	2117
r2349	yjm1252	XI	1992	2099
r2350	yjm1418	XI	2075	2117
r2351	yjm1252	XI	2236	9558
r2352	yjm1388	XI	2253	2326
r2353	yjm1389	XI	2491	2549
r2354	yjm1592	XI	2491	2549
r2355	yjm1341	XI	2503	2556
r2356	yjm189	XI	2503	2549
r2357	yjm244	XI	2503	2549
r2358	yjm1129	XI	2514	2556
r2359	yjm1199	XI	2514	2556
r2360	yjm1208	XI	2514	2555
r2361	yjm1242	XI	2514	2556
r2362	yjm1304	XI	2514	2556
r2363	yjm1338	XI	2514	2556
r2364	yjm1386	XI	2514	2556
r2365	yjm1400	XI	2514	2555
r2366	yjm1419	XI	2514	2555

region id	strain	chromosome	start	end
r2367	yjm1443	XI	2514	2556
r2368	yjm1447	XI	2514	2556
r2369	yjm1477	XI	2514	2556
r2370	yjm1479	XI	2514	2555
r2371	yjm1526	XI	2514	2556
r2372	yjm193	XI	2514	2549
r2373	yjm969	XI	2514	2549
r2374	yjm972	XI	2514	2556
r2375	yjm975	XI	2514	2556
r2376	yjm978	XI	2514	2556
r2377	yjm981	XI	2514	2556
r2378	yjm984	XI	2514	2556
r2379	yjm987	XI	2514	2556
r2380	yjm990	XI	2514	2556
r2381	yjm993	XI	2514	2556
r2382	yjm1078	XI	6948	7137
r2383	yjm1078	XI	7266	9558
r2384	yjm1252	XI	10101	15515
r2385	yjm1252	XI	62900	64351
r2386	yjm248	XI	63101	64328
r2387	yjm1252	XI	95151	96177
r2388	yjm1078	XI	257816	258700
r2389	yjm1252	XI	257816	258700
r2390	yjm248	XI	257816	258700
r2391	yjm689	XI	257816	264688
r2392	yjm1078	XI	258799	264688
r2393	yjm1252	XI	258799	264688
r2394	yjm248	XI	258799	264688
r2395	yjm1399	XI	326460	327352
r2396	yjm1383	XI	458300	458454
r2397	yjm1332	XI	458396	458454
r2398	yjm555	XI	458396	458454
r2399	yjm554	XI	458420	458454
r2400	yjm248	XI	512100	513325
r2401	yjm1078	XI	512970	513325
r2402	yjm1252	XI	512970	513325
r2403	yjm1078	XI	513673	524989
r2404	yjm1252	XI	513673	525080
r2405	yjm248	XI	513673	525080
r2406	yjm1383	XI	647290	647344
r2407	yjm554	XI	665790	665890
r2408	yjm627	XI	665974	666122
r2409	yjm1311	XI	666215	666303
r2410	yjm1208	XI	666272	666420
r2411	yjm975	XII	5669	5842
r2412	yjm450	XII	12014	12071
r2413	yjm1338	XII	47927	48365
r2414	yjm1399	XII	47927	48296
r2415	yjm1592	XII	47927	48296
r2416	yjm975	XII	47927	48365
r2417	yjm1434	XII	48091	48371
r2418	yjm1443	XII	48185	48296
r2419	yjm1434	XII	64367	64442
r2420	yjm1311	XII	64547	64613
r2421	yjm1399	XII	64553	64661
r2422	yjm1244	XII	64664	64733
r2423	yjm996	XII	64664	64691
r2424	yjm1129	XII	64667	64762
r2425	yjm1208	XII	64667	64762
r2426	yjm1242	XII	64667	64762
r2427	yjm1326	XII	64667	64762
r2428	yjm1332	XII	64667	64762
r2429	yjm1336	XII	64667	64762
r2430	yjm1341	XII	64667	64762
r2431	yjm1383	XII	64667	64762
r2432	yjm1387	XII	64667	64762
r2433	yjm1415	XII	64667	64762
r2434	yjm1417	XII	64667	64762
r2435	yjm1450	XII	64667	64762
r2436	yjm1477	XII	64667	64762
r2437	yjm1478	XII	64667	64762
r2438	yjm1526	XII	64667	64762
r2439	yjm1549	XII	64667	64762
r2440	yjm1615	XII	64667	64762
r2441	yjm189	XII	64667	64762
r2442	yjm326	XII	64667	64762
r2443	yjm428	XII	64667	64762
r2444	yjm453	XII	64667	64762
r2445	yjm555	XII	64667	64762
r2446	yjm682	XII	64667	64762
r2447	yjm683	XII	64667	64762
r2448	yjm693	XII	64667	64762
r2449	yjm969	XII	64667	64762
r2450	yjm978	XII	64667	64762
r2451	yjm984	XII	64667	64762
r2452	yjm987	XII	64667	64762
r2453	yjm990	XII	64667	64762
r2454	yjm1418	XII	64775	64889
r2455	yjm1417	XII	95934	96168
r2456	yjm1434	XII	95934	96039
r2457	yjm1447	XII	95934	96039
r2458	yjm450	XII	95934	96168
r2459	yjm1252	XII	285894	291887
