Quantitative insights into the role of copy number variants in adaptive evolution

by

Grace Avecilla

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Biology

New York University

May 2022

_____

David Gresham, PhD

Acknowledgements

There are so many people without whom I cannot imagine having done a PhD. I must thank my undergraduate mentors, Hojun Song, Eunsoo Kim, and Ken Fedorka, who gave me my first opportunities to work in labs and confidence in my scientific abilities. I would like to thank my PhD advisor, David Gresham, for his mentorship and support – I especially valued the balance between guidance and independence he provided, and his encouragement to pursue every opportunity. I'd like to thank my committee members, Mark Siegal, Jane Carlton, Michael Purugganan, and Molly Przeworski, for their support and advice about my scientific, career, and personal development over the years. I would like to thank all of the Gresham Lab members I've had the honor of working with: Darach Miller, Nathan Brandt, Siyu Sun, and Farah Abdul-Rahman who welcoming me into the lab and making it feel like a place I belong; Chris Jackson and Pieter Spielman for endless advice and commiseration; Marissa Knoll, Julie Chuong, Titir De, Ying Xie, Megan Hockman, Lauren Lashua, Angela Hickey, Ina Suresh, and Julia Matthews for support and discussion and generally making the lab a great place to work; my undergraduate mentees Lauren Brown and Miles Tran for their contributions to my work and enthusiasm about science; and especially Steff Lauer, for giving me such a wonderful example of what it means to be a scientist, a mentor, and a friend. The scientific community is enormous, and there are so many people who I have interacted with who have helped and influenced me in ways both large and small. Every conference I have been lucky enough to participate in helped me and my ideas grow and develop, and I thank all the scientists I have been fortunate to interact with at meetings over the years. I thank my collaborators, who taught me so much! Thank you to Joseph Schacherer for hosting me in Strasbourg for a wonderful summer, Elodie Caudal for teaching me transposon mutagenesis, and the rest of the Schacherer lab for being so welcoming. Thank you to Yoav Ram, who has taught me so much and is so encouraging, it

has been a great joy to work together these past years. Thank you to lineage tracking experts Sasha Levy, Gavin Sherlock, and Fangfei Li, without whom the insights of the research would have been lesser. Thank you to the Biology Department at NYU at large, faculty, staff, postdocs, researchers, and administrators, who all made this a place where I could and wanted to work. I would like to thank those who I taught with over the years, especially David Gresham, Mary Killilea, Eugene Plavskin, who empowered me to take an active role in course design and development. I have to thank all of the Biology PhD students I have overlapped with at NYU. I really enjoyed being a part of this community, and I especially appreciate the efforts so many people put into making it a community, especially those who put some much effort into the Graduate Biology Group, Graduate Women in Science, and the Respect is a Part of Research workshop. A very, very special thank you to everyone in my cohort, especially Katherine Johnson, Bianca Diaz, Logan Schachtner, Dan Pollack, and Porfirio Fernandez, for the support and commiseration, science discussions, gift exchanges, and many many many brunches and lunches and dinners and happy hours over the years. A special thanks to Peter Whitney, for all the stuff in the previous sentence and then a little more. Thank you to others at NYU who have made my time here better, especially the NYU chapter of SACNAS, our graduate worker union GSOC, and everyone else who is trying to make this place more welcoming and a better place to work for all. I would like to thank my family for supporting and encouraging my interest in science my whole life. I thank my father, Guillermo Avecilla, for talking seriously to a ninth grader about apoptosis over dinner; my mother, Josephine Avecilla, for offering to proofread abstracts about everything from insect immune systems to the yeast genome; my brother, Matthew Avecilla, for participating in Envirothon competitions with me and for knowing a lot more about wildlife than I do; all of my family for a million other things over the years. And finally, I must thank my little cats, Faun and Zara, for reminding me what is important – treats, naps, and being with those you care about.

Abstract

Copy number variation (CNV), the duplication or deletion of segments of DNA, is a ubiquitous form of genetic variation that contributes to rapid adaptation, gene family evolution, and disease. This doctoral thesis focuses on the role of CNVs in mediating evolution to novel environments, using the yeast *Saccharomyces cerevisiae* as a model organism. In chapter one, I summarize our understanding of evolutionary dynamics in asexual populations, and discuss the role of CNVs in evolution. In chapter two, I observed the dynamics of CNV at the locus *GAP1* locus during evolution in glutamine-limited chemostats using a fluorescent CNV reporter and lineage tracking barcodes, and discovered that hundreds to thousands of competing *GAP1* CNV lineages contribute to the rapid and repeatable rise of CNVs in evolving populations. In chapter three, I used simulation-based inference with neural networks to estimate the formation rate and selection coefficients of CNVs using the observed population level dynamics from chapter one. I found that *GAP1* CNVs are generated at high rates and have large selection coefficients, and validated the approach with inference from barcode lineage dynamics and empirical measurements of *GAP1* CNV fitness. In chapter four, I used transposon mutagenesis, transcriptome profiling, and fitness assays to investigate the genetic and functional effects of CNVs with different structures. I found that amplification confers novel mutational tolerance, and that CNVs with low fitness have genetic interactions with genes involved in translation and mitochondrial function. Furthermore, while amplification results in increased gene expression, some strains also exhibit dosage compensation. CNVs do not exhibit previously described gene expression signatures of aneuploidy, instead they downregulate genes involved in cellular respiration, nucleoside biosynthetic processes, and small molecule metabolism, and upregulate genes involved in transposition, nucleic acid metabolism, and siderophore transport, though to different degrees in each strain. The implications of this work and possible future directions are discussed in chapter five.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Understanding how genomes evolve is fundamental for understanding how life on earth came to be, and for predicting how organisms will evolve in the future. Organisms' genomes are composed of DNA which is made up of four bases - A, T, C, G - strung together in linear or circular chromosomes. Across the tree of life, genomes vary in sequence and in configuration and structure of segments of DNA. Genomes can evolve in many ways: through single nucleotide variants (SNV), when one single base is replaced with another; through rearrangements, where chromosomes break and the broken piece fuses to another chromosome, or reverses in orientation; through copy number variants (CNV), in which large portions of the DNA are deleted or duplicated; and through other mechanisms as well (Press et al. 2019).

Each of these mutations can have different effects on an organism's fitness - its ability to survive and reproduce. SNVs usually (but not always!) have smaller effects, whether positive or negative, than rearrangements or CNVs, because they affect less of the genome (Press et al. 2019). Moreover, the effect of a mutation may differ based on the environmental or genetic context. A mutation that is beneficial in a cold environment may be detrimental in a hot environment (Bennett and Lenski 2007), or a mutation that usually only slightly increases one's risk of cancer may dramatically increase the risk when another cancer causing mutation, such as a BRCA1 mutation, is present (Tutuncuoglu and Krogan 2019).

Each of these ways in which the genome evolves occur through different mechanisms, and therefore occur at different rates. The rate at which different types of mutations occur affects the probability of a mutation occurring, and therefore can have a very large impact on how a population evolves (Press et al. 2019; Ferenci and Maharjan 2015). Both SNVs and

CNVs may help harmful bacteria become resistant to specific antibiotics. Although CNVs occur frequently, it may take many mutations for bacteria to become fully resistant if each CNV has a small effect. By contrast, a single SNV may be sufficient to confer resistance, but if it occurs at a low rate, may be a less likely route to drug resistance (Schenk et al. 2022). Therefore consideration of both the rate, and effect, of mutations is of practical significance when prescribing antibiotic regimes (Andersson 2015).

My doctoral thesis research focuses specifically on the role of CNVs in mediating evolution to novel environments, using the yeast *Saccharomyces cerevisiae* as a model organism. In chapter two, I observe the dynamics of copy number variants at a locus using a variety of approaches, which allows us to investigate CNV dynamics at a range of perspectives, from comparing dynamics between populations, to lineages within populations, to characterizing CNV structures in single cells. In chapter three, I use simulation-based inference to estimate the formation rate and selection coefficients of CNVs using the observed population level dynamics from chapter one. In chapter four, I examine how CNVs with different structures interact with the genetic and physical environment, and try to understand if there are common effects associated with CNVs.

1.1 Evolutionary parameters and the dynamics of evolution

When one thinks of Darwinian evolution, one often thinks of the phrase, “survival of the fittest.” What does this mean? Darwin’s theory stated that individuals within populations have variable phenotypes. This phenotypic variation is heritable, and selection acts on these variable populations so that individuals with some phenotypic characteristics are able to differentially survive and reproduce, resulting in the “survival of the fittest.” This process of “survival of the fittest” results in evolution of populations, ultimately resulting in the diversity of life we see in the world (Darwin 1859). When people think about evolution, this emphasis is often on the “survival

of the fittest" aspect, or the selection aspect, but it is also important that there is variation in a population. We now know the basis of this heritable variation is genetic, and new genetic variation is created by mutation. A new mutation can potentially change the phenotype of an individual and allow that individual to succeed in a selective environment where it otherwise would not have. The rate and order in which these mutations arrive and add new variation to populations is as important an aspect of evolutionary dynamics as the selective events. Understanding how the arrival of new mutations can constrain and facilitate evolution and how selection acts on these mutations is essential for understanding evolutionary processes, and for predicting future evolution. These questions become both more interesting and more difficult to understand when we consider that different types of mutations have different characteristics. Here, I will focus on asexually reproducing microbial populations and use examples from evolution experiments, though much interesting work is being done on other types of populations and in other environments. This section does not comprehensively review the literature on the dynamics of evolution in microbial experimental evolution (a good review can be found here (Van den Bergh et al. 2018)), rather, my aim is to introduce some important terminology and recent insights, as well as give the reader a sense of the complexity of evolutionary dynamics and the importance of the underlying parameters.

1.1.1 The dynamics of evolution

Evolutionary change can be broken down into two parts: change in the distribution of phenotypes in a population and change in the allele frequencies in a population. While a change in one of these can cause a change in the other, this is not necessarily the case. However, selection ultimately acts on the phenotype. There are many phenotypes at many scales, from molecules (e.g., mRNA expression) to life history (e.g., number of offspring). Here, I will be focusing primarily on changes in allele frequencies and a summary phenotype, relative fitness,

which is the combined growth and reproductive success of an individual with a given genotype relative to an individual of another genotype.

A simple case of a change in allele frequencies is when positive selection occurs. If an individual acquires a mutation that has a beneficial fitness effect, that individual will have an advantage over others in the population and survive and reproduce at a greater rate, allowing the mutation to increase in frequency, or “sweep”, though the population. Conversely, if a mutation is deleterious, it will be kept at a low frequency or removed from the population by purifying selection. Selection, however, is not the only force that controls allele frequencies in a population. They may also change due to genetic drift which is a neutral process. Since drift is essentially sampling error, the effect of drift is dependent on population size. Thus, a mutation with no fitness effect can still change in frequency, and mutations can change in frequency in directions opposite of what would be expected from their fitness effects. We can thus predict the ultimate fate of a mutation based on the selection coefficient (or fitness effect) of the mutation and the size of the population.

The above scenarios consider a simple situation, in which a population is initially isogenic, a new mutation arises in a single individual, and the fate of that mutation is based on both the probability that the mutation survives drift and its selection coefficient. This would be the case in a so-called strong selection, weak mutation regime (J. H. Gillespie 1991) which occurs when the population size and mutation rate result in a low mutation supply, so that new mutations are sufficiently rare as to appear, and then sweep to fixation if beneficial or be purged if deleterious, before the next new mutation arrives.

There are many situations (probably most, in fact), in which things are not this simple. One such situation is when a population is sufficiently large that before a beneficial mutation fixes another individual in the population acquires a different beneficial mutation. If the second beneficial mutation has a higher selection coefficient (i.e., is more beneficial) than the first, it can

outcompete the first mutation which subsequently disappears from the population. This phenomenon, in which multiple beneficial lineages simultaneously compete with each other, is called clonal interference (Gerrish and Lenski 1998). Experimental evolution studies have shown that clonal interference is very common using a variety of methods. These include whole-genome, whole-population sequencing at multiple timepoints, in which the relationship between the dynamics of different alleles is used to infer lineage dynamics (Kvitek and Sherlock 2013; Hong and Gresham 2014a); divergence and non-linearity in trajectories of individuals with different phenotypic markers in the population (Frenkel, Good, and Desai 2014; Lang et al. 2013); and introducing random barcodes to individuals in the population and tracking the frequency of the barcodes over time (Levy et al. 2015; Lauer et al. 2018; Nguyen Ba et al. 2019). The dynamics of competing lineages in populations are often shown with Muller plots (Muller 1932). When mutation rates are sufficiently high, not only can multiple individuals in a population acquire mutations at very close to the same time, but a single individual may acquire multiple different mutations before any fix. In this way, in asexual populations, neutral or even deleterious mutations can rise to high frequencies in populations by “hitchhiking” along with very beneficial mutations (Lang et al. 2013; Gresham et al. 2008).

Clonal interference and hitchhiking can result in interesting and complex evolutionary dynamics, because we cannot think about the fate of each mutation in isolation, but must consider it in the context of other mutations in the same genome and other mutations in the population. One such dynamic is that of the “traveling wave,” in which an evolving population initially diversifies as many different mutations arise in different lineages (Hallatschek 2011). As beneficial mutations spread, the mean population fitness increases and filters out less fit lineages, which should reduce diversity. However, different individuals in the expanding beneficial lineages gain new mutations, some of which are beneficial, and thus diversity is both maintained and the “wave” of mutants of higher and higher fitness travels onward (Rouzine,

Brunet, and Wilke 2008; Nguyen Ba et al. 2019). Thus, a population can maintain high diversity while the mean population fitness continually increases. In this type of scenario, a mutation that is very beneficial may rise to high frequency but ultimately be lost from the population if a lineage that is initially less fit but still beneficial acquires subsequent beneficial mutation(s) that give it the ability to “leapfrog” over the other lineage (Gerrish and Lenski 1998; Nguyen Ba et al. 2019).

A further complication is that the effect of a mutation may differ when it arises in a lineage that already has a mutation compared with its effect in the ancestral lineage. We often assume that the effects of two mutations are simply the sum or product of the effect of each individual mutation, but in fact, mutations may interact with each other so that their combined effect results in lower or higher than expected fitness. This phenomenon is known as epistasis. Therefore, some evolutionary trajectories may be less likely than others because they require particular mutations to occur in a particular order (Blount, Lenski, and Losos 2018). While there are many types of epistasis (reviewed in (Domingo, Baeza-Centurion, and Lehner 2019)), one particularly interesting type is diminishing returns epistasis, which occurs when the effect of new beneficial mutations is smaller in more fit lineages, has frequently been observed in evolution experiments, and results in a decline in the rate of adaptation as evolution progresses (Good and Desai 2015; Kryazhimskiy et al. 2014; Wei and Zhang 2019; Khan et al. 2011).

1.1.2 Evolutionary dynamics depend on the parameters

As discussed above, mutations do not all arise at the same rate or have the same fitness effect. Generally, the majority of mutations are deleterious, many mutations are neutral or nearly neutral, and beneficial mutations are rare, though the relative frequencies of each of these categories, or the distribution of fitness effects (DFE), differs between species and environments. It is important to note that the DFE is context dependent – mutations that are

beneficial in one environment may be neutral or deleterious in another (i.e., exhibit pleiotropy) and mutations that are beneficial in one genetic background may not be in a different genetic background (i.e., exhibit epistasis). Despite these complications, it is still important to understand the DFE in order to understand quantitative genetic variation and predict evolutionary outcomes. Since highly deleterious mutations are quickly purged from populations by purifying selection and many mutations are effectively neutral, many studies have focused on defining the DFE for beneficial mutations. Early theoretical work predicted that the beneficial DFE would be exponential (J. H. Gillespie 1984; Orr 2003), which has mixed experimental support (Imhof and Schlotterer 2001; Kassen and Bataillon 2006; Rokyta et al. 2005, 2008). Recent studies using thousands of barcodes to track lineage dynamics in experimentally evolving populations have found evidence for a non-monotonic beneficial DFE (Levy et al. 2015; Nguyen Ba et al. 2019).

Additionally, it is important to note that mutation rates themselves can evolve. In experimental evolution hypermutators, which have mutations that cause defects in DNA repair or proofreading resulting in increased mutation rates, often arise (Sniegowski, Gerrish, and Lenski 1997; Raynes and Sniegowski 2014). While the fitness effect of increased mutation rates are generally negative, since most mutations are deleterious, hypermutators are more likely to generate rare highly beneficial mutations or to generate a combination of beneficial mutations, and can therefore hitchhike to high frequencies with the beneficial mutations they generate (Desai and Fisher 2011). Features of the genome and genomic context also impact mutation rate across the genome. There is a GC bias in mutation (Lynch 2010) and sites with neighboring GC pairs have much higher mutation rates than sites with other neighboring nucleotides (Sung et al. 2015). Other genomic features, such as repetitive elements, can also result in elevated mutation rates (Moxon, Bayliss, and Hood 2006; K. Zhou, Aertsen, and Michiels 2014).

In addition to multiple beneficial alleles at different loci contemporaneously increasing in frequency in a population, multiple alleles at the same locus can sweep through the population at the same time, in what is termed a “soft sweep”. Whereas soft sweeps were originally conceived of as occurring from standing genetic variation (Hermisson and Pennings 2005), when mutation supply is high and/or the mutational target is large, multiple alleles at the same locus can arise *de novo* in a population (Messer and Petrov 2013). What does this mean for evolutionary dynamics in asexual populations? At the locus of interest, recurrent mutation may result in mutation stacking, in which multiple different mutations occur at the same locus and the haplotype with the most beneficial mutations is the most fit. However, there can also be negative epistatic effects between alleles at the same locus or even mutually exclusive alleles (for example, a codon deletion would prevent nonsynonymous change at the codon) (Hermisson and Pennings 2017). When soft sweeps occur, different alleles at the same locus will be linked to different alleles in the rest of the genome, resulting in maintenance of diversity, unlike in hard sweeps, where selection for the single most fit haplotype reduces the diversity in the population. Maintenance of diversity may be beneficial for a population in a selective environment, as more evolutionary routes remain open to the population. It may also be beneficial when selection is relaxed, as different alleles may have different trade-offs in other environments (Hermisson and Pennings 2017). Importantly, the probability of soft sweep depends not only on the allelic mutation rate, but the genome-wide population mutation rate (Hermisson and Pennings 2017).

What happens when different types of mutations have different parameters? The rate of adaptation, or the increase in frequency of beneficial mutants, for each mutation type will depend on both the mutation rate and the selection coefficient. Recently, a study investigated how adaptation rate affects two traits evolving together using models and simulation (Gomez, Bertram, and Masel 2020). They found that when the rates of adaptation are the same for two traits when evolving separately, when evolving together, the trait with the higher selection

coefficient will be more successful due to clonal interference. However, when the traits have not only different mutation rates and selection coefficients, but different rates of adaptation on their own, when they evolve together the one with the higher rate of adaptation continues adapting and adaptation in the other trait “stalls”, even when clonal interference is present. That means that high mutation rates can “bias” evolution if the adaptation rate is higher, even if the other type of trait has a higher selection coefficient (Gomez, Bertram, and Masel 2020). A study in which researchers evolved bacteria with perturbed translation machinery found that adaptation initially occurred through mutations in the translation machinery (Venkataram et al. 2020). However, translation machinery adaptation soon stalled before it was restored to wild type function, and other modules instead adapted, suggesting the rate of adaptation for translation machinery was less than that for other modules. Relatedly, a recent study investigated how the contribution of mutation rates and fitness effects depend on population size in populations of bacteria adapting to an antibiotic (Schenk et al. 2022). They observed that in the larger population, mutations of large effect that occurred at a lower rate drove evolutionary dynamics, while in the small population high-rate mutations of smaller effect size drove the dynamics. While the populations were both large enough to exhibit clonal interference, the intensity of this effect was lower in the smaller population, leading to the observed differences. Thus, different classes of mutation, not just the overall mutation rate and DFE, make important contributions to evolutionary dynamics.

1.2 What is a CNV?

The focus of my thesis is a particular class of mutation called copy number variation (CNV). The working definition of CNV varies in the literature because our appreciation of genomic structural variation has quickly evolved as a vast diversity of genomes are characterized with ever-improving DNA sequencing technologies. While CNVs were initially

conceived as duplication or deletion of a putative functional unit of DNA, such as a gene, intron, exon, promoter, enhancer or other regulatory region (Ohno 1970; Lewis 1978), more recently CNV has been defined as an increase or decrease in the number of copies of a DNA sequence that range in size from a few bases to an entire chromosome. CNVs have previously been defined based on a range of minimal length of sequence that is duplicated or deleted, including >50 (Iafrate et al. 2004; Sebat et al. 2004; Feuk, Carson, and Scherer 2006; Korbel et al. 2007), >100 (F. Zhang et al. 2009), and >1000 base pairs (Feuk, Carson, and Scherer 2006; Itsara et al. 2009; Zarrei et al. 2015; F. Zhang et al. 2009). Historically, these definitions have been influenced by the resolution of available technologies used to detect CNVs; the first CNVs were detected by microscopy-based cytogenetic methods (Bridges 1936), array-based comparative genomic hybridization (array CGH) was common in the early 21st century, and currently short and long read comparative genome sequencing are being used to more accurately define the size and structure of CNVs (Iafrate et al. 2004; Sebat et al. 2004; Feuk, Carson, and Scherer 2006; Korbel et al. 2007). The above definitions have usually distinguished a CNV from an “indel”, or small insertion or deletion (< 1kb) (Werdyani et al. 2017; Scherer et al. 2007). Some have also postulated that indels form by different mechanisms than CNVs, such as polymerase slippage during replication (Scherer et al. 2007; Montgomery et al. 2013); however, CNVs can form by a variety of mechanisms including replication slippage (for more information about mechanisms of formation of CNVs see Section 1.4.1) making a definition based on formation mechanism problematic.

The important aspect of a CNV is in the name - CNV refers to some DNA sequence in a genome that varies in copy relative to other genomes (Pös et al. 2021). Thus, a CNV may exist in a genome in comparison to a reference or ancestral genome sequence, or in a population of individuals. This definition excludes structural variants that do not change the copy number, such as inversions. Another class of variation that is distinguished from CNVs are whole

genome duplications, which increase copy number but do not alter the relative copy number of any individual element in the genome. Using this definition, CNV is an umbrella term that encompasses a large number of variants with diverse structures, from small repeated sequences including microsatellites, to whole chromosome gain or loss (i.e. aneuploidy), to more complex structures including unbalanced translocations or amplifications with multiple inverted copies. A key defining feature of CNVs is that they comprise a difference in DNA content *with respect to the rest of the genome* thereby altering the stoichiometric relationship between individual genes or genetic elements and their encoded products.

1.3 Experimental evolution reveals CNVs as a major source of adaptation

Experimental evolution is a process by which scientists expose a population of organisms to a well defined selective pressure in the laboratory over the course of many generations and study evolutionary processes as they occur (Garland and Rose 2009; Bailey and Bataillon 2016; Kawecki et al. 2012). Experimental evolution subjects organisms to a strong selection over a short period of time, and CNVs have repeatedly been shown to be an important mechanism of adaptation in these regimes.

Many evolution experiments have been performed in microbes, because of their short generation, ease of maintenance, and ease of genetic analysis. Early evolution experiments in bacteria and phages showed tandem duplications are a common mechanism of adaptation (R. P. Anderson and Roth 1977). Many evolution experiments in microbes have shown that CNVs involving genes that encode for nutrient transporters often arise in nutrient limited conditions. The first studies to show this were of lac operon amplifications in *Escherichia coli* limited for lactose (Horiuchi, Tomizawa, and Novick 1962; Horiuchi, Horiuchi, and Novick 1963), *Saccharomyces cerevisiae* limited for phosphate (Hansche 1975), and *Salmonella typhimurium* limited for different carbon sources (Straus 1975; Sonti and Roth 1989). Subsequently, multiple,

independent studies have identified amplifications of the high affinity glucose-transporter genes *HXT6/7* in glucose-limited media (Brown, Todd, and Rosenzweig 1998a; Kao and Sherlock 2008; Gresham et al. 2008). *SUL1*, which encodes a high-affinity sulfur transporter, has also been reported as a frequent target of amplification under sulfur-limited conditions from independent trials (Payen et al. 2014; Gresham et al. 2008). The authors of one study report that CNVs at the *SUL1* locus are responsible for fitness increases as large as 50% over the ancestral strain (Gresham et al. 2008). More recently, specific amplification alleles for the corresponding limiting nutrient have been identified in the high-affinity proline transporter *PUT4*, the urea transporter *DUR3*, the allantoin permease *DAL4*, the general amino acid permease *GAP1* and the ammonia permease *MEP2* (Lauer et al. 2018; Gresham et al. 2008; Hong and Gresham 2014a). Importantly, amplifications may be single gene amplifications, or comprise long chromosomal segments encompassing many additional genes.

Evolution experiments have been performed with metazoan model organisms including the worm *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster*. In contrast to microbial evolution experiments where asexual propagation is routinely (though, not always) used, these organisms can reproduce sexually, leading to important insights for evolutionary processes. A study using *C. elegans* imposed 200 generations of selection for recovered fecundity after reduced productivity due to mutation accumulation and inbreeding. Duplications and deletions increased in frequency over time in many replicate populations, and CNVs were enriched for genes related to reproduction and development and often spanned the same region, suggesting strong positive selection (Farslow et al. 2015). Genomic analysis of a fly line reared in darkness for 1400 generations found about 150 putative CNVs. One verified CNV contained a 500 base pair deletion within *CG459*, a gene of unknown function whose mammalian homologues are involved in fatty acid metabolism in the mitochondria (Izutsu et al. 2012). Despite their importance and prevalence in *Drosophila* populations (Zichner et al. 2013),

CNVs are not as frequently addressed as single nucleotide variants in many evolution experiments (Burke et al. 2010; T. L. Turner et al. 2011; D. Zhou et al. 2011; Remolina et al. 2012; Jalvingh et al. 2014; Kang et al. 2016; Graves et al. 2017; Kezos et al. 2019; Phillips et al. 2018), and should be considered an active area of research for further study.

These evolution experiments have given important insights into the dynamics, repeatability, and mechanisms of CNV evolution. In many microbial experiments, CNVs arise early and repeatedly in response to strong selection (Lauer et al. 2018; Payen et al. 2014; Sun et al. 2012; Morgenthaler et al. 2019). This striking degree of parallelism, with CNVs identified early at high frequency, has been observed in experimental evolution in other systems including the algae *Chloralla variablis* co-evolving with a virus (Frickel et al. 2018), *Caenorhabditis elegans* (Farslow et al. 2015), and *Arabidopsis thaliana* (DeBolt 2010). In shorter experimental evolution (< 1000 generations), CNVs are frequently maintained at high frequency for the duration of the experiment (Sunshine et al. 2015; A. M. Selmecki et al. 2009; Lauer et al. 2018; Payen et al. 2014; Morgenthaler et al. 2019). In longer experiments (> 1000 generations), CNVs can be maintained in the population (Fisher et al. 2018), or replaced by other high-fitness mutations or revert to the ancestral state (Yona et al. 2012). Like other types of mutations (Blundell et al. 2019), CNV dynamics seem to be much more predictable in the early stages of evolution, and more stochastic later on (Lauer et al. 2018). Many different mechanisms of CNV formation have been observed to contribute to adaptive dynamics in evolution experiments (Lauer et al. 2018; Dunham et al. 2002; Todd et al. 2019; Schacherer et al. 2005), though the relative frequency of each of these mechanisms remains unknown, and is likely to be species, locus, and condition specific.

CNVs spanning the same region have been observed in many replicate populations during the same experiment, suggesting that CNVs containing specific genes are under strong positive selection (A. Selmecki, Forche, and Berman 2006; Farslow et al. 2015; Gresham et al.

2008). Several studies have performed in depth analysis to find the exact genes/features inside CNV that are under selection (A. Selmecki et al. 2008; Mount et al. 2018). However, this can be difficult and time consuming and often a single gene within the amplified region is inferred to be the element of the CNV under selection, based on the selective regime (Lauer et al. 2018; Hope et al. 2017; Gerstein et al. 2015). This is challenging as often CNVs arise that encompass many genes, which may be due to more than one gene being under selection, and/or because there are several mechanisms of CNV formation that result in recurrent mutation. Duplications and deletions with the exact same breakpoints in independent populations have been found in several experiments (Farslow et al. 2015; Lauer et al. 2018). This high rate of recurrent CNV formation suggests that mutation rate at the locus can have as large of an impact as selection (Farslow et al. 2015; Lipinski et al. 2011).

While empirical studies performed in laboratory settings are important for determining the role of CNVs in adaptive evolution, there are a few caveats. Natural environments are complex, and can fluctuate (for example, in temperature, predation rates, or nutrient content). Even subtle variations in the environment can cause selective pressures to vary, or can increase the consequences of antagonistic pleiotropy. This is in direct contrast to adaptive laboratory evolution and artificial selection such as domestication, where a single, strong selective pressure is applied. Furthermore, experimental evolution is predominantly performed in asexually reproducing microbes, often with very large population sizes. These characteristics which affect the mutation supply rate and eliminate the effect of recombination, which are important population genetic parameters. In a recent study using *Leishmania*, the authors detected whole-chromosomal aneuploidies as major drivers of adaptation during in vitro culture, but identified smaller CNVs from clinical isolates adapting in the field (Bussotti et al. 2018). In order to validate insights gleaned from experimental evolution, studies that examine the role of

CNVs in natural populations are integral to a comprehensive understanding of the role of CNVs in adaptation.

1.4 CNV formation

1.4.1 Different formation mechanisms give rise to different CNV structures

CNVs are formed through complex and diverse processes, but the molecular basis of these events is still being elucidated (see (Hastings, Lupski, et al. 2009; Reams and Roth 2015) for excellent reviews). Highly repetitive elements, including centromeres (reviewed in (Barra and Fachinetti 2018)), telomeres, and transposable elements, are often implicated in CNV formation, as they are prone to DNA breakage, and the repetitive nature of these elements facilitates homologous recombination. Non-allelic homologous recombination (NAHR) between repetitive sequences is a major driver of CNV formation. In prokaryotes, small insertion sequence (IS) elements are flanked by long terminal repeats (LTRs) (Siguier, Gourbeyre, and Chandler 2014; Brügger et al. 2002), which are also frequently found in the yeast genome (Carr, Bensasson, and Bergman 2012). Many eukaryotes have longer repetitive elements including segmental duplications, which are >1 kilobase in size and are dispersed throughout the genome (Eichler 2001). Extensive homology between repetitive sequences enables recombination and can lead to increases or decreases in copy number (Peng et al. 2015), even with long intervening distances between repeats (Todd et al. 2019).

Transposable elements can also mediate CNV formation through reverse transcription and insertion into the genome (reviewed in (Casola and Betrán 2017)). This mechanism of CNV formation is unique in that it can only create duplications, and not deletions (Schacherer et al. 2004; Ewing et al. 2013). Duplicates formed through retrotransposition lack introns and the promoter, contain a poly-A tract, and have low linkage disequilibrium with surrounding sequence (Schrider and Hahn 2010; Schacherer et al. 2004).

CNVs can be generated through replication errors (Koszul et al. 2004; Cardoso-Moreira, Arguello, and Clark 2012; L. Chen et al. 2015) that involve replication slippage (Ohye et al. 2014), template switching (Slack et al. 2006), sequence microhomology (Hastings, Ira, et al. 2009), and/or the generation of extrachromosomal circles or circular intermediates (Gresham et al. 2010; Brewer et al. 2011, 2015; Møller et al. 2015; Cohen and Segal 2009; K. M. Turner et al. 2017). Replication stress has been directly linked to increased generation of CNVs in human cells, including variants associated with disease and tumorigenesis (Durkin et al. 2008; Arlt et al. 2009). The extent of replication-mediated CNV formation may have been previously underestimated (Lauer et al. 2018), since many early studies focused on recurrent disease-related variants formed by NAHR or retro-transposition, which can be easier to detect and characterize.

Recurrent CNVs, which repeatedly occur in specific regions of the genome, typically underlie re-occurring germline mutations and human disease (Itsara et al. 2009; Girirajan, Campbell, and Eichler 2011). Current evidence suggests that CNVs are enriched in pericentromeric and subtelomeric chromatin (Zarrei et al. 2015), and that recurrent CNVs arise due to specific features of the neighboring genomic sequence including: repetitive elements (Farslow et al. 2015), tRNA genes (Bermudez-Santana et al. 2010), origins of replication (Di Rienzi et al. 2009), and replication fork barriers (Labib et al. 2007).

1.4.2 CNVs of different types may form and revert to the ancestral state at different rates

Gene duplications and deletions occur at a higher rate than SNVs. Early on, researchers studying gene duplications, including the bar mutation, noticed this phenomenon (Sturtevant 1925). A clever genetic screen in *E. coli* revealed that mutation rates were high and that cells with amplifications quickly rose to a high frequency (Cairns and Foster 1991; Hastings et al. 2000). Reported frequencies of duplications per locus per generation range from 10^{-2} to 10^{-6} in *E. coli* and *Salmonella (*R. P. Anderson and Roth 1977; Horiuchi, Horiuchi, and Novick 1963;

Reams et al. 2010; P. Anderson and Roth 1981; Starlinger 1977; Langridge 1969), 10^{-6} in yeast (Lynch et al. 2008), 10^{-4} to 10^{-6} in *Drosophila* (Gelbart and Chovnick 1979; Shapira and Finnerty 1986), and 10^{-5} to 10^{-7} in human sperm (Lam and Jeffreys 2006; D. J. Turner et al. 2008). Further discussion of CNV formation rates are in (Katju and Bergthorsson 2013).

CNVs formed by different mechanisms likely form at different rates. Each locus in the genome may have a different mutational spectrum, and these mutational spectra can be influenced by many factors, including local features, transcription, and the current environment. Many studies have shown that stress can lead to increases in genome-wide SNP and indel mutation rates in bacteria and yeast (Foster 2007; Galhardo, Hastings, and Rosenberg 2007; Shor, Fox, and Broach 2013), as well as multicellular organisms (reviewed in (Fitzgerald, Hastings, and Rosenberg 2017)). Rates of CNV formation have also been shown to be elevated under stress (Fitzgerald, Hastings, and Rosenberg 2017; Fitzgerald and Rosenberg 2019), and though most research has been undertaken in bacterial models, evidence is emerging that this is true in eukaryotes as well (Chain et al. 2019; Shewaramani et al. 2017). Active transcription units may play a role in elevating mutation rates and generating these hotspots (Thomas and Rothstein 1989; Skourti-Stathaki and Proudfoot 2014; Wilson et al. 2015). Increases in the rate of transcription lead directly to amplification of the rDNA and other loci (Jack et al. 2015; Hull et al. 2017). A recent study in yeast found that while overall mutation rates decrease with decreasing growth rate, the rate of CNV formation increases in slow growing cells (H. Liu and Zhang 2019). Additionally, in some environments (e.g., lithium chloride) there were concurrent increases in the rate of formation of aneuploidy and sub-chromosomal sized CNVs. However, in other environments (e.g., sodium chloride) an increase in the rate of formation of aneuploidy was observed, but not other types of CNV (H. Liu and Zhang 2019). In human cells, hypoxia, but not other stressors, induces site-specific amplification that is also commonly observed in tumors. Interestly, one of these CNV regions is syntenic in zebrafish cells and also undergoes

amplification in hypoxic conditions, while the other is not and is not amplified in hypoxia (Black et al. 2015). This implicates the role of chromosomal architecture in recurrent CNV formation. Indubitably, certain regions of the genome are more susceptible to mutation. Understanding the full repertoire of mechanisms that underlie CNV formation, how these differently contribute to the mutational spectrum by locus, and whether these processes can be directly stimulated by the environment are important open questions in the field.

Several mechanisms may change the mutational spectrum of CNVs by locus and environment. Studies in yeast have shown that rates of aneuploidy increase in response to reduced Hsp90 activity, due to the role of Hsp90 in kinetochore assembly (G. Chen et al. 2012). Reduced Hsp90 activity also results in increased transposon mobilization (Kaplan and Li 2012), which can lead to single gene duplications. Other studies have suggested that epigenetic marks may also change the mutational spectra between loci. MicroRNAs and histone methylation have both been shown to suppress site specific amplifications in human cells (Black et al. 2016; Mishra et al. 2018). An experimental evolution study in yeast selecting increased gene expression found that genes with nucleosome-free promoters achieved greater expression by duplication, while genes with dynamic promoter regions achieved greater expression through point mutations (Rosin et al. 2012). Ploidy can also contribute to changes in the mutational spectra, as diploid cells can mask deleterious mutations or reduce deleterious stoichiometric effects of CNV that haploids cannot (Fisher et al. 2018). Recently, a group simulated transcription/replication conflicts in yeast that resulted in CNVs or point mutations at different rates. The performed *in silico* evolution experiments, and found that populations with higher rates of CNV were more fit because duplications could mask inactivating point mutations and deletions could eventually purge copies that had been inactivated by point mutations, resulting in a lower genetic load, even at higher mutation rates (Colizzi and Hogeweg 2019). These

studies show how the relative rates of different types of mutations can affect evolutionary outcomes.

CNVs formed by different mechanisms may revert to the ancestral euploid state at different rates as well. Several studies have shown that in yeast and mammalian cells, aneuploidies revert readily to ancestral euploid state when selection for aneuploidy is released (Yona et al. 2012; G. Chen et al. 2012; A. Selmecki, Forche, and Berman 2006; Rosin et al. 2012). Tandem duplications have also been shown to readily collapse to the ancestral copy number when selection is released or when one copy gains a high fitness beneficial SNV (Morgenthaler et al. 2019). More complex amplifications may be more difficult to resolve back to ancestral states, though due to their repetitive sequence they can undergo recombination and remodeling (Morgenthaler, Fritts, and Copley 2022). Heterozygous deletions could revert back to ancestral state by recombination with the homologous chromosome. While in some cases (for example, aneuploids), the rate of CNV reversion may be approximately equal to the rate of formation, in other cases different mechanisms may act in formation and reversion, resulting in different rates and increased complexity in evolutionary dynamics

1.5 CNVs and fitness

1.5.1 CNVs mediate rapid adaptation through a variety of mechanisms

CNV formation can have immediate consequences for organismal fitness. Since CNVs typically encompass large regions of the genome, they can affect multiple protein-coding genes and regulatory regions simultaneously. Large duplications and deletions leading to increases or decreases in gene dosage can subsequently result in widespread protein abundance changes (Rice and McLysaght 2017a; Tang and Amon 2013). CNVs can affect neighboring loci, leading to changes in expression for genes outside the CNV boundary (Brooks et al. 2022; Molina et al. 2008; Merla et al. 2006). CNVs can also have effects in *trans*: by changing the expression of

distal transcripts (Gamazon, Nicolae, and Cox 2011), by affecting global levels of transcription (Henrichsen et al. 2009), and by changing the topology of chromatin organization (Lupiáñez, Spielmann, and Mundlos 2016; Lupiáñez et al. 2015; Franke et al. 2016; Spielmann, Lupiáñez, and Mundlos 2018).

We typically think of CNVs as protein-coding gene deletions or duplications, however, copy number changes in intergenic sequences have also been identified. CNV formation can result in position effects that disrupt or modify regulatory elements (Koszul et al. 2004; Chan et al. 2010). Promoter capture, where spatial re-arrangement of an amplified DNA segment leads to its regulation by a different promoter, has been observed repeatedly in diverse systems (Usakin et al. 2005; Adam, Dimitrijevic, and Schartl 1993; Whoriskey et al. 1987). Perhaps the most famous example of promoter capture is the Cit+ lineage in Richard Lenski's long term evolution experiment, in which citrate utilization evolved in *Escherichia coli* by duplication of the citrate transporter that resulted in one copy being controlled by a new promoter producing transcription of the transporter in the evolution environment and subsequent use of citrate as a carbon source (Blount, Borland, and Lenski 2008; Blount et al. 2012). Similarly, amplifications or deletions of enhancers and introns have been identified as key adaptive events in stickleback fish (Bell 1987; Chan et al. 2010), chickens (Wright et al. 2009), and rice (Wang et al. 2015).

In addition to the gene expression changes described above, CNVs can affect the fitness of an organism through other mechanisms. Gene duplications can increase fitness by buffering fluctuations in gene expression (Rodrigo and Fares 2018), masking deleterious mutations (Gu et al. 2003), or promoting heterozygote advantage (Sellis et al. 2016). While CNVs are often thought of as a substrate for future innovation through duplication and subsequent divergence, *de novo* CNVs can immediately provide new functionality. For example, gene duplications, deletions, and unbalanced translocations can lead to the formation of chimeric genes (Rippey et al. 2013; Mayo et al. 2017; Arguello et al. 2006; Aigner et al. 2013; Schrider et al. 2013).

Therefore, CNVs can drive important adaptive innovations during short-term evolutionary scenarios. Collectively, these findings demonstrate that a single class of mutation can provide a range of functional effects and adaptive phenotypes.

1.5.2 CNVs have costs

CNVs, as a class of mutation, can have variable copy number and size (resulting in variable numbers of genes within the CNV). As a result, there may be many differently structured CNV alleles at a specific locus within a population (Lauer et al. 2018). This is important because selection on CNV alleles is not only dependent on selection for increased or decreased dosage of a single gene, but the aggregate effect of the entire allele. In some systems, the co-duplication of adjacent genes specifically provides a fitness benefit (Reams and Neidle 2004). However, large CNVs are more often associated with fitness costs, which can be attributed to disruption of cellular homeostasis at multiple levels: inherent costs due to increases in genome size (Elde et al. 2012), changes to local and global gene expression (Sheltzer et al. 2012), increased translational capacity and changes to protein stoichiometry (Torres et al. 2007), or increased burden on protein degradation machinery (Torres et al. 2010; Stingele et al. 2012). In one study, there was a 0.15% reduction in fitness for every kilobase pair amplified in *E. coli* (Adler et al. 2014). However, in another study, there was no correlation between the size of the duplicated region and fitness reduction for the organism (Pettersson et al. 2009).

Many studies have investigated the basis of fitness costs of aneuploidies, primarily in yeast (recently reviewed in (Tsai and Nelliat 2019), but also in mammalian cells. The largest costs have been associated with overproduction of proteins on the duplicated chromosome, however, there is conflicting evidence as to whether this is driven by a few especially harmful genes or by mass action of all genes involved in the aneuploidy (Bonney, Moriya, and Amon 2015). The phenotypes resulting from this proteotoxic stress include cell cycle delays, DNA

damage, increased sensitivity to some drugs, lethality, and increased intracellular osmolarity leading to hypo-osmotic stress (Torres et al. 2007; Tsai et al. 2019; Sheltzer et al. 2011; Anders et al. 2009; M. Li et al. 2010).

One hypothesis that regards a few harmful genes as the basis of fitness costs associated with CNVs in the gene dosage balance hypothesis (also called the gene balance hypothesis or the dosage balance hypothesis) (Veitia 2004; Birchler and Veitia 2012; Papp, Pál, and Hurst 2003). This hypothesis proposes that the fitness cost of CNVs arises from stoichiometric imbalances in macromolecular complexes, signaling pathways, and protein-protein interactions. This imbalance may result in excess subunits or components that are not stabilized by their normal interacting partners, that are then susceptible to degradation, putting stress on the proteostasis system (Veitia, Bottani, and Birchler 2008). In the case of a deletion, it can shift the reaction to unproductive subcomplexes or change reaction speed, producing less of the complete product, which in turn affects other cellular processes (Birchler and Veitia 2012).

While some fitness costs of sub-chromosome arm sized CNVs probably also arise from large scale proteostatic imbalances, the basis of fitness trade-offs of the CNVs may be condition dependent, rather than a generalizable cost (Sunshine et al. 2015). These condition dependent costs appear to result from deleterious misexpression of a few proteins, though few studies have investigated the basis of fitness cost in “smaller” CNVs in depth. Since natural populations are usually not under a single static selective pressure, and experience a wide range of environmental conditions, understanding why and how CNVs can be deleterious in some conditions is important for understanding their evolution in nature.