r2460	yjm1342	XII	287376	287603
r2461	yjm1252	XII	296678	297908

region id	strain	chromosome	start	end
r2462	yjm1078	XII	385069	385174
r2463	yjm1252	XII	385069	385174
r2464	yjm248	XII	385069	385491
r2465	yjm1078	XII	385342	385491
r2466	yjm1252	XII	385342	385491
r2467	yjm248	XII	385672	385809
r2468	yjm248	XII	385931	386150
r2469	yjm1078	XII	385959	386052
r2470	yjm1252	XII	386003	386052
r2471	yjm248	XII	386985	387067
r2472	yjm1078	XII	402952	403120
r2473	yjm1252	XII	402952	403120
r2474	yjm248	XII	402952	403360
r2475	yjm1078	XII	540053	541766
r2476	yjm1252	XII	540053	541766
r2477	yjm1248	XII	658149	658456
r2478	yjm1342	XII	658149	658418
r2479	yjm1439	XII	658149	658456
r2480	yjm627	XII	658149	658456
r2481	yjm1447	XII	734398	734572
r2482	yjm1434	XII	811753	812488
r2483	yjm1434	XII	813019	813148
r2484	yjm1434	XII	814870	815191
r2485	yjm1388	XII	1012269	1012327
r2486	yjm1401	XII	1012269	1012327
r2487	yjm1419	XII	1012269	1012327
r2488	yjm1460	XII	1012269	1012327
r2489	yjm456	XII	1012269	1012327
r2490	yjm1078	XII	1034374	1052154
r2491	yjm1252	XII	1034374	1053561
r2492	yjm248	XII	1034374	1053924
r2493	yjm1078	XII	1053619	1053924
r2494	yjm1252	XII	1053619	1058097
r2495	yjm1078	XII	1057961	1059011
r2496	yjm248	XII	1057961	1059028
r2497	yjm975	XII	1059868	1060076
r2498	yjm1083	XII	1059908	1060077
r2499	yjm1129	XII	1059908	1060077
r2500	yjm1208	XII	1059908	1060127
r2501	yjm1242	XII	1059908	1060127
r2502	yjm1244	XII	1059908	1060077
r2503	yjm1307	XII	1059908	1060077
r2504	yjm1332	XII	1059908	1060077
r2505	yjm1336	XII	1059908	1060077
r2506	yjm1341	XII	1059908	1060077
r2507	yjm1342	XII	1059908	1060077
r2508	yjm1355	XII	1059908	1060077
r2509	yjm1356	XII	1059908	1060077
r2510	yjm1381	XII	1059908	1060127
r2511	yjm1387	XII	1059908	1060127
r2512	yjm1417	XII	1059908	1060077
r2513	yjm1433	XII	1059908	1060077
r2514	yjm1447	XII	1059908	1060077
r2515	yjm1450	XII	1059908	1060077
r2516	yjm1463	XII	1059908	1060077
r2517	yjm1477	XII	1059908	1060077
r2518	yjm1478	XII	1059908	1060077
r2519	yjm1526	XII	1059908	1060077
r2520	yjm1549	XII	1059908	1060077
r2521	yjm1574	XII	1059908	1060077
r2522	yjm189	XII	1059908	1060077
r2523	yjm193	XII	1059908	1060077
r2524	yjm244	XII	1059908	1060077
r2525	yjm270	XII	1059908	1060077
r2526	yjm320	XII	1059908	1060077
r2527	yjm451	XII	1059908	1060127
r2528	yjm453	XII	1059908	1060077
r2529	yjm554	XII	1059908	1060077
r2530	yjm627	XII	1059908	1060077
r2531	yjm969	XII	1059908	1060077
r2532	yjm978	XII	1059908	1060077
r2533	yjm981	XII	1059908	1060077
r2534	yjm984	XII	1059908	1060077
r2535	yjm987	XII	1059908	1060077
r2536	yjm990	XII	1059908	1060077
r2537	yjm996	XII	1059908	1060077
r2538	yjm1248	XII	1060014	1060077
r2539	yjm1439	XII	1060014	1060077
r2540	yjm195	XII	1060014	1060077
r2541	yjm450	XII	1060014	1060098
r2542	yjm1418	XIII	21606	21714
r2543	yjm1444	XIII	24981	25167
r2544	yjm1252	XIII	96172	100073
r2545	yjm1129	XIII	162715	162906
r2546	yjm1208	XIII	162715	162906
r2547	yjm1242	XIII	162715	162906
r2548	yjm1244	XIII	162715	162906
r2549	yjm1250	XIII	162715	162906
r2550	yjm1311	XIII	162715	162906
r2551	yjm1332	XIII	162715	162906
r2552	yjm1336	XIII	162715	162906
r2553	yjm1338	XIII	162715	162906
r2554	yjm1341	XIII	162715	162906
r2555	yjm1355	XIII	162715	162906
r2556	yjm1385	XIII	162715	162906