Since CNVs can confer substantial fitness costs, it has been proposed that they may occur transiently and are not an effective long-term solutions for organisms adapting to stressful

conditions (Yona et al. 2012). One proposed mechanism of alleviating fitness costs is the use of "genomic accordions," which involve expansions and contractions of genic arrays (Roth and Andersson 2012; Elde et al. 2012). Gene duplications occur at a high rate, and incremental increases in gene dosage improve cell growth such that cells with the duplication rise to high population frequency. Multiple gene copies, as well as many individuals with multiple copies, increase the likelihood of generating beneficial SNVs in the gene under selection (Sun et al. 2009). If these SNVs provide significant fitness benefits, selection on maintenance of multiple gene copies is relaxed, and the alternative copies can be subsequently lost. This phenomenon has been observed in a variety of systems: viruses adapting to host defenses (Elde et al. 2012), bacteria growing in lactose-limiting environments (Slechta et al. 2003), the evolution of antibiotic resistance (Pränting and Andersson 2011; Paulander, Andersson, and Maisnier-Patin 2010), and the evolution of metabolic enzymes (reviewed in (Copley 2012)). However, a recent study in *E. coli* evolving for 500-1000 generations under selection for improved function of an enzyme's weak secondary activity found that in all eight experimental populations the enzyme was amplified, a secondary mutation that improved the enzyme's function in an amplified copy that resulted in contraction of the amplification only occured in one population (Morgenthaler et al. 2019). Instead, the majority of populations (⅞) had adaptive mutations outside the enzyme under selection rise to high frequency after the initial amplification (Morgenthaler et al. 2019). This suggests that there are many ways to compensate for CNVs, and the evolutionary outcomes of CNV may be complex and dependent on the CNV structure and genetic background.

Chapter 2: Single-cell copy number variant detection reveals the dynamics and diversity of adaptation

This chapter is based on "Single-cell copy number variant detection reveals the dynamics and diversity of adaptation" by Stephanie Lauer, **Grace Avecilla**, Pieter Spealman, Gunjan Sethia, Nathan Brandt, Sasha F. Levy, and David Gresham, published in PLoS Biology (2018) (https://doi.org/10.1371/journal.pbio.3000069).

Below is the abstract, then excerpts of the introduction, results, discussion, and methods to which I contributed or which are important for subsequent chapters. The text has been edited for clarity and relevance. I contributed to generating data for Figure 2.3 and Table 2.2, performed experiments for and generated Figure 2.4, Figure 2.S1, Figure 2.S2, and Table 2.S1, and contributed to writing the corresponding sections 2.3.5 and 2.3.6.

2.1 Abstract

Copy number variants (CNVs) are a pervasive source of genetic variation and evolutionary potential, but the dynamics and diversity of CNVs within evolving populations remain unclear. Long-term evolution experiments in chemostats provide an ideal system for studying the molecular processes underlying CNV formation and the temporal dynamics with which they are generated, selected, and maintained. Here, we developed a fluorescent CNV reporter to detect de novo gene amplifications and deletions in individual cells. We used the CNV reporter in Saccharomyces cerevisiae to study CNV formation at the *GAP1* locus, which encodes the general amino acid permease, in different nutrient-limited chemostat conditions. We find that under strong selection, *GAP1* CNVs are repeatedly generated and selected during

the early stages of adaptive evolution, resulting in predictable dynamics. Molecular characterization of CNV-containing lineages shows that the CNV reporter detects different classes of CNVs, including aneuploidies, nonreciprocal translocations, tandem duplications, and complex CNVs. Despite *GAP1*'s proximity to repeat sequences that facilitate intrachromosomal recombination, breakpoint analysis revealed that short inverted repeat sequences mediate formation of at least 50% of *GAP1* CNVs. Analysis of 28 CNV breakpoints indicates that inverted repeats are typically 8 nucleotides in length and separated by 40 bases. The features of these CNVs are consistent with origin-dependent inverted-repeat amplification (ODIRA), suggesting that replication-based mechanisms of CNV formation may be a common source of gene amplification. We combined the CNV reporter with barcode lineage tracking and found that 10^2–10^4 independent CNV-containing lineages initially compete within populations, resulting in extreme clonal interference. However, only a small number (18–21) of CNV lineages ever constitute more than 1% of the CNV subpopulation, and as selection progresses, the diversity of CNV lineages declines. Our study introduces a novel means of studying CNVs in heterogeneous cell populations and provides insight into their dynamics, diversity, and formation mechanisms in the context of adaptive evolution.

2.2 Introduction

Copy number variants (CNVs) drive rapid adaptive evolution in diverse scenarios ranging from niche specialization to speciation and tumor evolution (Conant and Wolfe 2008; Zuellig and Sweigart 2018; Shlien and Malkin 2009; Stratton, Campbell, and Futreal 2009). CNVs, which include duplications and deletions of genomic segments, underlie phenotypic diversity in natural populations (Barreiro et al. 2008; Iskow et al. 2012; Clop, Vidal, and Amills 2012; Żmieńko et al. 2014; Greenblum, Carr, and Borenstein 2015; Zarrei et al. 2015), and provide a substrate for evolutionary novelty through modification of existing heritable material

(Ohno 1970; Lynch and Conery 2000; A. L. Hughes 1994; R. P. Anderson and Roth 1977). Beneficial CNVs are associated with defense against disease in plants, increased nutrient transport in microbes, and drug resistant phenotypes in parasites and viruses (Iantorno et al. 2017; Cowell et al. 2018; Dolatabadian et al. 2017; Elde et al. 2012; Greenblum, Carr, and Borenstein 2015). Despite the importance of CNVs for phenotypic variation, evolution and disease, the dynamics with which these alleles are generated and selected in evolving populations are not well understood.

Long term experimental evolution provides an efficient means of gaining insights into evolutionary processes using controlled and replicated selective conditions (Lenski et al. 1991; Good et al. 2017). Chemostats are devices that maintain cells in a constant nutrient-poor growth state using continuous culturing (Gresham and Dunham 2014). Nutrient limitation in chemostats provides a defined and strong selective pressure in which CNVs have been repeatedly identified as major drivers of adaptation. CNVs containing the gene responsible for transporting the limiting nutrient are repeatedly selected in a variety of organisms and conditions including *Escherichia coli* limited for lactose (Horiuchi, Horiuchi, and Novick 1963), *Salmonella typhimurium* in different carbon source limitations (Sonti and Roth 1989), and *Saccharomyces cerevisiae* in glucose-, phosphate-, sulfur- and nitrogen-limited chemostats (Hong and Gresham 2014a; Gresham et al. 2010; Payen et al. 2014; Gresham et al. 2008; Kao and Sherlock 2008; Hansche 1975; Brown, Todd, and Rosenzweig 1998b). CNVs confer large selective advantages and multiple, independent CNV alleles have been identified within experimental evolution populations (Payen et al. 2014; Gresham et al. 2008; Kvitek and Sherlock 2011; Gresham et al. 2010). These findings suggest that CNVs are generated at a high rate, but estimates differ greatly, ranging from 1×10^{-10} to 3.4×10^{-6} duplications per cell per division, with variation in CNV formation rates potentially differing between loci and/or condition (Dorsey et al. 1992; Lynch et al. 2008). A high rate of CNV formation suggests that multiple, independent CNV-containing

lineages may compete during adaptive evolution resulting in clonal interference, which is characteristic of large, evolving populations (Lang et al. 2013; J. M. Hughes et al. 2012; Maddamsetti, Lenski, and Barrick 2015; Kao and Sherlock 2008). However, the extent to which clonal interference among CNV-containing lineages influences the dynamics of adaptation is unknown.

The general amino acid permease gene, *GAP1*, is well suited to studying the role of CNVs in adaptive evolution. *GAP1* encodes a high-affinity transporter for all naturally occurring amino acids, and it is highly expressed in nitrogen-poor conditions (Grenson, Hou, and Crabeel 1970; Stanbrough and Magasanik 1995). We have previously shown that two classes of CNVs are selected at the *GAP1* locus in *S. cerevisiae* when a sole nitrogen source is provided: *GAP1* amplification alleles are selected in glutamine and glutamate-limited chemostats and *GAP1* deletion alleles are selected in urea- and allantoin-limited chemostats (Gresham et al. 2010; Hong and Gresham 2014a). *GAP1* CNVs are also found in natural populations. In the nectar yeast, *Metschnikowia reukaufii*, multiple tandem copies of *GAP1* result in a competitive advantage over other microbes when amino acids are scarce (Dhami, Hartwig, and Fukami 2016). As a target of selection in adverse environments in both experimental and natural populations, *GAP1* is a model locus for studying the dynamics and mechanisms underlying both gene amplification and deletion in evolving populations.

CNVs are generated by two primary classes of mechanisms: homologous recombination and DNA replication (Hastings, Lupski, et al. 2009; Reams and Roth 2015; Carvalho and Lupski 2016). DNA double strand breaks (DSBs) are typically repaired by homologous recombination and do not result in CNV formation. However, non-allelic homologous recombination (NAHR) can generate CNVs when the incorrect repair template is used, which occurs more often with repetitive DNA sequences such as transposable elements and long terminal repeats (LTRs) (Stankiewicz and Lupski 2002). During DNA replication, stalled and broken replication forks can

re-initiate DNA replication through processes including break-induced replication (BIR), microhomology-mediated break-induced replication (MMBIR), and fork stalling and template switching (FoSTes) (J. A. Lee, Carvalho, and Lupski 2007; Hastings, Ira, et al. 2009; Payen et al. 2008). BIR is driven by homologous sequences, whereas MMBIR relies on shorter stretches of sequence homology. Recently, origin-dependent inverted-repeat amplification (ODIRA) has been identified as a novel mechanism underlying amplification of the *SUL1* locus in yeast (Brewer et al. 2011, 2015). ODIRA is mediated by short inverted repeat sequences that facilitate ligation of the leading and lagging strands following regression of the replication fork during DNA synthesis. ODIRA is hypothesized to involve the formation of an extrachromosomal circular intermediate that replicates independently and therefore requires an origin of replication within the amplified region. Subsequent integration of the circle into the original locus via homologous recombination results in an inverted triplication. Extrachromosomal circular DNA is common in yeast (Møller et al. 2015), can drive tumorigenesis (K. M. Turner et al. 2017), and may represent a rapid and reversible mechanism of generating adaptive CNVs (Møller, Andersen, and Regenberg 2013; Cohen and Segal 2009). Previously, we found that some *GAP1* amplifications are extrachromosomal circular elements. We hypothesized that $GAP1^{circle}$ alleles are generated as a result of NAHR between flanking LTRs resulting in their excision from the chromosome (Gresham et al. 2010). Identifying the mechanisms underlying CNV formation is required for understanding the roles of CNVs in evolutionary processes and human disease.

A key limitation to the study of CNVs in evolving populations is the challenge of identifying them at low frequencies in heterogeneous populations. CNVs are typically detected using molecular methods including qPCR, Southern blotting, DNA microarrays and sequencing (Gresham et al. 2010; Payen et al. 2014; Hong and Gresham 2014a). However, using any of these methods, *de novo* CNVs are undetectable in a heterogeneous population until present at high frequency (e.g. >50%). This precludes analysis of the early dynamics with which CNVs

emerge and compete in evolving populations. As CNVs usually comprise genomic regions that include multiple neighboring genes (Hong and Gresham 2014a), we hypothesized that CNVs could be identified on the basis of increased expression of a constitutively expressed fluorescent reporter gene inserted adjacent to a target gene of interest. A major benefit of this approach is that it detects CNVs independently of whole genome sequencing, enabling a high-resolution and efficient assay of CNV dynamics with single-cell resolution in evolving populations.

In this study, we constructed strains containing a fluorescent CNV reporter adjacent to *GAP1* in *S. cerevisiae* and performed evolution experiments in different selective environments using chemostats. The CNV reporter allowed us to visualize selection of CNVs at the *GAP1* locus in real time with unprecedented temporal resolution. We find that CNV dynamics occur in two distinct phases: CNVs are selected early during adaptive evolution and quickly rise to high frequencies, but the subsequent dynamics are complex. We find that *GAP1* CNVs are diverse in size and copy number, and can be generated by a range of processes including aneuploidy, non-reciprocal translocations and tandem duplication by NAHR. Nucleotide resolution analysis of *GAP1* CNV breakpoints revealed that CNV formation is mediated by short, interrupted inverted repeats for half of the resolvable cases, suggesting that replication-based mechanisms also underlie gene amplification at the *GAP1* locus. The presence of inverted repeats, in combination with a replication origin and inverted triplication, is consistent with *GAP1* CNV formation through ODIRA. ODIRA may be a major source of *de novo* CNVs in yeast, as these breakpoint features also characterize CNVs at an additional locus identified in our study, *DUR3*. To determine the underlying structure of the CNV subpopulation, we generated a lineage-tracking library using random DNA barcodes. FACS-based fractionation of CNV lineages and barcode sequencing identified hundreds to thousands of individual CNV lineages within populations, consistent with a high CNV supply rate and extreme clonal interference.

Together, our results show that CNVs are generated repeatedly by diverse processes, resulting in predictable dynamics, but that the long term fate of CNV-containing lineages in evolving populations is shaped by clonal interference and additional variation.

2.3 Results

2.3.1 Protein fluorescence increases proportionally with gene copy number

We sought to construct a reporter for CNVs that occur at a given locus of interest. Based on previous studies (Suzuki et al. 2011; Gruber et al. 2012; Kafri et al. 2016; Steinrueck and Guet 2017), we hypothesized that CNVs that alter the number of copies of a constitutively expressed fluorescent protein gene would facilitate single cell detection of *de novo* copy number variation. To test the feasibility of this approach, we constructed haploid *S. cerevisiae* strains isogenic to the reference strain (S288c) with one or two copies of a constitutively expressed GFP variant mCitrine (Griesbeck et al. 2001), and diploid strains with 1-4 copies of mCitrine, integrated into the genome.

Flow cytometry analysis confirmed that additional copies of mCitrine produce quantitatively distinct distributions of protein fluorescence (**Figure 2.1A**). Haploid cells with two copies of mCitrine have higher fluorescence than those with a single copy and there is minimal overlap between the distributions of fluorescent signal in the two strains. Normalization of the fluorescent signal by forward scatter, which is correlated with cell size, shows that the concentration of fluorescent protein is proportional to the ploidy normalized copy number of the mCitrine gene (i.e. one copy in a haploid results in a signal equivalent to two copies in a diploid and two copies in a haploid results in a signal similar to four copies in a diploid). Thus, the cell size-normalized fluorescent signal, or concentration, accurately reports on the number of copies of the fluorescent gene in single cells. Therefore, integrating a constitutively expressed

fluorescent protein gene proximate to an anticipated target of selection functions as a CNV reporter for tracking gene amplifications and deletions in evolving populations (**Figure 2.1B**).

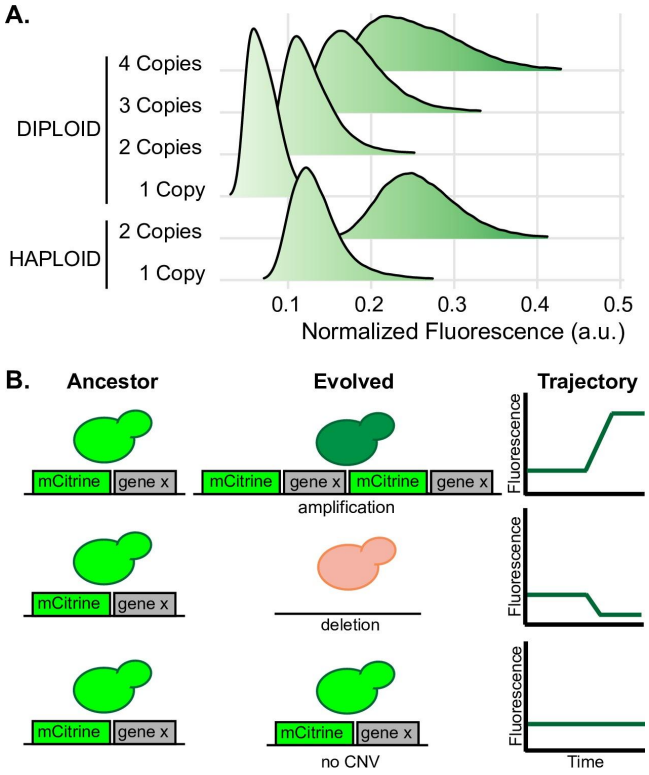

Figure 2.1. Fluorescent protein signal is proportional to gene copy number. (**A**) Protein fluorescence increases with increasing copies of the mCitrine gene. We determined the fluorescence of haploid and diploid cells containing variable numbers of a constitutively expressed mCitrine gene integrated at either the *HO* locus and/or the dubious ORF, *YLR123C*. The two copy diploid is heterozygous at both loci. Each distribution was estimated using 100,000 single cell measurements normalized by forward scatter. (**B**) Schematic representation of how the fluorescent reporter enables CNV detection in heterogeneous evolving populations through quantitative changes in protein fluorescence.

2.3.2 A CNV reporter tracks the dynamics of *GAP1* CNVs in real time

Previous work has shown that spontaneous *GAP1* amplifications are positively selected when glutamine is the sole limiting nitrogen source during evolution experiments in chemostats (Gresham et al. 2010). *GAP1* copy number amplifications result in increased amino-acid transporters on the plasma membrane, providing cells with a selective advantage when nitrogen is scarce (Gresham et al. 2010; Hong and Gresham 2014a). Conversely, *GAP1* deletions

provide a fitness benefit and are selected in urea-limited conditions (Gresham et al. 2010). Thus, the use of different nitrogen sources in nitrogen-limited chemostats enables the study of both *GAP1* amplification and deletion, making it an ideal system for studying the dynamics of CNV selection in evolving populations.

We constructed a haploid strain containing a mCitrine CNV reporter located 1,118 bases upstream of the *GAP1* start codon to ensure that the native regulation of *GAP1* was unaffected (Stanbrough and Magasanik 1996). We inoculated the *GAP1* CNV reporter strain into 9 glutamine-limited chemostats and included two control populations: one containing a single copy of the mCitrine CNV reporter at a neutral locus (one copy control) and one containing two copies of the mCitrine CNV reporter at two neutral loci (two copy control). All populations were maintained in continuous mode (dilution rate = 0.12 culture volumes/hr; population doubling time = 5.8 hours) for 267 generations over 65 days. We sampled each of the 32 populations every 8 generations and used flow cytometry to measure fluorescence of 100,000 cells per sample.

Experimental evolution in a glutamine-limited chemostat resulted in clear increases in fluorescence in individual cells containing the *GAP1* CNV reporter by generation 79 (**Figure 2.2A**). By contrast, populations containing one or two copies of mCitrine at neutral loci exhibited stable fluorescence for the duration of the experiment (**Figure 2.2A**). Maintenance of protein fluorescence in one and two copy control populations is consistent with the absence of a detectable fitness cost associated with one or two copies of the CNV reporter in glutamine-limited chemostats, which we confirmed using competition assays. Analysis of six additional independent populations evolving in glutamine-limited chemostats showed qualitatively similar dynamics of single-cell fluorescence over time (**Figure 2.S2A**). To summarize the dynamics of CNVs in evolving populations, we determined the median normalized fluorescence in each population at each time point. The fluorescent signal of the *GAP1* CNV reporter increases during selection in all populations evolving in glutamine-limited

chemostats (**Figure 2.2B**), consistent with the *de novo* generation and selection of CNVs at the *GAP1* locus in all 9 populations.

To quantify the proportion of cells containing a *GAP1* duplication, we used one and two copy control strains to define flow cytometry gates. We found that the fluorescence of control strains varied slightly, which may be indicative of either instrument variation or changes in cell physiology and morphology during the experiment as suggested by systematic changes in forward scatter with time. Using a conservative method to classify individual cells containing *GAP1* amplifications, we find that *GAP1* amplification alleles are selected with remarkably reproducible dynamics in the nine glutamine-limited populations (**Figure 2.2C**). CNVs are predominantly duplications (two copies), but quantification of fluorescence suggests that many cells contain three or more copies of the *GAP1* locus.

We quantified the dynamics of CNVs in each population evolved in glutamine-limited chemostats using metrics defined by Lang et al. (Lang, Botstein, and Desai 2011). CNVs are detected by generation 70-75 (average = 72.8) in all 9 populations (T_{up}) (**Table 2.1**). To estimate the fitness of all CNV lineages relative to the mean population fitness, we calculated S_{up}, the rate of increase in the abundance of the CNV subpopulation. The average relative fitness of the CNV subpopulation is 1.077 (S_{up}) and CNV alleles are at frequencies greater than 75% in all populations by 250 generations (**Table 2.1**). Thus, in all replicated glutamine-limited selection experiments, *GAP1* amplifications emerge early, increase in frequency rapidly, and are maintained in each population throughout the selection.

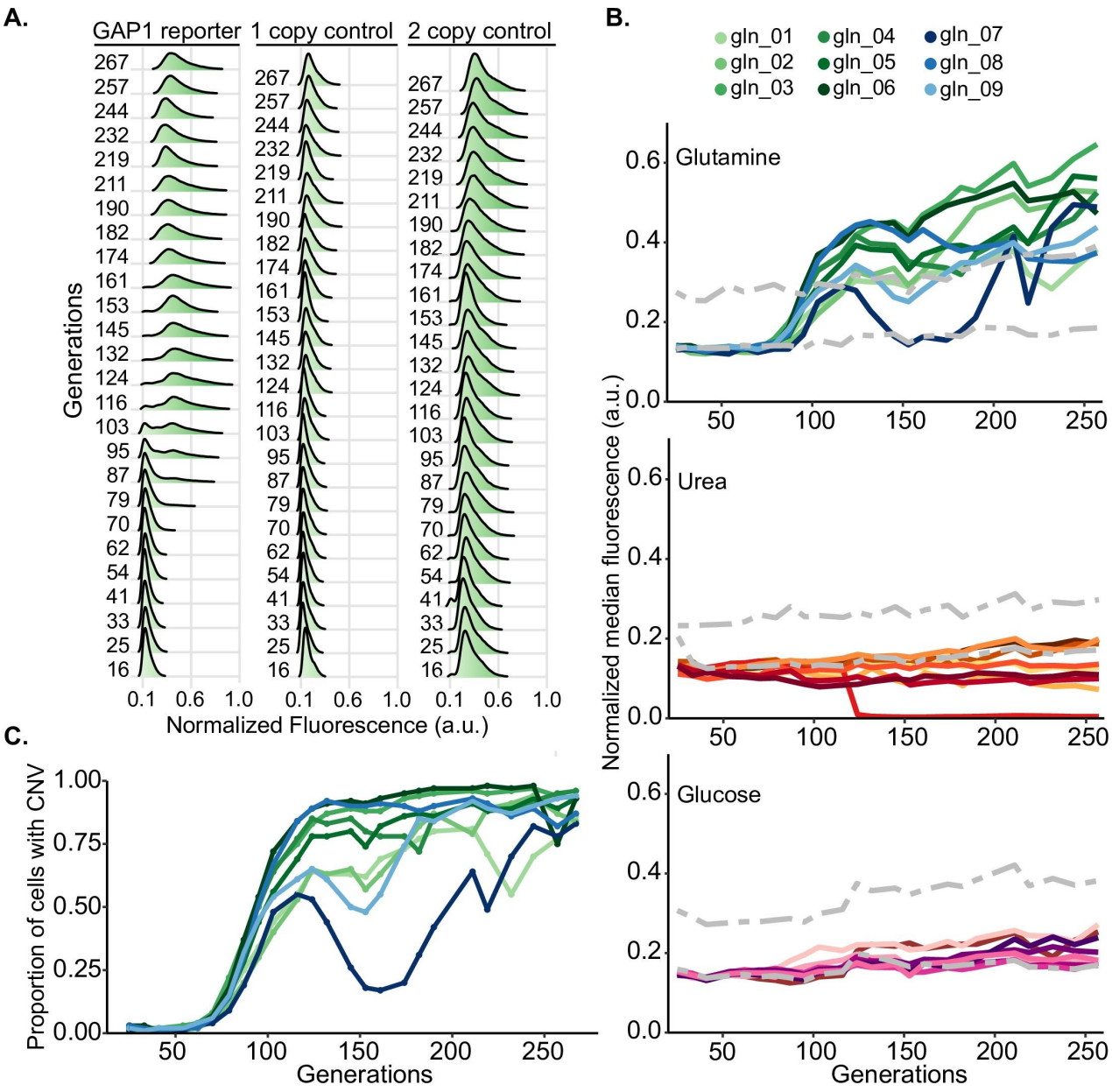

Figure 2.2. Dynamics of *GAP1* CNVs in evolving populations. (**A**) Normalized distributions of single-cell fluorescence over time for a representative *GAP1* CNV reporter strain and one and two copy control strains evolving in glutamine-limited chemostats. Single cell fluorescence is normalized by the forward scatter measurement of the cell. (**B**) Normalized median fluorescence for each population evolving in glutamine- (n = 9), urea- (n= 9) and glucose- (n = 8) limited chemostats. The fluorescence of the one and two copy control strains is plotted for reference (grey dotted lines). (**C**) Estimates of the proportion of cells with *GAP1* amplifications over time for nine glutamine-limited populations containing the *GAP1* CNV reporter.

Table 2.1. Summary statistics of *GAP1* CNV dynamics in glutamine-limited chemostats. T_{up} is the number of elapsed generations before CNVs are reliably detected (>7% frequency, see **methods**). S_{up} is the rate of increase in CNV abundance during the initial expansion of the CNV subpopulation (**S1 Text**). The frequency of CNVs in the population at generation 150 and generation 250, when genome sequencing was performed, is also reported.

Population	T_{up}	$1 + S_{up} \pm SE$	g150%	g250%
gln_01	70	1.066 ± 0.0038	62	77
gln_02	75	1.071 ± 0.0034	57	87
gln_03	70	1.071 ± 0.0037	88	94
gln_04	70	1.079 ± 0.0036	80	95
gln_05	75	1.077 ± 0.0041	74	89
gln_06	70	1.082 ± 0.0043	91	75
gln_07	75	1.094 ± 0.0048	18	78
gln_08	75	1.090 ± 0.0052	90	82
gln_09	75	1.066 ± 0.0050	48	93
AVG ± STD	72.8 ± 2.6	1.077 ± 0.01	68 ± 24	86 ± 8

GAP1 CNVs undergo two distinct phases of population dynamics. The initial dynamics with which CNV subpopulations emerge and increase in frequency are highly reproducible in independent evolving populations. However, after 125 generations, the trajectories of the CNV subpopulation in the different replicate populations diverge. Many populations maintain a high frequency of *GAP1* amplification alleles, but in some populations they decrease in frequency. In one population, *GAP1* CNV alleles are nearly lost from the population before subsequently increasing to an appreciable frequency (gln_07).

2.3.3 *GAP1* CNV alleles are diverse within and between replicate populations

Based on prior studies (Payen et al. 2014; Hong and Gresham 2014a), we hypothesized that multiple CNV alleles exist within each population. To characterize the diversity of *GAP1* CNVs, we isolated a total of 29 clones containing increased fluorescence from glutamine-limited chemostats at 150 and 250 generations for whole genome sequencing. We used read depth to calculate *GAP1* copy number and to estimate CNV boundaries (**Figure 2.3A**). We find that

GAP1 copy number estimated by sequencing read depth correlates with the fluorescent signal for individual clones (**Figure 2.3B**), indicating that fluorescent signal is predictive of copy number. In 3 clones, we find increased read depth across the entirety of chromosome XI consistent with aneuploidy. Thus, the CNV reporter is able to detect aneuploid chromosomes as well as subchromosomal CNVs.

We identified diverse *GAP1* CNVs between and within populations (**Figure 2.3C**). In the majority of populations (6/9) different clones had different CNVs. For example, in population gln_01 at generation 150, we identified a large *GAP1* CNV that includes the entire right arm of chromosome XI and another clone that was aneuploid for chromosome XI. At generation 250, clones isolated from population gln_01 have CNV alleles that are distinct from each other and from those observed at generation 150. Clones from the 8 additional glutamine-limited populations show evidence for CNV diversity within and between the two time points analyzed (**Figure 2.3C**) suggesting the presence of multiple CNV lineages within evolving populations. Furthermore, the diversity of *GAP1* CNVs indicates that they are not predominantly formed through a recurrent mechanism as might be anticipated by the presence of proximate repetitive elements.

We used pulsed-field gel electrophoresis and Southern blotting to confirm CNV structures. Using *GAP1* and *CEN11* probes for Southern blotting, we identified size shifts in some samples consistent with the large CNVs (>140 kilobases) we identified in several clones. Interestingly, in some cases, we identified two discrete bands in our *GAP1* Southern blot, indicating that the additional copies of *GAP1* were not contained on chromosome XI. The *GAP1* Southern also provided further evidence for the *GAP1* deletion in a clone isolated from urea-limitation. Importantly, while control populations evolving in glutamine-limited chemostats did not show evidence for *GAP1* CNVs on the basis of fluorescence, sequence and Southern blotting analysis identified *GAP1* amplifications in lineages isolated from these populations. As

one and two copy control strains do not have the *GAP1* CNV reporter, this suggests that *GAP1* CNV formation and selection is not affected by the reporter. Moreover, we find no evidence that the molecular features of *GAP1* CNVs are affected by the presence of the CNV reporter.

We determined the fitness of *GAP1* CNV-containing clones using pair-wise competitive fitness assays in glutamine-limited chemostats (**Figure 2.3C)**. Four independent competition assays with the ancestral strain containing the *GAP1* CNV reporter showed no significant differences in fitness compared to the isogenic non-fluorescent reference strain. The majority of evolved clones (18/28) have higher relative fitness than the ancestor, indicating that *GAP1* CNVs typically confer large fitness benefits. Several clones have neutral (8/28) or lower (2/28) relative fitness, which indicates that either 1) the fitness effect of *GAP1* CNVs may be context-specific or 2) not all *GAP1* CNVs confer a fitness benefit.

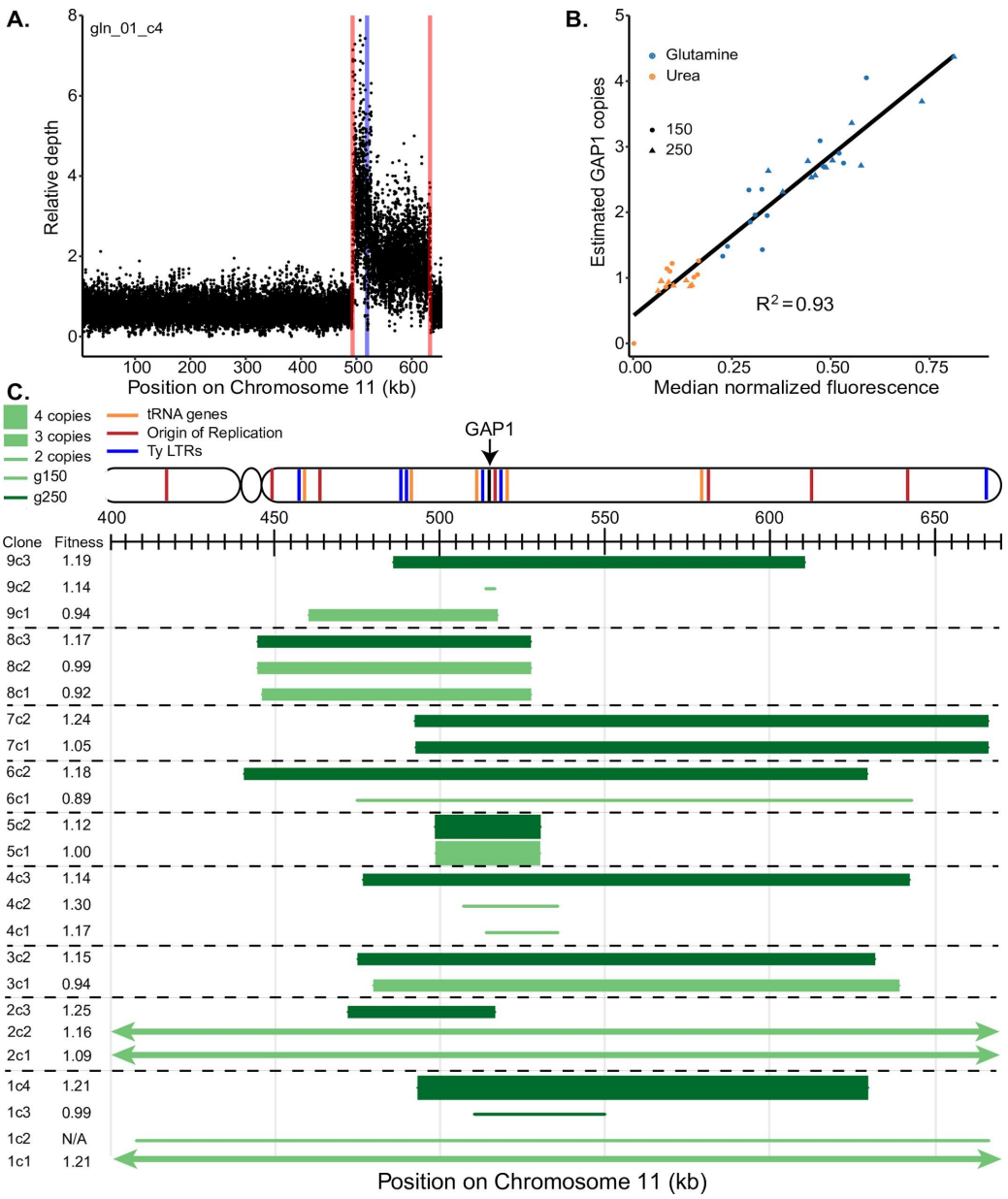

Figure 2.3. Diversity and fitness effects of *GAP1* CNVs. (**A**) Representative sequence read depth plot from a glutamine-limited clone (gln_01_c4). The nucleotide coordinates of *GAP1* in our CNV reporter strain are chromosome XI: 518438-520246 (blue line). Estimated breakpoint boundaries are shown in red. Read depth was normalized to the average read depth on chromosome XI. Reads at each nucleotide position were randomly downsampled for presentation purposes. (**B**) Read depth based estimates of *GAP1* copy number are positively correlated with median fluorescence of glutamine-limited clones, indicating that fluorescence is informative about the copy number of *de novo* CNVs. (**C**) Schematic representation of CNVs identified in clones isolated from glutamine-limited populations. The relative fitness of each clone is also indicated. Copy number and CNV boundaries were estimated using read depth. This schematic is simplified for presentation purposes: the reported copy number refers specifically to the *GAP1* coding sequence and does not necessarily reflect copy number throughout the entire CNV, which may vary.

2.3.4 CNV breakpoints are characterized by short, interrupted inverted repeats

We developed a breakpoint detection pipeline that integrates information from read depth, discordant reads and split reads. To define the breakpoint sequence, we performed *de novo* assembly using split reads and aligned the resulting contig against the reference genome. We analyzed 29 lineages containing *GAP1* CNVs and inferred the underlying mechanisms for 19 (66%) of them on the basis of copy number and breakpoint sequences. Of the 19 *GAP1* CNVs that can be reliably resolved, 3 are the result of aneuploidies and 2 are the result of non-reciprocal interchromosomal translocations. Translocations were confirmed using pulsed-field gel electrophoresis and Southern blot analysis, which clearly shows that the second copy of *GAP1* is located on a different chromosome. Southern blotting also indicates that an additional 3 *GAP1* CNVs are the result of partial (i.e. segmental) aneuploidies, which include the chromosome XI centromere (*CEN11*) but are smaller than the ancestral chromosome XI (**S5 Fig**). At least 4 *GAP1* CNVs appear to be the result of a tandem duplication mediated by non-allelic homologous recombination (NAHR). For two of these CNVs, novel junction sequences were obtained that included a hybrid sequence composed of half of each flanking long terminal repeat (*YKRCdelta11*/*YKRCdelta12*), similar to our previous report (Gresham et al. 2010).

For 12 out of 29 (41%) *GAP1* CNVs, we identified a pair of short, interrupted, inverted repeats proximate to at least one breakpoint. We were able to resolve breakpoints at both ends of the CNV for 12 of the 20 CNVs. Analysis of these breakpoints indicates that inverted repeat sequences range in length from 4-24 base pairs and are typically separated by 40 base pairs. Microhomology at breakpoint junctions is characteristic of replication-based CNV formation, including microhomology-mediated break-induced replication (MMBIR) and origin-dependent inverted-repeat amplification (ODIRA). ODIRA has several other requirements including the presence of at least one replication origin within the CNV, an internal inversion, and an odd copy

number. The identification of inverted sequence relative to the reference at all identified breakpoint junctions is consistent with an inverted structure. We find that 6/29 GAP1 CNVs meet these criteria and thus are likely the result of ODIRA. In cases where the CNV lacks an odd copy number we cannot reliably infer the mechanism.

2.3.5 Lineage tracking reveals extensive clonal interference among CNV lineages

The reproducible dynamics of CNV lineages observed during glutamine-limited experimental evolution may be due to two non-exclusive reasons: either 1) a high supply rate of *de novo* CNVs or 2) pre-existing CNVs in the ancestral population. In both scenarios, a single CNV or multiple, competing CNVs may underlie the reproducible dynamics. Sequence analysis of clonal lineages suggests at least two, and as many as four, CNV lineages may co-exist in populations (**Figure 2.3**); however, genome sequencing is uninformative about the total number of lineages for two key reasons. First, the recurrent formation of CNVs confounds distinguishing CNVs that are identical by state from those that are identical by descent. Second, CNVs that arise *de novo* may subsequently diversify over time resulting in distinct alleles that are derived from a common event.

To quantify the number, relationship and dynamics of individual CNV lineages, we constructed a lineage tracking library using random DNA barcodes (Levy et al. 2015). We constructed a library of ~80,000 unique barcodes (**Figure 2.S1**) in the background of the *GAP1* CNV reporter and performed six independent replicate experiments in glutamine-limited chemostats. Real time monitoring of CNV dynamics using the *GAP1* CNV reporter recapitulated the dynamics of our original experiment (**Figure 2.4A**, **Figure 2.S2A** and **Table 2.S1)** although CNV lineages appeared significantly earlier in these populations (T_{up}; t-test p-value < 0.01). As the lineage tracking strain was independently derived from the strain used in our original experiment, these results indicate that selection of *GAP1* CNVs in glutamine-limited chemostats is reproducible and independent of genetic background.

To quantify individual lineages, we isolated the subpopulation containing CNVs from two populations (bc01 and bc02) at multiple timepoints (generations 70, 90, 150, and 270). Isolation of the CNV subpopulation was performed by fluorescence activated cell sorting (FACS) using gates based on one and two copy control populations (**Figure 2.4A**). We sequenced barcodes from the CNV subpopulation at each time point and determined the number of unique lineages ((Zhao et al. 2017) and **methods**). To account for variation in the purity of the FACS-isolated CNV subpopulation, we analyzed individual clones using a flow cytometer. Using these data, we estimated a false positive rate, which we find varies between time points (**Figure 2.S2B** and **methods**), and applied this correction to barcode counts (**Table 2.2**).

Table 2.2. Estimation of CNV lineages in evolving populations across time. We determined the number of *GAP1* CNV containing lineages by correcting the number of identified barcodes by the estimated false positive rate associated with CNV isolation using FACS. High confidence *GAP1* CNV lineages are defined as those that are found at two or more consecutive timepoints.

Population	Generation	Number of detected barcodes	False positive rate (FP)	FP corrected barcode count	Barcodes identified at >1 time point
bc01	70	9650	0.27	7067	891
bc01	90	1064	0.09	973	891
bc01	150	136	0.04	131	131
bc01	270	79	0.04	76	38
bc02	70	7243	0.27	5305	2676
bc02	90	5851	0.09	5351	2710
bc02	150	606	0.04	583	162
bc02	270	29	0.04	28	22

Strikingly, we detect thousands of independent *GAP1* CNV lineages at generation 70 indicating that a large number of independent *GAP1* CNVs are generated and selected in the early stages of the evolution experiments **(Figure 2.4B)**. Applying a conservative false positive correction, we identified 7,067 *GAP1* CNV lineages in bc01 and 5,305 *GAP1* CNV lineages in bc02 at generation 70 **(Table 2.2)**. If we only consider lineages detected in the CNV subpopulation at multiple time points, we identify 891 CNV lineages in bc01 and 2,676 CNV

lineages at generation 70 **(Table 2.2)**. Thus, between 10^2-10^4 independent CNV lineages initially compete within each population that are on the order of ~10^8 cells. The overall diversity of CNV lineages decreases with time, consistent with decreases in lineage diversity observed in other evolution experiments (Levy et al. 2015; Blundell et al. 2019). By generation 270, we detect only 76 CNV lineages in bc01 and 28 CNV lineages in bc02. To determine the dominant lineages in each population, we identified barcodes that reached greater than 1% frequency in the CNV subpopulation in at least one time point: 21 independent lineages are found at greater than 1% frequency in bc01 and 18 independent lineages are found at greater than 1% frequency in bc02 (**Figure 2.4B**). These results indicate the presence and persistence of multiple *GAP1* CNVs across hundreds of generations of selection during which there is a continuous reduction in the overall diversity of CNV lineages.

Although CNVs rise to high frequencies in both populations (**Figure 2.4A**), the composition of competing CNV lineages is dramatically different: in bc02, a single lineage dominates the population by generation 150 (**Figure 2.4B**), whereas in bc01, there is much greater diversity at later time points. In both populations, several CNV lineages that comprise a large fraction of the CNV subpopulation at early generations (generations 70, 90, or 150) are extinct by generation 270. Thus, within populations, individual CNV lineages do not increase in frequency with uniform dynamics despite the consistent and reproducible dynamics of the entire CNV subpopulations (**Figure 2.2A** and **Figure 2.4A**). Differences in fitness between individual CNV lineages, possibly as a result of variation in copy number, CNV size and secondary adaptive mutations, are likely to contribute to these dynamics.