region id	strain	chromosome	start	end
r2557	yjm1387	XIII	162715	162906
r2558	yjm1415	XIII	162715	162906
r2559	yjm1417	XIII	162715	162906
r2560	yjm1477	XIII	162715	162906
r2561	yjm1526	XIII	162715	162906
r2562	yjm1574	XIII	162715	162906
r2563	yjm1615	XIII	162715	162906
r2564	yjm189	XIII	162715	162894
r2565	yjm244	XIII	162715	162906
r2566	yjm326	XIII	162715	162906
r2567	yjm450	XIII	162715	162906
r2568	yjm453	XIII	162715	162906
r2569	yjm681	XIII	162715	162906
r2570	yjm682	XIII	162715	162906
r2571	yjm683	XIII	162715	162906
r2572	yjm689	XIII	162715	162906
r2573	yjm693	XIII	162715	162839
r2574	yjm972	XIII	162715	162906
r2575	yjm975	XIII	162715	162906
r2576	yjm978	XIII	162715	162906
r2577	yjm981	XIII	162715	162906
r2578	yjm984	XIII	162715	162906
r2579	yjm987	XIII	162715	162906
r2580	yjm990	XIII	162715	162906
r2581	yjm993	XIII	162715	162906
r2582	yjm996	XIII	162715	162906
r2583	yjm1326	XIII	216320	216367
r2584	yjm1450	XIII	216320	216367
r2585	yjm683	XIII	216320	216367
r2586	yjm1078	XIII	280531	282239
r2587	yjm1078	XIII	336129	344768
r2588	yjm1252	XIII	336129	356015
r2589	yjm248	XIII	336129	339666
r2590	yjm1078	XIII	344815	348503
r2591	yjm1202	XIII	481230	481314
r2592	yjm1078	XIII	509571	509985
r2593	yjm1252	XIII	509571	510465
r2594	yjm248	XIII	509571	509985
r2595	yjm1078	XIII	510197	510465
r2596	yjm248	XIII	510197	510465
r2597	yjm1078	XIII	510492	518522
r2598	yjm1252	XIII	510492	518522
r2599	yjm248	XIII	510492	518522
r2600	yjm1078	XIII	534979	535381
r2601	yjm1341	XIII	599963	600011
r2602	yjm1386	XIII	599963	600004
r2603	yjm693	XIII	599963	600011
r2604	yjm1341	XIII	600077	600343
r2605	yjm244	XIII	600077	600343
r2606	yjm693	XIII	600170	600214
r2607	yjm1341	XIII	600440	600542
r2608	yjm1386	XIII	600440	600542
r2609	yjm244	XIII	600440	600542
r2610	yjm693	XIII	600440	600542
r2611	yjm1444	XIII	609659	609800
r2612	yjm248	XIII	613049	625296
r2613	yjm1273	XIII	798591	798753
r2614	yjm1402	XIII	798591	798846
r2615	yjm1434	XIII	798591	798846
r2616	yjm1573	XIII	798591	798846
r2617	yjm1402	XIII	799780	799876
r2618	yjm1573	XIII	799780	799876
r2619	yjm1402	XIII	807182	807254
r2620	yjm1573	XIII	807182	807254
r2621	yjm1248	XIII	873884	874139
r2622	yjm1273	XIII	873884	874139
r2623	yjm1338	XIII	873884	874139
r2624	yjm1342	XIII	873884	874139
r2625	yjm1399	XIII	873884	874139
r2626	yjm1400	XIII	873884	874139
r2627	yjm1402	XIII	873884	874139
r2628	yjm1418	XIII	873884	874139
r2629	yjm1434	XIII	873884	874139
r2630	yjm1439	XIII	873884	874139
r2631	yjm1443	XIII	873884	874139
r2632	yjm1447	XIII	873884	874139
r2633	yjm1479	XIII	873884	874139
r2634	yjm1573	XIII	873884	874139
r2635	yjm1311	XIII	873895	874139
r2636	yjm1326	XIII	873895	874139
r2637	yjm1401	XIII	874034	874139
r2638	yjm1244	XIII	874091	874139
r2639	yjm1356	XIII	874091	874139
r2640	yjm972	XIII	874091	874139
r2641	yjm975	XIII	874091	874139
r2642	yjm981	XIII	874091	874139
r2643	yjm984	XIII	874091	874139
r2644	yjm987	XIII	874091	874139
r2645	yjm990	XIII	874091	874139
r2646	yjm1133	XIII	874095	874139
r2647	yjm1336	XIII	874095	874139
r2648	yjm451	XIII	874095	874139
r2649	yjm453	XIII	874095	874139
r2650	yjm681	XIII	874095	874139
r2651	yjm1078	XIII	879845	880565

region id	strain	chromosome	start	end
r2652	yjm248	XIII	879845	880565
r2653	yjm1078	XIII	885915	886054