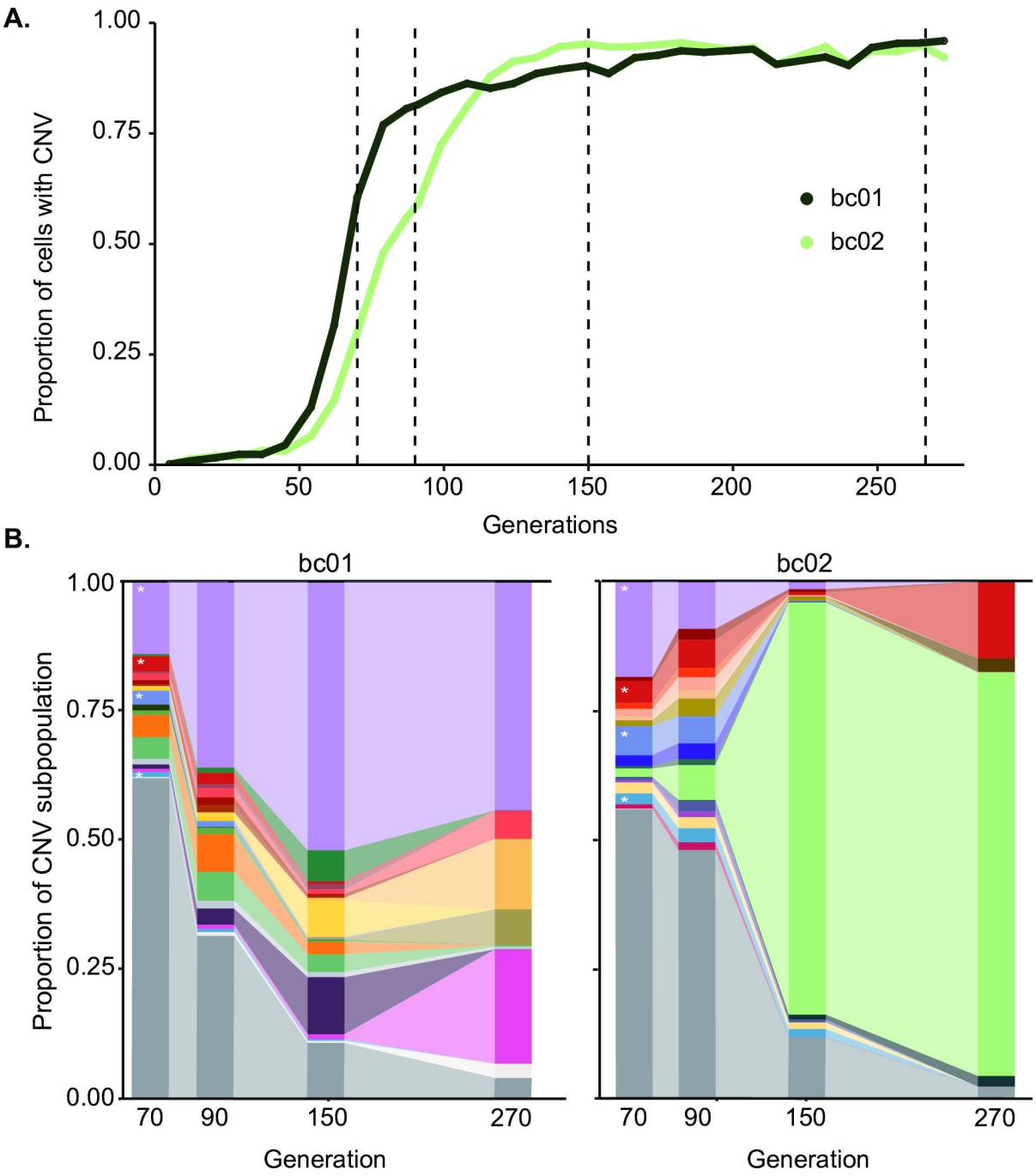

Figure 2.4 Lineage tracking reveals extensive clonal interference among CNV-containing lineages. (**A**) We used fluorescence-activated cell sorting (FACS) to fractionate cells containing *GAP1* CNVs from two populations at four time points (dashed black lines) and performed barcode sequencing. (**B**) Using a sample- and time point-specific false positive correction, we identified 7067, 973, 131, and 76 barcodes in one population, bc01 (left), and 5305, 5351, 583, and 28 barcodes in another population, bc02 (right), at generations 70, 90, 150 and 270 respectively. Each barcode found at >1% frequency in at least one time point is represented by a unique color in the plot, for a total of 21 barcodes in bc01 and 18 barcodes in bc02. All other lineages that are never detected at >1% frequency are shown in grey. Lineages denoted by a * are found at >1% frequency in both populations.

2.3.6 CNV subpopulations comprise *de novo* and pre-existing CNV alleles

To distinguish the contribution of pre-existing genetic variation (i.e. CNVs introduced to the population before chemostat inoculation) and *de novo* variation (i.e. CNVs introduced to the population following chemostat inoculation) to CNV lineage dynamics, we assessed whether barcodes were shared between CNV lineages in independent populations. We identified four barcodes at greater than 1% frequency that are common to both populations (**Figure 2.4B**). At generation 70, one of these barcodes (indicated in light purple) was present at 14% and 19% in bc01 and bc02, respectively. We find that the barcode for this lineage was over-represented in the ancestral unselected population (an initial frequency of 0.014%, which is one order of magnitude greater than the average starting frequency of 0.0011%; **Figure 2.S1**). Although there is a possibility that *de novo* CNVs formed independently in this barcode lineage, it is more likely that this lineage contained a pre-existing CNV in the ancestral population. While this lineage represented a sizable fraction of the CNV subpopulation in both replicate populations, it was only maintained at high frequency in one of them (bc01). Only one of the four pre-existing CNV lineages persists throughout the experiment in both populations. By contrast, in each population, we identified 17 and 14 unique high frequency CNV lineages that are most likely new CNVs. These results indicate that both pre-existing CNVs and *de novo* CNVs that arise during glutamine limitation contribute to adaptive evolution.

2.4 Discussion

Copy number variants are an important class of genetic variation and adaptive potential. In this study, we sought to understand the short-term fate of CNVs as they are generated and selected in evolving populations. Previous work from our laboratory and others has shown that the defined, strong selective conditions of a chemostat provides an ideal system for studying

CNVs. We used nitrogen limitation to establish conditions that select for amplification and deletion of the gene *GAP1,* which encodes the general amino acid permease, in *S. cerevisiae*.

2.4.1 A *GAP1* CNV reporter reveals the dynamics of selection

To determine the dynamics with which CNVs are selected at the *GAP1* locus, we inserted a constitutively expressed fluorescent gene adjacent to *GAP1* and tracked changes in single cell fluorescence over time. While one and two copy control strains with *mCitrine* at neutral loci maintain a steady fluorescent signal over 250 generations of selection, all glutamine-limited populations with the *GAP1* CNV reporter show increased fluorescence by generation 75. Importantly, the structure and breakpoints of CNVs within and between populations are different, indicating independent formation of CNVs. Control strains were inoculated independently, and have different genetic backgrounds, but also form CNVs at the *GAP1* locus as determined by whole genome sequencing and Southern blot analysis. These data indicate that *GAP1* CNVs are positively selected early and repeatedly in glutamine-limited environments.

While the majority of evolved clones with *GAP1* CNVs (18/28) have higher relative fitness in glutamine-limited chemostats compared to the ancestor, several clones have neutral (8/28) or lower (2/28) relative fitness. CNV-containing clones were selected on the basis of increased fluorescence, which does not necessarily mean the clone had higher fitness than the ancestor. The fitness effect of a CNV within the chemostat environment is context-specific, and may depend on factors such as frequency-dependent selection. In addition, if *GAP1* CNVs are generated at a high rate as we have hypothesized, neutral or deleterious CNVs could be present for several generations before these lineages are purged from the population or acquire additional adaptive mutations.

2.4.2 Inference of CNV formation mechanisms

Whole genome sequencing of *GAP1* CNV lineages isolated on the basis of increased fluorescence uncovered a wide range of CNV structures within and between populations. We found cases in which distinct alleles were identified within populations at different time points and cases in which we identified the same CNV allele 100 generations later. *GAP1* CNV alleles are 105 kilobases on average, but can include the entire right arm of chromosome XI (260 kilobases).

Our reporter detects increases in gene copy number that result from a variety of processes including aneuploidy, non-reciprocal translocation, tandem duplication, and complex copy number variants including inverted triplications. The ability to track and isolate these diverse gene amplifications allows us to enumerate the frequency of each type and characterize the mechanisms underlying their formation. Combining our approach with molecular techniques allowed us to further understand the nature of these *GAP1* CNVs. Three particularly interesting *GAP1* CNV-containing clones appear to have partial (i.e. segmental) aneuploidies that encompass centromere XI. As the presence of two centromeres in one chromosome is extremely unlikely, it is plausible that these exist as independent, supernumerary chromosomes (Natesuntorn et al. 2015). Similar adaptive rearrangements occur in other yeast species: isochromosome formation, potentially mediated by the presence of inverted repeats, has been observed during treatment of *Candida albicans* with antifungal drugs (A. Selmecki, Forche, and Berman 2006). The use of a CNV reporter should facilitate determination of the frequency with which these and other complex mechanisms give rise to CNVs at a given locus.

We identified 9 *GAP1* CNVs containing breakpoints that comprise closely-spaced inverted repeat sequences. Of these, the majority (14/17) also had an odd copy number, and contained an origin of replication consistent with the ODIRA mechanism (Brewer et al. 2011, 2015). Our results suggest that replication-based mechanisms may be a major source of gene

amplification in yeast. This is consistent with increasing evidence for replication-based CNV formation in diverse organisms including yeast, mice, and humans (Feng Zhang et al. 2009; Ottaviani, LeCain, and Sheer 2014; Arlt et al. 2012; Sakofsky et al. 2015).

2.4.3 Clonal interference underlies CNV dynamics

By combining a CNV reporter with lineage tracking, we identified a surprisingly large number of independent CNV lineages. Whereas clonal isolation and sequencing suggested at least four independent lineages within populations, lineage tracking indicates that hundreds to thousands of individual CNV lineages emerge within less than 100 generations. Most of these lineages do not achieve high frequency, as we identified only 18-21 lineages present at >1% frequency in the CNV subpopulation. The number of independent CNV lineages we identified is remarkable. Although we have attempted to account for technical factors that may inflate this number, unanticipated aspects of barcode transformation and library construction, cell sorting, and barcode sequencing and identification may impact this estimation. Conversely, the exact number of CNV lineages may be underestimated, as the unselected barcode library was not maximally diverse and each unique barcode was shared by multiple founding cells.

While we found lineages that were common to both populations (at least one of which is likely to contain a pre-existing CNV), ancestral CNV lineages do not drive the evolutionary dynamics. Pre-existing CNV lineages have different dynamics in each population, and do not prevent the emergence of unique *de novo* CNV lineages. This demonstrates that the ultimate fate of a CNV lineage depends on multiple factors, and a high frequency at an early generation does not guarantee that a lineage will persist in the population. Thus, CNV dynamics result from pre-existing and *de novo* variation and are characterized by extensive clonal interference and replacement among competing CNV lineages.

The large number of CNV lineages identified in our study indicates that they occur at a high rate. Recent studies have suggested that adaptive mutations may be stimulated by the environment. Stress can lead to increases in genome-wide mutation rates in both bacteria and yeast (Foster 2007; Galhardo, Hastings, and Rosenberg 2007; Shor, Fox, and Broach 2013) and replicative stress can lead directly to increased formation of CNVs (L. Chen et al. 2015; Wilson et al. 2015). Other groups have proposed an interplay between transcription and CNV generation, and that active transcription units might even be “hotspots” of CNV formation (Thomas and Rothstein 1989; Skourti-Stathaki and Proudfoot 2014; Aguilera and Gaillard 2014). These hotspots, often designated as common fragile sites, may occur in long, late replicating genes, with large inter-origin distances (Wilson et al. 2015). Local transcription at the rDNA locus leads to rDNA amplification, and is thought to be regulated in response to the environment (Jack et al. 2015; Mansisidor et al. 2018). Transcription of the *CUP1* locus in response to environmental copper leads to promoter activity that further destabilizes stalled replication forks and generates CNVs (Hull et al. 2017). Given the high level of *GAP1* transcription in nitrogen limited chemostats (Airoldi et al. 2016) it is tempting to speculate that this condition may promote the formation of *GAP1* CNVs. Further studies are required to understand the full extent of processes that underlie CNV formation at the *GAP1* locus and how these different mechanisms may contribute to the fitness and overall success of CNV lineages.

The frequency of *GAP1* CNVs can be attributed to a combination of factors including: a high mutation supply rate due in part to the large chemostat population size (~10^8), the strength of selection, and the fitness benefit typically conferred by *GAP1* amplification. Together, these factors contribute to an early, deterministic phase, during which CNVs are formed at a high rate and thousands of lineages with CNVs rapidly increase in frequency. During a second phase, the dynamics are more variable as competition from different types of adaptive lineages, and additional acquired variation, influence evolutionary trajectories of individual CNV lineages. This

phenomenon has recently been observed in other evolution experiments, where early events are driven by multiple competing single-mutant lineages (Blundell et al. 2019), but later dynamics are influenced by stochastic factors and secondary mutations (Levy et al. 2015).

The high degree of clonal interference observed among a single class of adaptive mutations may have important implications for adaptive evolution. CNVs are alleles of large effect that can simultaneously change the dosage of multiple protein-coding genes and subsequently lead to changes in cell physiology. Epistatic relationships between CNVs and other adaptive mutations could therefore dramatically alter the fitness landscape (Kvitek and Sherlock 2011). Additionally, CNVs can confer a fitness benefit *per se* but also serve to increase the amount of DNA in the genome that can accumulate mutations. Therefore, CNVs can potentially increase the rate of adaptive evolution by increasing the target size for adaptive mutations. In this study, we found evidence for polymorphisms within individual CNVs and potential epistasis between SNVs and CNV alleles, two phenomena that require further exploration as we continue to define the role of CNVs in driving rapid adaptive evolution.

2.5 Conclusion

The combined use of a fluorescent CNV reporter and barcode lineage tracking provides unprecedented insight into this important class of mutation. Previous studies have tracked specific mutations and their fitness effects (Lang, Botstein, and Desai 2011), but ours is the first single-cell based approach to identify an entire class of mutations and follow evolutionary trajectories with high resolution. While barcode tracking alone provides information about the number of adaptive lineages and their fitness effects, the CNV reporter enables us to specifically determine the number of unique CNV events. In addition, the reporter provides an estimate of the total proportion of CNVs in the population, which we can use to inform our understanding of lineage dynamics. Using these tools, we have shown that CNVs are generated at a high rate

through diverse mechanisms including homologous recombination and replication-based errors. These processes lead to the formation of many distinct CNV alleles segregating within populations. One limitation of our approach is that a complex copy number variant could be the product of multiple, independent events (for example, a duplication followed by a subsequent triplication). Evolution experiments that start with a pre-existing CNV would be informative for studying how CNVs diversify when maintained under selection.

Our results demonstrate an important role for CNVs in driving rapid adaptive evolution in microbial populations, but could be broadly applicable to plants, animals, and humans. Our system provides a facile means for studying the molecular processes underlying CNV generation as well as evolutionary aspects of CNVs including: whether there are fundamental differences in CNV formation and selection at different loci, the impact of a high rate of CNV formation on the evolutionary dynamics of other adaptive lineages, how CNVs are maintained or refined over longer evolutionary timescales, how CNVs interact with other adaptive mutations to influence fitness landscapes, whether there are consequences and tradeoffs in alternative environments, and how the formation of CNVs impacts gene expression and genome architecture. Extension of this method is likely to be useful for addressing additional fundamental questions regarding the evolutionary and pathogenic role of CNVs in diverse systems.

2.6 Methods

2.6.1 Strains and media

We used FY4 and FY4/5, haploid and diploid derivatives of the reference strain S288c, for all experiments. To generate fluorescent strains, we performed high efficiency yeast transformation (Gietz and Schiestl 2007b) with an mCitrine gene under control of the constitutively expressed *ACT1* promoter (*ACT1pr::mCitrine::ADH1term*) and marked by the

KanMX G418-resistance cassette (*TEFpr::KanMX::TEFterm*). The entire construct, which we refer to as the mCitrine CNV reporter, is 3,375 base pairs. For control strains, the mCitrine reporter was integrated at two neutral loci: *HO* (*YDL227C*) on chromosome IV and the dubious ORF, *YLR123C* on chromosome XII. Diploid control strains containing 3 and 4 copies of the mCitrine CNV reporter were generated using a combination of backcrossing and mating. We constructed the *GAP1* CNV reporter by integrating the mCitrine construct at an intergenic region 1,118 base pairs upstream of *GAP1* (integration coordinates, chromosome XI: 513945-517320). PCR and Sanger sequencing were used to confirm integration of the *GAP1* CNV reporter at each location. Transformants were subsequently backcrossed and sporulated, and the resulting segregants were genotyped.

For the purpose of lineage tracking, we constructed a strain containing a landing pad and the *GAP1* CNV reporter by segregation analysis after mating the original *GAP1* CNV reporter strain to a landing pad strain (derived from BY4709) (Levy et al. 2015). As the kanMX cassette is present at two loci in this cross, we performed tetrad dissection and identified four spore tetrads that exhibited 2:2 G418 resistance. A segregant with the correct genotype (G418 resistant, ura-) was identified and confirmed using a combination of PCR and fluorescence analysis. We introduced a library of random barcodes by transformation and selection on SC-ura plates (Levy et al. 2015). We plated an average of 500 transformants on 200 petri plates and estimated 78,000 independent transformants.

Nitrogen limiting media contained 800 μM nitrogen and 1 g/L $CaCl_2$-$2H_2O$, 1 g/L of NaCl, 5 g/L of $MgSO_4$-$7H_2O$, 10 g/L KH_2PO_4, 2% glucose and trace metals and vitamins as previously described (Hong and Gresham 2014a).

2.6.2 Long-term experimental evolution

We inoculated the *GAP1* CNV reporter strain into 20mL ministat vessels (Miller et al. 2013) containing either glutamine-, urea-, or glucose-limited media. Control populations

containing either one or two copies of the CNV reporter at neutral loci (*HO* and *YLR123C*) were also inoculated in ministat vessels for each media condition. Ministats were maintained at 30°C in aerobic conditions and diluted at a rate of 0.12 hr^{-1} (corresponding to a population doubling time of 5.8 hours). Steady state populations of 3 x 10^8 cells were maintained in continuous mode for 270 generations (65 days). Every 30 generations, we archived 2 mL population samples at -80°C in 15% glycerol.

2.6.3 Flow cytometry sampling and analysis

To monitor the dynamics of CNVs, we sampled 1mL from each population every ~8 generations. We performed sonication to disrupt any cellular aggregates and immediately analyzed the samples on an Accuri flow cytometer, measuring 100,000 cells per population for mCitrine fluorescence signal (excitation = 516nm, emission = 529nm, filter = 514/20nm), cell size (forward scatter) and cell complexity (side scatter). We generated a modified version of our laboratory flow cytometry pipeline for this analysis (https://github.com/GreshamLab/flow), which uses the R package *flowCore (Ellis et al. 2016)*. We used forward scatter height (FSC-H) and forward scatter area (FSC-A) to filter out doublets, and FSC-A and side scatter area (SSC-A) to filter debris. We quantified fluorescence for each cell and divided this value by the forward scatter measurement for the cell to account for differences in cell size. To determine population frequencies of cells with zero, one, two, and three plus copies of *GAP1*, we used one and two copy control strains grown in glutamine-limited chemostats to define gates and perform manual gating. We used a conservative gating approach to reduce the number of false positive CNV calls by first manually drawing a liberal gate for the one copy control strain, followed by a non-overlapping gate for the two copy control strain.

2.6.4 Isolation and analysis of evolved clones

Clonal isolates were obtained from each glutamine- and urea-limited population at generation 150 and generation 250. We isolated clones by plating cells onto rich media (YPD) and randomly selecting individual colonies. We inoculated each clone into 96 well plates containing the limited media used for evolution experiments and analyzed them on an Accuri flow cytometer following 24 hours of growth. We compared fluorescence to unevolved ancestral strains and evolved 1 and 2 copy controls grown under the same conditions, and chose a subset of clones for whole genome sequencing.

To measure the fitness coefficient of evolved clones, we performed pairwise competitive fitness assays in glutamine-limited chemostats using the same, glutamine-limited conditions as our evolution experiments (Hong and Gresham 2014a). We co-cultured our fluorescent evolved strains with a non-fluorescent, unevolved reference strain (FY4). We determined the relative abundance of each strain every 2-3 generations for approximately 15 generations using flow cytometry. We performed linear analysis of the natural log of the ratio of the two genotypes against time and estimated the fitness, and associated error, relative to the ancestral strain.

2.6.5 Quantifying the number of CNV lineages

We inoculated the lineage tracking library into 20mL ministat vessels (Miller et al. 2013) containing glutamine-limited media. Control populations containing either zero, one or two copies of the *GAP1* CNV reporter at neutral loci (*HO* and *YLR123C*) were also inoculated in ministat vessels for each media condition. Control populations did not contain lineage tracking barcodes. Ministat vessels were maintained and archived as above. Samples were taken for flow cytometry every ~8 generations and analyzed as previously described.

We used fluorescence activated cell sorting (FACS) to isolate the subpopulation of cells containing two or more copies of the mCitrine CNV reporter using a FACSAria. We defined our gates using zero, one, and two copy mCitrine control strains sampled from ministat vessels at

the corresponding timepoints: 70, 90, 150, and 265 generations. Depending on the sample, we isolated 500,000-1,000,000 cells with increased fluorescence, corresponding to two or more copies of the reporter. We grew the isolated subpopulation containing CNVs for 48 hours in glutamine-limited media and performed genomic DNA extraction using a modified Hoffman-Winston protocol (Hoffman and Winston 1987). We verified FACS isolation of true CNVs by isolating clones from subpopulations sorted at generation 70, 90, and 150 (sorted from all lineage tracking populations, bc01-06) and performing independent flow cytometry analysis using an Accuri. We estimated the average false positive rate of CNV isolation at each time point as the percent of clones from a population with FL1 less than one standard deviation above the median FL1 in the one copy control strain. Only subpopulations with fluorescence measurements for at least 25 clones were included in calculations of false positive rate.

We performed a sequential PCR protocol to amplify DNA barcodes and purified the products using a Nucleospin PCR clean-up kit (Levy et al. 2015). We quantified DNA concentrations by qPCR before balancing and pooling libraries. DNA libraries were sequenced using a paired-end (2x150) protocol on an Illumina MiSeq 300 Cycle v2. Standard metrics were used to assess data quality (Q30 and %PF). However, the reverse read failed due to over-clustering, so all analyses were performed only using the forward read. We used the Bartender algorithm with UMI handling to account for PCR duplicates and to cluster sequences with merging decisions based solely on distance except in cases of low coverage (<500 reads/barcode), for which the default cluster merging threshold was used (Zhao et al. 2017). Clusters with a size less than four or with high entropy (>0.75 quality score) were discarded. We estimated the relative abundance of barcodes using the number of unique reads supporting a cluster compared to total library size.

2.7 Supplemental Material

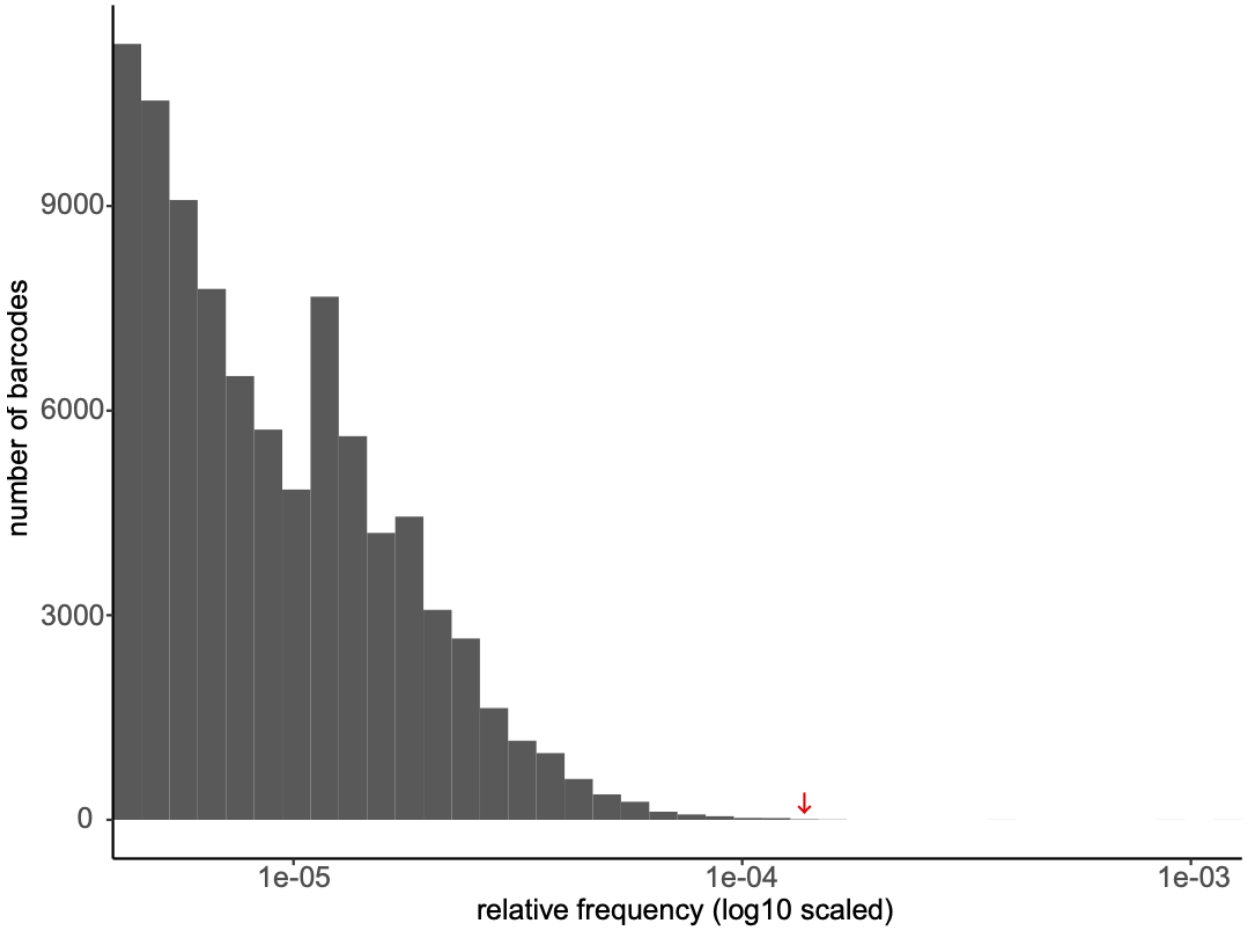

Figure 2.S1. **Distribution of barcode counts in ancestral populations**. We determined the distribution of read counts supporting each unique barcode in the ancestral population, after filtering out low confidence clusters. The relative frequencies of barcodes vary by over an order of magnitude and we observe a long tail with a few barcodes significantly overrepresented in the ancestral population. The red arrow indicates an overrepresented barcode in the ancestral population that was identified in the CNV subpopulation in both independent barcoded evolution experiments (indicated in purple in **Fig 5B**). This distribution is consistent with that found in other barcode lineage tracking experiments (Levy et al. 2015).

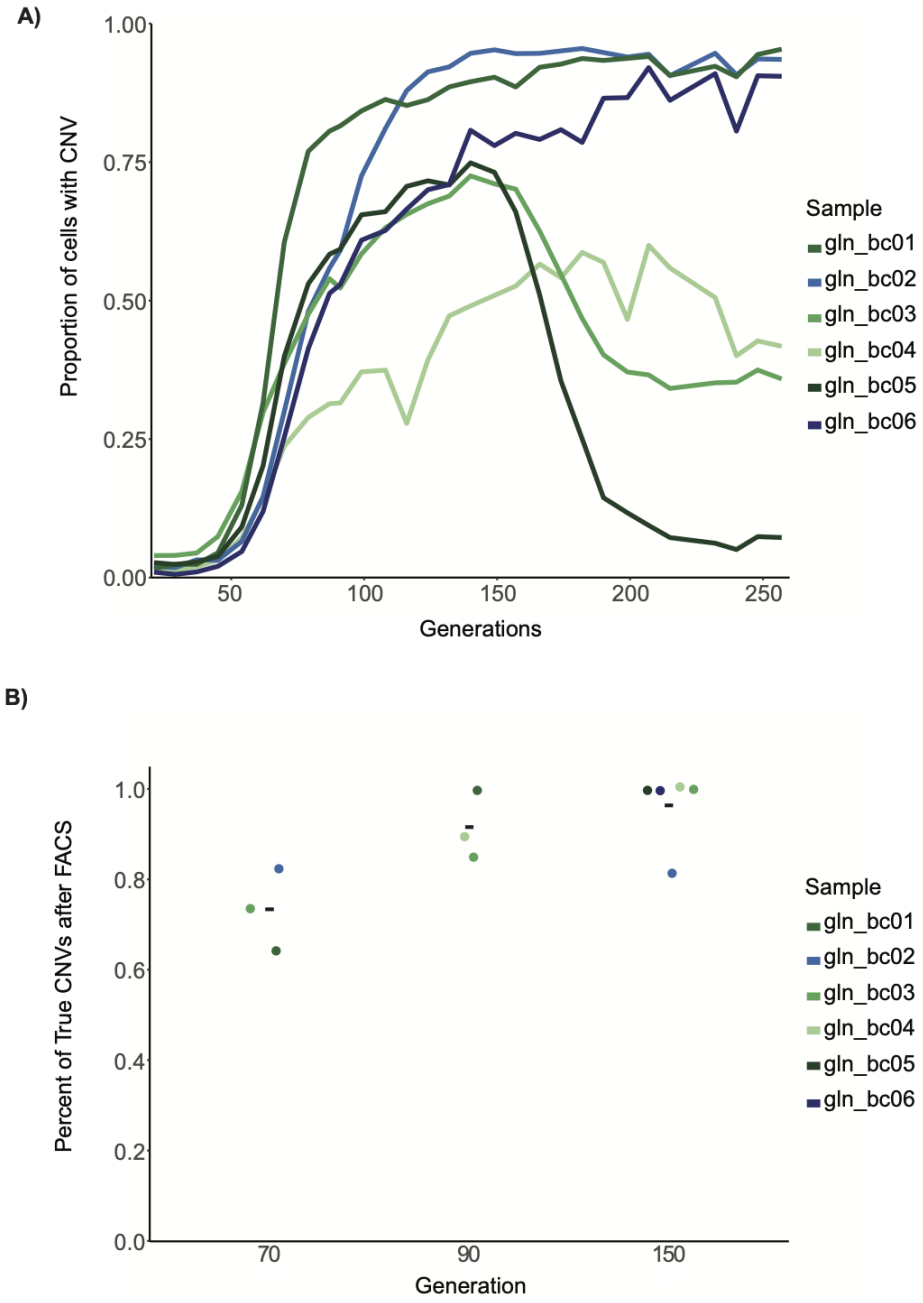

Figure 2.S2 Identification of barcoded *GAP1* CNV-lineages in evolving populations. (**A**) *GAP1* CNV dynamics in barcoded populations assayed using a CNV reporter. **(B)** Estimation of true positive rate of CNV isolation by FACS at generations 70, 90, and 150. CNV subpopulations were isolated by FACS at each timepoint and clones isolated by plating for single colonies. The percent of cells containing a CNV in the fractionated subpopulation was estimated using at least 25 clones. A one copy control strain was used to define gates.

Sample ID	(Tup) Time of initial detection	1 + Sup ± SE	95% CI Sup	95% CI Sup	Sup Range	Max %	(Tmax) Time to max frequency	g70 %	g90 %	g150 %	g270 %	Monotonic
gln_bc01	50	1.124 ± 0.0052	0.1096	0.1387	37-79	96%	273	61%	82 %	90%	96%	yes
gln_bc02	54	1.073 ± 0.0024	0.0676	0.0781	29-124	96%	182	30%	59 %	95%	92%	yes
gln_bc03	41	1.065 ± 0.0046	0.0537	0.0762	29-87	73%	140	39%	52 %	71%	44%	no
gln_bc04	50	1.069 ± 0.0046	0.0574	0.0796	21-79	60%	207	24%	32 %	48%	38%	no
gln_bc05	50	1.08 ± 0.0059	0.0659	0.0950	29-87	75%	140	40%	59 %	73%	6%	no
gln_bc06	58	1.096 ± 0.0035	0.0872	0.1042	29-87	94%	265	25%	53 %	78%	92%	yes

Table 2.S1 Summary statistics for *GAP1* CNV dynamics, determined using the *GAP1* CNV reporter, in replicated evolution experiments using lineage tracking libraries. Summary statistics are defined as in Table 1.

Chapter 3: Simulation-based inference of evolutionary parameters from adaptation dynamics using neural networks

This chapter is based on "Simulation-based inference of evolutionary parameters from adaptation dynamics using neural networks" by **Grace Avecilla**, Julie N. Chuong, Fangfei Li, Gavin Sherlock, David Gresham, and Yoav Ram, which is posted on bioRxiv (https://www.biorxiv.org/content/10.1101/2021.09.30.462581v1) and has been accepted for publication at PLoS Biology.

I performed all analysis in collaboration with Yoav Ram, except for that shown in Figure 3.7C (lineage tracking). I contributed to generating data for Figure 3.7. I generated all figures and tables, and wrote the manuscript text with editing from David Gresham and Yoav Ram.

3.1 Abstract

The rate of adaptive evolution depends on the rate at which beneficial mutations are introduced into a population and the fitness effects of those mutations. The rate of beneficial mutations and their expected fitness effects is often difficult to empirically quantify. As these two parameters determine the pace of evolutionary change in a population, the dynamics of adaptive evolution may enable inference of their values. Copy number variants (CNVs) are a pervasive source of heritable variation that can facilitate rapid adaptive evolution. Previously, we developed a locus-specific fluorescent CNV reporter to quantify CNV dynamics in evolving populations maintained in nutrient-limiting conditions using chemostats. Here, we use CNV adaptation dynamics to estimate the rate at which beneficial CNVs are introduced through *de novo* mutation and their fitness effects using simulation-based Bayesian likelihood-free inference approaches. We tested the suitability of two evolutionary models: a standard

Wright-Fisher model and a chemostat model. We evaluated two likelihood-free inference algorithms: the well-established *Approximate Bayesian Computation with Sequential Monte Carlo* (ABC-SMC) algorithm, and the recently developed *Neural Posterior Estimation* (NPE) algorithm, which applies an artificial neural network to directly estimate the posterior distribution. By systematically evaluating the suitability of different inference methods and models we show that NPE has several advantages over ABC-SMC and that a Wright-Fisher evolutionary model suffices in most cases. Using our validated inference framework, we estimate the CNV formation rate at the *GAP1* locus in the yeast *Saccharomyces cerevisiae* to be $10^{-4.7}$ - 10^{-4} CNVs per cell division, and a fitness coefficient of 0.04 - 0.1 per generation for *GAP1* CNVs in glutamine-limited chemostats. We experimentally validated our inference-based estimates using two distinct experimental methods - barcode lineage tracking and pairwise fitness assays - that provide independent confirmation of the accuracy of our approach. Our results are consistent with a beneficial CNV supply rate that is 10-fold greater than the estimated rates of beneficial single-nucleotide mutations, explaining the outsized importance of CNVs in rapid adaptive evolution. More generally, our study demonstrates the utility of novel neural-network-based likelihood-free inference methods for inferring the rates and effects of evolutionary processes from empirical data with possible applications ranging from tumor to viral evolution.

3.2 Introduction

Evolutionary dynamics are determined by the supply rate of beneficial mutations and their associated fitness effect. As the combination of these two parameters determines the overall rate of adaptive evolution, experimental methods are required for separately estimating them. The fitness effects of beneficial mutations can be determined using competition assays (Gallet et al. 2012; Ram et al. 2019) and mutation rates are typically estimated using mutation accumulation or Luria-Delbrück fluctuation assays (Gallet et al. 2012; Kondrashov and

Kondrashov 2010). An alternative approach to estimating both the rate and effect of beneficial mutations entails quantifying the dynamics of adaptive evolution and using statistical inference methods to find parameter values that are consistent with the dynamics (Sousa et al. 2013; Hegreness et al. 2006; Barrick et al. 2010; Nguyen Ba et al. 2019). Approaches to measure the dynamics of adaptive evolution, quantified as changes in the frequencies of beneficial alleles, have become increasingly accessible using either phenotypic makers (Lang, Botstein, and Desai 2011) or high-throughput DNA sequencing (Torada et al. 2019). Thus, inference methods using adaptation-dynamics data hold great promise for determining the underlying evolutionary parameters.

Fitness effects of beneficial mutations are not constant, but comprise a portion of a distribution of fitness effects (DFE). Determining the parameters of the DFE in a given condition is a central goal of evolutionary biology. Typically, beneficial mutations can occur at multiple loci and thus variance in the DFE reflects genetic heterogeneity. However, in some scenarios a single locus is the dominant gene in which beneficial mutations occur, such as the case of mutations in the *β*-lactamase gene underlying *β*-lactam antibiotic resistance or in *rpoB* underlying rifampicin resistance in bacteria (Weinreich et al. 2006; MacLean and Buckling 2009). In this case different mutations at the same locus confer differential beneficial effects resulting in a locus specific DFE. Typically, a DFE of beneficial mutations encompasses both allelic and locus heterogeneity.

Copy number variants (CNVs) are defined as deletions or amplifications of genomic sequences. Due to their high rate of formation and strong fitness effects, they can underlie rapid adaptive evolution in diverse scenarios ranging from niche adaptation to speciation (Zuellig and Sweigart 2018; Dhami, Hartwig, and Fukami 2016; K. M. Turner et al. 2017; Geiger, Cox, and Mann 2010; Stratton, Campbell, and Futreal 2009). In the short term, CNVs may provide immediate fitness benefits by altering gene dosage. Over longer evolutionary timescales, CNVs

can provide the raw material for the generation of evolutionary novelty through diversification of different gene copies (M.-C. Harrison et al. 2021). As a result, CNVs are common in human populations (Barreiro et al. 2008; Iskow et al. 2012; Zarrei et al. 2015), domesticated and wild populations of animals and plants (Ramirez et al. 2014; Clop, Vidal, and Amills 2012; Żmieńko et al. 2014), pathogenic and non-pathogenic microbes (Greenblum, Carr, and Borenstein 2015; Nair et al. 2008; Iantorno et al. 2017; Dulmage et al. 2018), and viruses (Gao et al. 2017; Rezelj, Levi, and Vignuzzi 2018; Elde et al. 2012). CNVs can be both a driver and a consequence of cancers (reviewed in (Ben-David and Amon 2020)).

Although critically important to adaptive evolution, our understanding of the dynamics and reproducibility of CNVs in adaptive evolution is poor. Specifically, key evolutionary properties of CNVs, including their rate of formation and fitness effects, are largely unknown. As with other classes of genomic variation, CNV formation is a relatively rare event, occurring at sufficiently low frequencies to make experimental measurement challenging. Estimates of *de novo* CNV rates are derived from indirect and imprecise methods, and even when genome-wide mutation rates are directly quantified by mutation accumulation studies and whole-genome sequencing, estimates depend on both genotype and condition (Kondrashov and Kondrashov 2010) and vary by orders of magnitude (Y. O. Zhu et al. 2014; R. P. Anderson and Roth 1977; Horiuchi, Horiuchi, and Novick 1963; Reams et al. 2010; P. Anderson and Roth 1981; Sharp et al. 2018; Sui et al. 2020; H. Liu and Zhang 2019).

Fitness effects of CNVs vary depending on gene content, genetic background and the environment. In evolution experiments in many systems, CNVs arise repeatedly in response to strong selection (Lauer et al. 2018; Payen et al. 2014; Sun et al. 2012; Farslow et al. 2015; Morgenthaler et al. 2019; Frickel et al. 2018; DeBolt 2010; Todd and Selmecki 2020; Sunshine et al. 2015), consistent with strong beneficial fitness effects. Several of these studies measured fitness of clonal isolates containing CNVs, and reported selection coefficients ranging from -0.11

to 0.6 (Payen et al. 2014; Lauer et al. 2018; Sunshine et al. 2015). However, the fitness of lineages containing CNVs varies between isolates even within studies, which could be due to additional heritable variation or to differences in fitness between different types of CNVs (e.g. aneuploidy vs. single-gene amplification).

Due to the challenge of empirically measuring rates and effects of beneficial mutations across many genetic backgrounds, conditions, and types of mutations, researchers have attempted to infer these parameters from population-level data using evolutionary models and Bayesian inference (Hegreness et al. 2006; Barrick et al. 2010; Harari et al. 2018). This approach has several advantages. First, model-based inference provides estimations of interpretable parameters and the opportunity to compare multiple models. Second, the degree of uncertainty associated with a point estimate can be quantified. Third, a posterior distribution over model parameters allows exploration of parameter combinations that are consistent with the observed data, and posterior distributions can provide insight into certain relationships between parameters (Gonçalves et al. 2020). Fourth, posterior predictions can be generated using the model and either compared to the data or used to predict the outcome of differing scenarios.

Standard Bayesian inference requires a likelihood function, which gives the probability of obtaining the observed data given some values of the model parameters. However, for many evolutionary models, such as the Wright-Fisher model, the likelihood function is analytically and/or computationally intractable. Likelihood-free simulation-based Bayesian inference methods that bypass the likelihood function, such as *Approximate Bayesian Computation* (ABC; (Sunnåker et al. 2013)), have been developed and used extensively in population genetics (Beaumont, Zhang, and Balding 2002), ecology and epidemiology (Tanaka et al. 2006; Beaumont 2010), cosmology (Jennings and Madigan 2017), as well as experimental evolution (Bank et al. 2014; Blanquart and Bataillon 2016; Barrick et al. 2010; Sousa et al. 2013; Harari,

Ram, and Kupiec 2018). The simplest form of likelihood-free inference is rejection-ABC (Tavaré et al. 1997; Pritchard et al. 1999), in which model parameter proposals are sampled from a prior distribution, simulations are generated based on those parameter proposals, and simulated data are compared to empirical observations using a summary and distance function. Proposals that generate simulated data with a distance less than a defined tolerance threshold are considered samples from the posterior distribution and can therefore be used for its estimation. Efficient sampling methods have been introduced, namely MCMC (Marjoram et al. 2003) and SMC (Sisson, Fan, and Tanaka 2007), that iteratively select proposals based on previous parameters samples so that regions of the parameter space with higher posterior density are explored more often. A shortcoming of ABC is that it requires summary and distance functions, which may be difficult to choose appropriately and compute efficiently, especially when using high-dimensional or multi-modal data, although methods have been developed to address this challenge (Blum and François 2010; Csilléry, François, and Blum 2012; Beaumont, Zhang, and Balding 2002).

Recently, new inference methods have been introduced that directly approximate the likelihood or the posterior density function using *deep neural density estimators*—artificial neural networks that approximate density functions. These methods, which have recently been used in neuroscience (Gonçalves et al. 2020), population genetics (Flagel, Brandvain, and Schrider 2019), and cosmology (Alsing et al. 2019), forego the summary and distance functions, can use data with higher dimensionality, and perform inference more efficiently (Gonçalves et al. 2020; Alsing et al. 2019; Cranmer, Brehmer, and Louppe 2020). However, neural network-based inference methods have not previously been applied to experimental evolution.

Despite being originally developed to analyze population-genetic data, e.g. to infer parameters of the coalescent model (Tavaré et al. 1997; Pritchard et al. 1999; Sisson, Fan, and Tanaka 2007; Marjoram et al. 2003), likelihood-free methods have only been used in a small number of experimental evolution studies. Hegreness et al (Hegreness et al. 2006) estimated

the rate and mean fitness effect of beneficial mutations in *E. coli*. They performed 72 replicates of a serial-dilution evolution experiment, starting with equal frequencies of two strains that differ only in a fluorescent marker in a putatively neutral location and allowed them to evolve over 300 generations. Following the marker frequencies, they estimated from each experimental replicate two summary statistics: the time when a beneficial mutation starts to spread in the population and the rate at which its frequency increases. They then ran 500 simulations of an evolutionary model using a grid of model parameters to produce a theoretical distribution of summary statistics. Finally, they used the one-dimensional Kolmogorov-Smirnov distance between the empirical and theoretical summary-statistic distributions to assess the inferred parameters. Barrick et al (Barrick et al. 2010) also inferred the rate and mean fitness effect from similar serial-dilution experiments using a different evolutionary model implemented with a τ-leap stochastic simulation algorithm. They used the same summary statistics but applied the two-dimensional Komogorov-Smirnov distance function to better account for dependence between the summary statistics. Moura de Sousa et al (Moura de Sousa, Campos, and Gordo 2013) also focused on evolutionary experiments with two neutral markers. Their model included three parameters: the beneficial mutation rate, and the two parameters of a Gamma distribution for the fitness effects of beneficial mutations. They introduced a new summary statistic that uses both the marker frequency trajectories and the population mean fitness trajectories (measured using competition assays). They summarized these data by creating histograms of the frequency values and fitness values for each of six time-points. This resulted in 66 summary statistics necessitating the application of a regression-based method to reduce the dimensionality of the summary statistics and achieve greater efficiency (Moura de Sousa, Campos, and Gordo 2013; Csilléry, François, and Blum 2012). A simpler approach was taken by Harari et al (Harari et al. 2018), who used a rejection-ABC approach to estimate a single model parameter, the endoreduplication rate, from evolutionary experiments with yeast. They used the

frequency dynamics of three genotypes (haploid and diploid homozygous and heterozygous at the *MAT* locus) without a summary statistic. The distance between the empirical results and 100 simulations was computed as the mean absolute error. These prior studies point to the potential of simulation-based inference.