r2654	yjm248	XIII	885915	886054
r2655	yjm1078	XIII	886356	886534
r2656	yjm248	XIII	886356	886534
r2657	yjm1242	XIII	908628	908659
r2658	yjm1244	XIII	908628	908659
r2659	yjm1338	XIII	908628	908659
r2660	yjm1356	XIII	908628	908659
r2661	yjm1387	XIII	908628	908659
r2662	yjm1450	XIII	908628	908659
r2663	yjm1526	XIII	908628	908659
r2664	yjm683	XIII	908628	908659
r2665	yjm975	XIII	908628	908659
r2666	yjm987	XIII	908628	908659
r2667	yjm555	XIII	917862	917936
r2668	yjm248	XIV	69	605
r2669	yjm1443	XIV	872	1814
r2670	yjm1399	XIV	1341	1712
r2671	yjm1273	XIV	4390	4885
r2672	yjm1399	XIV	4506	4885
r2673	yjm1443	XIV	4506	4885
r2674	yjm1460	XIV	6828	6915
r2675	yjm1549	XIV	7002	7061
r2676	yjm1385	XIV	7009	7100
r2677	yjm1307	XIV	12970	14135
r2678	yjm450	XIV	14165	14509
r2679	yjm1342	XIV	15857	15918
r2680	yjm1385	XIV	15857	16585
r2681	yjm1386	XIV	15857	16567
r2682	yjm1443	XIV	15857	16800
r2683	yjm451	XIV	15857	16567
r2684	yjm1129	XIV	15954	16444
r2685	yjm1338	XIV	15954	16444
r2686	yjm1415	XIV	15954	16444
r2687	yjm1450	XIV	15954	23458
r2688	yjm1463	XIV	15954	16444
r2689	yjm1526	XIV	15954	16444
r2690	yjm326	XIV	15954	23458
r2691	yjm627	XIV	15954	16444
r2692	yjm682	XIV	15954	16444
r2693	yjm683	XIV	15954	16444
r2694	yjm993	XIV	15954	16444
r2695	yjm1311	XIV	16089	16444
r2696	yjm1381	XIV	16089	16444
r2697	yjm1399	XIV	16089	18045
r2698	yjm189	XIV	16089	16444
r2699	yjm1078	XIV	16723	21525
r2700	yjm1129	XIV	16723	21634
r2701	yjm1133	XIV	16723	21400
r2702	yjm1202	XIV	16723	21400
r2703	yjm1242	XIV	16723	21525
r2704	yjm1244	XIV	16723	21525
r2705	yjm1250	XIV	16723	21525
r2706	yjm1252	XIV	16723	21525
r2707	yjm1311	XIV	16723	21525
r2708	yjm1326	XIV	16723	21400
r2709	yjm1332	XIV	16723	21525
r2710	yjm1336	XIV	16723	21525
r2711	yjm1338	XIV	16723	21525
r2712	yjm1341	XIV	16723	21525
r2713	yjm1356	XIV	16723	21525
r2714	yjm1381	XIV	16723	21400
r2715	yjm1383	XIV	16723	21525
r2716	yjm1387	XIV	16723	21525
r2717	yjm1415	XIV	16723	22816
r2718	yjm1417	XIV	16723	21525
r2719	yjm1433	XIV	16723	21525
r2720	yjm1463	XIV	16723	21525
r2721	yjm1477	XIV	16723	21525
r2722	yjm1526	XIV	16723	21634
r2723	yjm1527	XIV	16723	21525
r2724	yjm1549	XIV	16723	22089
r2725	yjm1574	XIV	16723	21948
r2726	yjm189	XIV	16723	21525
r2727	yjm244	XIV	16723	21525
r2728	yjm248	XIV	16723	21525
r2729	yjm270	XIV	16723	21525
r2730	yjm320	XIV	16723	21400
r2731	yjm453	XIV	16723	21525
r2732	yjm456	XIV	16723	21525
r2733	yjm541	XIV	16723	21400
r2734	yjm554	XIV	16723	21400
r2735	yjm555	XIV	16723	21400
r2736	yjm627	XIV	16723	21525
r2737	yjm682	XIV	16723	22816
r2738	yjm683	XIV	16723	21634
r2739	yjm693	XIV	16723	21525
r2740	yjm969	XIV	16723	21525
r2741	yjm972	XIV	16723	21525
r2742	yjm975	XIV	16723	21525
r2743	yjm978	XIV	16723	21525
r2744	yjm981	XIV	16723	21525
r2745	yjm984	XIV	16723	21525
r2746	yjm987	XIV	16723	21525

region id	strain	chromosome	start	end
r2747	yjm990	XIV	16723	21525
r2748	yjm993	XIV	16723	21634
r2749	yjm996	XIV	16723	21525
r2750	yjm1385	XIV	17164	25172
r2751	yjm450	XIV	17164	19191
r2752	yjm1418	XIV	17834	22020
r2753	yjm450	XIV	20017	24615
r2754	yjm681	XIV	21300	25136
r2755	yjm1248	XIV	24220	33539
r2756	yjm1439	XIV	24220	33539
r2757	yjm1573	XIV	24220	33539
r2758	yjm1447	XIV	24615	33650
r2759	yjm1450	XIV	24615	28674
r2760	yjm326	XIV	24615	31526
r2761	yjm1418	XIV	25368	28601