Previously, we developed a fluorescent CNV reporter system in the budding yeast, *Saccharomyces cerevisiae,* to quantify the dynamics of *de novo* CNVs during adaptive evolution (Lauer et al. 2018). Using this system we quantified CNV dynamics at the *GAP1* locus, which encodes a general amino acid permease, in nitrogen-limited chemostats for over 250 generations in multiple populations. We found that *GAP1* CNVs reproducibly arise early and sweep through the population. By combining the *GAP1* CNV reporter with barcode lineage tracking and whole-genome sequencing we found that 10^2–10^4 independent CNV-containing lineages comprising diverse structures compete within populations.

In this study, we estimate the formation rate and fitness effect of *GAP1* CNVs. We tested both ABC-SMC (Klinger, Rickert, and Hasenauer 2018) and a neural density estimation method, NPE (Tejero-Cantero et al. 2020), using a classical Wright-Fisher model (Otto and Day 2007) and a chemostat model (Dean 2005). Using simulated data we tested the utility of the different evolutionary models and inference methods. We find that NPE has better performance than ABC-SMC. Although a more complex model has improved performance, the simpler and more computationally efficient Wright-Fisher model is appropriate in most scenarios. We validated our approach by comparison to two different experimental methods: lineage tracking and pairwise fitness assays. We estimate that in glutamine-limited chemostats, beneficial *GAP1* CNVs are introduced at a rate of $10^{-4.7}$ -10^{-4} per cell division, and have a selection coefficient of 0.04 - 0.1 per generation. NPE is likely to be a useful method for inferring evolutionary parameters across a variety of scenarios, including tumor and viral evolution, providing a powerful approach for combining experimental and computational methods.

3.3 Results

In a previous experimental evolution study, we quantified the dynamics of *de novo* CNVs in nine populations using a prototrophic yeast strain containing a fluorescent *GAP1* CNV reporter. (Lauer et al. 2018). Populations were maintained in glutamine-limited chemostats for over 250 generations and sampled every 8-20 generations (25 time points in total) to determine the proportion of cells containing a *GAP1* CNV using flow cytometry (populations gln_01-gln_09 **Figure 3.1A**). In the same study, we also performed two replicate evolution experiments using the fluorescent *GAP1* CNV reporter and lineage-tracking barcodes quantifying the proportion of the population with a *GAP1* CNV at 32 time points (populations bc01-bc02 in **Figure 3.1A**) (Lauer et al. 2018) . We used interpolation to match timepoints between these two experiments (**Figure 3.S1**) resulting in a dataset comprising the proportion of the population with a *GAP1* CNV at 25 timepoints in 11 replicate evolution experiments. In this study, we tested whether the observed dynamics of CNV-mediated evolution provide a means of inferring the underlying evolutionary parameters.

3.3.1 Overview of evolutionary models

We tested two models of evolution: the classical Wright-Fisher model (Otto and Day 2007) and a specialized chemostat model (Dean 2005). Previously, it has been shown that a single effective selection coefficient may be sufficient to model evolutionary dynamics in populations undergoing adaptation (Hegreness et al. 2006). Therefore, we focus on beneficial mutations and assume a single selection coefficient for each class of mutation. In both models, we start with an isogenic population in which *GAP1* CNV mutations occur at a rate δ_C and other beneficial mutations occur at rate δ_B (**Figure 3.1B**). In our simulations, cells can acquire only a single beneficial mutation, either a CNV at *GAP1* or some other beneficial mutation (i.e. SNV, transposition, diploidization, or CNV at another locus). In all simulations (except for sensitivity

analysis, see *Inference from empirical GAP1 dynamics*), the mutation rate of beneficial mutations other than *GAP1* CNVs was fixed at $\delta_B=10^{-5}$ per genome per cell division and the selection coefficient was fixed at $s_B=0.001$, based on estimates from previous experiments using yeast in several conditions (Venkataram et al. 2016; Joseph and Hall 2004; Hall et al. 2008). Our goal was to infer the *GAP1* CNV mutation rate, δ_C, and *GAP1* CNV selection coefficient, s_C.

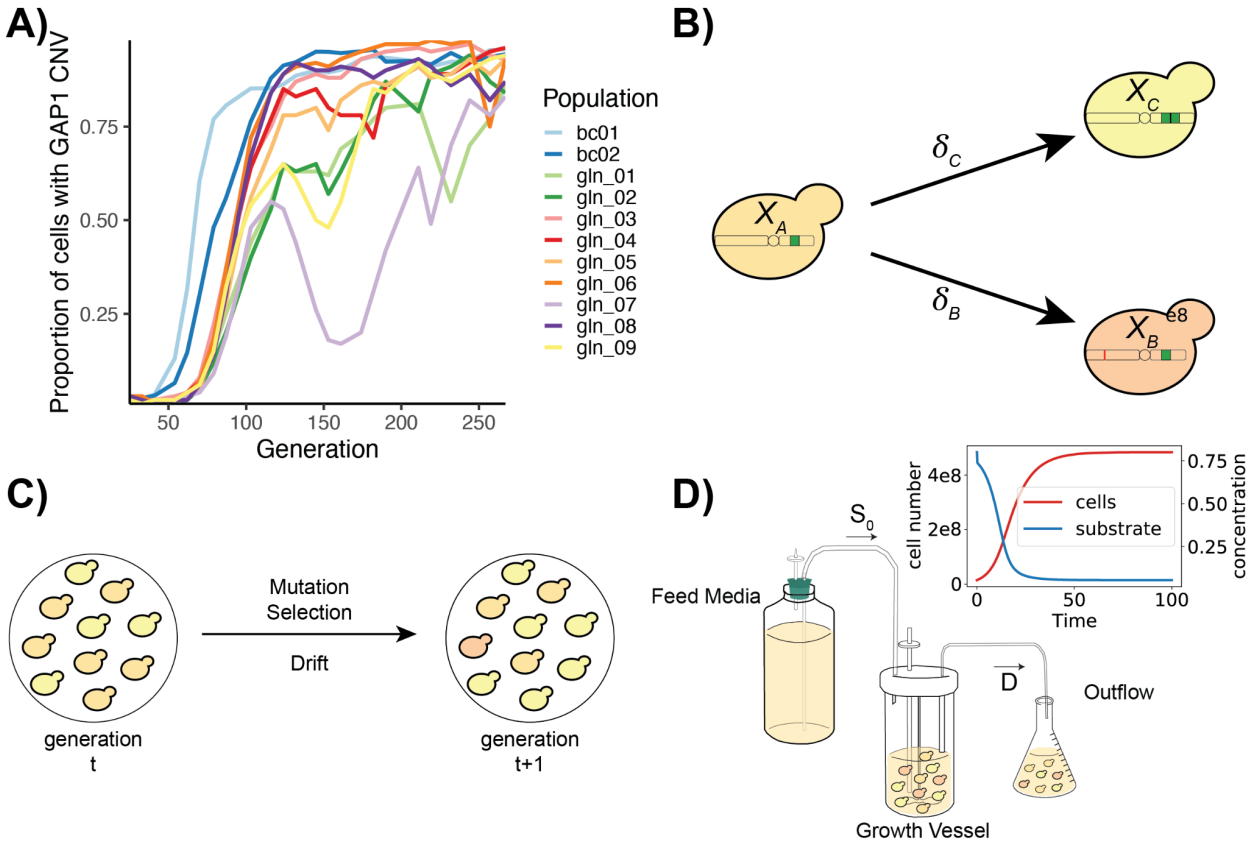

Figure 3.1. Empirical data and evolutionary models. A) Estimates of the proportion of cells with *GAP1* CNVs for eleven *S. cerevisiae* populations containing either a fluorescent *GAP1* CNV reporter (gln_01 - gln_09) or a fluorescent *GAP1* CNV reporter and lineage tracking barcodes (bc01 and bc02) evolving in glutamine-limited chemostats, from (Lauer et al. 2018). **B)** In our models, cells with the ancestral genotype (X_A) can give rise to cells with a *GAP1* CNV (X_C) or other beneficial mutation (X_B) at rates δ_C and δ_B, respectively. **C)** The Wright-Fisher model has discrete, non-overlapping generations and a constant population size. Allele frequencies in the next generation change from the previous generation due to mutation, selection, and drift. **D)** In the chemostat model, medium containing a defined concentration of a growth limiting nutrient (S_0) is added to the culture at a constant rate. The culture, containing cells and medium, is removed by continuous dilution at rate *D*. Upon inoculation, the number of cells in the growth vessel increases and the limiting-nutrient concentration decreases until a steady state is reached (red and blue curves in inset). Within the growth vessel, cells grow in continuous, overlapping generations undergoing mutation, selection, and drift.

The two evolutionary models have several unique features. In the Wright-Fisher model the population size is constant and each generation is discrete. Therefore, genetic drift is efficiently modeled using multinomial sampling (**Figure 3.1C**). In the chemostat model (Dean 2005), fresh medium is added to the growth vessel at a constant rate and medium and cells are removed from the growth vessel at the same rate resulting in continuous dilution of the culture (**Figure 3.1D**). Individuals are randomly removed from the population through the dilution process, regardless of fitness, in a manner analogous to genetic drift. In the chemostat model, we start with a small initial population size and a high initial concentration of the growth-limiting nutrient. Following inoculation, the population size increases and the growth-limiting nutrient concentration decreases until a steady state is attained that persists throughout the experiment. As generations are continuous and overlapping in the chemostat model, we use the Gillespie algorithm with τ-leaping (D. T. Gillespie 2001) to simulate the population dynamics. Growth parameters in the chemostat are based on experimental conditions during the evolution experiments (Lauer et al. 2018) or taken from the literature (**Table 3.1**).

3.3.2 Overview of inference strategies

We tested two likelihood-free Bayesian methods for joint inference of the *GAP1* CNV mutation rate and the *GAP1* CNV fitness effect: Approximate Bayesian Computation with Sequential Monte Carlo (ABC-SMC) (Sisson, Fan, and Tanaka 2007) and Neural Posterior Estimation (NPE) (Lueckmann et al. 2017; Greenberg, Nonnenmacher, and Macke 2019; Papamakarios and Murray 2016). We used the proportion of the population with a *GAP1* CNV at 25 time points as the observed data (**Figure 3.1A)**. For both methods, we defined a log-uniform prior distribution for the CNV mutation rate ranging from 10^{-12} to 10^{-3} and a log-uniform prior distribution for the selection coefficient ranging from 10^{-4} to 0.4.

We applied ABC-SMC (**Figure 3.2A**), implemented in the Python package *pyABC* (Klinger, Rickert, and Hasenauer 2018). We used an adaptively weighted Euclidean distance

function to compare simulated data to observed data. Thus, the distance function adapts over the course of the inference process based on the amount of variance at each time point (Prangle 2017). The number of samples drawn from the proposal distribution (and therefore number of simulations) is adapted at each iteration of the ABC-SMC algorithm using the adaptive population strategy, which is adapted based on the shape of the current posterior distribution (Klinger and Hasenauer 2017). We applied bounds on the maximum number of samples used to approximate the posterior in each iteration; however, the total number of samples (simulations) used in each iteration is greater because not all simulations are accepted for posterior estimation (see **Methods**). For each observation, we performed ABC-SMC with multiple iterations until either the acceptance threshold ($\varepsilon = 0.002$) was reached or until 10 iterations had been completed. We performed inference on each observation independently three times. Although we refer to different observations belonging to the same "training set", a different ABC-SMC procedure must be performed for each observation.

We applied NPE (**Figure 3.2B**), implemented in the Python package *sbi* (Tejero-Cantero et al. 2020), and tested two specialized normalizing flows as density estimators: a *masked autoregressive flow* (MAF) (Papamakarios, Pavlakou, and Murray 2017) and a *neural spline flow* (NSF) *(Durkan et al. 2019)*. The normalizing flow is used as a density estimator to "learn" an amortized posterior distribution, which can then be evaluated for specific observations. Thus, amortization allows for evaluation of the posterior for each new observation without the need to re-train the neural network. To test the sensitivity of our inference results on the set of simulations used to learn the amortized posterior, we trained three independent amortized networks with different sets of simulations generated from the prior distribution and compared our resulting posterior distributions for each observation. We refer to inferences made with the same amortized network as having the same "training set."

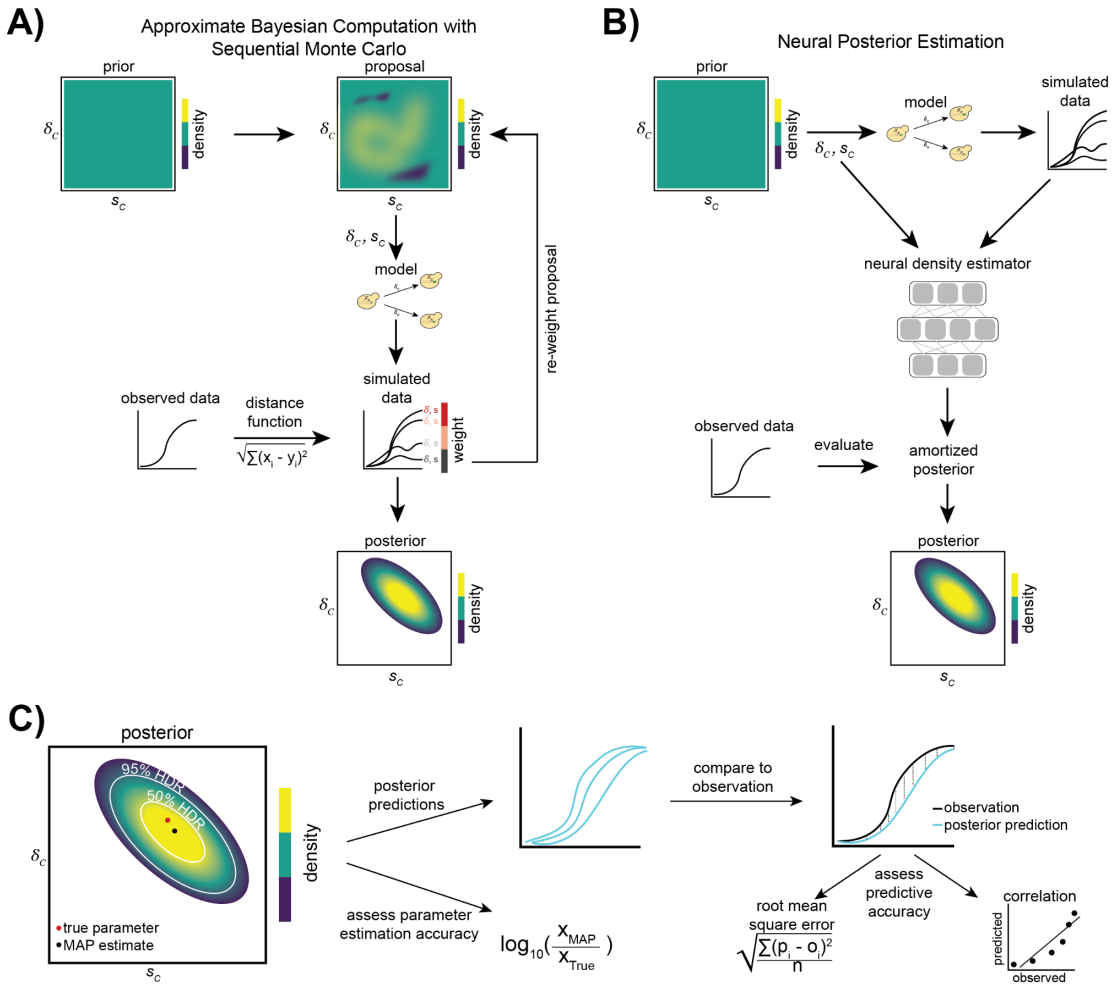

Figure 3.2. Inference methods and performance assessment. A) When using Approximate Bayesian Computation with Sequential Monte Carlo (ABC-SMC), in the first iteration a proposal for the parameters δ_C (*GAP1* CNV mutation rate) and s_C (*GAP1* CNV selection coefficient) is sampled from the prior distribution. Simulated data are generated using either a Wright-Fisher or chemostat model and the current parameter proposal. The distance between the simulated data and the observed data is computed, and the proposed parameters are weighted by this distance. These weighted parameters are used to sample the proposed parameters in the next iteration. Over many iterations, the weighted parameter proposals provide an increasingly better approximation of the posterior distribution of δ_C and s_C (adapted from (Cranmer, Brehmer, and Louppe 2020)). **B)** In Neural Posterior Estimation (NPE), simulated data are generated using parameters sampled from the prior distribution. From the simulated data and parameters, a density-estimating neural network learns the joint density of the model parameters and simulated data (the "amortized posterior"). The network then evaluates the conditional density of model parameters given the observed data, thus providing an approximation of the posterior distribution of δ_C and s_C (adapted from (Cranmer, Brehmer, and Louppe 2020) and (Gonçalves et al. 2020).) **C)** Assessment of inference performance. The 50% and 95% highest density regions (HDRs) are shown on the joint posterior distribution with the true parameters and the *maximum a posteriori* (MAP) parameter estimates. We compare the true parameters to the estimates by their log ratio. We also generate posterior predictions (sampling 50 parameters from the joint posterior distribution and using them to simulate frequency trajectories, ρ_i), which we compare to the observation, o_i, using the root mean square error (RMSE) and the correlation coefficient.

3.3.3 NPE outperforms ABC-SMC

To test the performance of each inference method and evolutionary model, we generated 20 simulated synthetic observations for each model (Wright-Fisher or chemostat) over four combinations of CNV formation rates and selection coefficients, resulting in 40 synthetic observations (i.e., five simulated observations per combination of model, δ_C, and s_C). We refer to the parameters that generated the synthetic observation as the "true" parameters. For each synthetic observation we performed inference using each method three times. Inference was performed using the same evolutionary model as that used to generate the observation. We found that NPE using NSF as the density estimator was superior to NPE using MAF, and therefore we report results using NSF in the main text (results using MAF are in **Figure 3.S2**).

For each inference method we plotted the joint posterior distribution with the 50% and 95% highest density regions (HDR) (Kruschke 2014) demarcated (**Figure 3.2C**, **Supplementary Files**). The true parameters are expected to be covered by these HDRs at least 50% and 95% of the time, respectively. We also computed the marginal 95% highest density intervals (HDI) (Kruschke 2014) using the marginal posterior distributions for the *GAP1* CNV selection coefficient and *GAP1* CNV formation rate. We found that the true parameters were within the 50% HDR in half or more of the tests (averaged over three training sets) across a range of parameter values with the exception of ABC-SMC applied to the Wright-Fisher model when the *GAP1* CNV formation rate (δ_C=10^{-7}) and selection coefficient (s_C=0.001) were both low (**Figure 3.3A**). The true parameters were within the 95% HDR in 100% of tests (**Supplementary Files**). The width of the HDI is informative about the degree of uncertainty associated with the parameter estimation. The HDIs for both fitness effect and mutation rate tend to be smaller when inferring with NPE compared to ABC-SMC, and this advantage of NPE is more pronounced when the CNV formation rate is high (δ_C=10^{-5}) (**Figure 3.3B-C**).

We computed the *maximum a posteriori* (MAP) estimate of the *GAP1* CNV formation rate and selection coefficient by determining the mode (i.e. argmax) of the joint posterior distribution, and computed the log-ratio of the MAP relative to the true parameters. We find that the MAP estimate is close to the true parameter (i.e. the log-ratio is close to zero) when the selection coefficient is high (s_C=0.1), regardless of the model or method, and much of the error is due to the mutation rate estimation error (**Figure 3.3D**). Generally, the MAP estimate is within an order of magnitude of the true parameter (i.e. the log-ratio is less than one), except when the formation rate and selection coefficient are both low (δ_C=10^{-7}, s_C=0.001); in this case the formation rate was under-estimated up to four-fold and the selection coefficient was slightly over-estimated (**Figure 3.3D**). In some cases there are substantial differences in log-ratio between training sets using NPE; however, this variation in log-ratio is usually less than the variation in the log-ratio when performing inference with ABC-SMC. Overall, the log-ratio tends to be closer to zero (i.e estimate close to true parameter) when using NPE (**Figure 3.3D**).

We performed posterior predictive checks by simulating *GAP1* CNV dynamics using the MAP estimates as well as 50 parameter values sampled from the posterior distribution (**Supplementary Files**). We computed both the root mean squared error (RMSE) and the correlation coefficient between posterior predictions and the observation to measure the prediction accuracy (**Figure 3.3E, Figure 3.S3**). We find that the RMSE posterior predictive accuracy of NPE is similar to, or better than, that of ABC-SMC (**Figure 3.3E**). The predictive accuracy quantified using correlation was close to 1 for all cases except when *GAP1* CNV formation rate and selection coefficient are both low (s_C=0.001 and δ_C=10^{-7}) (**Figure 3.S3**).

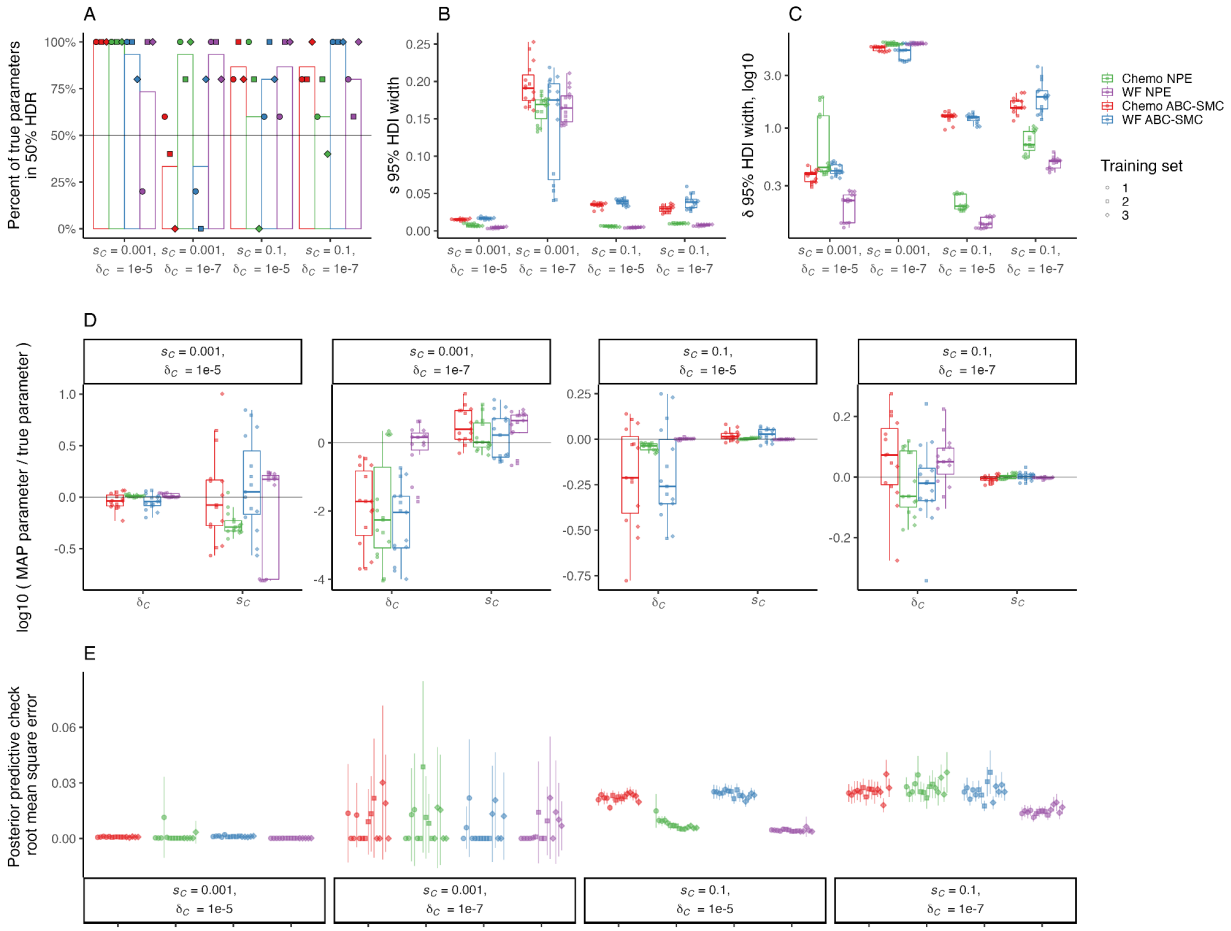

Figure 3.3. Performance assessment of inference methods using simulated synthetic observations. The figure shows the results of inference on five simulated synthetic observations using either the Wright-Fisher (WF) or chemostat (Chemo) model per combination of fitness effect s_C and mutation rate δ_C. Simulations and inference were performed using the same model. For NPE, each training set corresponds to an independently amortized posterior distribution trained on a different set of 100,000 simulations, with which each synthetic observation was evaluated to produce a separate posterior distribution. For ABC-SMC, each training set corresponds to independent inference procedures on each observation with a maximum of 10,000 total simulations accepted for each inference procedure and a stopping criteria of 10 iterations or ε <= 0.002, whichever occurs first. **A)** The percent of true parameters covered by the 50% HDR of the inferred posterior distribution. The bar height shows the average of three training sets. Horizontal line marks 50%. **B-C)** Distribution of widths of 95% highest density interval (HDI) of the posterior distribution of the fitness effect s_C (**B**) and CNV mutation rate δ_C (**C**), calculated as the difference between the 97.5 percentile and 2.5 percentile, for each separately inferred posterior distribution. **D)** Log-ratio of MAP estimate to true parameter for s_C and δ_C. Note the different y-axis ranges. Gray horizontal line represents a log-ratio of zero, indicating an accurate MAP estimate. **E)** Mean and 95% confidence interval of RMSE of 50 posterior predictions compared to the synthetic observation from which the posterior was inferred.

We performed model comparison using both AIC (Akaike information criterion), computed using the MAP estimate, and WAIC (widely applicable information criterion),

computed over the entire posterior distribution (Gelman et al. 2013). Lower values imply higher predictive accuracy and a difference of 2 is considered significant (**Figure 3.S4**) (Kass and Raftery 1995). We find similar results for both criteria: NPE with either model have similar values, though the value for Wright-Fisher is sometimes slightly lower than the value for the chemostat model. When s_C=0.1, the value for NPE is consistently and significantly lower than for ABC-SMC. When δ_C=10^{-5} and s_C=0.001, the value for NPE with the Wright-Fisher model is significantly lower than that for ABC-SMC, while the NPE with the chemostat model is not. The difference between any combination of model and method was insignificant for δ_C=10^{-7} and s_C=0.001. Therefore, NPE is similar or better than ABC-SMC using either evolutionary model and for all tested combinations of *GAP1* CNV formation rate and selection coefficient, and we further confirmed the generality of this trend using the Wright-Fisher model and eight additional parameter combinations (**Figure 3.S5).**

We performed NPE using 10,000 or 100,000 simulations to train the neural network and found that increasing the number of simulations did not substantially reduce the MAP estimation error, but did tend to decrease the width of the 95% HDIs for both parameters (**Figure 3.S6**). Similarly, we performed ABC-SMC with per observation maximum accepted parameter samples (i.e. “particles” or “population size”) numbers of 10,000 and 100,000, which correspond to increasing number of simulations per inference procedure, and found that increasing the budget decreases the widths of the 95% HDIs for both parameters (**Figure 3.S6**). Overall, amortization with NPE allowed for more accurate inference using fewer simulations corresponding to less computation time (**Figure 3.S7**).

3.3.4 The Wright-Fisher model is suitable for inference using chemostat dynamics

Whereas the chemostat model is a more precise description of our evolution experiments, both the model itself and its computational implementation have some drawbacks. First, the model is a stochastic continuous-time model implemented using the τ-leap method (D.

T. Gillespie 2001). In this method, time is incremented in discrete steps and the number of stochastic events that occur within that time step is sampled based on the rate of events and the system state at the previous time step. For accurate stochastic simulation, event rates and probabilities must be computed at each time step, and time steps must be sufficiently small. This incurs a heavy computational cost as time steps are considerably smaller than one generation, which is the time step used in the simpler Wright-Fisher model. Moreover, the chemostat model itself has additional parameters compared to the Wright-Fisher model, which must be experimentally measured or estimated.

The Wright-Fisher model is more general and more computationally efficient than the chemostat model (**Table 3.S1**). Therefore, we investigated if it can be used to perform accurate inference with NPE on synthetic observations generated by the chemostat model. By assessing how often the true parameters were covered by the HDRs, we found that the Wright-Fisher is a good-enough approximation of the full chemostat dynamics when selection is weak (s_C = 0.001) (**Figure 3.S8**), and it performs similarly to the chemostat model in parameter estimation accuracy (**Figure 3.4A-B**). The Wright-Fisher is less suitable when selection is strong (s_C = 0.1), as the true parameters are not covered by the 50% or 95% HDR (**Figure 3.S8**). Nevertheless, estimation of the selection coefficient remains accurate, and the difference in estimation of the formation rate is less than an order of magnitude, with a 3-5-fold overestimation (MAP log-ratio between 0.5 and 0.7) (**Figure 3.4C-D**).

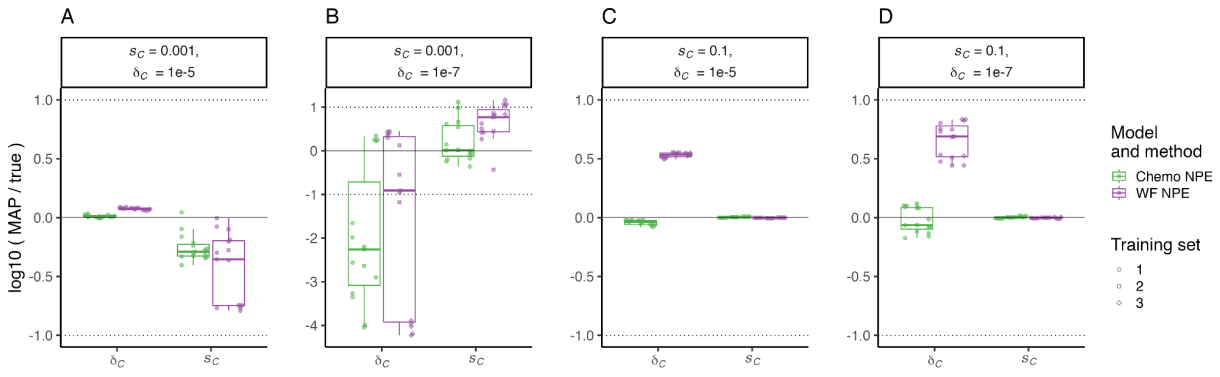

Figure 3.4. Inference with Wright-Fisher model from chemostat dynamics. The figure shows results of inference using NPE and either the Wright-Fisher (WF) or chemostat (Chemo) model on five simulated synthetic observations generated using the chemostat model for different combinations of fitness effect s_C and formation rate δ_C. Boxplots and markers show the log-ratio of MAP estimate to true parameters for s_C and δ_C. Horizontal solid line represents a log-ratio of zero, indicating an accurate MAP estimate; dotted lines indicate an order of magnitude difference between the MAP estimate and the true parameter.

3.3.5 Inference using a set of observations

Our empirical dataset includes eleven biological replicates of the same evolution experiment. Differences in the dynamics between independent replicates may be explained by an underlying distribution of fitness effects (DFE) rather than a single constant selection coefficient. It is possible to infer the DFE using all experiments simultaneously. However, inference of distributions from multiple experiments presents several challenges, common to other mixed-effects or hierarchical models (X. A. Harrison et al. 2018). Alternatively, individual values inferred from individual experiments could provide an approximation of the underlying DFE.

To test these two alternative strategies for inferring the DFE, we performed simulations in which we allowed for variation in the selection coefficient of *GAP1* CNVs for each population in a set of observations. We sampled eleven selection coefficients from a Gamma distribution with shape and scale parameters α and β, respectively, and an expected value $E(s) = \alpha\beta$ (Moura de Sousa, Campos, and Gordo 2013), and then simulated a single observation for each sampled selection coefficient. As the Wright-Fisher model is a suitable approximation of the

chemostat model (**Figure 3.4**), we used the Wright-Fisher model both for generating our observation sets and for parameter inference.

For the observation sets, we used NPE to either infer a single selection coefficient for each observation or to directly infer the Gamma distribution parameters α and β from all eleven observations. When inferring eleven selection coefficients, one for each observation in the observation set, we fit a Gamma distribution to eight of the eleven inferred values (**Figure 3.5**, green lines). When directly inferring the DFE, we used a uniform prior for α from 0.5 to 15 and a log-uniform prior for β from 10^{-3} to 0.8. We held out three experiments from the set of eleven and used a three-layer neural network to reduce the remaining eight observations to a five-feature summary statistic vector, which we then used as an embedding net (Tejero-Cantero et al. 2020) with NPE to infer the joint posterior distribution of α, β, and δ_C (**Figure 3.5**, blue lines). For each observation set, we performed each inference method three times, using different sets of eight experiments to infer the underlying DFE.

We used Kullback–Leibler divergence to measure the difference between the true DFE and inferred DFE, and find that the inferred selection coefficients from the single experiments capture the underlying DFE as well or better than direct inference of the DFE from a set of observations for both $\alpha = 1$ (an exponential distribution) and $\alpha = 10$ (sum of ten exponentials) (**Figure 3.5, Figure 3.S9)**. The only exception we found is when $\alpha = 10,\ E(s) = 0.001$, and δ_C=10^{-5} (**Figure 3.S9, Table 3.S2**). We assessed the performance of inference from a set of observations using out-of-sample posterior predictive accuracy (Gelman et al. 2013) and found that inferring α and β from a set of observations results in lower posterior predictive accuracy compared to inferring s_C from a single observation (**Figure 3.S10**). Therefore, we conclude that estimating the DFE through inference of individual selection coefficients from each observation is superior to inference of the distribution from multiple observations.

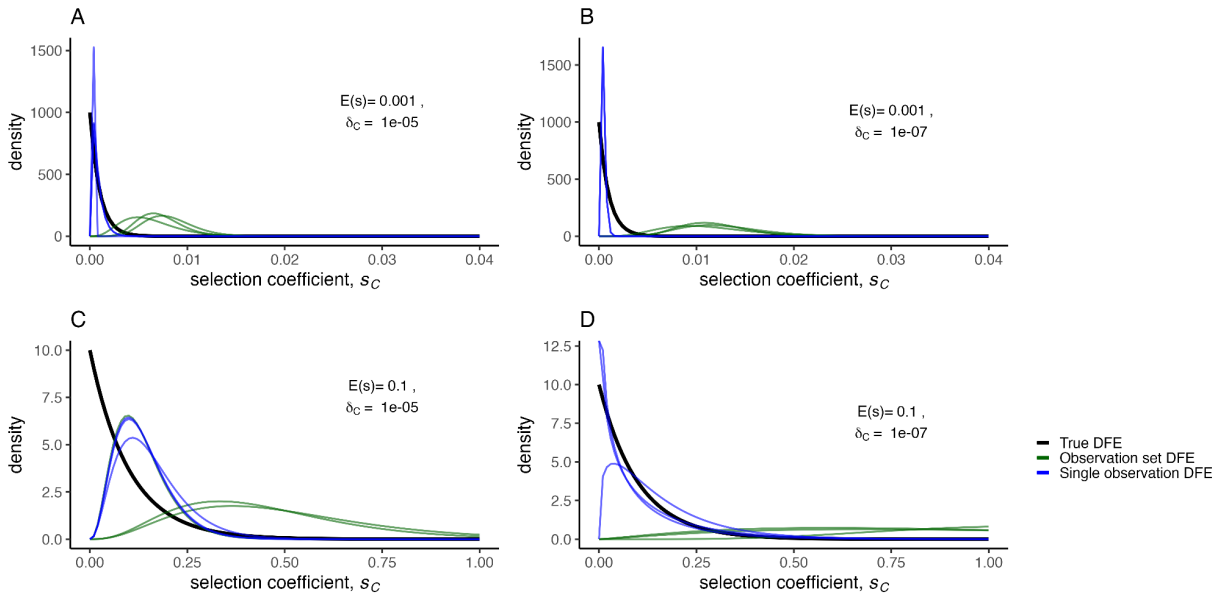

Figure 3.5. Inference of the distribution of fitness effects. A set of eleven simulated synthetic observations was generated from a Wright-Fisher model with CNV selection coefficients sampled from an exponential (Gamma with $\alpha = 1$) distribution of fitness effects (true DFE; black curve). The MAP DFEs (observation set DFE, green curves) were directly inferred using three different subsets of eight out of eleven synthetic observations. We also inferred the selection coefficient for each individual observation in the set of eleven separately, and fit a Gamma distribution (single observation DFE, blue curves) to sets of eight inferred selection coefficients. All inferences were performed with NPE using the same amortized network to infer a posterior for each set of eight synthetic observations or each single observation. **A**) weak selection, high formation rate, **B)** weak selection, low formation rate, **C**) strong selection, high formation rate, **D**) strong selection, low formation rate.

3.3.6 Inference from empirical evolutionary dynamics

To apply our approach to empirical data we inferred *GAP1* CNV selection coefficients and formation rates using eleven replicated evolutionary experiments in glutamine-limited chemostats (Lauer et al. 2018) (**Figure 3.1A**) using NPE with both evolution models. We performed posterior predictive checks, drawing parameter values from the posterior distribution, and found that *GAP1* CNV were predicted to increase in frequency earlier and more gradually than is observed in our experimental populations (**Figure 3.S11**). This discrepancy is especially apparent in experimental populations that appear to experience clonal interference with other beneficial lineages (i.e. gln07, gln09). Therefore, we excluded data after generation 116, by which point CNVs have reached high frequency in the populations but do not yet exhibit the

non-monotonic and variable dynamics observed in later time points, and performed inference. The resulting posterior predictions are more similar to the observations in initial generations (average MAP RMSE for the eleven observations up to generation 116 is 0.06 when inference excludes late time points versus 0.13 when inference includes all time points). Furthermore, the overall RMSE (for observations up to generation 267) was not significantly different (average MAP RMSE is 0.129 and 0.126 when excluding or including late time points, respectively; **Figure 3.S12**). Restricting the analysis to early time points did not dramatically affect estimates of *GAP1* CNV selection coefficient and formation rate, but it did result in less variability in estimates between populations (i.e. independent observations) and some reordering of populations' selection coefficients and formation rate relative to each other (**Figure 3.S13**). Thus, we focused on inference using data prior to generation 116.

The inferred *GAP1* CNV selection coefficients were similar regardless of model, with the range of MAP estimates for all populations between 0.04 and 0.1, whereas the range of inferred *GAP1* CNV formation rates was somewhat higher when using the Wright-Fisher model, $10^{-4.1}$-$10^{-3.4}$, compared to the chemostat model, $10^{-4.7}$ -10^{-4} (**Figure 3.6A-B**). While there is variation in inferred parameters due to the training set, variation between observations (replicate evolution experiments) is higher than variation between training sets (**Figure 3.6A-C**). Posterior predictions using the chemostat model, a fuller depiction of the evolution experiments, tend to have slightly lower RMSE than predictions using the Wright-Fisher model (**Figure 3.6C**). However, predictions using both models recapitulate actual *GAP1* CNV dynamics, especially in early generations (**Figure 3.6D**).

To test the sensitivity of these estimates, we also inferred the *GAP1* CNV selection coefficient and formation rate using the Wright-Fisher model in the absence of other beneficial mutations (δ_B=0), and for nine additional combinations of other beneficial mutation selection coefficient s_B and formation rate δ_B (**Figure 3.S14**). In general, perturbations to the rate and

selection coefficient of other beneficial mutations did not alter the inferred *GAP1* CNV selection coefficient or formation rate. We found a single exception: when both the formation rate and fitness effect of other beneficial mutations is high (s_B=0.1 and δ_B=10^{-5}), the *GAP1* CNV selection coefficient was approximately 1.6-fold higher and the formation rate was approximately 2-fold lower (**Figure 3.S14**); however, posterior predictions were poor for this set of parameter values (**Figure 3.S15**) suggesting these values are inappropriate.

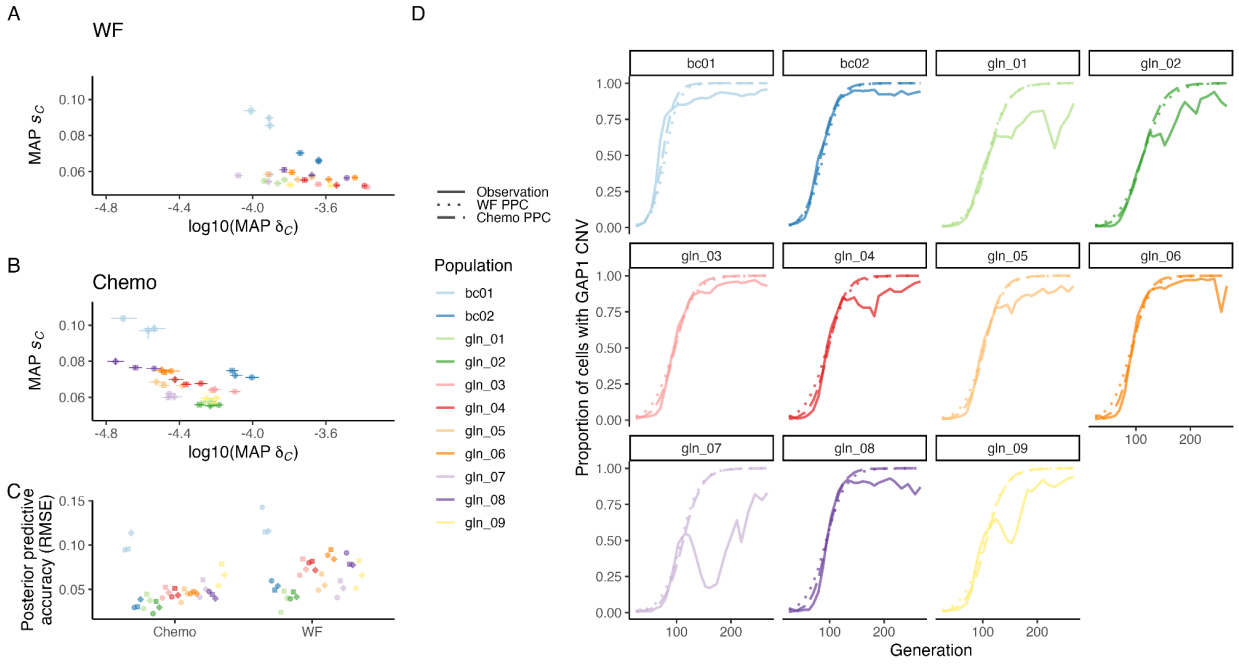

Figure 3.6. Inference of CNV formation rate and fitness effect from empirical evolutionary dynamics. The inferred MAP estimate and 95% highest density intervals (HDI) for fitness effect s_C and formation rate δ_C, using the **(A)** Wright-Fisher (WF) or **(B)** chemostat (Chemo) model and NPE for each experimental population from (Lauer et al. 2018). Inference performed with data up to generation 116, and each training set (marker shape) corresponds to an independent amortized posterior distribution estimated with 100,000 simulations. **C)** Mean and 95% confidence interval for RMSE of 50 posterior predictions compared to empirical observations up to generation 116. **D)** Proportion of the population with a *GAP1* CNV in the experimental observations (solid lines) and in posterior predictions using the MAP estimate from one of the training sets shown in panels A and B with either the Wright-Fisher (dotted line) or chemostat (dashed line) model. Mutation rate and fitness effect of other beneficial mutations set to 10^{-5} and 10^{-3}, respectively.