r2762	yjm1199	XIV	26952	33539
r2763	yjm1273	XIV	26952	33539
r2764	yjm1385	XIV	26952	33539
r2765	yjm1400	XIV	26952	34395
r2766	yjm1402	XIV	26952	33539
r2767	yjm1434	XIV	26952	33539
r2768	yjm1479	XIV	26952	36501
r2769	yjm450	XIV	26952	34650
r2770	yjm451	XIV	26952	33539
r2771	yjm1450	XIV	29546	31526
r2772	yjm1415	XIV	29974	31526
r2773	yjm682	XIV	29974	31526
r2774	yjm1342	XIV	32092	33539
r2775	yjm1450	XIV	32092	37854
r2776	yjm326	XIV	32092	37854
r2777	yjm993	XIV	32092	34650
r2778	yjm1478	XIV	32200	34192
r2779	yjm1415	XIV	32284	34650
r2780	yjm682	XIV	32284	37854
r2781	yjm1381	XIV	33229	34395
r2782	yjm1418	XIV	34133	37890
r2783	yjm1447	XIV	34133	37890
r2784	yjm1385	XIV	34171	37890
r2785	yjm1401	XIV	34585	37890
r2786	yjm1208	XIV	77909	78036
r2787	yjm1248	XIV	77909	78036
r2788	yjm1311	XIV	77909	78036
r2789	yjm1338	XIV	77909	78036
r2790	yjm1386	XIV	77909	78036
r2791	yjm1388	XIV	77909	78036
r2792	yjm1389	XIV	77909	78036
r2793	yjm1401	XIV	77909	78036
r2794	yjm1419	XIV	77909	78036
r2795	yjm1439	XIV	77909	78036
r2796	yjm1443	XIV	77909	78036
r2797	yjm1444	XIV	77909	78036
r2798	yjm1460	XIV	77909	78036
r2799	yjm1592	XIV	77909	78036
r2800	yjm1615	XIV	77909	78036
r2801	yjm195	XIV	77909	78036
r2802	yjm428	XIV	77909	78036
r2803	yjm451	XIV	77909	78036
r2804	yjm450	XIV	102346	102432
r2805	yjm1415	XIV	102352	102449
r2806	yjm1399	XIV	102397	102506
r2807	yjm1078	XIV	234226	236523
r2808	yjm1252	XIV	234226	236523
r2809	yjm248	XIV	234226	236523
r2810	yjm1443	XIV	253581	253851
r2811	yjm1478	XIV	567826	567980
r2812	yjm450	XIV	567832	567980
r2813	yjm682	XIV	567832	567980
r2814	yjm1450	XIV	704279	704500
r2815	yjm681	XIV	704279	704500
r2816	yjm1478	XIV	704401	704500
r2817	yjm1444	XIV	750739	750914
r2818	yjm1450	XIV	750742	751384
r2819	yjm326	XIV	750742	751384
r2820	yjm682	XIV	750742	750914
r2821	yjm1248	XIV	751003	751158
r2822	yjm1386	XIV	751003	751257
r2823	yjm1400	XIV	751003	751158
r2824	yjm1439	XIV	751003	751257
r2825	yjm1443	XIV	751003	751257
r2826	yjm1447	XIV	751003	751158
r2827	yjm1479	XIV	751003	751158
r2828	yjm451	XIV	751003	751257
r2829	yjm1273	XIV	751147	751257
r2830	yjm1381	XIV	751147	751257
r2831	yjm1401	XIV	751147	751257
r2832	yjm1418	XIV	751147	751257
r2833	yjm1573	XIV	751147	751257
r2834	yjm627	XIV	751147	751257
r2835	yjm682	XIV	751147	751257
r2836	yjm1444	XIV	751263	751384
r2837	yjm1447	XIV	751263	751384
r2838	yjm681	XIV	751263	751384
r2839	yjm1450	XIV	751717	752755
r2840	yjm326	XIV	752121	753127
r2841	yjm1447	XIV	752146	752308

region id	strain	chromosome	start	end
r2842	yjm1418	XIV	752156	752308
r2843	yjm1381	XIV	752157	752308
r2844	yjm1386	XIV	752157	752308
r2845	yjm1400	XIV	752157	752308
r2846	yjm1401	XIV	752157	752308
r2847	yjm1444	XIV	752157	752536
r2848	yjm1460	XIV	752157	752308
r2849	yjm681	XIV	752157	753068
r2850	yjm1385	XIV	752168	752368
r2851	yjm1400	XIV	752365	753039
r2852	yjm1401	XIV	752365	753039
r2853	yjm1447	XIV	752365	753039
r2854	yjm1444	XIV	752826	753068
r2855	yjm1386	XIV	752932	753063
r2856	yjm1418	XIV	752932	753039
r2857	yjm1443	XIV	752932	753039
r2858	yjm1479	XIV	752932	753039
r2859	yjm689	XIV	752932	753039
r2860	yjm1400	XIV	754574	754630
r2861	yjm1444	XIV	754574	754621
r2862	yjm1479	XIV	754574	754630
r2863	yjm1273	XIV	754578	754621
r2864	yjm195	XIV	754578	754621