3.3.7 Experimental confirmation of fitness effects inferred from adaptive dynamics

To experimentally validate the inferred selection coefficients, we used lineage tracking to estimate the distribution of fitness effects (Levy et al. 2015; Nguyen Ba et al. 2019; Aggeli, Li,

and Sherlock, n.d.). We performed barseq on the entire evolving population at multiple time points and identified lineages that did and did not contain *GAP1* CNVs (**Figure 3.7A**). Using barcode trajectories to estimate fitness effects ((Levy et al. 2015); see **Methods**), we identified 1,569 out of 80,751 lineages (1.94%) as adaptive in the bc01 population. A total of 1,513 (96.4%) adaptive lineages have a *GAP1* CNV (**Figure 3.7A**).

As a complementary experimental approach, selection coefficients can be directly measured using competition assays by fitting a linear model to the log-ratio of the *GAP1* CNV strain and ancestral strain frequencies over time (**Figure 3.7B**). Therefore, we isolated *GAP1* CNV containing clones from populations bc01 and bc02, determined their fitness (**Methods),** and combined these estimates with previously reported selection coefficients for *GAP1* CNV containing clones isolated from populations gln01-gln09 (Lauer et al. 2018) to define the DFE.

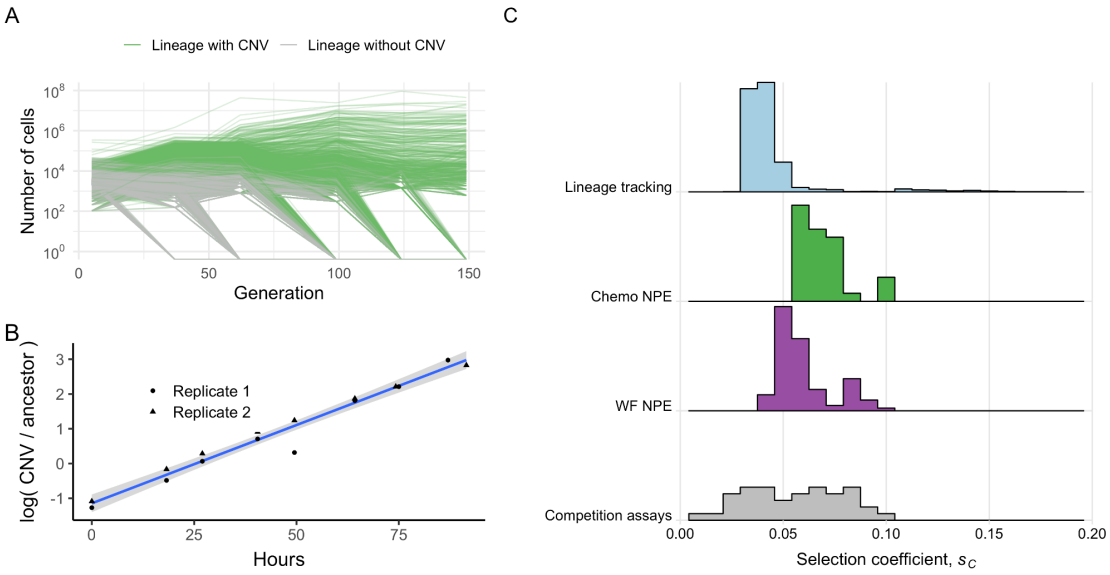

Figure 3.7. Comparison of DFE inferred using NPE, lineage-tracking barcodes, and competition assays. A) Barcode-based lineage frequency trajectories in experimental population bc01. Lineages with (green) and without (gray) *GAP1* CNVs are shown. **B)** Two replicates of a pairwise competition assay for a single *GAP1* CNV containing lineage isolated from an evolving population. The selection coefficient for the clone is estimated from the slope of the linear model (blue line) and 95% CI (gray). **C)** The distribution of fitness effects for all beneficial *GAP1* CNVs inferred from eleven populations using NPE and the Wright-Fisher (WF; purple) and chemostat (Chemo; green) models compared with the DFE inferred from barcode frequency trajectories in the bc01 population (light blue) and the DFE inferred using pairwise competition assays with different *GAP1* CNV containing clones (gray).

The DFE for adaptive *GAP1* CNV lineages in bc01 inferred using lineage-tracking barcodes and the DFE from pairwise competition assays share similar properties to the distribution inferred using NPE from all experimental populations (**Figure 3.7C**). Thus, our inference framework using CNV adaptation dynamics is a reliable estimate of the DFE estimated using laborious experimental methods that are gold-standards in the field.

3.4 Discussion

In this study we tested the application of simulation-based inference for determining key evolutionary parameters from observed adaptive dynamics in evolution experiments. We focused on the role of CNVs in adaptive evolution using experimental data in which we quantified the population frequency of *de novo* CNVs at a single locus using a fluorescent CNV reporter. The goal of our study was to test a new computational framework for simulation-based, likelihood-free inference, compare it to the state of the art method, and apply it to estimate the *GAP1* CNV selection coefficient and formation rates in experimental evolution using glutamine-limited chemostats.

Our study yielded several important methodological findings. Using synthetic data we tested two different model-based algorithms for joint inference of evolutionary parameters, the effect of different evolutionary models on inference performance, and how best to determine a distribution of fitness effects using multiple experiments. We find that the neural-network-based algorithm NPE outperforms ABC-SMC regardless of evolutionary model. Although a more complex evolutionary model better describes the evolution experiments performed in chemostats, we find that a standard Wright-Fisher model can be a sufficient approximation for inference using NPE. However, the inferred *GAP1* CNV formation rate under the Wright-Fisher model is higher than under the chemostat model (**Figure 3.6A-B**), which is consistent with the overprediction of formation rates using the Wright-Fisher model for inference when an

observation is generated by the chemostat model and selection coefficients are high (**Figure 3.4C-D**). This suggests that Wright-Fisher is not the best-suited model to use in all real-world cases, in particular if many beneficial CNVs turn out to have strong selection coefficients. Finally, although it is possible to perform joint inference on multiple independent experimental observations to infer a distribution of fitness effects, we find that inference performed on individual experiments and post-facto estimation of the distribution more accurately captures the underlying distribution of fitness effects.

Previous studies that applied likelihood-free inference to results of evolutionary experiments differ from our study in various ways (Hegreness et al. 2006; Barrick et al. 2010; Harari et al. 2018). First, they used serial-dilution rather than chemostat experiments. Second, most focused on all beneficial mutations, whereas we categorize beneficial mutations into two categories: *GAP1* CNVs and all other beneficial mutations; thus, they used an evolutionary model with a single process generating genetic variation whereas our study includes two such processes, but focuses inference on our mutation type of interest. Third, we used two different evolutionary models: the Wright-Fisher model, a standard model in evolutionary genetics, and a chemostat model. The latter is more realistic but also more computationally demanding. Fourth and importantly, previous studies applied relatively simple rejection-ABC methods (Hegreness et al. 2006; Barrick et al. 2010; Moura de Sousa, Campos, and Gordo 2013; Harari et al. 2018). We applied two modern approaches: ABC with sequential Monte Carlo sampling (Sisson, Fan, and Tanaka 2007), which is a computationally efficient algorithm for Bayesian inference, using an adaptive distance function (Prangle 2017); and NPE (Lueckmann et al. 2017; Greenberg, Nonnenmacher, and Macke 2019; Papamakarios and Murray 2016) with NSF (Durkan et al. 2019). NPE approximates an amortized posterior distribution from simulations. Thus, it is more efficient than ABC-SMC, as it can estimate a posterior distribution for new observations without requiring additional training. This feature is especially useful when a more computationally

demanding model is better (e.g., the chemostat model when selection coefficients are high). Our study is the first, to our knowledge, to use neural density estimation to apply likelihood-free inference to experimental evolution data.

Our application of simulation-based inference yielded new insights into the role of CNVs in adaptive evolution. We estimated *GAP1* CNV formation rate and selection coefficient from empirical population-level adaptive evolution dynamics and found that *GAP1* CNVs form at a rate of $10^{-3.5}$-$10^{-4.5}$ per generation (approximately 1 in 10,000 cell divisions) and have selection coefficients of 0.05-0.1 per generation. We experimentally validated our inferred fitness estimates using barcode lineage tracking and pairwise competition assays and showed that simulation-based inference is in good agreement with the two different experimental methods. The formation rate that we have determined for *GAP1* CNVs is remarkably high. Locus-specific CNV formation rates are extremely difficult to determine and fluctuation assays have yielded estimates ranging from 10^{-12} to 10^{-6} (Lynch et al. 2008; H. Zhang et al. 2013; Schacherer et al. 2005, 2007; Dorsey et al. 1992). Mutation accumulation studies have yielded genome-wide CNV rates of about 10^{-5} (Y. O. Zhu et al. 2014; Sharp et al. 2018; Sui et al. 2020), which is an order of magnitude lower than our locus specific formation rate. We posit two possible explanations for this high rate: 1) CNVs at the *GAP1* locus may be deleterious in most conditions, including the putative non-selective conditions used for mutation-selection experiments, and therefore underestimated in mutation accumulation assays due to negative selection; and 2) under nitrogen-limiting selective conditions, in which *GAP1* expression levels are extremely high, a mechanism of induced CNV formation may operate that increases the rate at which they are generated, as has been shown at other loci in the yeast genome (Hull et al. 2017; Whale et al. 2021). Empirical validation of the inferred rate of *GAP1* CNV formation in nitrogen-limiting conditions requires experimental confirmation.

This simulation-based inference approach can be readily extended to other evolution experiments. In this study we performed inference of parameters for a single type of mutation. This approach could be extended to infer the rates and effects of multiple types of mutations simultaneously. For example, instead of assuming a rate and selection coefficient for other beneficial mutations and performing ex post facto analyses looking at the sensitivity of inference of *GAP1* CNV parameters in other beneficial mutation regimes, one could simultaneously infer parameters for both of these types of mutations. As shown using our barcode-sequencing data, many CNVs arise during adaptive evolution, and previous studies have shown that CNVs have different structures and mechanisms of formation (Lauer et al. 2018; Hong and Gresham 2014a). Inferring a single effective selection coefficient and formation rate is a current limitation of our study that could be overcome by inferring rates and effects for different classes of CNVs (e.g, aneuploidy vs tandem duplication). Inspecting conditional correlations in posterior distributions involving multiple types of mutations has the potential to provide insights into how interactions between different classes of mutations shape evolutionary dynamics.

The approach could also be applied to CNV dynamics at other loci, in different genetic backgrounds, or in different media conditions. Ploidy and diverse molecular mechanisms likely impact CNV formation rates. For example, rates of aneuploidy, which result from nondisjunction errors, are higher in diploid yeast than haploid yeast, and chromosome gains are more frequent than chromosome losses (Sharp et al. 2018). There is considerable evidence for heterogeneity in the CNV rate between loci, as factors including local sequence features, transcriptional activity, genetic background, and the external environment may impact the mutation spectrum. For example, there is evidence that CNVs occur at a higher rate near certain genomic features, such as repetitive elements (Farslow et al. 2015), tRNA genes (Bermudez-Santana et al. 2010), origins of replication (Di Rienzi et al. 2009), and replication fork barriers (Labib et al. 2007).

Furthermore, this approach could be used to infer formation rates and selection coefficients for other types of mutations in different asexually reproducing populations; the empirical data required is simply the proportion of the population with a given mutation type over time, which can efficiently be determined using a phenotypic marker, or similar quantitative data such as whole-genome whole-population sequencing. Evolutionary models could be extended to more complex evolutionary scenarios including changing population sizes, fluctuating selection, and changing ploidy and reproductive strategy, with an ultimate goal of inferring their impact on a variety of evolutionary parameters and predicting evolutionary dynamics in complex environments and populations. Applications to tumor evolution and viral evolution are related problems that are likely amenable to this approach.

3.5 Methods

All source code and data for performing the analyses and reproducing the figures is available at https://doi.org/10.17605/OSF.IO/E9D5X. Code is also available at https://github.com/graceave/cnv_sims_inference.

3.5.1 Evolutionary models

We modeled the adaptive evolution from an isogenic asexual population with frequencies X_A of the ancestral (or wild-type) genotype, X_C of cells with a *GAP1* CNV, and X_B of cells with a different type of beneficial mutation. Ancestral cells can gain a *GAP1* CNV or another beneficial mutation at rates δ_C and δ_B, respectively. Therefore, the frequencies of cells of different genotypes after mutation are

$$x^{\dagger}_{A} = (1 - \delta_{B} - \delta_{C})x_{A},$$

$$x^{\dagger}_{B} = x_{A}\delta_{B} + x_{B},$$

$$x^{\dagger}_{C} = x_{A}\delta_{C} + x_{C}$$

For simplicity, this model neglects cells with multiple mutations, which is reasonable for short timescales, such as those considered here.

In the discrete time Wright-Fisher model, the change in frequency due to natural selection is modeled by

$$x^*_i = \frac{w_k x_i}{\bar{w}}, \qquad \bar{w} = \sum_{i \in \{A,B,C\}} w_i x_i,$$

where w_i is the relative fitness of cells with genotype i , and $\bar{w}$ is the population mean fitness relative to the ancestral type. Relative fitness is related to the selection coefficient by

$$w_i = 1 + s_i,\ i = B, C$$

The change in due random genetic drift is given by

$$n_i = Multinomial(N,\ (x^*_A,\ x^*_B,\ x^*_C)), \qquad x'_i = \frac{n_i}{N}$$

where N is the population size. In our simulations $N=3.3 \times 10^8$, the effective population size in the chemostat populations in our experiment (see **Determining effective population size in the chemostat**).

The chemostat model starts with a population size 1.5×10^{-7} and the concentration of the limiting nutrient in the growth vessel, *S*, is equal to the concentration of that nutrient in the fresh media, S_0. During continuous culture, the chemostat is continuously diluted as fresh media flows in and culture media and cells are removed at rate *D*. During the initial phase of growth, the population size grows and the limiting nutrient concentration is reduced until a steady state is attained at which the population size and limiting nutrient concentration are maintained indefinitely. We extended the model for competition between two haploid clonal populations for a single growth-limiting resource in a chemostat from (Dean 2005) to three populations such that

$$\frac{dx_A}{dt} = x_A\left(\frac{r_A S}{S+k_A} - D\right),$$

$$\frac{dx_B}{dt} = x_B\left(\frac{(r_B)S}{S+k_B} - D\right),$$

$$\frac{dx_C}{dt} = x_C\left(\frac{(r_C)S}{S+k_C} - D\right),$$

$$\frac{dS}{dt} = \left(S_0 - S\right)D - \frac{x_A r_A S}{(S+k_A)Y_A} - \frac{x_B r_B S}{(S+k_B)Y_B} - \frac{x_C r_C S}{(S+k_C)Y_C}$$

Y_i is the culture yield of strain i per mole of limiting nutrient. r_A is the Malthusian parameter, or intrinsic rate of increase, for the ancestral strain, and in the chemostat literature is frequently referred to as μ_{max}, the maximal growth rate. The growth rate in the chemostat, μ, depends on the the concentration of the limiting nutrient with saturating kinetics $\mu = \frac{\mu_{max} S}{k_s + S}$. k_i is the substrate concentration at half-maximal μ. r_C and r_B are the Malthusian parameters for strains with a CNV and strains with an other beneficial mutation, respectively, and are related to the ancestral Malthusian parameter and selection coefficient by (Chevin 2011)

$$s_i = \frac{r_i - r_A}{r_A} ln2,\ i = B, C\ .$$

The values for the parameters used in the chemostat model are in Table 3.1.

We simulated continuous time in the chemostat using the Gillespie algorithm with τ-leaping. Briefly, we calculate the rates of ancestral growth, ancestral dilution, CNV growth, CNV dilution, other mutant growth, other mutant dilution, mutation from ancestral to CNV, and mutation from ancestral to other mutant. For the next time interval τ we calculated the number of times each event occurs during the interval using the Poisson distribution. The limiting substrate concentration is then adjusted accordingly. These steps repeat until the desired number of generations is reached.

For the chemostat model, we began counting generations after 48 hours, which is approximately the amount of time required for the chemostat to reach steady state, and when we began recording generations in (Lauer et al. 2018).

Table 3.1. Chemostat parameters

Parameter	Value	Source
$k_A=k_B=k_C$	0.103 mM	Airoldi et al. 2016 https://doi.org/10.1091/mbc.E14-05-1013
$Y_A=Y_B=Y_C$	32,445,000 cells/mL/mM nitrogen	Airoldi et al. 2016 https://doi.org/10.1091/mbc.E14-05-1013
Expected *S* at steady state	Approximately 0.08 mM	Airoldi et al. 2016 https://doi.org/10.1091/mbc.E14-05-1013
μ_{max}	0.35 hr^{-1}	Cooper TG (1982) Nitrogen metabolism in Saccharomyces cerevisiae
D	0.12 hr^{-1}	Lauer et al. 2018
S_0	0.8 mM	Lauer et al. 2018
Expected cell density at steady state	Approximately $2.5x10^7$ cells/mL	Lauer et al. 2018
Doubling time	5.8 hours	Lauer et al. 2018

3.5.2 Determining the effective population size in the chemostat

In order to determine the effective population size in the chemostat, and thus the population size to use in with the Wright-Fisher model, we determined the conditional variance of the allele frequency in the next generation *p'* given the frequency in the current generation *p* in the chemostat. To do this, we simulated a chemostat population with two neutral alleles with

frequencies *p* and *q* (*p+q=1*), which begin at equal frequencies, *p=q*. We allowed the simulation to run for 1,000 generations, recording the frequency *p* at every generation, excluding the first 100 generations to ensure the population is at steady state. We then computed the conditional variance *Var(p'|p)* in each generation and estimated the effective population size as (where *t=900* is the total number of generations) by (Crow and Kimura 1970):

$$N_e = \frac{p(1-p)}{\frac{1}{t}\sum_{t} var(p'|p)}.$$

The estimated effective population size in our chemostat conditions is 3.3×10^8, which is approximately two thirds of the census population size *N* when the chemostat is at steady state.

3.5.3 Inference methods

For inference using single observations, we used the proportion of the population with a *GAP1* CNV at 25 time points as our summary statistics and defined a log-uniform prior for the mutation rate ranging from 10^{-12} to 10^{-3} and a log-uniform prior for the selection coefficient from 10^{-4} to 0.4.

For inference using sets of observation, we used a uniform prior for α from 0.5 to 15, a log-uniform prior for β from 10^{-3} to 0.8, and a log-uniform prior for the mutation rate ranging from 10^{-12} to 10^{-3}. For use with NPE, we used a three layer sequential neural network with linear transformations in each layer and Rectified Linear Unit as the activation functions to encode the observation set into five summary statistics, which we then used as an embedding net with NPE.

We applied ABC-SMC implemented in the Python package *pyABC* (Klinger, Rickert, and Hasenauer 2018). For inference using single observations we used an adaptively weighted Euclidean distance function with the root mean square deviation as the scale function. For inference using a set of observations, we used the squared Euclidean distance as our distance metric. We used 100 samples from the prior for initial calibration before the first round, and a

maximum acceptance rate of either 10,000 or 100,000 for both single observations and observation sets (i.e.10,000 single observations or 10,000 sets of 11 observations). For the acceptance rate of 10,000, we started inference with 100 samples, had a maximum of 1,000 accepted samples per round, and a maximum of ten rounds. For the acceptance rate of 100,000, we started inference with 1,000 samples, had a maximum of 10,000 accepted samples per round, and a maximum of ten rounds. The exact number of samples from the proposal distribution during each round of sampling were adaptively determined based on the shape of the current posterior distribution (Klinger and Hasenauer 2017). For inference of the posterior for each observation, we performed multiple rounds of sampling until either we reached the acceptance threshold $\varepsilon <= 0.002$ or ten rounds were performed.

We applied NPE implemented in the Python package *sbi* (Tejero-Cantero et al. 2020) using a *Masked Autoregressive Flow* (*MAF*) (Papamakarios, Pavlakou, and Murray 2017) or a a *neural spline flow (NSF) (Durkan et al. 2019)* as a conditional density estimator that learns an amortized posterior density for single observations. We used either 10,000 or 100,000 simulations to train the network. To test the dependence of our results on the set of simulations used to learn the posterior, we trained three independent amortized networks with different sets of simulations generated from the prior and compared our resulting posterior distributions for each observation.

3.5.4 Assessment of performance of each method with each model

To test each method, we simulated five populations for each combination of the following CNV mutation rates and fitness effects: s_C=0.001 and δ_C=10^{-5}; s_C=0.1 and δ_C=10^{-5}; s_C=0.001 and δ_C=10^{-7}; s_C=0.1 and δ_C=10^{-7}, for both the Wright-Fisher model and the chemostat model, resulting in 40 total simulated observations. We independently inferred the CNV fitness effect and mutation rate for each simulated observation three times.

We calculated the MAP estimate by first estimating a Gaussian kernel density estimate (KDE) using *SciPy* (*scipy.stats.gaussian_kde)* (Virtanen et al. 2020) with at least 1,000 parameter combinations and their weights drawn from the posterior distribution. We then found the maximum of the KDE (using *scipy.optimize.minimize* with the Nelder-Mead solver). We calculated the 95% highest density intervals for the MAP estimate of each parameter using *pyABC* (*pyabc.visualization.credible.compute_credible_interval*) (Klinger, Rickert, and Hasenauer 2018).

We performed posterior predictive checks by simulating CNV dynamics using the MAP estimate as well as 50 parameter values sampled from the posterior distribution. We calculated root mean square error (RMSE) and correlation to measure agreement of the 50 posterior predictions with the observation, and report the mean and 95% confidence intervals for these measures. For inference on sets of observations, we calculated the RMSE and correlation coefficient between the posterior predictions and each of the three held out observations, and report the mean and 95% confidence intervals for these measures over all three held out observations.

We calculated *Akaike's information criteria* (AIC) using the standard formula

$$AIC = -2log(p(y|\hat{\theta})) + 2k$$

where $\hat{\theta}$ is the MAP estimate, $k = 2$ is the number of inferred parameters, *y* is the observed data, and *p* is the inferred posterior distribution. We calculated *Watanabe-Akaike information criterion* or *widely applicable information criterion* (WAIC) according to both commonly used formulas:

$$WAIC1 = -2\sum_{i=1}^{n} log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^s)\right) + 2\sum_{i=1}^{n}\left(log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^s)\right) - \frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^s)\right)$$

$$WAIC2 = -2\sum_{i=1}^{n} log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^s)\right) + 2\sum_{i=1}^{n} V_{s=1}^{S}\left(logp(y_i|\theta^s)\right)$$

where S is the number of draws from the posterior distribution, θ^s is a sample from the posterior, and $V_{s=1}^{S}$ is the posterior sample variance.

3.5.6 Pairwise competitions

We isolated CNV-containing clones from the populations on the basis of fluorescence, and performed pairwise competitions between each clone and an unlabeled ancestral (FY4) strain. We also performed competitions between the ancestral *GAP1* CNV reporter strain, with and without barcodes. To perform the competitions, we grew fluorescent *GAP1* CNV clones and ancestral clones in glutamine-limited chemostats until they reached steady state (Lauer et al. 2018). We then mixed the fluorescent strains with the unlabeled ancestor in a ratio of approximately 1:9, and performed competitions in the chemostats for 92 hours or about 16 generations, sampling approximately every 2-3 generations. For each time point, at least 100,000 cells were analyzed using an Accuri flow cytometer to determine the relative abundance of each genotype. Previously, we established that the ancestral *GAP1* CNV reporter has no detectable fitness effect compared to the unlabeled ancestral strain (Lauer et al. 2018). However, the *GAP1* CNV reporter with barcodes does appear to have a slight fitness cost associated with it, therefore, we took slightly different approaches to determine the selection coefficient relative to the ancestral state depending on whether or not a *GAP1* CNV containing clone was barcoded. If a clone was not barcoded, we detemined relative fitness using linear regression of the log-ratio of the frequency of the two genotypes against the number of elapsed hours. If a clone was barcoded, relative fitness was computed using linear regression of the log-ratio of the frequencies of the barcoded *GAP1*-CNV-containing clone and the unlabeled ancestor, and the log-ratio of the frequencies of the unevolved barcoded *GAP1* CNV reporter

ancestor to the unlabeled ancestor against the number of elapsed hours, adding an additional interaction term for the evolved versus ancestral state. We converted relative fitness from per hour to generation by dividing by the natural log of two.

3.5.7 Barcode sequencing

In our prior study, populations with lineage tracking barcodes and the *GAP1* CNV reporter were evolved in glutamine-limited chemostats (Lauer et al. 2018), and whole population samples were periodically frozen in 15% glycerol. To extract DNA, we thawed pelleted cells using centrifugation and extracted genomic DNA using a modified Hoffman-Winston protocol, preceded by incubation with zymolyase at 37°C to enhance cell lysis (Hoffman and Winston 1987). We measured DNA quantity using a fluorometer, and used all DNA from each sample as input to a sequential PCR protocol to amplify DNA barcodes which were then purified using a Nucleospin PCR clean-up kit, as described previously(Levy et al. 2015; Lauer et al. 2018).

We measured fragment size with an Agilent TapeStation 2200 and performed qPCR to determine the final library concentration. DNA libraries were sequenced using a paired-end 2 × 150 bp protocol on an Illumina NovaSeq 6000 using an XP workflow. Standard metrics were used to assess data quality (Q30 and %PF). We used the Bartender algorithm with UMI handling to account for PCR duplicates and to cluster sequences with merging decisions based solely on distance except in cases of low coverage (<500 reads/barcode), for which the default cluster merging threshold was used [69]. Clusters with a size less than 4 or with high entropy (>0.75 quality score) were discarded. We estimated the relative abundance of barcodes using the number of unique reads supporting a cluster compared to total library size. Raw sequencing data is available through the SRA, BioProject ID PRJNA767552.

3.5.8 Detecting adaptive lineages in barcoded clonal populations

To detect spontaneous adaptive mutations in a barcoded clonal cell population that is evolved for over time, we used a Python-based pipeline (Li and Sherlock, in prep; https://github.com/FangfeiLi05/PyFitMut) based on a previously developed theoretical framework (Levy et al., 2015). The pipeline identifies adaptive lineages and infers their fitness effects and establishment time. In a barcoded population, a lineage refers to cells that share the same DNA barcode. For each lineage in the barcoded population, beneficial mutations continually occur at a total beneficial mutation rate Ub, with fitness effect s, which results in a certain spectrum of fitness effects of mutations μ(s). If a beneficial mutant survives random drift and becomes large enough to grow deterministically (exponentially), we say that the mutation carried by the mutant has established. Here, we use Wright fitness s, which is defined as average number of additional t offspring of a cell per generation, that is, $n(t) = n(0)\cdot(1 + s)$, with n(t) being the total number of cells at generation t (can be non-integers). Briefly, for each lineage, assuming that the lineage is adaptive (i.e., a lineage with a beneficial mutation occurred and established), then estimates of the fitness effect and establishment time of each lineage are made by random initialization, and the expected trajectory of each lineage is estimated and compared to the measured trajectory. Fitness effect and establishment time estimates are iteratively adjusted to better fit the observed data until an optimum is reached. At the same time, the expected trajectory of the lineage is also estimated assuming that the lineage is neutral. Finally, Bayesian inference is used to determine whether the lineage is adaptive or neutral. An accurate estimation of the mean fitness is necessary to detect mutations and quantify their fitness effects, but the mean fitness is a quantity that cannot be measured directly from the evolution. Rather, it needs to be inferred through other variables. Previously, the mean fitness was estimated by monitoring the decline of neutral lineages (Levy et al., 2015). However, this method fails when there is an insufficient number of neutral lineages as a result of low

sequencing read depth. Here, we instead estimate the mean mean fitness using an iterative method. Specifically, we first initialize the mean fitness of the population as zero at each sequencing time point, then we estimate the fitness effect and establishment time for adaptive mutations, then we recalculate the mean fitness with the optimized fitness and establishment time estimates, repeating the process for several iterations until the mean fitness converges. We established the improved the accuracy of the method using simulated data (Li and Sherlock, in prep).

3.6 Supplemental Material

Supplementary Files. Assessing inference method performance on single experiments. This is a zip folder containing the results of inference on single observations. Each file in the folder is named with the following naming convention: Model_Method_FlowType_SimulationBudget_InferenceSet_all.pdf. Each file contains 20 pages, each page corresponding to one of the 20 simulated synthetic observations. For NPE, each file corresponds to one training set (each observation was evaluated on a single amortized posterior was used) each page contains 6 panels, from top left to bottom right: a description of parameters used to generate the synthetic observation, a plot of the synthetic observation, the marginal posterior distribution for CNV formation rate, a plot of the posterior predictive check, the joint posterior distribution, and the marginal posterior distribution for CNV selection coefficient. For ABC-SMC, each page contains 8 panels from top left to bottom right: a description of parameters used to generate the synthetic observation, a plot of the synthetic observation, the effective sample size for each iteration of inference, epsilon values for each iteration of inference, the marginal posterior distribution for CNV formation rate for each iteration of inference, a plot of the posterior predictive check, the final joint posterior distribution, and the marginal posterior distribution for CNV selection coefficient for each iteration of inference. When the starting particle size = 100, the simulation budget was 10,000; when starting particle size = 1000, the simulation budget was 100,000. Supplementary files can be found at OSF: https://osf.io/e9d5x/

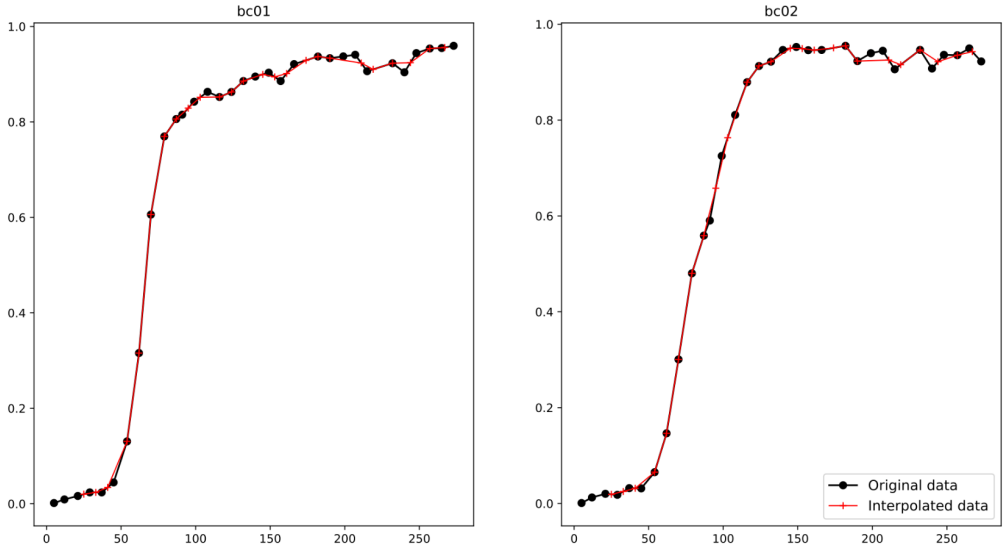

Figure 3.S1. Interpolation for bc01 and bc02. Populations gln01-gln09 and bc01-bc02 have different timepoints - the gln populations have 25 timepoints in total, whereas the bc populations have 32 timepoints in total. Of these, 12 of the timepoints are the same in both populations. To match the timepoints in the gln populations we interpolated from the two nearest timepoints in the bc populations (using pandas.DataFrame.interpolate('values')). This way, we can use the same data (same timepoints) for inference for all 11 populations so that we can use the same amortized NPE posterior to infer parameters for both gln populations and bc populations. Original bc data is shown as black dots, the matched data, with interpolated timepoints, is shown as red crosses.

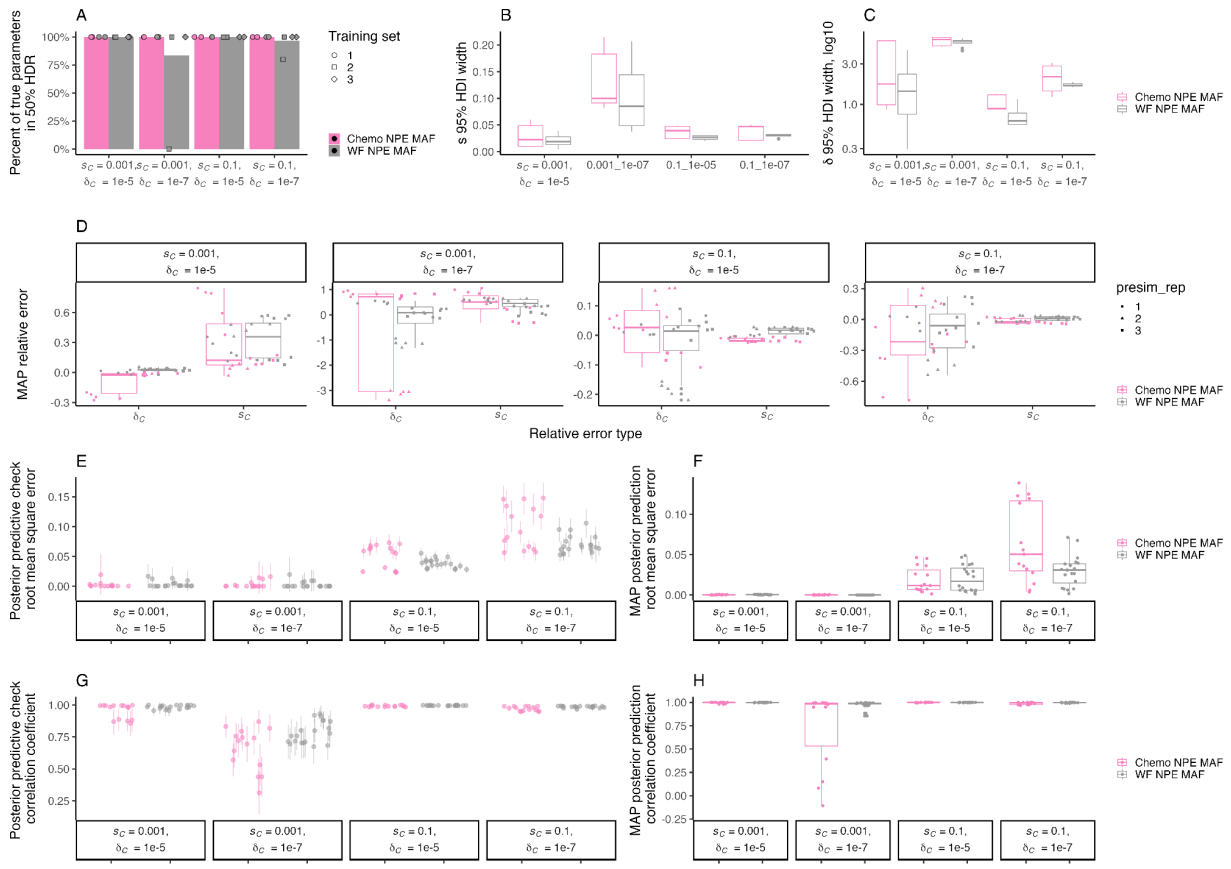

Figure 3.S2. Performance assessment of NPE with MAF using single simulated synthetic observations. These show the results of inference on five simulated synthetic observations generated using either the Wright-Fisher (WF) or chemostat (Chemo) model (and inference performed with the same model) per combination of fitness effect s_C and mutation rate δ_C. Here we show the results of performing one training set with NPE with MAF using 100,000 simulations for training and using the same amortized network to infer a posterior for each replicate synthetic observation. **A)** Percentage of true parameters within the 50% HDR. **B)** Distribution of widths of the fitness effect s_C 95% highest density interval (HDI) calculated as the difference between the 97.5 percentile and 2.5 percentile, for each inferred posterior distribution. **C)** Distribution of the number of orders of magnitude encompassed by the mutation rate δ 95% HDI, calculated as difference of the base 10 logarithms of the 97.5 percentile and 2.5 percentile, for each inferred posterior distribution. **D)** Log ratio MAP estimate as compared to true parameters for s_C and δ_C. Note that each panel has a different y axis. **E)** Mean and 95% confidence interval for RMSE of 50 posterior predictions as compared to the synthetic observation for which inference was performed. **F)** RMSE of posterior prediction generated with MAP parameters as compared to the synthetic observation for which inference was performed. **G)** Mean and 95% confidence interval for correlation coefficient of 50 posterior predictions compared to the synthetic observation for which inference was performed. **H)** Correlation coefficient of posterior prediction posterior prediction generated with MAP parameters compared to the synthetic observation for which inference was performed.

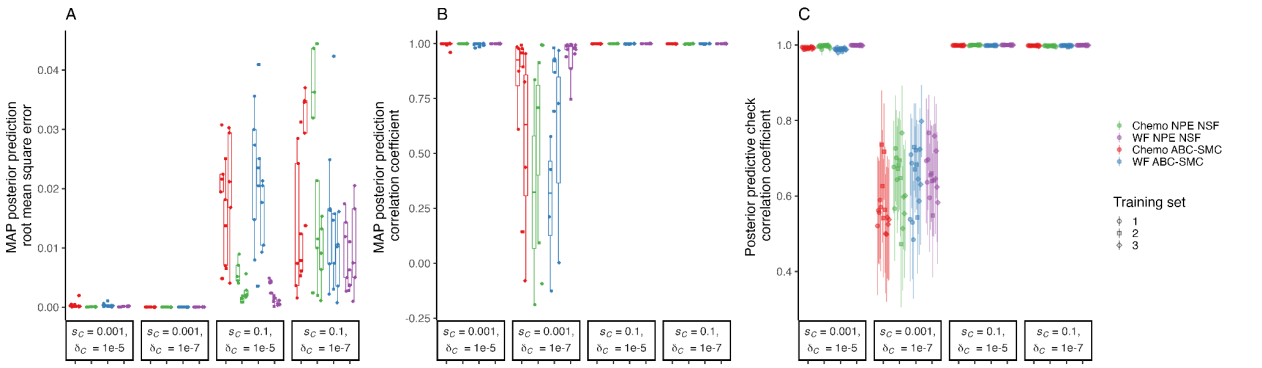

Figure 3.S3. NPE with the Wright-Fisher model performs as well or better than other combinations of model and method. Results of inference on five simulated single synthetic observations generated using either the Wright-Fisher (WF) or chemostat (Chemo) model (and inference performed with the same model) per combination of fitness effect s_C and mutation rate δ_C. Here we show the results of performing training with NPE with NSF using 100,000 simulations for training and using the same amortized network to infer a posterior for each replicate synthetic observation, or ABC-SMC when the training budget was 10,000. **A)** RMSE (lower is better) of posterior prediction generated with MAP parameters as compared to the synthetic observation on which inference was performed. **B)** Correlation coefficient (higher is better) of posterior prediction generated with MAP parameters compared to the synthetic observation on which inference was performed. **C)** Mean and 95% confidence interval for correlation coefficient (higher is better) of 50 posterior predictions (sampled from the posterior distribution) compared to the synthetic observation on which inference was performed.

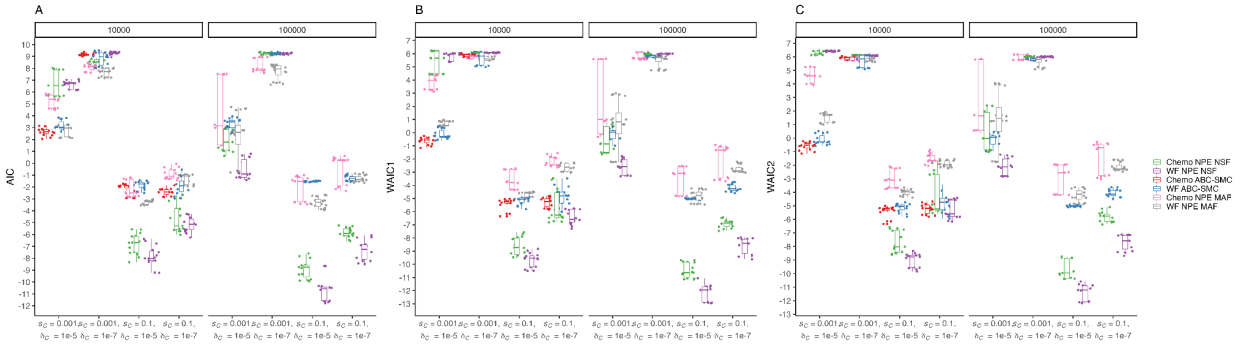

Figure 3.S4. NPE and WF have the lowest information criteria. WAIC and AIC (lower is better) of models fitted on single synthetic observations using either the Wright-Fisher (WF) or chemostat (Chemo) model and either ABC-SMC or NPE for different combinations of fitness effect s_C and mutation rate δ_C with simulation budgets of 10,000 or 100,000 simulations per inference procedure (facets). We were unable to complete ABC-SMC with the chemostat model (red) when the training budget was 100,000 within a reasonable time frame.

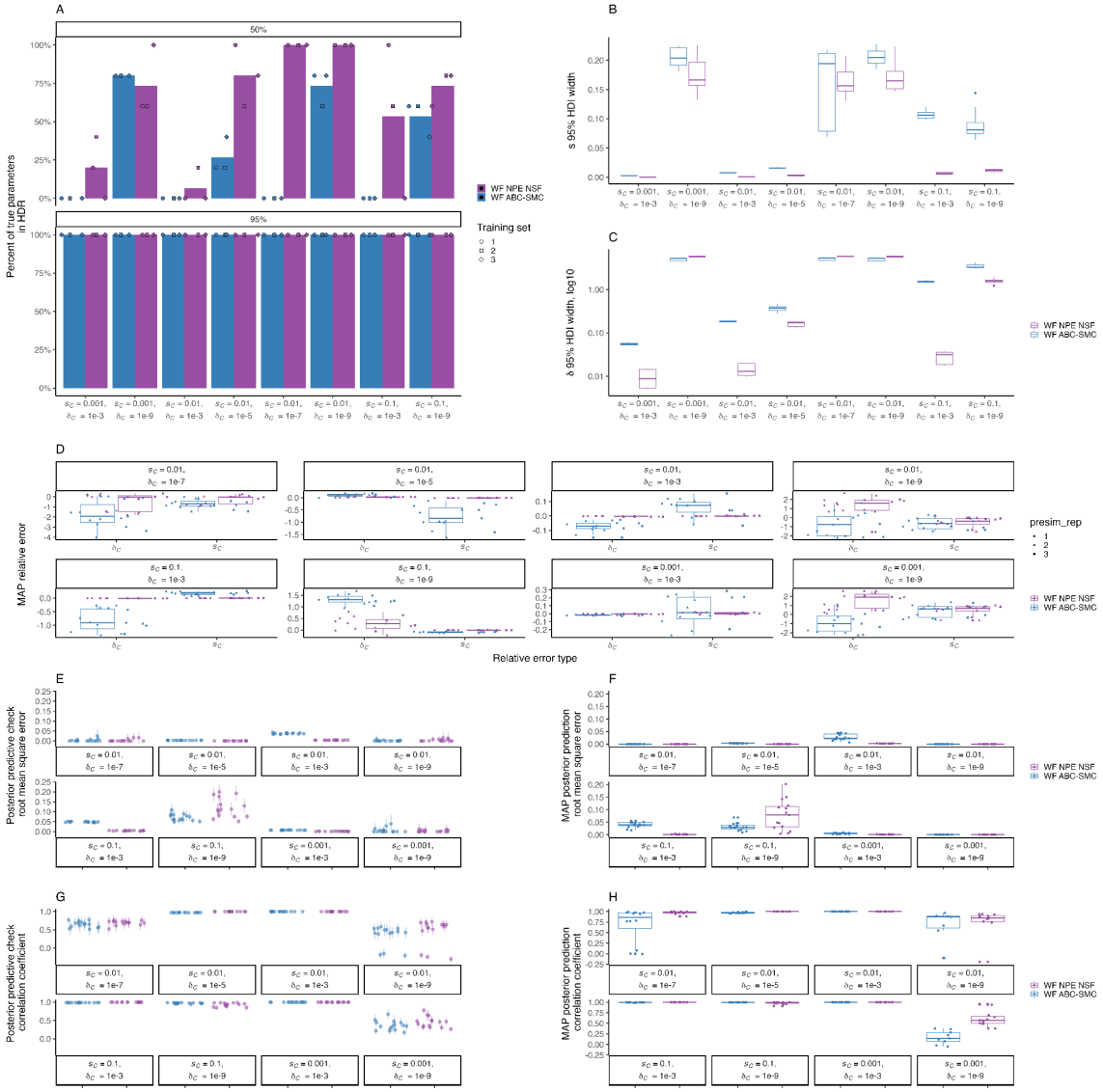

Figure 3.S5. NPE performs similar to or better than ABC-SMC for eight additional parameter combinations. The figure shows the results of inference on five simulated synthetic observations using the Wright-Fisher (WF) model per combination of fitness effect s_C and mutation rate δ_C. Simulations and inference were performed using the same model. For NPE, each training set corresponds to an independently amortized posterior distribution trained on a different set of 100,000 simulations, with which each synthetic observation was evaluated to produce a separate posterior distribution. For ABC-SMC, each training set corresponds to independent inference procedures on each observation with a maximum of 100,000 total simulations accepted for each inference procedure and a stopping criteria of 10 iterations or ε <= 0.002. **A**) The percent of true parameters within the 50% or 95% HDR of the inferred posterior distribution. Bar height shows the average of three training sets. **B-C**) Distribution of widths of 95% highest density interval (HDI) of the posterior distribution of the fitness effect s_C (**B**) and CNV mutation rate δ_C (**C**), calculated as the difference between the 97.5 percentile and 2.5 percentile, for each inferred posterior distribution. **D**) Log-ratio (relative error) of MAP estimate to true parameter. Note the different y-axis ranges. **E**) Mean and 95% confidence interval for RMSE of 50 posterior predictions as compared to the synthetic observation for which inference was performed. **F**) RMSE of posterior prediction generated with MAP parameters as compared to the synthetic observation for which inference was performed. **G**) Mean and 95% confidence interval for correlation coefficient of 50 posterior predictions compared to the synthetic observation. **H**) Correlation coefficient of posterior prediction posterior prediction generated with MAP parameters compared to the synthetic observation for which inference was performed.

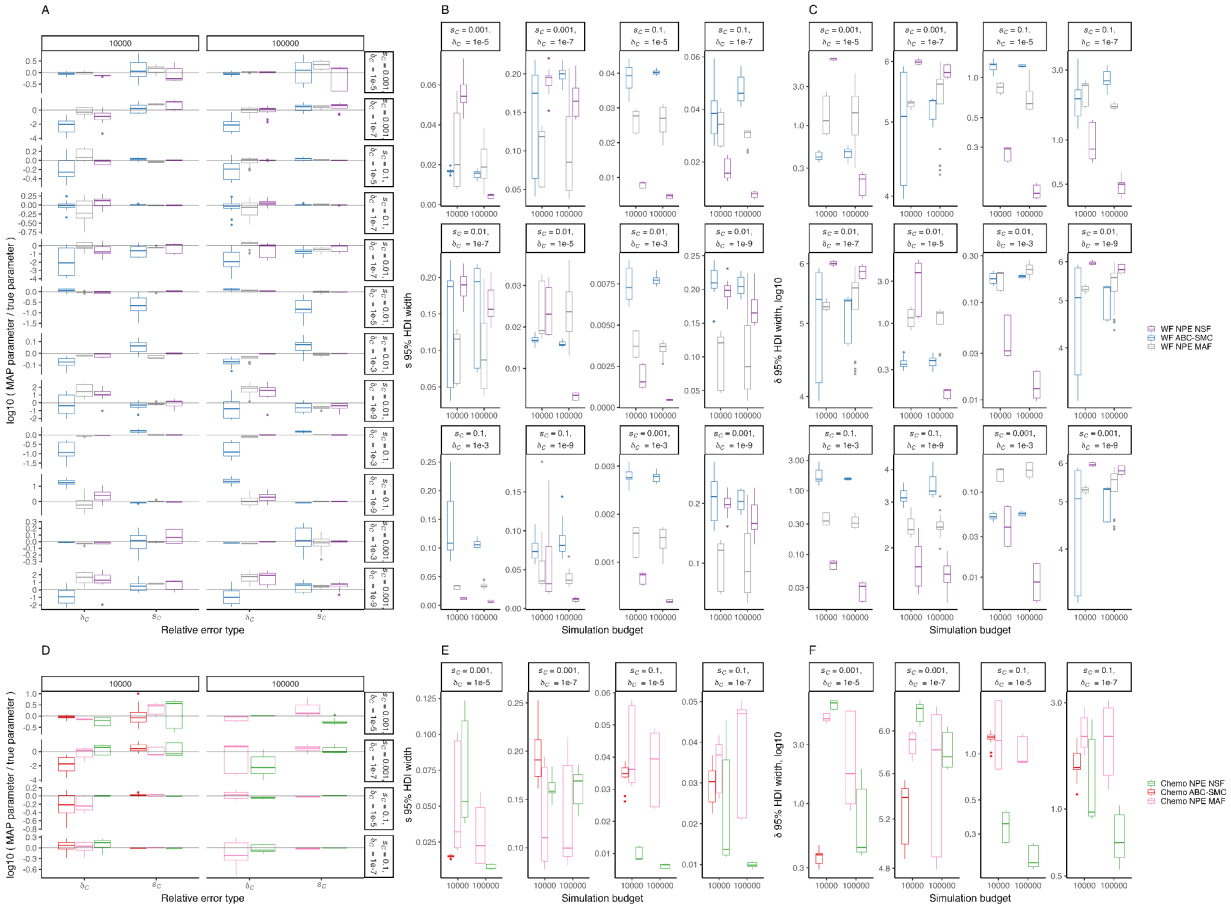

Figure 3.S6. Effect of simulation budget on relative error of MAP estimate and width of HDIs. For NPE, amortized posteriors were estimated using either 10,000 or 100,000 simulations, with which each synthetic observation was evaluated to produce a separate posterior distribution. For ABC-SMC, a posterior was independently inferred for each observation with a maximum of 10,000 or 100,000 total simulations accepted and a stopping criteria of 10 iterations or ε <= 0.002, whichever occurs first. The grey lines in (**A, D**) indicates a relative error of zero (i.e., no difference between MAP parameters and true parameters). (**D, E, F**) We were unable to complete ABC-SMC with the chemostat model (red) when the training budget was 100,000 within a reasonable time frame.

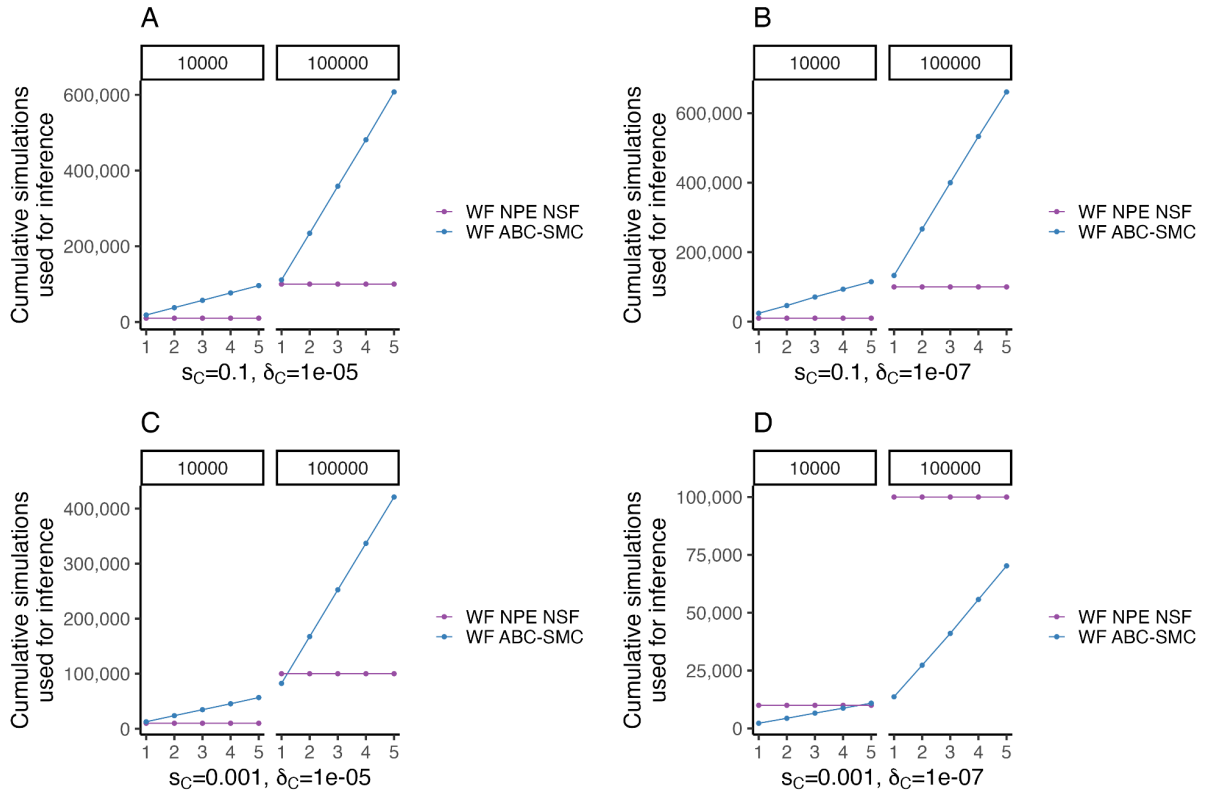

Figure 3.S7. The cumulative number of simulations needed to estimate posterior distributions for multiple observations. The x axis shows the number of replicate simulated synthetic observations for a combination of parameters and the y axis shows the cumulative number of simulations needed to infer posteriors for an increasing number of observations *(*see *Overview of inference strategies* for more details), for observations with different combinations of CNV selection coefficient s_C and CNV formation rate δ_C (**A-D**). Each facet represents a total simulation budget for NPE, or the maximum number of accepted simulations for ABC-SMC. Since NPE uses amortization, a single amortized network is trained with 10,000 or 100,000 simulations, and that network is then used to infer posteriors for each observation (note that a single amortized network was used to infer posteriors for all parameter combinations.) For ABC-SMC, each observation requires a separate inference procedure to be performed individually, and not all generated simulations are accepted for posterior estimation; therefore, the number of simulations used for a single observation may be more than the acceptance threshold, and the number of simulations needed increases with the number of observations for which a posterior is inferred.

Table 3.S1. Wall time to run one simulation. Running time for a single Wright-Fisher simulation or a single chemostat simulation for each of the following parameter combinations on a 2019 MacBook Pro operating Mac OS Catalina 10.15.7 with a 2.6 GHz 6-Core Intel Core i7 processor.

Model	δ_c	s_c	**Wall time (seconds)**
Wright-Fisher	0.1	10^{-5}	0.0136
Chemostat	0.1	10^{-5}	10.8608
Wright-Fisher	0.1	10^{-7}	0.0099
Chemostat	0.1	10^{-7}	11.2390
Wright-Fisher	0.001	10^{-5}	0.0108
Chemostat	0.001	10^{-5}	11.3547
Wright-Fisher	0.001	10^{-7}	0.0086
Chemostat	0.001	10^{-7}	10.6964

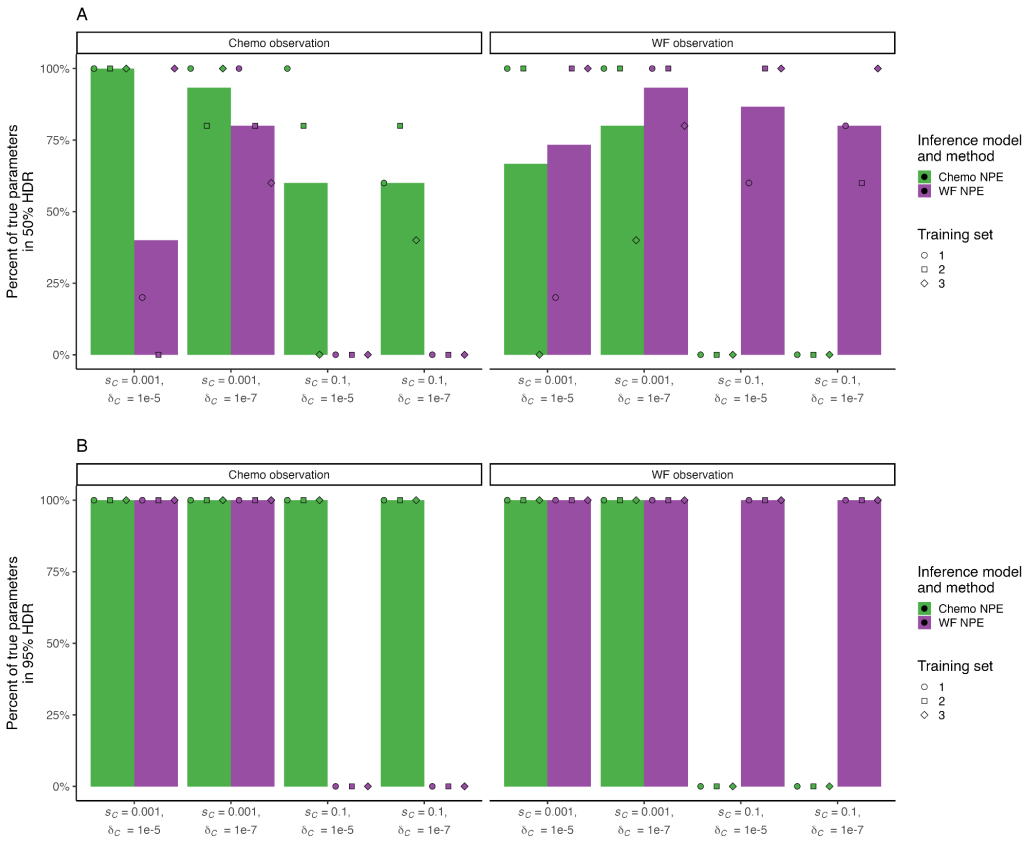

Figure 3.S8. Results of inference on five simulated synthetic observations generated using either the Wright-Fisher (WF) or chemostat (Chemo) model per combination of fitness effect s_C and mutation rate δ_C. We performed inference on each synthetic observation using both models. For NPE, each training set corresponds to an independent amortized posterior trained with 100,000 simulations, with which each synthetic observation was evaluated. **A**) Percentage of true parameters within the 50% HDR. The bar height shows the average of three training sets**. B**) Percentage of true parameters within the 95% HDR. The bar height shows the average of three training sets.

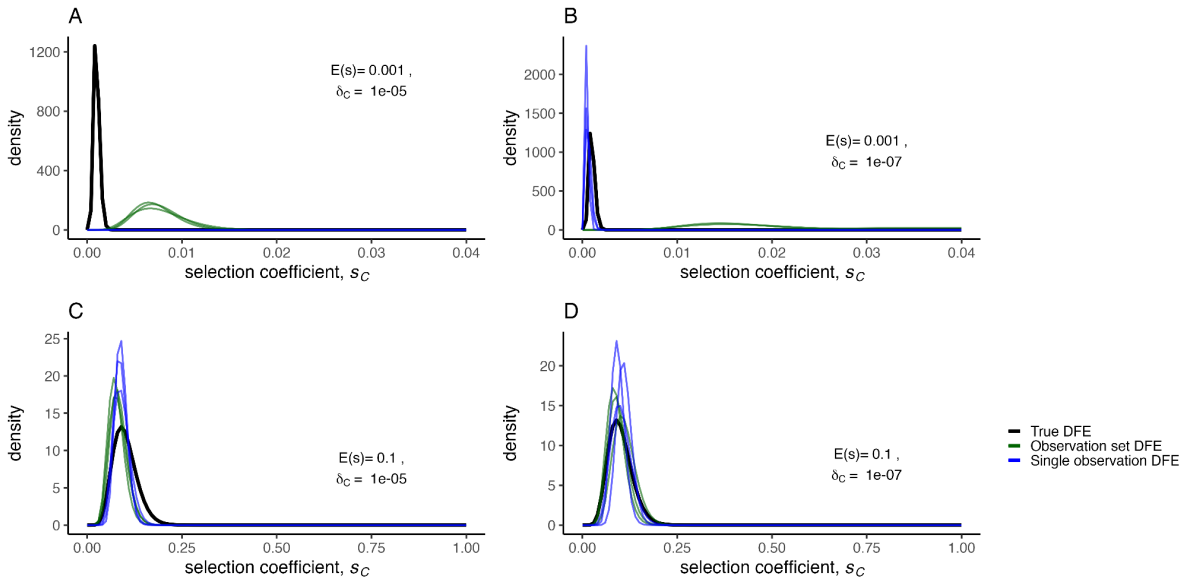

Figure 3.S9. A set of eleven simulated synthetic observations was generated from a Wright-Fisher model with CNV selection coefficients sampled from an Gamma distribution where $\alpha = 10$ of fitness effects (DFE) (black curve) . The MAP DFEs (blue curves) were directly inferred using three different subsets of eight out of eleven synthetic observations. We also inferred the selection coefficient for each observation in the set of eleven individually, and fit Gamma distributions to sets of eight inferred selection coefficients (green curves). All inferences were performed with NPE using the same amortized network to infer a posterior for each set of eight synthetic observations or each single observation.

Table 3.S2. Kullback–Leibler divergence for Gamma distributions fit from single inferred selection coefficients versus the true underlying DFE, or for directly inferred Gamma distributions versus the true underlying DFE.

KL divergence: Gamma fit from single inferred s_C	KL divergence: α and β directly inferred from set of observations	Observation set name	True α	True β	True δ_C
1359.8	572810.04	WF_shape1_scale0.001_mut5	1.0	0.001	1E-05
856459.8	200872.69	WF_shape1_scale0.001_mut5	1.0	0.001	1E-05
1338.0	644533.9	WF_shape1_scale0.001_mut5	1.0	0.001	1E-05
38967.7	664134.98	WF_shape1_scale0.001_mut7	1.0	0.001	1E-07
6522383.1	854560.75	WF_shape1_scale0.001_mut7	1.0	0.001	1E-07
38652.8	372597.74	WF_shape1_scale0.001_mut7	1.0	0.001	1E-07
233.8	1200.59	WF_shape1_scale0.1_mut5	1.0	0.1	1E-05
230.7	220.63	WF_shape1_scale0.1_mut5	1.0	0.1	1E-05
233.0	1161.7	WF_shape1_scale0.1_mut5	1.0	0.1	1E-05
9.4	5151.41	WF_shape1_scale0.1_mut7	1.0	0.1	1E-07
6.6	1255.33	WF_shape1_scale0.1_mut7	1.0	0.1	1E-07
21.7	1130.75	WF_shape1_scale0.1_mut7	1.0	0.1	1E-07
2079309.0	381627.51	WF_shape10_scale0.0001_mut5	10.0	0.0001	1E-05
1719636.6	562084.78	WF_shape10_scale0.0001_mut5	10.0	0.0001	1E-05
2125542.6	543314.78	WF_shape10_scale0.0001_mut5	10.0	0.0001	1E-05
32299.9	1124713.46	WF_shape10_scale0.0001_mut7	10.0	0.0001	1E-07
133767.3	818178.69	WF_shape10_scale0.0001_mut7	10.0	0.0001	1E-07
51454.3	993824.56	WF_shape10_scale0.0001_mut7	10.0	0.0001	1E-07
336.3	123.01	WF_shape10_scale0.01_mut5	10.0	0.01	1E-05
231.7	274.56	WF_shape10_scale0.01_mut5	10.0	0.01	1E-05
74.1	134.88	WF_shape10_scale0.01_mut5	10.0	0.01	1E-05
334.6	49.24	WF_shape10_scale0.01_mut7	10.0	0.01	1E-07
228.3	25.18	WF_shape10_scale0.01_mut7	10.0	0.01	1E-07
22.0	66.22	WF_shape10_scale0.01_mut7	10.0	0.01	1E-07

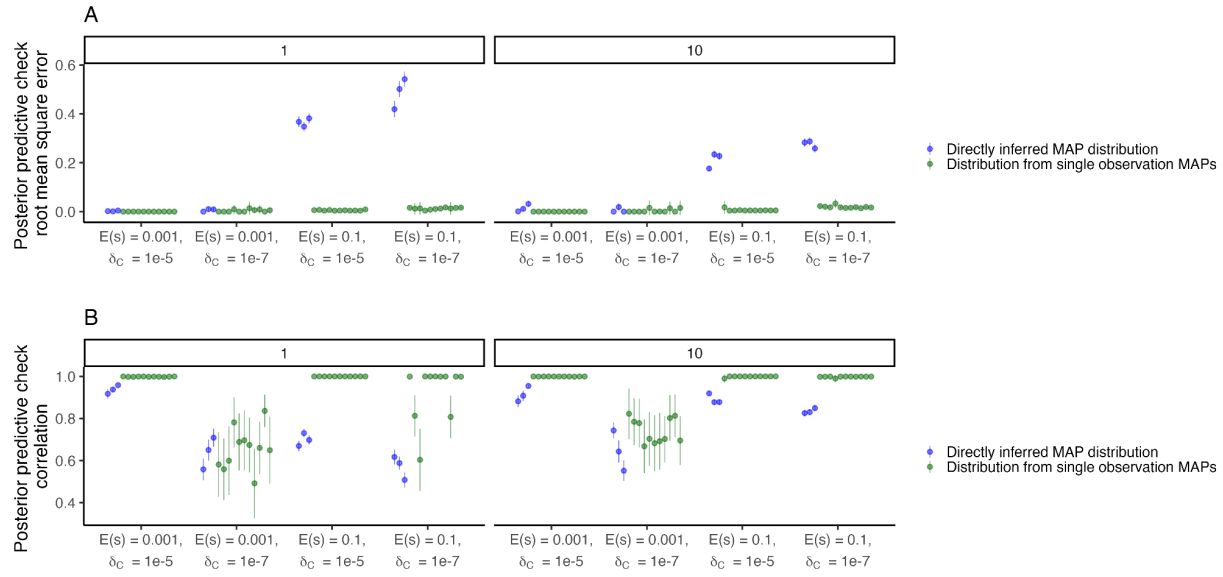

Figure 3.S10. Out-of-sample posterior predictive accuracy using root mean square error (**A**) or correlation (**B**) using three held out observations when α and β are directly inferred from the other eight observations, for $\alpha = 1$ or $\alpha = 10$ (facets).

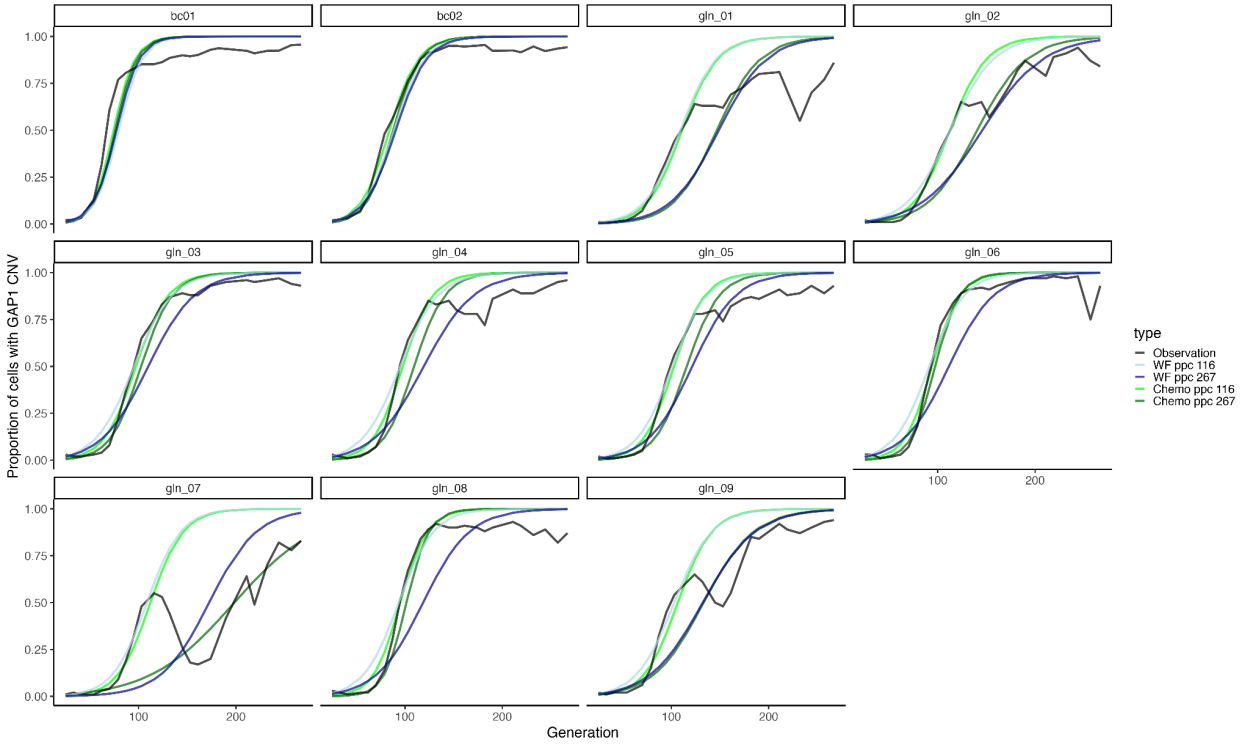

Figure 3.S11. Proportion of the population with a *GAP1* CNV in the experimental observations (black) and in posterior predictions using the MAP estimate shown in panels A and B with either the Wright-Fisher (WF) or chemostat (Chemo) model. Inference was performed with all data up to generation 267 (WF ppc 267, Chemo ppc 267), or excluding data after generation 116 (WF ppc 116, Chemo ppc 116). Mutation rate and fitness effect of other beneficial mutations set to 10^{-5} and 10^{-3}, respectively.

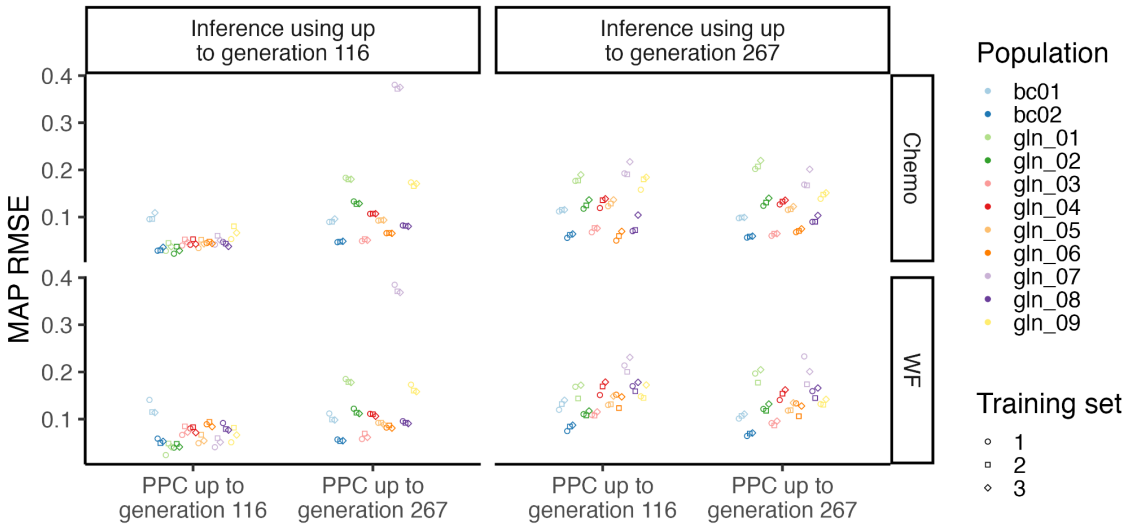

Figure 3.S12. MAP predictions have lower error when inference is performed using only up to generation 116, and are most accurate for the first 116 generations. MAP posterior prediction root mean square error (RMSE) when inference was performed excluding data after generation 116 (left) or using all data up to generation 267 (right). RMSE was calculated using either the first 116 generations, or using up to generation 267 (x-axis).

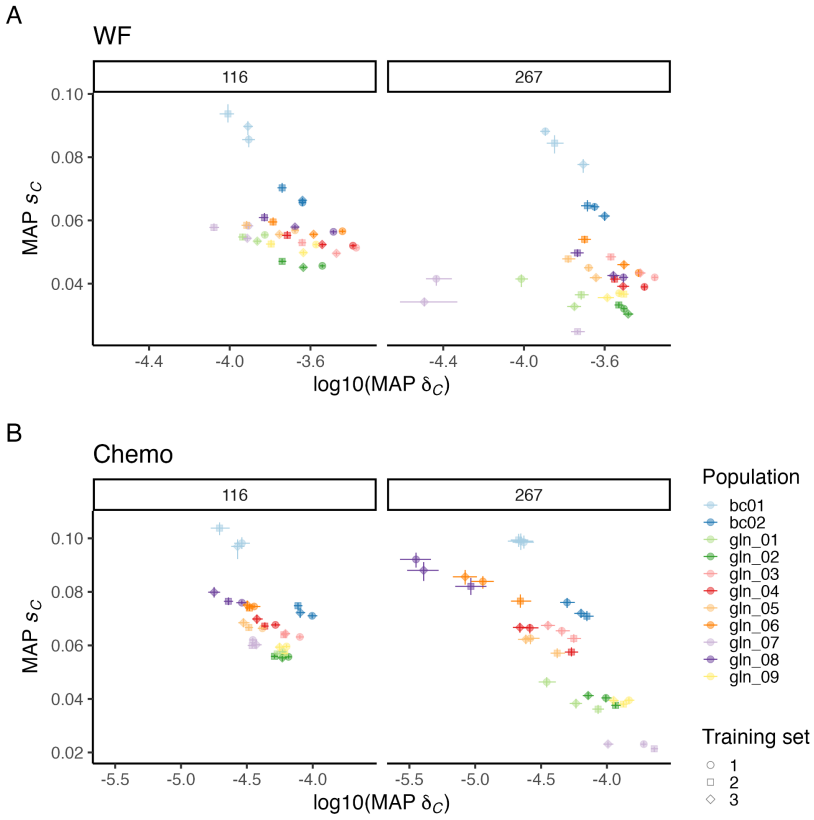

Figure 3.S13. The inferred MAP estimate and 95% highest density intervals (HDI) for fitness effect s_C and formation rate δ_C, using the **(A)** Wright-Fisher (WF) or **(B)** chemostat (Chemo) model and NPE for each experimental population from Lauer et al. (2018). Inference was either performed with data up to generation 116 or with all data, up to generation 267 (facets). Each training set corresponds to three independent amortized posterior distributions estimated with 100,000 simulations.

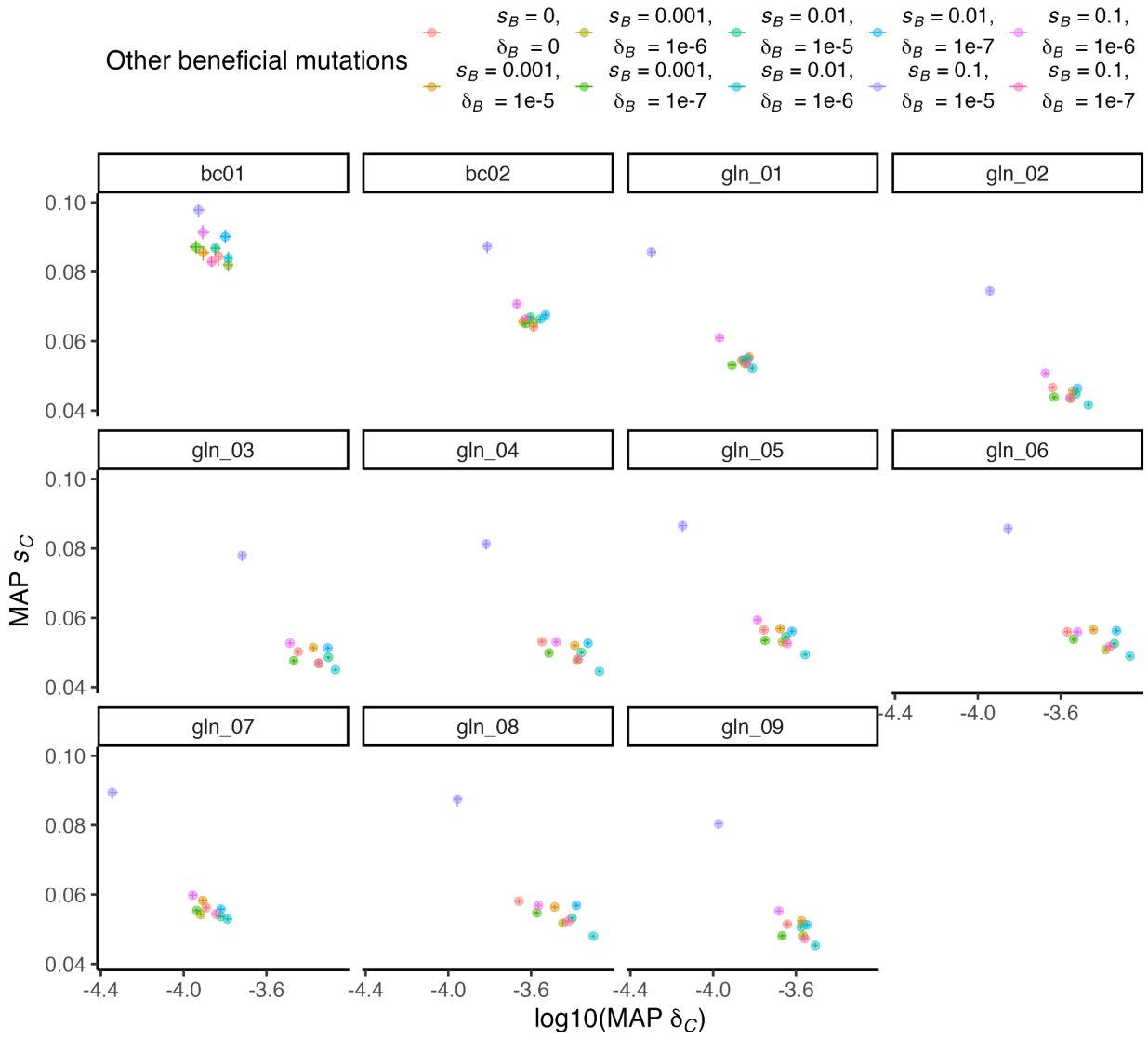

Figure 3.S14. Sensitivity analysis. *GAP1* CNV formation rate and selection coefficient inferred using NPE with the Wright-Fisher model does not change considerably when other beneficial mutations have different selection coefficients s_B and formation rates δ_B, except when both s_B and δ_B are high (purple).

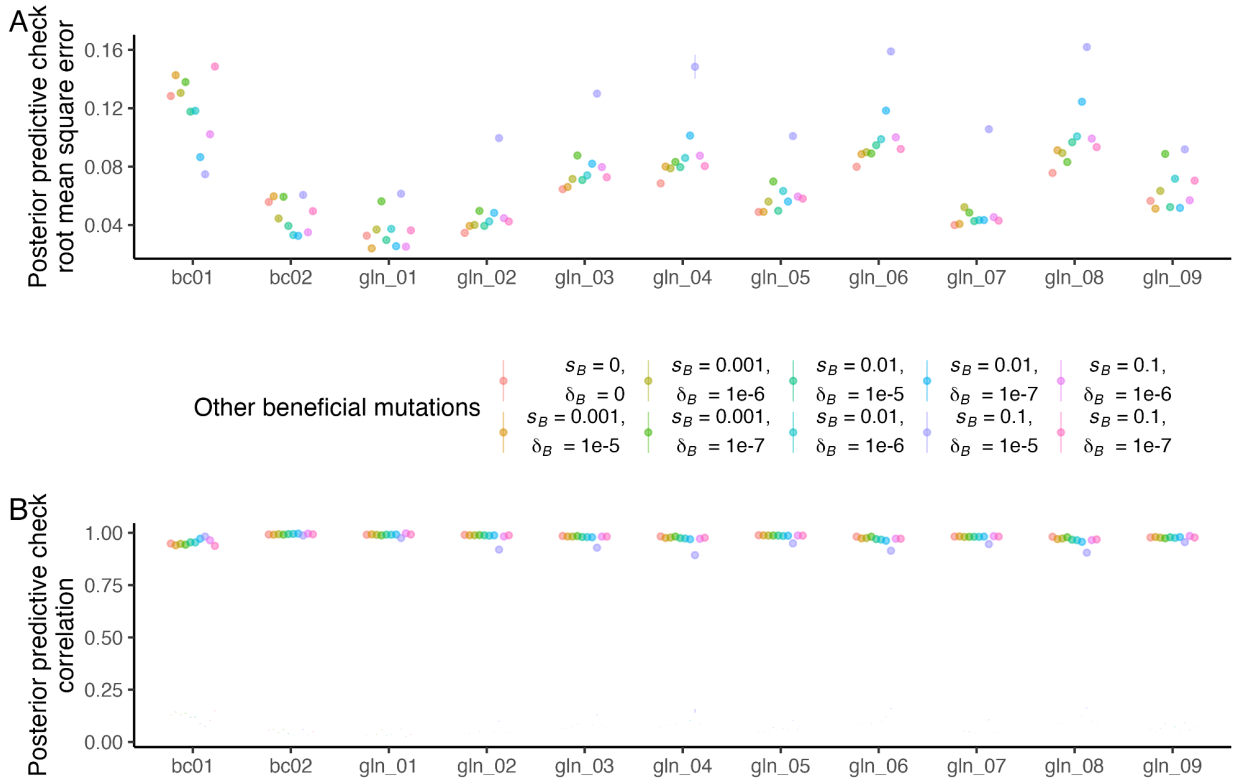

Figure 3.S15. Mean and 95% confidence interval for RMSE (**A**) and correlation (**B**) of 50 posterior predictions compared to empirical observations up to generation 116, using posterior distributions inferred when other beneficial mutations have different selection coefficients s_B and formation rates δ_B.

Chapter 4: Effects of diverse CNV structures on genetic interactions and mRNA expression

This chapter is based on "Effects of diverse CNV structures on genetic interactions and mRNA expression" by **Grace Avecilla**, Julia Matthews, Elodie Caudal, Joseph Schacherer, and David Gresham, which is in preparation and will be submitted to a peer-reviewed journal for publication.

I performed transposon mutagenesis experiments with assistance from Elodie Caudal, assisted Julia Matthews in performing RNAseq, and performed all other new experiments. I performed all analysis, generated all figures and tables, and wrote the manuscript text with editing from David Gresham.

4.1 Abstract

Copy number variants (CNVs), comprising duplications and deletions of existing genomic content, contribute to rapid evolutionary adaptation, but can also confer deleterious effects, and cause disease. Whereas the effects of amplifying individual genes or whole chromosomes (i.e., aneuploidy) have been studied extensively, much less is known about the genetic and functional effects of CNVs of varying sizes and structures. Here, we investigated seven *Saccharomyces cerevisiae* strains isolated from evolution experiments in glutamine-limited chemostats that have CNVs of variable structures all of which contain multiple copies of the gene *GAP1*. We find that despite being beneficial in glutamine-limited chemostats, CNVs result in decreased fitness compared with the euploid ancestor in rich media. We used transposon mutagenesis to investigate mutational tolerance and genetic interactions with CNVs. We find that CNVs confer mutational tolerance to essential genes and result in new genetic interactions. Some, but not all CNV strains have more insertions than the euploid in genes

related to translation, and fewer insertions in genes related to mitochondrial function. We profiled the transcriptome of each CNV, and find that although amplified genes have increased expression, dosage compensation may be occurring in some strains. We find that CNV strains do not exhibit previously described transcriptional signatures of aneuploidy, the environmental stress response or common aneuploidy gene-expression. Instead, CNV strains tend to downregulate genes involved in cellular respiration, nucleoside biosynthetic processes, and small molecule metabolism, and upregulate genes involved in transposition, nucleic acid metabolic processes, and siderophore transport, though to different degrees in each strain. Our study reveals the extent to which local and distal mutational tolerance is modulated by CNVs with implications for genome evolution and diseases with common CNVs, such as cancer.

4.2 Introduction

Evolution occurs through changes to an organism's genome and selection on the functional effects of these changes. Genomes can evolve in many ways, including through single nucleotide changes, structural rearrangements, and deletion and duplication of segments of DNA. Duplication of segments of DNA, a type of copy number variation (CNV), is a major force in rapid adaptive evolution and genome evolution. In the short term, amplification of genes can result in increased expression which provides a selective advantage facilitating rapid adaptive evolution, (Kondrashov 2012; Myhre et al. 2013). In the long term, amplification of genes may relax selective constraints, allowing accumulation of mutations on the additional gene copies and gene evolution through subfunctionalization or neofunctionalization (Freeling, Scanlon, and Fowler 2015; Innan and Kondrashov 2010; Ohno 1970). Rapid adaptation through gene amplification is prevalent throughout the tree of life. In particular, gene amplification has been shown to mediate rapid adaptation to a variety of selective pressures from nutrient limitation to antibiotics in natural and experimental populations of microbes (Lauer et al. 2018;

A. M. Selmecki et al. 2009; Todd and Selmecki 2020; Hong and Gresham 2014b; Dhami, Hartwig, and Fukami 2016; Nair et al. 2008; Pränting and Andersson 2011; Paulander, Andersson, and Maisnier-Patin 2010; Gresham et al. 2008). Gene amplification is also common in cancers, and can promote tumorigenesis (Ben-David and Amon 2020). Oncogene amplification confers enhanced proliferation properties to cells driving their aberrant growth. Thus, understanding the evolutionary, genetic, and functional consequences of CNVs is of central importance. The budding yeast, *Saccharomyces cerevisiae*, has been extensively used as a model to study the effects of amplifying genes on cellular state, fitness, and genetics.

Copy number variants can range from small duplications and deletions to the gain or loss of whole chromosomes, known as aneuploidy. Previous studies have investigated the effect of amplifying individual genes primarily using plasmid libraries with native or inducible promoters (Moriya 2015). These studies have found that in commonly used laboratory strains around 10-20% of genes are deleterious when overexpressed and 0-5% are beneficial (Ascencio et al. 2021; Sopko et al. 2006; Arita et al. 2021; Douglas et al. 2012). These effects are dependent on genetic background as a recent study found significant variation in the number of genes that are deleterious when overexpressed in 15 genetically diverse yeast lineages (Robinson et al. 2021). Fitness effects of gene amplification tend to be dependent on both the particular gene amplified and the environmental context, though most amplified genes are neutral regardless of environment (Ascencio et al. 2021; Payen et al. 2016). Amongst these studies, there is conflicting evidence for the Dosage Balance hypothesis hypothesis, which predicts that genes involved in complexes or with many interactions are more likely to be deleterious when overexpressed due to stoichiometric imbalances (Birchler and Veitia 2012; Rice and McLysaght 2017b); some studies find that genes that are deleterious when overexpressed are enriched for protein complexes and protein interactions (Robinson et al. 2021; Makanae et al. 2013), whereas others studies do not (Ascencio et al. 2021; Sopko et al. 2006; Arita et al. 2021). Single

gene overexpression libraries have also been used to identify synthetic dosage lethal interactions with gene deletions (Douglas et al. 2012; Sopko et al. 2006; C. Liu et al. 2009) in which an overexpressed gene is deleterious in the background of a gene knock out.