r2865	yjm1252	XIV	773555	775841
r2866	yjm1252	XIV	775973	776072
r2867	yjm1129	XIV	776429	776460
r2868	yjm1387	XIV	776429	776460
r2869	yjm1526	XIV	776429	776460
r2870	yjm978	XIV	776429	776460
r2871	yjm993	XIV	776429	776460
r2872	yjm1386	XIV	781348	781522
r2873	yjm1447	XIV	781371	781522
r2874	yjm1450	XIV	781405	781522
r2875	yjm1401	XIV	781470	781522
r2876	yjm1311	XIV	781472	781522
r2877	yjm1338	XIV	781472	781522
r2878	yjm195	XIV	781472	781522
r2879	yjm244	XIV	781472	781522
r2880	yjm1400	XIV	781504	781567
r2881	yjm1479	XIV	781504	781567
r2882	yjm1415	XIV	781508	781567
r2883	yjm1248	XIV	781567	781647
r2884	yjm1386	XIV	781567	781688
r2885	yjm1439	XIV	781567	781647
r2886	yjm1450	XIV	781832	782018
r2887	yjm1244	XIV	781838	781952
r2888	yjm1133	XIV	781851	781952
r2889	yjm1326	XIV	781851	781952
r2890	yjm1433	XIV	781851	781949
r2891	yjm1443	XIV	781851	781970
r2892	yjm1526	XIV	781851	781952
r2893	yjm456	XIV	781851	781949
r2894	yjm987	XIV	781851	781952
r2895	yjm993	XIV	781851	782215
r2896	yjm450	XIV	782524	782586
r2897	yjm1402	XV	192	230
r2898	yjm996	XV	313	828
r2899	yjm1341	XV	14287	26534
r2900	yjm1307	XV	14988	15099
r2901	yjm1399	XV	17553	17897
r2902	yjm1199	XV	17718	18216
r2903	yjm1388	XV	17718	18216
r2904	yjm1401	XV	17718	18216
r2905	yjm1574	XV	17718	18216
r2906	yjm1592	XV	17718	18216
r2907	yjm689	XV	17718	18216
r2908	yjm1399	XV	19388	19425
r2909	yjm1399	XV	19561	19910
r2910	yjm1399	XV	20126	20456
r2911	yjm1244	XV	20853	20945
r2912	yjm1381	XV	20853	20945
r2913	yjm1388	XV	20853	20945
r2914	yjm1400	XV	20853	20945
r2915	yjm1401	XV	20853	20945
r2916	yjm1479	XV	20853	20945
r2917	yjm1574	XV	20853	20945
r2918	yjm1592	XV	20853	20945
r2919	yjm1190	XV	21163	23588
r2920	yjm451	XV	22620	26172
r2921	yjm682	XV	23661	23741
r2922	yjm1444	XV	24691	24827
r2923	yjm1444	XV	25317	25427
r2924	yjm1447	XV	25745	26132
r2925	yjm451	XV	26252	26522
r2926	yjm693	XV	26486	26642
r2927	yjm451	XV	30052	30208
r2928	yjm451	XV	30448	31222
r2929	yjm1387	XV	30468	30541
r2930	yjm1477	XV	31260	31300
r2931	yjm451	XV	32118	39973
r2932	yjm1399	XV	35145	39973
r2933	yjm1399	XV	40195	43517
r2934	yjm451	XV	40195	42548
r2935	yjm1399	XV	43585	49402
r2936	yjm451	XV	45131	48160

region id	strain	chromosome	start	end
r2937	yjm1383	XV	75940	76504
r2938	yjm1399	XV	75940	76504
r2939	yjm326	XV	75940	76504
r2940	yjm682	XV	75940	76504
r2941	yjm1383	XV	76654	78552
r2942	yjm1399	XV	76654	78552
r2943	yjm326	XV	76654	78552
r2944	yjm682	XV	76654	78552
r2945	yjm1250	XV	88074	91410
r2946	yjm1250	XV	91693	91998
r2947	yjm1418	XV	113389	113431
r2948	yjm1250	XV	160138	160621
r2949	yjm1355	XV	160138	160189
r2950	yjm1383	XV	160138	160419
r2951	yjm1387	XV	160138	160419
r2952	yjm1417	XV	160138	160471
r2953	yjm1252	XV	163017	163942
r2954	yjm1399	XV	226865	226934
r2955	yjm1273	XV	227867	227967
r2956	yjm1402	XV	227867	227939
r2957	yjm1463	XV	227867	227967
r2958	yjm1573	XV	227867	227939
r2959	yjm1400	XV	227871	227945
r2960	yjm1434	XV	227871	227939
r2961	yjm1443	XV	227871	227967
r2962	yjm456	XV	227871	227967
r2963	yjm1078	XV	288075	289895
r2964	yjm1252	XV	288075	289895
r2965	yjm248	XV	288075	289895
r2966	yjm682	XV	299707	299764
r2967	yjm244	XV	463538	463618
r2968	yjm1078	XV	483905	485606
r2969	yjm1399	XV	523610	523803
r2970	yjm1133	XV	622025	622069
r2971	yjm1190	XV	622025	622059