Aneuploidy is common in strains of yeast that are not laboratory adapted (Peter et al. 2018; Hose et al. 2015; Gallone et al. 2016; Y. O. Zhu, Sherlock, and Petrov 2016; Scopel et al. 2021), and these aneuploids grow similarly to their euploid counterparts (Hose et al. 2015; Gasch et al. 2016). Aneuploids also frequently arise in evolution experiments and are associated with increased fitness (Sunshine et al. 2015; Lauer et al. 2018; Hong and Gresham 2014b; Rancati et al. 2008; Yona et al. 2012; Gresham et al. 2008). However, these adaptive aneuploids may exhibit antagonistic pleiotropy such that they are deleterious in other environments (Sunshine et al. 2015; Linder et al. 2017). Seminal studies of one laboratory strain, W303, found that aneuploids grow more slowly than euploids, regardless of karyotype (Torres et al. 2007; Sheltzer et al. 2012; Beach et al. 2017), exhibit a transcriptional signature characteristic of the yeast environmental stress response (ESR) (Torres et al. 2007; Sheltzer et al. 2012), and result in a variety of cellular stresses, including proteotoxic, metabolic, and mitotic stress (reviewed in (J. Zhu et al. 2018)). These effects may also be background dependent as a recent study mapped differences in aneuploidy tolerance between W303 and wild yeast strains to a single gene, *SSD1*, which has a truncating mutation in W303 (Hose et al. 2020). *SSD1* is a RNA-binding translational regulator, whose targets include mitochondrial transcripts. Loss of *SSD1* function results in defects in mitochondrial function and proteostasis that enhance sensitivity to aneuploidy (Hose et al. 2020). In addition to observing different fitness effects, studies of aneuploids in different genetic backgrounds have found differing results in transcriptomic dosage compensation, ESR , and proteotoxic stress (Muenzner et al., n.d.; Torres et al. 2007; Pavelka et al. 2010; Larrimore et al. 2020; Dephoure et al. 2014; Hose et al. 2015; Gasch et al. 2016; J. Zhu et al. 2018).

Whereas numerous studies have investigated the effects of single gene amplifications and aneuploidy, little is known about the effects of CNVs that vary in size and can have complex structures. One study sought to study the fitness effects of a diverse set of synthetic amplicons extending from the telomere and ranging in size from 0.4 to 1,000 kb across the genome in diploid yeast and measured their fitness in three conditions (Sunshine et al. 2015). Through comparison to single-gene amplifications (Payen et al. 2016) they found that the distribution of fitness effects for telomeric amplicons was broader than that of single gene amplifications. Notably, they also found that, of the telomere amplified regions that affected fitness, 94% had condition-dependent effects. However, much is still unknown if there are common fitness effects, genetic interactions, or transcriptomic states associated with CNVs more generally.

In this study, we investigated seven strains containing CNVs with variable structures that all contain the gene *GAP1*, that were previously isolated from evolution experiments in glutamine-limited chemostats (Lauer et al. 2018). We found that despite having fitness greater than or equal to the ancestral euploid in glutamine-limited chemostats, most CNV-containing lineages have fitness defects in rich media with galactose as a carbon source. We used transposon mutagenesis to investigate genetic interactions with CNVs, and found both common and strain specific interactions. We investigated how CNVs alter the transcriptome, and found that while amplified genes do have increased mRNA expression, some strains appear to exhibit dosage compensation. We did not observe previously described transcriptional signatures of aneuploidy in CNV strains; instead, we find that CNV-containing strains tend to have decreased expression of genes involved in respiration, nucleoside biosynthetic processes, and small molecule metabolism, and increased expression of genes involved in transposition, nucleic acid metabolic processes, and siderophore transport. Taken together, our experiments suggest there are both common and strain specific interactions and transcriptional responses that affect fitness in yeast with CNVs.

4.3 Results

4.3.1 *GAP1* CNVs confer variable fitness effects

Previously, we performed experimental evolution using the yeast *Saccharomyces cerevisiae* in glutamine-limited chemostats for approximately 270 generations (Lauer et al. 2018). The yeast strain (a haploid derivative of S288c) used to inoculate the evolution experiments contained a fluorescent CNV reporter adjacent to the general amino acid permease gene, *GAP1.* We isolated clones from the evolution experiment with *GAP1* CNVs on the basis of increased fluorescence. Using whole genome sequencing and pulsed-field gel electrophoresis, we defined the structure of the *GAP1* CNVs (Lauer et al. 2018). A subset of representative CNV strains, as well as the euploid *GAP1* CNV reporter ancestral strain, were used in this study (**Figure 4.1A; Table 4.S1**). The CNV strains range in *GAP1* copy number from two (aneu), to three (trip1, trip2, trip3, trip4, iso), to four (quad) copies; in the number of amplified genes ranging from 18 to 334; and in the total amount of amplified DNA from ~103,000 to ~670,000 additional nucleotides. The CNV strains have a variety of structures, including an aneuploid (aneu), inverse triplications characteristic of origin dependent inverse triplication (ODIRA; trip1, trip2, trip3, trip4), four copies with inversions (quad), and an isochromosome (iso) comprising a centromere and mirror image of the short arm of chromosome XI. In addition to CNVs, each strain contains a small number of unique nucleotide variants compared to the ancestor (**Table 4.S1**).

All CNV strains have fitness greater than or equal to the ancestral euploid strain in the glutamine-limited environment in which they evolved (**Figure 4.1B**). However, the CNV strains grow slower than the euploid strain in a different environment: yeast-peptone-galactose (YPGal) batch culture (**Figure 1C**). The fitness benefit in the environment in which they evolved and the fitness deficit in the alternative environment differ between strains. Fitness benefits and costs do

not correlate with the number of additional bases or the number of open reading frames in the CNV region (**Figure 4.S1**).

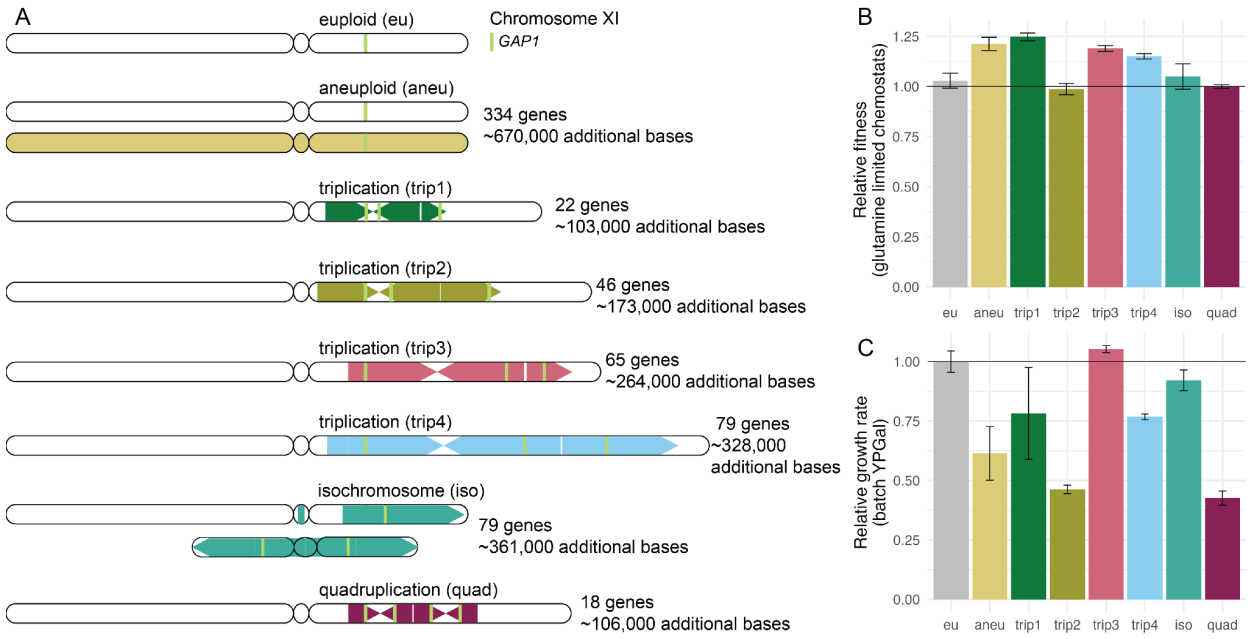

Figure 4.1. Strains with *GAP1* CNVs differ in structure and fitness. A) We previously evolved a euploid *S. cerevisiae* strain in glutamine-limited chemostats and isolated seven strains that have CNVs on Chromosome XI that include *GAP1*. The hypothesized structure (Lauer et al. 2018) of each *GAP1* CNV is diagrammed; the amplified region is shown as a colored block with arrows. Arrows pointing right represent copies that maintain their original orientation, whereas arrows pointing left represent copies that are inverted. The number of genes amplified and the approximate number of additional bases (quantified as the product of the copy number and size of the amplified CNV region) are annotated. **B)** The fitness of evolved strains containing *GAP1* CNVs was determined by pairwise competition experiments with a nonfluorescent reference strain in glutamine-limited chemostats. Error bars are 95% confidence intervals for the slope of the linear regression. **C)** Average and standard deviation (error bars) growth rate relative to the ancestral, euploid strain in YPGal batch culture. Horizontal black lines in **B** and **C** denote the ancestral euploid fitness.

4.3.2 Transposon mutagenesis reveals tolerance to mutation

We sought to investigate the genetic impact of CNVs in a high-throughput manner. Previous studies using transposon mutagenesis in bacteria and yeast have shown that transposon insertion density reflects tolerance to mutation and the is an efficient means of identifying genomic regions essential for cell survival in a specific environment or genetic background (Michel et al. 2017; Guo et al. 2013; Grech et al. 2019; Segal et al. 2018; Gale et al. 2020; Levitan et al. 2020). We generated *Hermes* insertion libraries in each CNV strain and in

two independent replicates of the ancestral euploid strain using modifications of published methods (Gangadharan et al. 2010; Caudal et al. 2021) (**Figure 4.2A**; **Methods**). Briefly, *Hermes* transposition was induced in YP Galactose (YPGal) media using batch cultures undergoing serial transfer, and transposition events were selected using an antibiotic marker. Insertion sites were identified by targeted PCR, followed by library preparation and deep sequencing (**Methods**). Unique insertion sites and the number of reads per insertion site were identified using a custom bioinformatic pipeline (**Methods**). As our sequencing and analysis pipeline cannot differentiate sequence reads that result from unique priming events from PCR duplicates we quantified the number of unique insertion sites per gene, unless otherwise noted. The nine libraries exhibited variation in the number of unique insertion sites which scaled with the total number of reads sequenced (**Figure 4.S2; Table 4.S2**). To normalize for differences in sequencing depth, we determined the number of insertions per million reads for analyses (**Methods**).

We compared our transposon insertion data to a list of essential genes generated in YPD (Winzeler et al. 1999), and to relative fitness measurements of genes grown on YPGal (Costanzo et al. 2021), that were defined using complete open reading frame deletions. In all strains, essential genes have fewer insertions than non-essential genes (**Figure 4.S3**), and that conditionally essential genes (i.e. GAL genes) are depleted in insertions, confirming that transposon insertion density is a reliable predictor of sequence tolerance to disruptive mutation in CNV strains.

4.3.3 Gene amplification increases mutational target size

We investigated how gene amplification affects insertion density by considering only coding sequences (which we refer to as genes) within the CNV region for each CNV strain, and comparing insertion density with all genes on chromosome XI in the euploid replicates. We find that in CNV strains, amplified genes have a higher insertion frequency than in the euploid

(Welch's two sample t-test, $p<0.0001$), consistent with increased target size resulting in increased mutation frequency. Essential genes (Winzeler et al. 1999) have significantly fewer insertions than non-essential genes in the euploid strain and in unamplified genes in the CNV strains (**Figure 4.S4A,** Welch's two sample t-test, $p<0.0001$). By contrast, no difference is observed in mutation frequency between amplified essential and non-essential genes in the CNV strains (Welch's two sample t-test, $p>0.01$) (**Figure 4.2B**). Using regression analysis, we observed that the mean number of insertions in a gene in the euploid strains predicts the number of insertions in a CNV strain, but the slope of the regression line is less than that expected on the basis of copy number (**Figure 4.2C)**. This is likely due to a compositional data effect, as the slope of regression performed on unamplified genes in most CNV strains is slightly less than one (**Figure 4.S4B**).

The number of insertions in some amplified genes is not well predicted by the model (i.e., they have large residuals); when the number of insertions is higher than the predicted value, this may indicate a gene that is particularly sensitive to amplification (**Figure 4.2C**). For example, one such gene, *UTH1*, has been shown to lead to cell death when overexpressed in the W303 genetic background (Camougrand et al. 2003). *UTH1* is a mitochondrial protein involved in regulating both mitochondrial biogenesis and degradation (Camougrand et al. 2004), and is regulated by *SSD1*, whose loss of function is associated with fitness defects in aneuploid strains of W303 (Hose et al. 2020). Interestingly, though *UTH1* is amplified in all CNV strains except trip1, it only has large residuals in four of these strains (**Figure 4.2C**). Conversely, significantly reduced insertional frequency may reflect an advantage due to increased copy number. For example, *YKR005C* is depleted in expected mutation frequency in two strains.

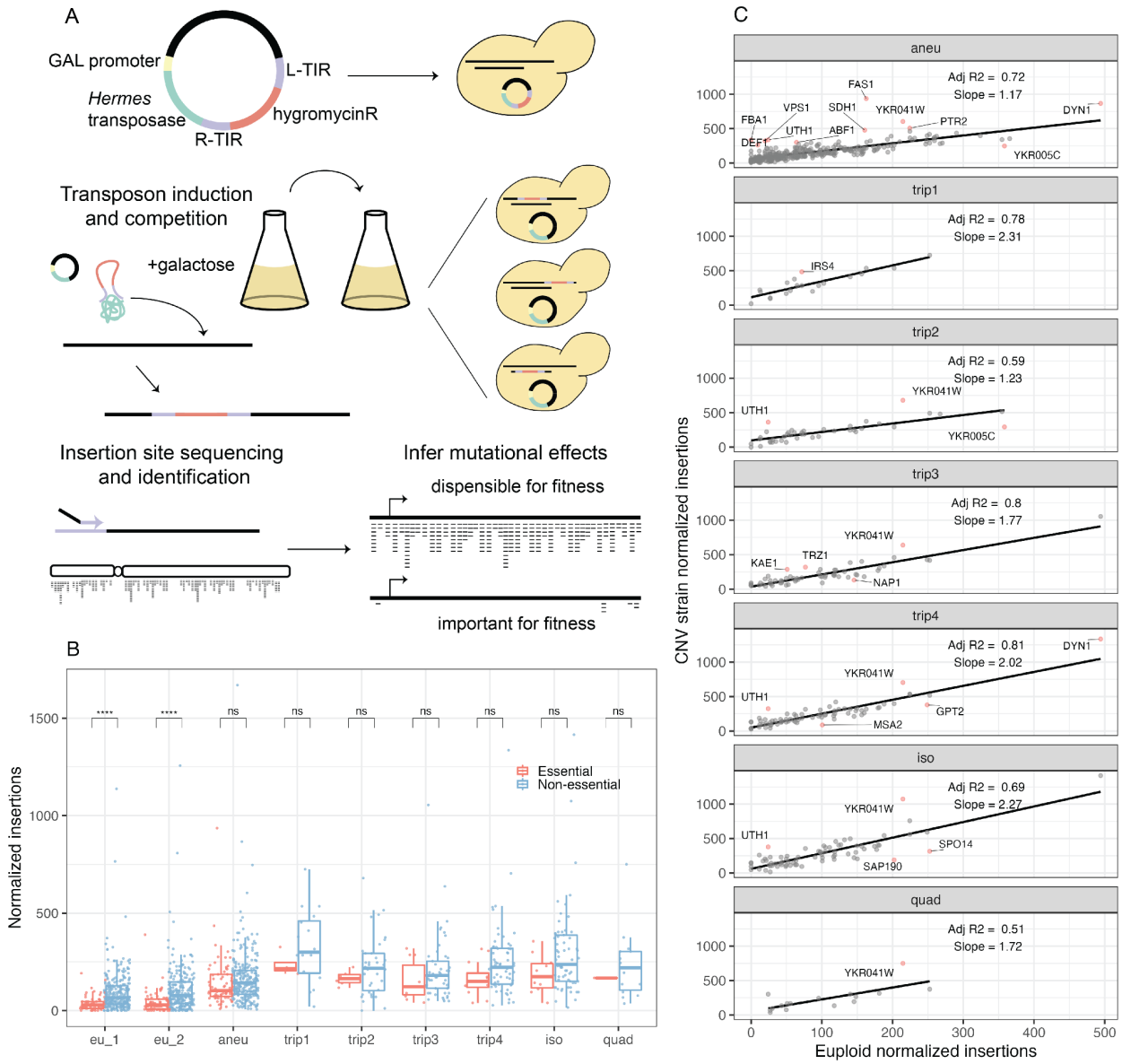

Figure 4.2. Profiling mutation tolerance using insertional mutagenesis. A) Plasmids containing the hermes transposase regulated by the GALS (truncated GAL1) promoter and a hygromycin resistance gene flanked by the hermes terminal inverted repeats (TIR) were transformed into each yeast strain. Upon addition of galactose, the transposase is expressed, and the hygromycin resistance gene flanked by the TIRs is excised from the plasmid and inserted in the yeast genome. DNA is extracted, digested with restriction enzymes, and circularized. Insertion sites are identified by inverse PCR and amplicon sequencing. Mutational tolerance is inferred by the number of unique insertion sites at a given region of the genome. **B)** Boxplots of unique insertion sites per gene, with individual genes plotted as points, for essential (red) and non-essential (blue) genes (Winzeler et al. 1999). All genes on Chromosome XI are displayed for the euploid replicates, eu_1 and eu_2. For the CNV strains, only genes that are amplified within the CNV region are shown. P-values are indicated by the following: ns: $p > 0.01$; ****: $p < 0.0001$. **C)** Linear regression models fit to the normalized number of insertions per amplified gene in CNV strains (y-axis) and the mean number of normalized insertions per gene in the euploid replicates (x-axis). Genes with residuals more than two standard deviations away from the mean residual value are highlighted in red.

4.3.4 CNVs result in common and strain specific genetic interactions

Genes that have no insertions events may be essential and intolerant of mutation, or may have no insertions due to chance. To establish a genome-wide view of differences in mutational tolerance between CNV strains and the euploid strain, we first identified 327 genes that have no insertions in either replicate of the euploid strain. Of these, 136 (42%) have previously been annotated as essential or as having low fitness in galactose. We define this set of genes as “euploid intolerant”. Many of these had insertions in one or more of the CNV strains and seven euploid intolerant genes had insertions in all CNV strains (**Figure 4.3A**). These seven genes have all been previously annotated as essential or as having low fitness in galactose. Although four of these genes were amplified in one or more CNV strains, none are amplified in all CNV strains (**Figure 4.3A**), suggesting that mutational tolerance in the CNV strains is not simply attributable to increased target size. We also looked for genes that had no insertions in any CNV strain but did have insertions in both replicates of the euploid strain: we identified one gene, *RRN10*. However, it only had one insertion in eu_1 and two insertions in eu_2, so this is unlikely to be meaningful. This suggests that CNVs do not result in novel genetic vulnerabilities.

To quantitatively assess genetic interactions between the CNV and all other genes throughout the genome we quantified differential insertion frequency, using the number of unique insertion sites per coding sequence, between each CNV strain and the euploid replicates. To assess the global trend we performed gene set enrichment analysis (GSEA) using the ranked list of fold change in number of insertions (**Figure 4.3B**). We found that three strains, aneu, quad, and trip2, have an increased mutational tolerance in genes annotated with terms related to translation and mitochondrial gene expression, and decreased mutational tolerance for genes with functions in the aerobic electron transport chain. We also observed enrichment for terms that are unique to individual strains. For example, the isochromosome (iso) exhibits

decreased tolerance for mutations in genes with functions in the mitotic cell cycle and nuclear division (**Figure 4.3B**).

We identified individual genes with differences in insertional tolerance in CNV strains compared to the ancestral euploid strain (**Figure 4.3C**). We see a similar trend as in the gene set enrichment analysis. The aneu, quad, and trip2 CNV strains all have significantly more insertions than the euploid ancestor in *BMH1*, which is involved in many processes including regulation of mitochondrial-nuclear signaling (Z. Liu et al. 2003), carbon metabolism (Dombek, Kacherovsky, and Young 2004), as well as transcription and chromatin organization (Kumar 2017; Jain, Janning, and Neumann 2021). These strains also have significantly fewer insertions in genes encoded by the mitochondrial genome, including *COX1*. Most genes that are significant in one strain tend to have similar trends in other CNV strains, with few exceptions (**Figure 4.S5A**).

Differences in insertion tolerance in genes that are not contained within the CNV reflect differential genetic interactions. To confirm this we generated complete deletions of the coding sequence of *BMH1* in all strains except trip1, for which we could not obtain a transformant, and measured growth rates of the single and double mutants in YPGal (**Figure 4.3D**). We find that deletion of *BMH1* does not significantly affect growth rate in aneu, quad, and trip2, but does result in reduced growth rate in other strains. We calculated the genetic interaction of *BMH1* with the CNV for each strain (Mani et al. 2008), and confirmed positive interactions for these three strains, consistent with transposon insertion profiles (**Figure 4.3E**, **Figure 4.S5B**).

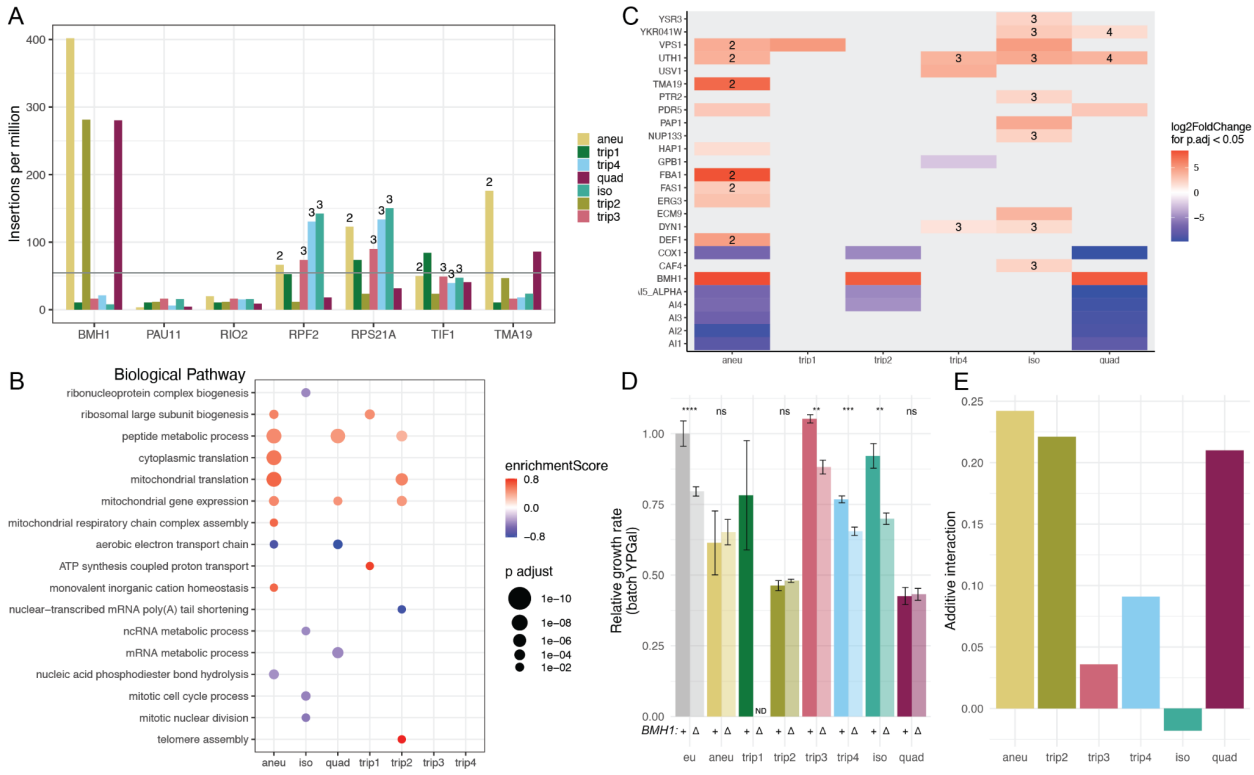

Figure 4.3. CNV strains have common and allele specific genetic interactions. A) Seven genes have no insertions in either replicate of the euploid strain, whereas insertions are identified in these genes in all CNV strains. The gray line represents the median insertions per million per gene across all strains. Numbers indicate the copy number if a gene is contained within the CNV. **B)** Enriched GO terms identified using Gene Set Enrichment Analysis (GSEA). GSEA was applied to a ranked gene list based on log2 fold changes obtained in differential analysis comparing each CNV insertion profile to the euploid insertion profiles, with the false discovery rate (FDR) for enriched terms set to 0.05. Terms with adjusted p-value < 0.05 are shown (circle size). Positive enrichment scores (red) indicate functions that have increased insertions in the CNV strain. Negative enrichment scores (blue) indicate functions that have decreased insertion frequencies in the CNV strain. **C)** Significant genes (p.adjust<0.05) from differential analysis comparing each CNV insertion profile to the euploid insertion profiles. Positive log2 Fold Change values have more insertions in CNV strains than euploid strains, whereas negative log2 Fold Change values indicate genes with fewer insertions in the CNV strain than euploid strains. If a gene is amplified the copy number is annotated. **D)** Average and standard deviation (error bars) of growth rate relative to the ancestral, euploid strain in YPGal batch culture. P-values from two-sample t-test are indicated by the following: ns: not significant; *: p < 0.05; **: p < 0.01; ***: p < 0.001; ****: p < 0.0001. **E)** Additive genetic interaction (epsilon = double - single*single) for CNV and *BMH1* double mutants, calculated from growth rates shown in (**D**).

4.3.5 Amplified genes have increased RNA expression

To test how gene expression impacts mutational tolerance in CNV lineages we performed RNAseq in triplicate on each euploid and CNV strain growing in YPGal, and quantified gene expression in each CNV strain relative to the euploid ancestor. First, we investigated genes encoded on chromosome XI for evidence of dosage compensation within the

CNV region. We found that in each CNV strain, amplified genes have significantly higher mRNA expression than in the euploid ancestor (t-test $p < 0.0001$, **Figure 4.4A**), and expression in amplified genes is highly correlated with euploid expression (**Figure 4.S6**). However, for the aneu, trip3, iso, and quad, the mean increase in mRNA expression is less than the expectation based on CNV copy number (**Figure 4.4B, Table 4.S3, 95% CI**), suggesting that dosage compensation may operate in these strains. Genes on chromosome XI that are not amplified do not differ significantly from the euploid strain in gene expression (t-test $p > 0.05$), with the exception of trip1 and the isochromosome. Trip1 appears to have increased expression of genes that are on the right arm of chromosome XI but not within the CNV itself, which may be caused by transcriptional neighborhood effects of that particular CNV structure (**Figure 4.4A, Table 4.S3**) (Brooks et al. 2022). The isochromosome appears to have increased expression extending past the left boundary to the centromere, which could be caused by transcriptional neighborhood effects, although we cannot rule out misidentification of the CNV boundary (**Figure 4.4A, Table 4.S3**).

To test the relationship between gene expression and mutational tolerance we compared the log2 fold change of transposon insertion frequency with the log2 fold change for mRNA expression for each CNV strain compared to the euploid strain. If CNV burden is related to a general cost associated with increased expression of all amplified genes, or the “mass action” of the CNV, we expect that the fold change in transposon insertions would positively correlate with the fold change in mRNA expression. Indeed, we observe a positive relationship between the increase in insertion frequency and increase in mRNA expression for the aneuploid strain (Pearson’s $r=0.13$, $p=0.0237$), but no correlation in any of the other strains ($p > 0.05$) (**Figure 4.4C**). This supports the hypothesis that adverse effects of CNVs do not stem from the “mass action” of increased expression of all genes in the CNV, but from a few critical genes that are deleterious even when slightly overexpressed (Bonney, Moriya, and Amon 2015).

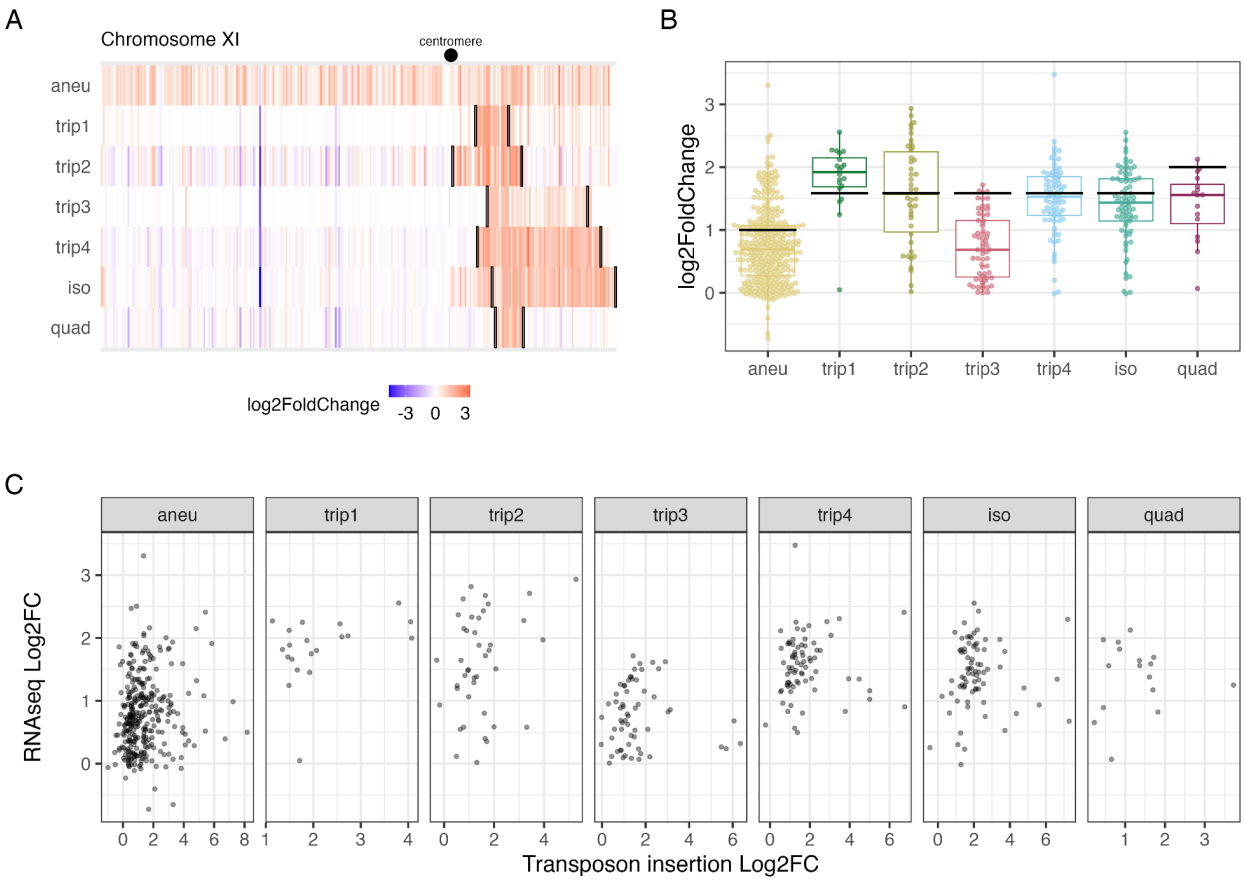

Figure 4.4. Amplified genes result in increasedmRNA expression but are not associated with increased mutation frequency. A) Replicate averaged log2 fold change of all genes on chromosome XI, ordered as on the chromosome, in each CNV strain compared to the euploid ancestor. Black lines denote CNV boundaries. **B)** Replicate averaged log2 fold change mRNA expression compared to the euploid for genes that are amplified in each CNV strain. Black lines indicate expected log2 fold change based on CNV copy number. **C)** Log2 fold change of mutational frequency within the CNV for each strain relative to the euploid CNV mRNA expression relative to the euploid for all amplified genes in each CNV strain.

4.3.6 CNV strains do not exhibit transcriptional signatures of aneuploidy

Previous studies of a laboratory strain of yeast (W303) report a transcriptomic signature of aneuploidy independent of which chromosome is duplicated (Torres et al. 2007; Terhorst et al. 2020), that is characteristic of the yeast environmental stress response (ESR) (Gasch et al. 2000) and comprises 868 genes. The expression of genes in the ESR are correlated with growth rate (Brauer et al. 2008) and several studies have shown that strains with higher degrees of aneuploidy (i.e. more additional base pairs) exhibit lower growth rates and stronger ESR (Torres et al. 2007; Terhorst et al. 2020). We compared the ESR gene expression profiles

in our CNV strains and a previous study which identified the ESR as a response to aneuploidy (Torres et al. 2007). Surprisingly, we find a significant negative correlation for all strains except trip3, with the strength of the anticorrelation decreasing as growth rate increases (**Figure 4.5A, Figure 4.S7**). We ruled out that this was due to aberrant behavior of the euploid strain. Torres et al. also profiled the gene expression of aneuploid yeast while controlling for growth rate by growth in a chemostat; ESR gene expression in this data set is significantly positively correlated with for all CNV strains except trip3, with the strength of the correlation decreasing as growth rate increases (**Figure 4.5B, Figure 4.S8**).

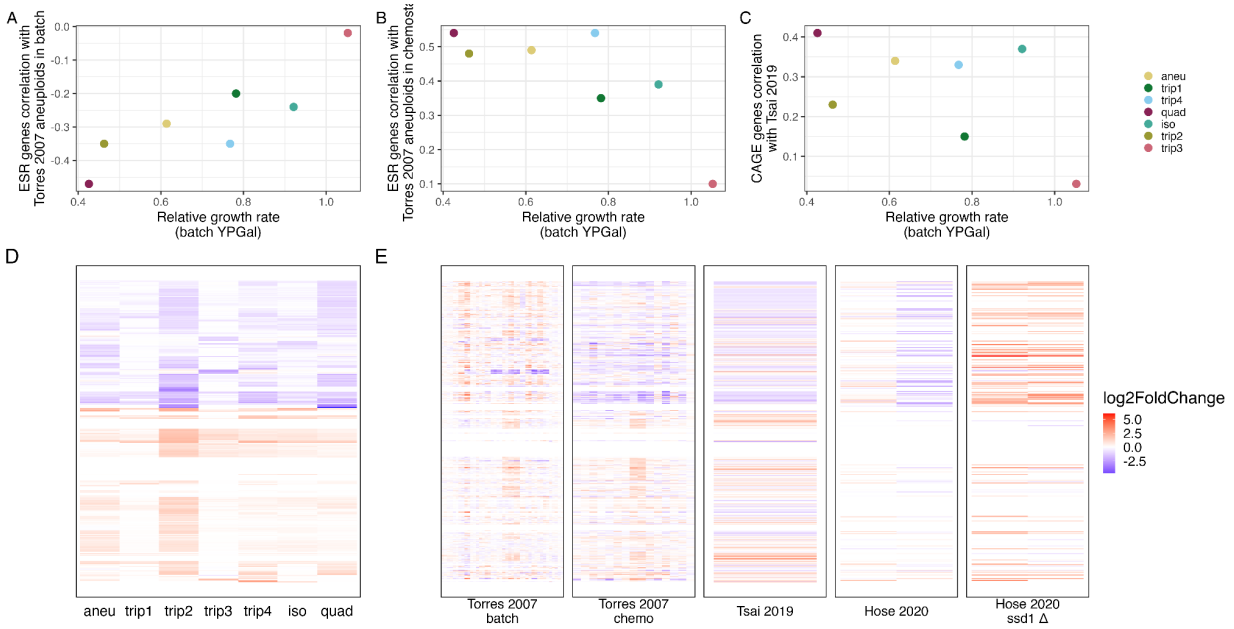

Figure 4.5. Global gene expression signatures in CNV strains are distinct from aneuploidy effects. We compared the mean growth rate of CNV strains in YPGal to **A)** the pearson correlation between log2 fold change in mRNA expression in CNV strains vs euploids in this study and the mean log2 fold change in mRNA expression in for aneuploids vs euploids grown in batch culture in Torres et al. 2007 for 798 ESR genes for which we have complete data, **B)** pearson correlation between log2 fold change in mRNA expression in CNV strains vs euploids in this study and the mean log2 fold change in mRNA expression in for aneuploids vs euploids grown in growth-rate controlled chemostats in Torres et al. 2007 for 801 ESR genes for which we had complete data, **C)** pearson correlation between log2 fold change in mRNA expression in CNV strains vs euploids in this study and the log2 fold change in mRNA expression in for aneuploids vs euploids in Tsai et al. 2019 for the 215 CAGE genes for which we had complete data. **D)** Log2 mRNA expression for 436 genes (rows) significantly differentially expressed in at least one CNV strain versus the euploid strain. **E)** Data corresponding to genes from **(D)** from Torres et al. 2007 aneuploids in batch, Torres et al. 2007 aneuploids in chemostat, Tsai et al. 2019, Hose et al. 2020 wild aneuploid strains, and Hose et al. wild aneuploid strains with ssd1 deleted. The former four are compared to closely related euploids, the aneuploids with ssd1 deletions are compared to their wild-type aneuploid counterparts.

Recently, a common aneuploidy gene-expression (CAGE) signature across yeast aneuploid for many different chromosomes that is similar to the transcriptional response to hypo-osmotic shock was defined in a derivative of the S228c genetic background (Tsai et al. 2019). With the exception of trip3, expression of CAGE genes is moderately positively correlated between Tsai et al. and our CNV strains (**Figure 4.5C, Figure 4.S9**). Interestingly, expression of CAGE genes in Tsai et al. is also positively correlated with expression of CAGE genes in growth rate controlled aneuploid strains (**Figure 4.S10**).

4.3.7 Genome-wide gene expression effects of CNVs

From our differential analysis, we identified 436 genes, 341 of which are not located on chromosome XI, that had significantly altered expression in one or more CNV strains compared to the euploid strain (log2 fold change > 1.5, Benjamini and Hochberg adjusted $p < 0.05$, **Figure 4.5D**). Of the significant genes, 73 are ESR genes and 13 are in the CAGE signature. These genes have two major clusters; genes that have decreased expression are involved in cellular respiration, nucleoside biosynthetic processes, and small molecule metabolism, and genes that have increased expression are involved in transposition, nucleic acid metabolic processes, and siderophore transport (hypergeometric test $p < 0.0001$). Similarly, wild yeast strains that are tolerant of aneuploidy exhibit down-regulation of mitochondrial ribosomal proteins and genes involved in respiration, and upregulation of oxidoreductases (Hose et al. 2015). The clusters and enrichment patterns remain even when excluding genes on chromosome 11. Additionally, we performed GSEA on the log2 fold change for genes in each CNV strain compared to the euploid. Generally, we see similar enrichment as in **Figure 4.5D**, with some variation in particular terms between strains.

There are nine genes that have significantly different expression than the euploid in all strains and are not on chromosome XI (**Figure 4.S11**). Three genes have increased expression: two are retrotransposons and the third, *RGI2*, is a protein of unknown function that is involved in

energy metabolism under respiratory conditions (Domitrovic et al. 2010). Repressed genes include the paralogs *MRH1* and *YRO2*, both of which localize to the mitochondria (Jörg Reinders et al. 2007; Joerg Reinders et al. 2006); *OPT2*, an oligopeptide transporter (Wiles et al. 2006); *YGP1*, a cell wall-related secretory glycoprotein (Destruelle, Holzer, and Klionsky 1994); and two proteins of unknown function, *PNS1* and *RTC3*.

We compared the genes that are significantly differentially expressed in one or more CNV strains to the data generated from aneuploid strains in Torres 2007, Tsai 2019, and Hose 2020 (**Figure 4.5E**). As with comparison to the ESR and CAGE signatures, we see that our strains are more similar to the aneuploids grown in chemostats in Torres et al. 2007 and the aneuploids in Tsai 2019 than the aneuploids growing in batch culture and exhibiting the ESR. Hose et al. compared gene expression in strains aneuploid wild yeast that are tolerant of CNV to their euploid counterparts and to aneuploid wild yeast with *SSD1* deleted. The gene expression profiles of the wild-type aneuploid wild yeast more closely resemble our strains than the aneuploid *SSD1* mutants, which are more similar to the aneuploids grown in batch culture in Torres 2007 (**Figure 4.5E**).

4.3.8 Low fitness is associated with mitochondrial dysfunction

Common genetic interactions and mRNA expression signature in CNV strains appear to be linked to mitochondrial function and translation, with strains that exhibit stronger profiles having lower fitness. *BMH1***,** which exhibits positive genetic interactions in most CNV strains (**Figure 4.3E**), is a negative regulator of retrograde signaling (da Cunha, Torelli, and Kowaltowski 2015). We hypothesized that this interaction might occur if mitochondria are dysfunctional and therefore retrograde signaling is activated. However, *CIT2* and *DLD3*, which are robustly upregulated in the canonical retrograde response (da Cunha, Torelli, and

Kowaltowski 2015), do not have significantly different expression from the euploid in any CNV strain.

Yeast simultaneously ferment and respire galactose (Fendt and Sauer 2010). Therefore, to test whether CNV lineages have impaired mitochondrial function we tested growth in the presence of carbonyl-cyanide 3-chlorophenylhydrazone (CCCP), a mitochondrial uncoupling agent. We found that treating with CCCP nearly abolished growth in two of the three strains that showed strongest signals differential mitochondrial function, trip2 and quad, whereas the reduction in growth in other strains was similar to that of the euploid (**Figure 4.6A**).

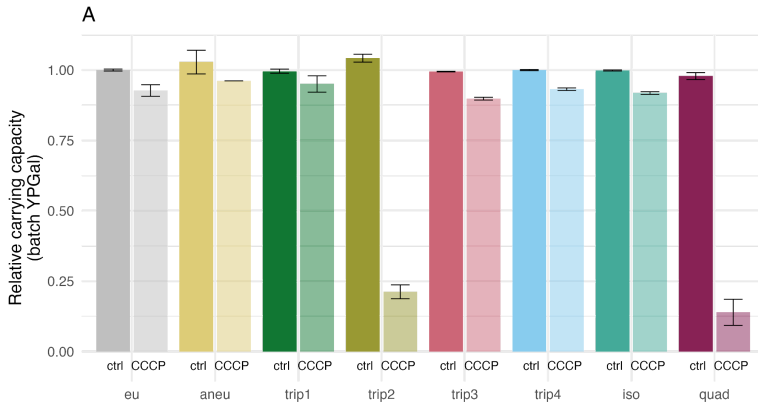

Figure 4.6. Growth response to treatment with CCCP. A) Average and standard deviation (error bars) carrying capacity (i.e. maximum optical density) relative to the ancestral, euploid strain in YPGal batch culture in either control condition or with 25 μM CCCP.

4.4 Discussion

In this study, we sought to understand the effect of diverse CNVs on genetic interactions and transcriptomic state. Though investigations of evolutionary trajectories and combinations of mutations that arise in evolution experiments have suggested that epistasis between CNVs and other mutations is an important contributor to evolutionary dynamics (Pavani et al. 2021; Lauer et al. 2018), few studies have systematically investigated genetic interactions with CNVs (Dodgson et al. 2016). We find that amplification results in relaxed selection against mutation in essential genes. Additionally, we find both CNV specific genetic interactions and interactions

that are shared by several strains. To gain further insight into the effects of CNV on the cells, we performed RNAseq. Whereas amplification results in increased mRNA expression compared to the euploid, we also observe dosage compensation in several strains. Consistent with a recent study in the same genetic background as our strains (S288c), we do not find activation of the ESR in various aneuploids (Larrimore et al. 2020), nor did we observe the CAGE response (Tsai et al. 2019). Instead, CNVs tend to have increased expression of genes involved in transposition, nucleic acid metabolic processes, and siderophore transport, and decreased expression of those involved in cellular respiration, nucleoside biosynthetic processes, and small molecule metabolism, though the extent to which the expression differed from the euploid varied between CNV strains.