r2972	yjm1199	XV	622025	622069
r2973	yjm1304	XV	622025	622059
r2974	yjm1326	XV	622025	622065
r2975	yjm320	XV	622025	622069
r2976	yjm450	XV	622025	622065
r2977	yjm541	XV	622025	622069
r2978	yjm554	XV	622025	622069
r2979	yjm681	XV	622025	622059
r2980	yjm683	XV	622025	622069
r2981	yjm689	XV	622025	622069
r2982	yjm1190	XV	622406	622500
r2983	yjm554	XV	622406	622500
r2984	yjm689	XV	622406	622500
r2985	yjm1133	XV	622415	622500
r2986	yjm1199	XV	622415	622500
r2987	yjm1304	XV	622415	622500
r2988	yjm1326	XV	622415	622500
r2989	yjm320	XV	622415	622500
r2990	yjm428	XV	622415	622500
r2991	yjm450	XV	622415	622500
r2992	yjm541	XV	622415	622500
r2993	yjm681	XV	622415	622500
r2994	yjm683	XV	622415	622500
r2995	yjm1133	XV	623420	623468
r2996	yjm1190	XV	623420	623468
r2997	yjm1199	XV	623420	623468
r2998	yjm1304	XV	623420	623468
r2999	yjm1326	XV	623420	623468
r3000	yjm320	XV	623420	623468
r3001	yjm450	XV	623420	623468
r3002	yjm541	XV	623420	623468
r3003	yjm554	XV	623420	623468
r3004	yjm681	XV	623420	623468
r3005	yjm683	XV	623420	623468
r3006	yjm689	XV	623420	623468
r3007	yjm1133	XV	624147	624290
r3008	yjm1190	XV	624147	624290
r3009	yjm1199	XV	624147	624290
r3010	yjm1304	XV	624147	624290
r3011	yjm1326	XV	624147	624290
r3012	yjm320	XV	624147	624290
r3013	yjm428	XV	624147	624191
r3014	yjm450	XV	624147	624290
r3015	yjm541	XV	624147	624290
r3016	yjm554	XV	624147	624290
r3017	yjm681	XV	624147	624290
r3018	yjm683	XV	624147	624290
r3019	yjm689	XV	624147	624290
r3020	yjm1078	XV	960088	980700
r3021	yjm1252	XV	960088	980700
r3022	yjm248	XV	960088	980700
r3023	yjm1078	XV	1049614	1050940
r3024	yjm1252	XV	1049614	1050940
r3025	yjm248	XV	1049614	1050940
r3026	yjm1078	XV	1051792	1052359
r3027	yjm248	XV	1051792	1052359
r3028	yjm1307	XV	1057193	1057412
r3029	yjm1381	XV	1057311	1057418
r3030	yjm195	XV	1069613	1069833
r3031	yjm1444	XV	1069719	1069800

region id	strain	chromosome	start	end
r3032	yjm1450	XV	1069719	1069785
r3033	yjm1248	XV	1069776	1070049
r3034	yjm1439	XV	1069776	1070034
r3035	yjm1387	XV	1069925	1070103
r3036	yjm1417	XV	1069925	1070103
r3037	yjm1477	XV	1069925	1070103
r3038	yjm555	XV	1069925	1070103
r3039	yjm1418	XV	1069962	1070103
r3040	yjm1383	XV	1070456	1070543
r3041	yjm1418	XV	1070456	1070543
r3042	yjm554	XV	1070456	1070543
r3043	yjm1083	XV	1070472	1070537
r3044	yjm1190	XV	1070472	1070535
r3045	yjm1208	XV	1070472	1070537
r3046	yjm1248	XV	1070472	1070545
r3047	yjm1273	XV	1070472	1070537
r3048	yjm1304	XV	1070472	1070535
r3049	yjm1388	XV	1070472	1070535
r3050	yjm1399	XV	1070472	1070537
r3051	yjm1401	XV	1070472	1070537
r3052	yjm1419	XV	1070472	1070535
r3053	yjm1434	XV	1070472	1070535
r3054	yjm1439	XV	1070472	1070535
r3055	yjm1443	XV	1070472	1070545
r3056	yjm1447	XV	1070472	1070535
r3057	yjm1573	XV	1070472	1070545
r3058	yjm195	XV	1070472	1070545
r3059	yjm271	XV	1070472	1070545
r3060	yjm320	XV	1070472	1070537
r3061	yjm326	XV	1070472	1070537
r3062	yjm450	XV	1070472	1070535
r3063	yjm456	XV	1070472	1070545
r3064	yjm470	XV	1070472	1070535
r3065	yjm683	XV	1070472	1070537
r3066	yjm1450	XV	1070482	1070537
r3067	yjm1250	XV	1070493	1070543
r3068	yjm555	XV	1070493	1070543
r3069	yjm1248	XV	1073359	1074236
r3070	yjm1401	XV	1073359	1074128
r3071	yjm1439	XV	1073359	1074236
r3072	yjm1399	XV	1073378	1074236
r3073	yjm1447	XV	1073378	1074128
r3074	yjm627	XV	1073378	1073838
r3075	yjm689	XV	1073378	1073838
r3076	yjm987	XV	1073378	1074348
r3077	yjm990	XV	1073378	1073838
r3078	yjm993	XV	1073378	1073838
r3079	yjm1444	XV	1073436	1074128
r3080	yjm1083	XV	1073791	1073838
r3081	yjm244	XV	1073791	1073838
r3082	yjm1244	XV	1073870	1074236
r3083	yjm1273	XV	1073870	1074236
r3084	yjm1307	XV	1073870	1074236
r3085	yjm1336	XV	1073870	1074128
r3086	yjm1388	XV	1073870	1074236
r3087	yjm1415	XV	1073870	1074310
r3088	yjm1417	XV	1073870	1074236
r3089	yjm1443	XV	1073870	1074236
r3090	yjm1477	XV	1073870	1074236
r3091	yjm1549	XV	1073870	1074069
r3092	yjm1574	XV	1073870	1074348
r3093	yjm244	XV	1073870	1074069
r3094	yjm453	XV	1073870	1074236
r3095	yjm554	XV	1073870	1074236
r3096	yjm555	XV	1073870	1074236
r3097	yjm627	XV	1073870	1074310
r3098	yjm689	XV	1073870	1074069
r3099	yjm969	XV	1073870	1074310
r3100	yjm975	XV	1073870	1074310
r3101	yjm984	XV	1073870	1074123
r3102	yjm990	XV	1073870	1074236
r3103	yjm993	XV	1073870	1074236
r3104	yjm1450	XV	1073876	1074123
r3105	yjm456	XV	1073876	1074236
r3106	yjm1342	XV	1073975	1074123
r3107	yjm1190	XV	1074058	1074236
r3108	yjm1202	XV	1074058	1074236
r3109	yjm1381	XV	1074058	1074236
r3110	yjm320	XV	1074058	1074236
r3111	yjm326	XV	1074058	1074348
r3112	yjm541	XV	1074058	1074236
r3113	yjm682	XV	1074058	1074236
r3114	yjm693	XV	1074058	1074236
r3115	yjm1401	XV	1074147	1074348
r3116	yjm1447	XV	1074147	1074256
r3117	yjm1336	XV	1074175	1074348
r3118	yjm1549	XV	1074175	1074245
r3119	yjm689	XV	1074175	1074236
r3120	yjm450	XV	1074853	1074940
r3121	yjm1133	XV	1080247	1080575
r3122	yjm1133	XV	1080965	1081006
r3123	yjm1307	XV	1082378	1082470
r3124	yjm1463	XV	1082378	1082648
r3125	yjm1252	XV	1082405	1082452
r3126	yjm1439	XV	1082439	1082497

region id	strain	chromosome	start	end
r3127	yjm1307	XV	1082517	1082648
r3128	yjm975	XV	1082959	1083016
r3129	yjm1574	XV	1089487	1089536
r3130	yjm1388	XVI	6681	6729
r3131	yjm554	XVI	8019	8184
r3132	yjm1402	XVI	97098	97164
r3133	yjm975	XVI	97608	98055
r3134	yjm1252	XVI	333928	334671
r3135	yjm1252	XVI	336339	336369
r3136	yjm1252	XVI	489612	494471
r3137	yjm1342	XVI	572037	572115
r3138	yjm1388	XVI	572040	572115
r3139	yjm627	XVI	572060	572115
r3140	yjm1078	XVI	811220	811403
r3141	yjm248	XVI	811220	811403
r3142	yjm1078	XVI	817521	841738
r3143	yjm248	XVI	817521	841786
r3144	yjm1355	XVI	839114	839225
r3145	yjm681	XVI	839115	839225
r3146	yjm1311	XVI	868998	869166
r3147	yjm1399	XVI	869136	869184

VITA

Anne Elizabeth Clark grew up in Burien, Washington. She attended Harvey Mudd College in Claremont, California for her undergraduate studies, where she contributed to research projects on simulating negative cooperativity and quantifying cataract surgical rates in sub-Saharan Africa. She also worked with Dr. Roni Rosenfeld at Carnegie Mellon University on forecasting the daily incidence of influenza through the TECBio summer REU. She completed her senior thesis on the role of horizontally transferred genes in biofilm production in *E. coli* under the advising of Professor Dan Stoebel, graduating in 2013 with a B.S. in Mathematical and Computational Biology. She enrolled in the Genome Sciences doctoral program at the University of Washington in 2013, eventually joining the lab of Dr. Joshua Akey.