We have demonstrated that transposon mutagenesis is a powerful tool to investigate genetic interactions genome-wide in strains with large and complex mutations. Unlike synthetic genetic array (SGA) analysis, which is commonly used to investigate genetic interactions, transposon mutagenesis does not require mating the query strain to the deletion collection. Transposon mutagenesis therefore avoids a some of the issues that are encountered using SGA: the deletion collection has some inaccuracies (Giaever and Nislow 2014; Ben-Shitrit et al. 2012), other mutations including aneuploidy (T. R. Hughes et al. 2000), and the gene expression of non-target genes is sometimes impacted by the deletion of neighboring genes (Ben-Shitrit et al. 2012; Baryshnikova and Andrews 2012) which can result in false positive or false negative genetic interactions. Furthermore, the requirement to mate the query strain to the deletion collection means that genetic interactions identified are in a diploid and potentially hybrid (if the query strain and deletion collection are different) background, which further complicates the interpretation of results (Dodgson et al. 2016). Despite overcoming some of these challenges, transposon mutagenesis also has some shortcomings. Transposon insertion efficiency can differ between genetic backgrounds (Caudal et al. 2021) and ability to detect interactions is

dependent on the number of insertions identified. Our experimental design induces transposition in galactose for four days, which may be enough time for additional mutations to also accumulate, reducing the number of conditions in which genetic interactions can be examined and power of the experiment to detect loci which tolerate mutation.

A question that naturally arises from our study is why do different CNV structures result in heterogeneous fitness effects, genetic interactions, and transcriptional responses? One reason may be the particular composition of the CNV. Both the copy number and the particular genes amplified likely play a role in the fitness effect. For example, the aneuploid was much less fit in YPGal than several of the other CNV strains. Previous work has shown that the amplification of the left arm of chromosome XI has negative fitness effects in other conditions (Sunshine et al. 2015) - the aneuploid is the only strain with the left arm amplified in our experiment, and that could be part of the basis of the fitness consequences. Additionally, CNV strains used here were each isolated from an evolution experiment, and other mutations were identified. While it seems unlikely that all mutations would modify the effect of the CNV, some could. For example, tri257, which has fitness, genetic interactions, and transcriptome similar to the ancestor in YPGal, has a mutation causing a premature stop in *SSK2* (**Table 4.S1**). *SSK2* is a MAP kinase kinase kinase of *HOG1* signaling pathway, that controls osmoregulation, which may attenuate stress from CNV if it is, as Tsai et al. found in aneuploids, similar to hypo-osmotic stress. Further studies in CNVs of various structures encompassing different regions and in isogenic backgrounds may help to disentangle these factors.

A large-scale analysis of aneuploidy across over 1,000 *S. cerevisiae* isolates showed that genetic background alone (rather than ecology) could predict aneuploidy prevalence (Scopel et al. 2021), and several studies have shown that tolerance to aneuploidy varies across genetic backgrounds (Hose et al. 2015; Gasch et al. 2016; Hose et al. 2020; Larrimore et al. 2020). This understanding leads us to important questions: is there a common response to

aneuploidy and more generally CNV in the genetic backgrounds that do not well tolerate them? Can we predict which genetic backgrounds will be able to tolerate CNV or be sensitive to CNV? Hose et al. found that aneuploidy sensitivity in the laboratory strain W303 resulted from synergistic defects in mitochondrial function and proteostasis. Interestingly, our results also point to mitochondrial function and translation as important and the least fit strains in our study are particularly sensitive to mitochondrial stress. The laboratory strain S288c, which was used in this study, has a hypomorphic allele of *HAP1 (Gaisne et al. 1999)*, which is a heme-responsive transcriptional activator of genes involved in respiration (L. Zhang and Hach 1999) (notably, genes with increased expression in CNV strains include siderophores, which chelate iron). Across many genetically distinct strains of yeast, genes involved in aerobic respiration and the electron transport chain vary more than any other category during growth in glucose-limited chemostats (Skelly et al. 2013), and genes involved in mitochondrial function have continuous variation in fitness effects across different isolates (Caudal et al. 2021). It would be instructive to study the relationship between mitochondrial function and CNV tolerance.

4.5 Methods

4.5.2 Strains

The euploid ancestral *GAP1* CNV reporter and the evolved *GAP1* CNV strains were previously described and characterized in Lauer et al. 2018, and are haploid derivatives of the reference strain S288C (and more specifically, FY4/5) with a constitutively expressed *mCitrine* gene and KanMX G418-resistance cassette inserted 1,118 base pairs upstream of *GAP1.* This construct is referred to as the *GAP1* CNV reporter. The CNV strains are clonal isolates that evolved for 150 or 250 generations in glutamine limited chemostats (Lauer et al. 2018).

Each strain was transformed with pSG36_HygMX using the EZ-Yeast™ Transformation Kit (MP Biomedicals, cat #2100200). Transformants were recovered on YPG agar + 200 µg/mL

Hygromycin B. A single colony was picked from the plate of transformants to perform each transposon mutagenesis experiment. Separate transformation and colony selection was performed for each replicate of the euploid.

To generate *BMH1* mutants, we performed high-efficiency yeast transformation into frozen competent yeast cells for each strain (Gietz and Schiestl 2007a) with an *mCherry* gene under control of the constitutively expressed *ACT1* promoter (*ACT1pr::mCherry::ADH1term*) and marked by the HphMX Hygromycin B-resistance cassette (*TEFpr::HygR::TEFterm)*. The plasmid DGP363, containing this construct, was used as template for PCR using primers containing the same *BMH1*-specific targeting homology, and transformation resulted in a complete deletion of the *BMH1* open reading frame. Transformants were recovered on YPD agar + 400 µg/mL G418 + 200 µg/mL Hygromycin B, and *BMH1* deletion positive transformants were confirmed using *BMH1* specific primers and a HygR primer. We verified that *mCitrine* copy number remained unchanged and *mCherry* fluorescence using a Cytek Aurora flow cytometer.

4.5.2 Growth curves

For each experiment, we inoculated three colonies per strain into 3-5 mL YPGal, and grew them overnight at 30℃. In triplicate per original colony, we back diluted 5 µL of culture into 195 µL fresh YPGal or YPGal with 25 µM carbonyl-cyanide 3-chlorophenylhydrazone in a Costar Round Bottom 96 well plate (Ref 3788). We treated the lid with 0.05% Triton X-100 in 20% ethanol to prevent condensation (Brewster 2003). We collected OD600 data over approximately 48 hours using a Tecan Spark with the following parameters: Temperature control: On; Target temperature: 30 [°C]; Kinetic Loop; Kinetic cycles: 530; Interval time: Not defined; Mode: Shaking; Shaking (Double Orbital) Duration: 240 [s]; Shaking (Double Orbital) Position: Current; Shaking (Double Orbital) Amplitude: 2 [mm]; Shaking (Double Orbital) Frequency: 150 [rpm]; Mode: Absorbance; Measurement wavelength: 600 [nm]; Number of flashes: 10; Settle time: 50 [ms]; Mode: Fluorescence Top Reading; Excitation: Monochromator;

Excitation wavelength: 497 [nm]; ExcitationBandwidth: 30 [nm]; Gain: Calculated From: B5 (50%); Mirror: AUTOMATIC; Number of flashes: 30; Integration Time: 40 [µs]; Lag time: 0 [µs]; Settle time: 0 [µs]; Z-Position mode: From well B5.

Using growthcurver (Sprouffske and Wagner 2016), we fit the OD600 data to a logistic equation, using the value of the parameter *r* as the intrinsic growth rate of the population and the parameter *k* as the carrying capacity. We checked for and discarded outliers by examining OD curves and histograms of sigma (goodness of fit). We normalized each growth rate and carrying capacity to that of the ancestral euploid *GAP1* CNV reporter grown in the same plate.

4.5.3 Transposon mutagenesis

A single transformant for each strain was used to inoculate a 30 mL YPD + 200 µg/mL Hygromycin B, and incubated approximately 24 hours at 30℃ with agitation, until OD5. To induce transposition, the culture was then diluted to OD0.05 in YPGalactose + 200 µg/mL Hygromycin B to a final volume of 50 mL, and incubated 24 hours at 30℃ with agitation. The culture was diluted to 0.05 in 50 mL YPGalactose + 200 µg/mL Hygromycin B and incubate 24 hours three more times, for a total of four serial transfers in YPGalactose + 200 µg/mL Hygromycin B. The culture was pelleted by centrifugation for five minutes at 4000 rpm, the supernatant removed, then resuspended to OD0.5 in 50 mL YPD and incubated 24 hours at 30℃ with agitation, then diluted again to OD0.5 in 50 mL YPD and incubated 24 hours at 30℃ with agitation, to release selection to maintain pSG36_HygMX. The cultures were then diluted to OD0.5 in 100 mL YPD + 200 µg/mL Hygromycin B and incubated 24 hours at 30℃ with agitation to select for cells with the transposon in the genome. The final culture was pelleted by centrifugation for five minutes at 4000 rpm, the supernatant removed, resuspended with 1 mL sterile water, split into four 250 µL aliquots, and pelleted for two minutes at 8000 rpm. The supernatant was removed and cell pellets were frozen at -20℃ for storage until DNA extraction was performed.

4.5.4 Insertion site sequencing

DNA was extracted from cell pellets using the MasterPure™ Yeast DNA Purification Kit (Lucigen, cat #MPY80200), with an additional initial incubation with zymolase at 37°C to enhance cell lysis, and using a Glycogen/Sodium Acetate/Ethanol DNA precipitation (Green and Sambrook 2016). For each sample, 2 µg of DNA was digested with 50 units of DpnII and 5 µL NEBuffer™ DpnII (NEB, cat #R0543L), in a total volume of 50 µL; and 2 µg of DNA was digested with 50 units of NlaIII and 5 µL CutSmart® Buffer (NEB, cat #R0125L), in a total volume of 50 µL, for 16 hours at 37℃. The reactions were heat inactivated, then circularized by ligation in the same tube with 25 Weiss units T4 Ligase and 40 µL T4 ligase buffer (Thermo Scientific cat #EL0011) for 6 hr at 22°C, in a volume of 400 µL. Circularized DNA was precipitated using a Glycogen/Sodium Acetate/Ethanol DNA precipitation (Green and Sambrook 2016). Inverse PCRs for each sample and digestion were performed with primers Hermes_F and Hermes_R with 0.5 µL of each circularized DNA sample per reaction. PCR was performed with DreamTaq (ThermoFisher cat #EP0701), with the following program: 2 min at 95°C followed by 32 cycles of 30 s at 95°C, 30 s at 57.6°C, 3 min at 72°C, and a final extension step of 10 min at 72°C. The PCRs products were confirmed on 2% agarose gels, and the concentration was quantified using Qubit™ dsDNA BR Assay Kit.

Library preparation and sequencing were performed using two different library preparation methods and sequencing set ups.

BGI

For each sample (1728, 1736, and 1740) and digestion 35 PCR were performed as described above and the PCR products were pooled and cleaned using a Glycogen/Sodium Acetate/Ethanol DNA precipitation (Green and Sambrook 2016). For each sample, at least 6 µg at minimum 30 ng/µl was then sent to the BGI (Beijing Genomics Institute) for library preparation

and sequenced using a paired-end (2 x 100) protocol on a Illumina Hi-Seq 4000 or DNBseq platform.

NYC

For each sample (all) and digestion 4 PCR were performed as described above and the PCR products were pooled by sample and cleaned using a Glycogen/Sodium Acetate/Ethanol DNA precipitation (Green and Sambrook 2016). Five ng of each PCR product pool was used as input into a modified Nextera XT library preparation. To increase library complexity, for each sample, two tagmentation reactions were performed. PCR to enrich for fragments with hermes sequence and add an i5 adaptor were performed on the tagmented DNA using NPM Buffer, primers Nextera_hermes_enrichment and Nextera_i7_enrichment, and the following program: 3 min at 72°C, then 30 s at 95°C, followed by 9 cycles of 10 s at 95°C, 30 s at 55°C, 30 s at 72°C, and a final extension step of 5 min at 72°C. The reactions were pooled by sample, cleaned using AmPure XP beads, and resuspended in 20 µL of molecular grade water, which was used as input for an indexing and library amplification PCR. Each sample was indexed with an i7 index from the Nextera XT kit, and amplification of the i5 end was performed with primer i5_amp (which contains no i5 index), using the 2X KAPA PCR master mix (Roche cat #KK2611), and the same program described for the PCR after tagmentation.PCR cleanup and size selection was performed with AmPure XP beads. The fragment size of each library was measured with an Agilent TapeStation 2200 and qPCR was performed to determine the library concentration. The libraries were pooled at equimolar concentrations, and sequenced using a single-end (1 x 150) protocol on an Illumina NextSeq 500. Libraries were prepared once, but sequenced in two consecutive sequencing runs for increased coverage.

4.5.5 Transposon insertion sequencing site identification and annotation

Using cutadapt v1.16 (Martin 2011) with the expected hermes TIR sequence on the 5' end were identified, and the TIR was trimmed. If the TIR was followed by plasmid sequence, these reads were discarded. For reads sequenced at BGI (paired end sequencing), the read with the TIR sequence was identified and its mate was discarded. For reads sequenced at NYC (Nextera based prep, single end sequencing), Nextera transposase sequences were identified and removed. Reads with a length less than 20 bases after all cleaning steps were discarded, and the remaining reads were checked for quality using fastqc v0.11.8 ("Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data" n.d.). Reads were aligned to the modified reference genome using bwa mem v.0.7.15 (Heng Li and Durbin 2010) and BAMs were generated with samtools v1.9 (H. Li et al. 2009). Samples prepared and sequenced by more than one method had high Pearson correlations (0.85-0.94) in the number of unique insertions identified per gene, and therefore were combined into a single BAM file before performing downstream analysis. For the majority of the analyses, BAMs were combined by sample, for ease of processing and to prevent redundant insertion site identification. BAMs were parsed with a custom python script which identifies the first base of the read as the position of the insertion. The script output all unique insertion positions and the number of reads per insertion position. Positions were annotated using bedtools v2.26.0 (Quinlan and Hall 2010) and a custom GFF containing amended annotations for the custom genome (available on OSF [OSF LINK HERE]). All analyses use unique insertion positions, and do not take into account the number of reads per unique insertion position. Uniquely identified insertion sites are supported by an average of 18.6 sequencing reads. The libraries have between 85,327 and 329,624 unique insertion sites identified, with an average of 176,664 insertion sites, corresponding to approximately one insertion per 69 bases in the yeast genome (NCBI R64 assembly**; STable 2**). We normalize for differences in sequencing depth by

calculating insertions per million: number of unique insertion sites per feature/(total unique insertion sites/1,000,000). We do not normalize for gene length, as we are comparing genes between strains, not within strains. All code used for analysis can be found on GitHub https://github.com/graceave/hermes_analysis.

4.5.6 Genetic interaction analysis

To quantitatively investigate genetic interactions using the transposon sequencing data, we performed differential analysis using DESeq2 version 1.30.1 (Love, Huber, and Anders 2014), using the number of insertions per gene and comparing each CNV strain to the two euploid replicates. We used clusterProfiler version 3.18.1 (Guangchuang Yu et al. 2012) to perform fast gene set enrichment analysis (Korotkevich et al., n.d.) using the ranked log2 fold change in insertions generated by DESeq2 and GO terms were summarized by semantic similarity (G. Yu et al. 2010) then by hand for clarity.

To calculate genetic interactions based on growth rates, we first calculate the relative fitness of each single mutant by:

$$W_{mutant} = \frac{m_{mutant}}{m_{wild-type}},$$

where m is the intrinsic growth rate of the strain (parameter *r* from logistic equation used to fit growth curves). We then calculated the expected fitness of the double mutant using either the additive model:

$$E(W_{xy}) = W_x + W_y + 1,$$

Or the multiplicative model:

$$E(W_{xy}) = W_x \times W_y.$$

We then calculate the genetic interaction:

$$\varepsilon = W_{xy} - E(W_{xy})$$ (Mani et al. 2008).

4.5.7 RNA sequencing

For RNA sequencing, we grew overnight cultures from three replicate colonies per strain in 5 mL YPGal, then 2 mL (euploid, trip3) or 5 mL (other strains) of overnight culture was pelleted and subsequently resuspended in 5 mL fresh YPGal. The cultures were allowed to grow for three hours in fresh YPGal before harvesting cells by vacuum filtration and fixing immediately in liquid nitrogen, so that all cultures were harvested while cells were proliferating. RNA was extracted and purified using a hot acid phenol/chloroform and Phase Lock Gels as described in (Neymotin, Athanasiadou, and Gresham 2014). Samples were enriched for polyadenylated RNA using the Lexogen Poly(A) RNA Selection Kit V1.5 (cat. # 157.96) and stranded RNAseq libraries were prepared using the Lexogen CORALL Total RNA-Seq Library Prep Kit (cat. # 095.96) according to the manufacturer's protocol. The libraries were pooled at equimolar concentrations, and sequenced using a paired-end (2 x 150) protocol on an Illumina NextSeq 500. The resulting fastqs were trimmed, aligned, and UMI deduplicated, and coverage per feature was calculated using an in-house pipeline which can be found at https://greshamlab.bio.nyu.edu/wp-content/uploads/2021/11/Windchime_pipeline.nb_.html. Coverage per feature correlation between replicates was high, with the exception of one replicate of the quadruplication, which was excluded from further analysis. Trip3 also only had two replicates, as library preparation failed for one replicate.

4.6 Supplemental Material

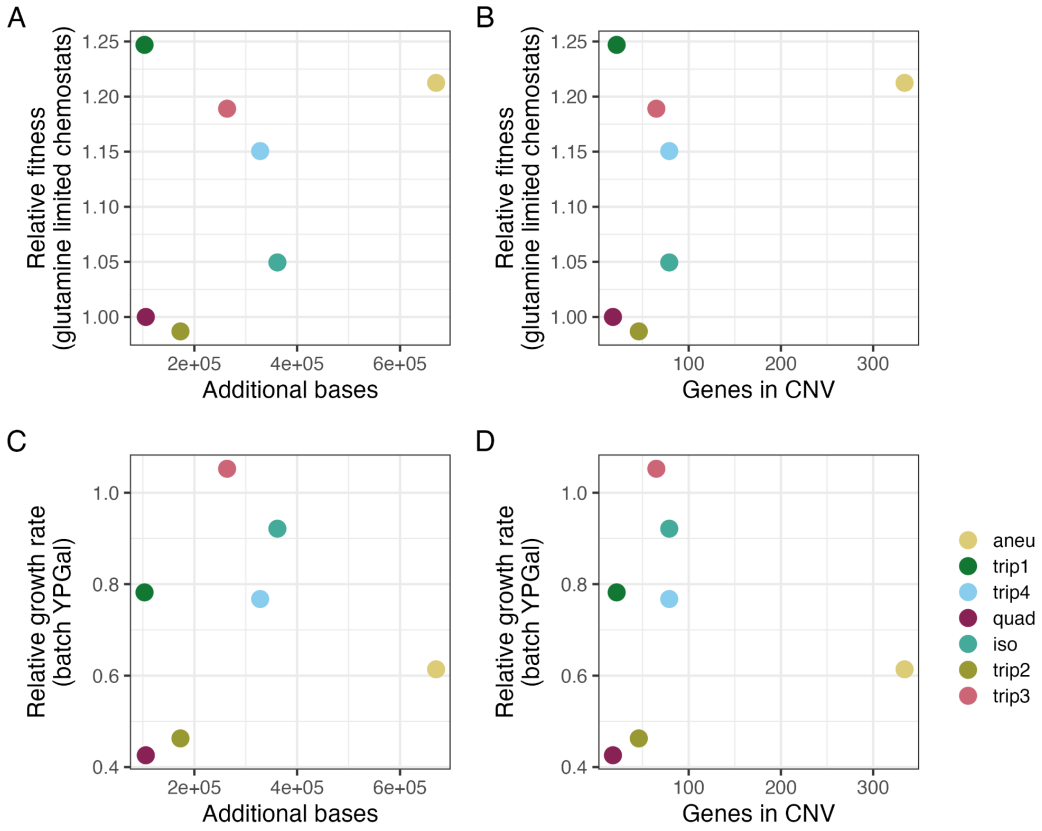

Figure 4.S1 There is no relationship between CNV size and relative fitness. A-B) The fitness of evolved strains containing *GAP1* CNVs was determined by pairwise competition experiments with a nonfluorescent, unevolved reference strain in glutamine-limited chemostats. **C-D**) Average growth rate of *GAP1* CNVs relative to the ancestral, euploid strain in YPGal batch culture.

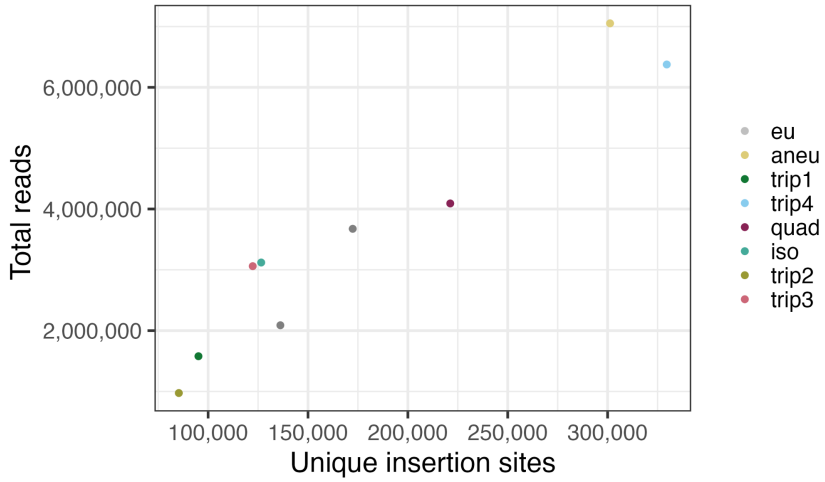

Figure 4.S2 The number of unique insertion sites scales with the number of reads sequenced. The total number of unique insertion sites identified per library increases with the total number of reads sequenced (using all methods and sequencing runs).

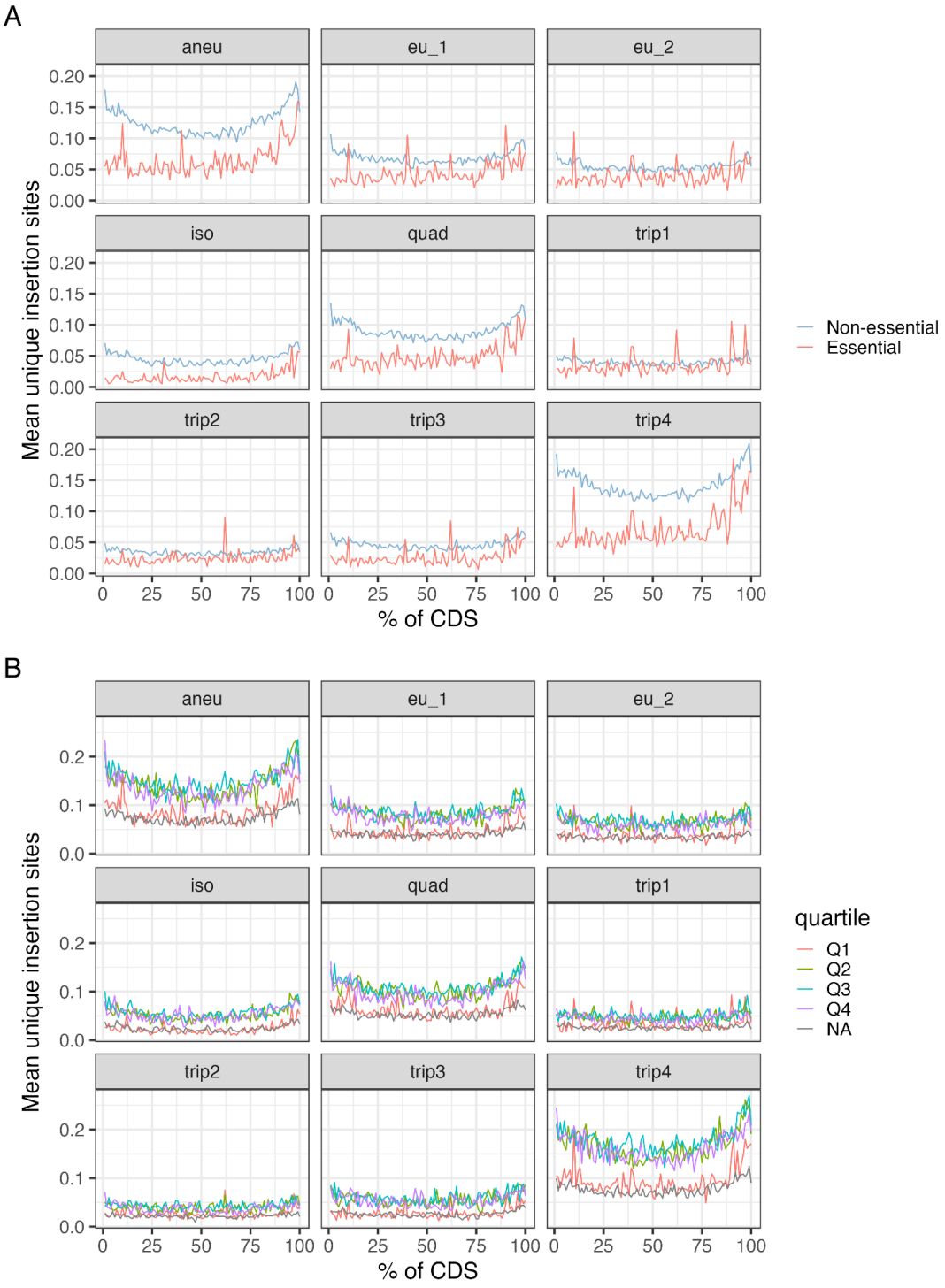

Figure 4.S3. There are fewer insertions in essential genes and genes whose deletion results in low fitness in YPGal. A) We grouped genes into those that had been previously annotated as essential or non-essential by deletion and measurement of growth on rich media (yeast peptone dextrose) (Winzeler et al. 1999). **B)** We grouped genes into four quartiles based on relative fitness measurements on rich media with 2% galactose from 3704 viable deletion mutant strains and 782 temperature-sensitive (TS) alleles (Costanzo et al. 2021). The first quartile (Q1, red) contains genes whose deletion causes the greatest measurable fitness defects, with relative fitness between 0.053 and 0.896. There was no relative fitness obtained for 21 genes (presumably there was no growth), these are marked NA (grey).

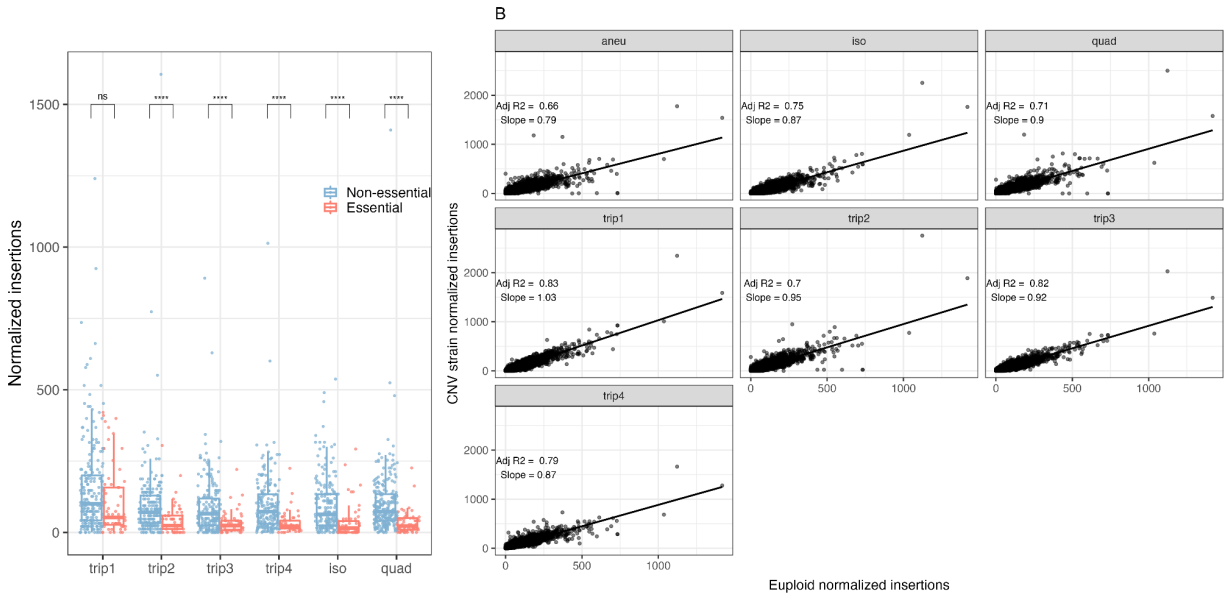

Figure 4.S4. Transposon insertions in non-amplified genes. A) Boxplots of unique insertion sites per gene, with individual genes plotted as points, for essential (red) and non-essential (blue) genes (Winzeler et al. 1999). All genes on Chromosome XI that are not within the CNV boundaries are shown. P-values from Welch's t-test are indicated by the following: ns: p > 0.01; ****: p < 0.0001. **B)** Linear regression was used to fit the normalized insertions per non-amplified gene in CNV strains (y-axis) to the mean number of normalized insertions per gene in the euploid replicates (x-axis), genome-wide. Adjusted p-values and slope from linear regression are annotated.

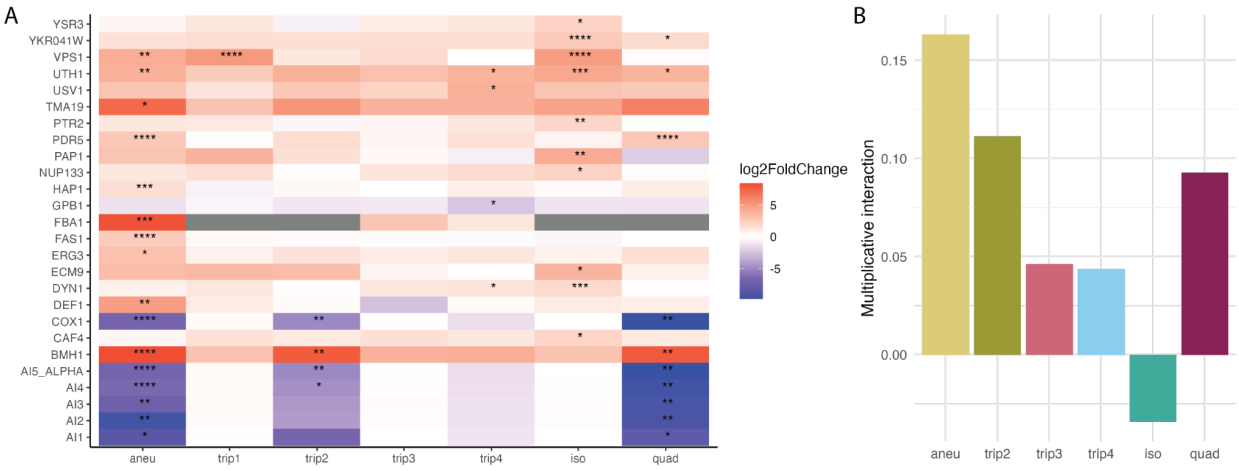

Figure 4.S5 Genetic interactions of CNV strains. A) Genes that have significantly different insertions in CNV strains versus euploid. Genes which were significant for at least one CNV strain, from differential analysis comparing each CNV insertion profile to the euploid insertion profiles. Positive log2FoldChange values have more insertions in CNV strains than euploid strains, while negative log2FoldChange have fewer insertions in CNV strains than euploid strains. P-values adjusted with the Benjamini and Hochberg method: *: p < 0.05; **: p < 0.01; ***: p < 0.001; ****: p < 0.0001. **B)** Multiplicative genetic interaction for CNV and *BMH1* double mutants. Calculated from growth rates shown in Figure 4.3D.

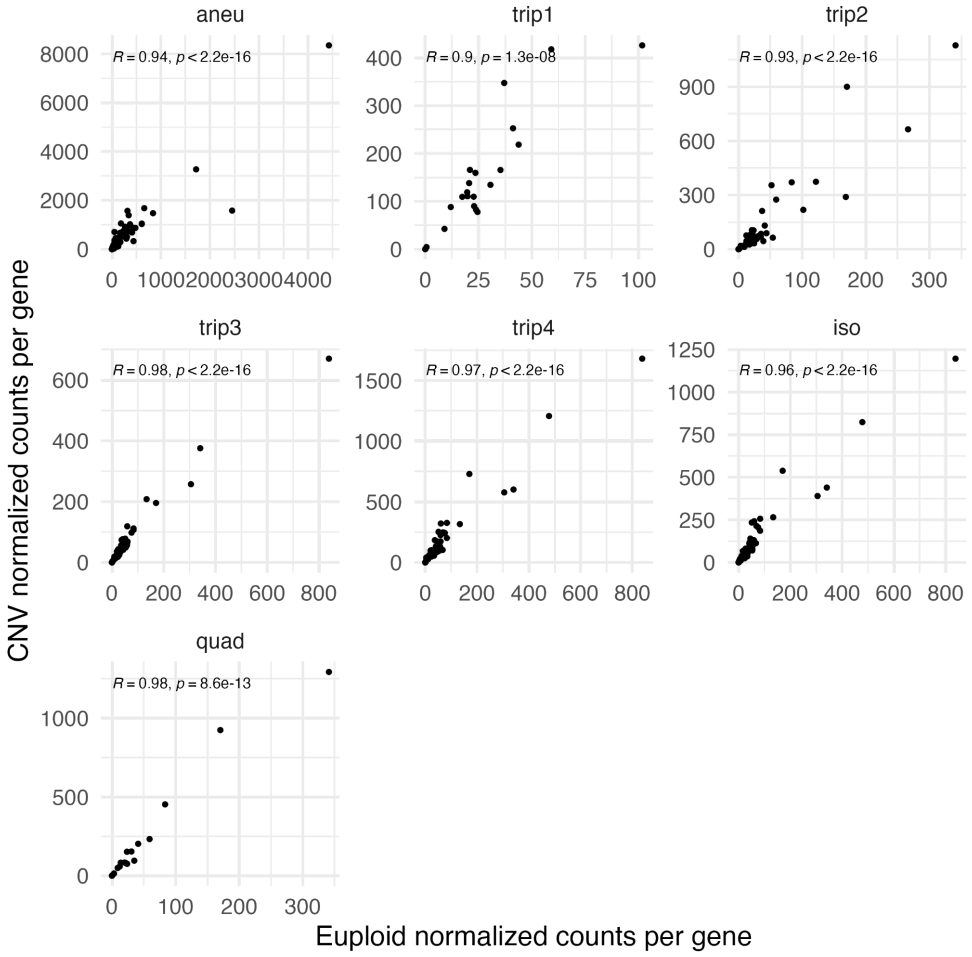

Figure 4.S6 mRNA expression of amplified genes is highly correlated with euploid expression. For each CNV, the subset of genes within the CNV boundaries are shown. Pearson's correlation coefficient and corresponding p-value are annotated.

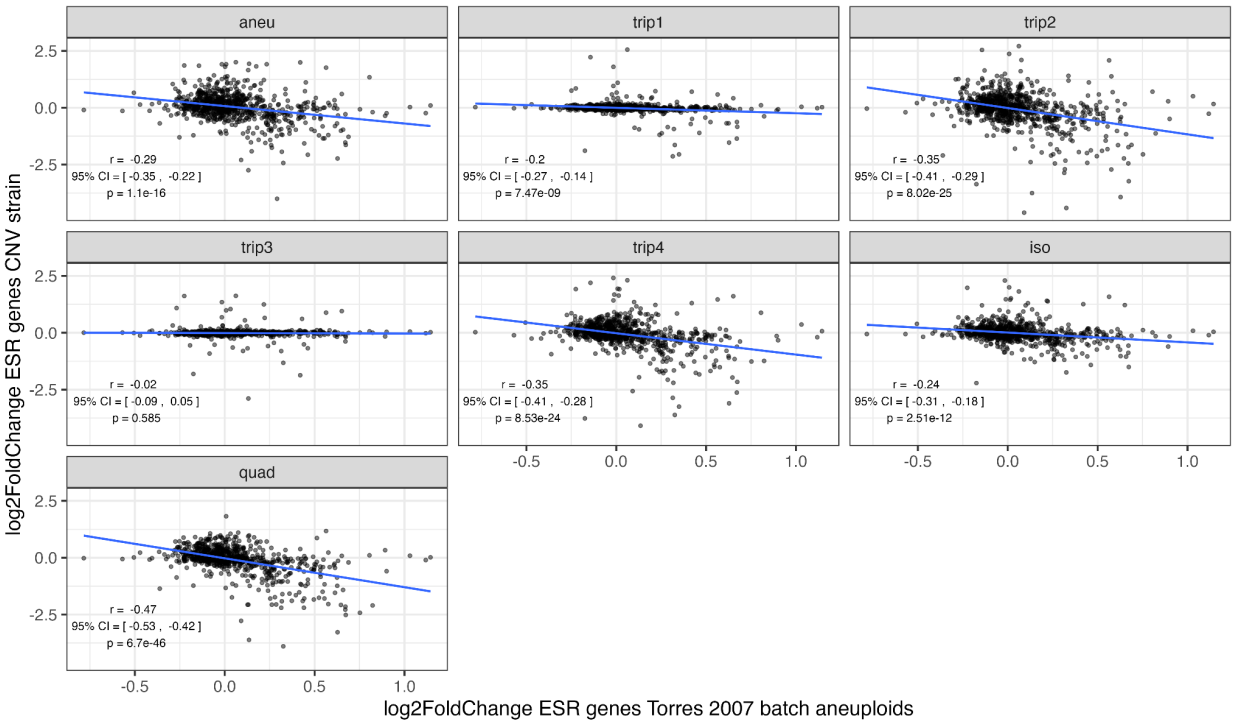

Figure 4.S7 Pearson correlation between CNV strains and Torres 2007 aneuploids grown in batch culture for ESR genes. Log2 fold change in mRNA expression comparing CNV or aneuploid strain to euploid strain. The data from Torres is the mean for all aneuploid strains measured.

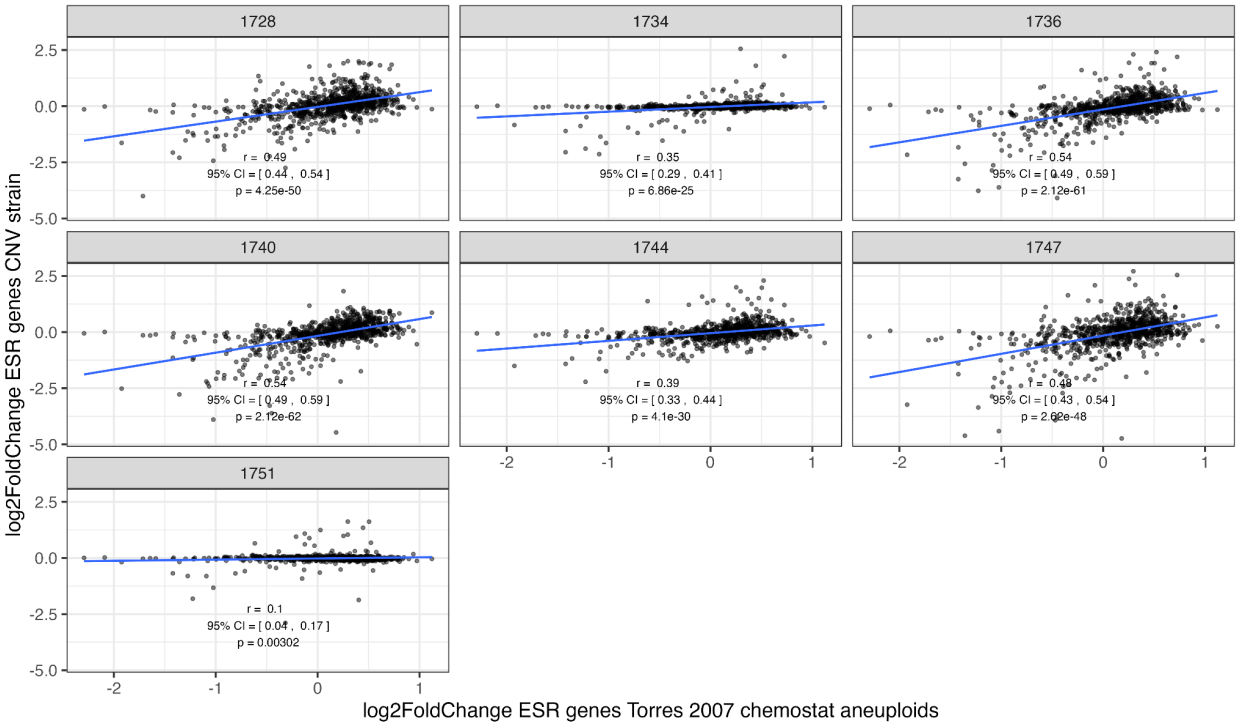

Figure 4.S8 Pearson correlation between CNV strains and Torres 2007 aneuploids grown in chemostats for ESR genes. Log2 fold change in mRNA expression comparing CNV or aneuploid strain to euploid strain. The data from Torres is the mean for all aneuploid strains measured.

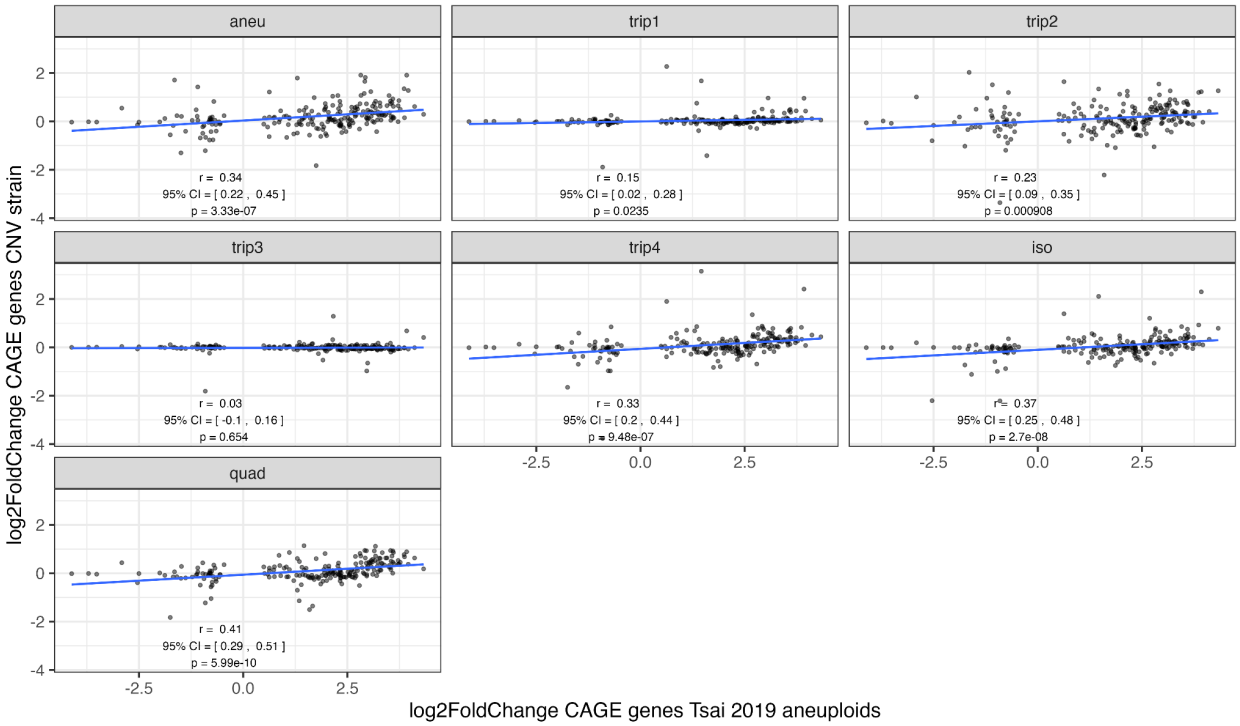

Figure 4.S9 Pearson correlation between CNV strains and Tsai 2019 aneuploids for CAGE genes. Log2 fold change in mRNA expression comparing CNV or aneuploid strain to euploid strain.

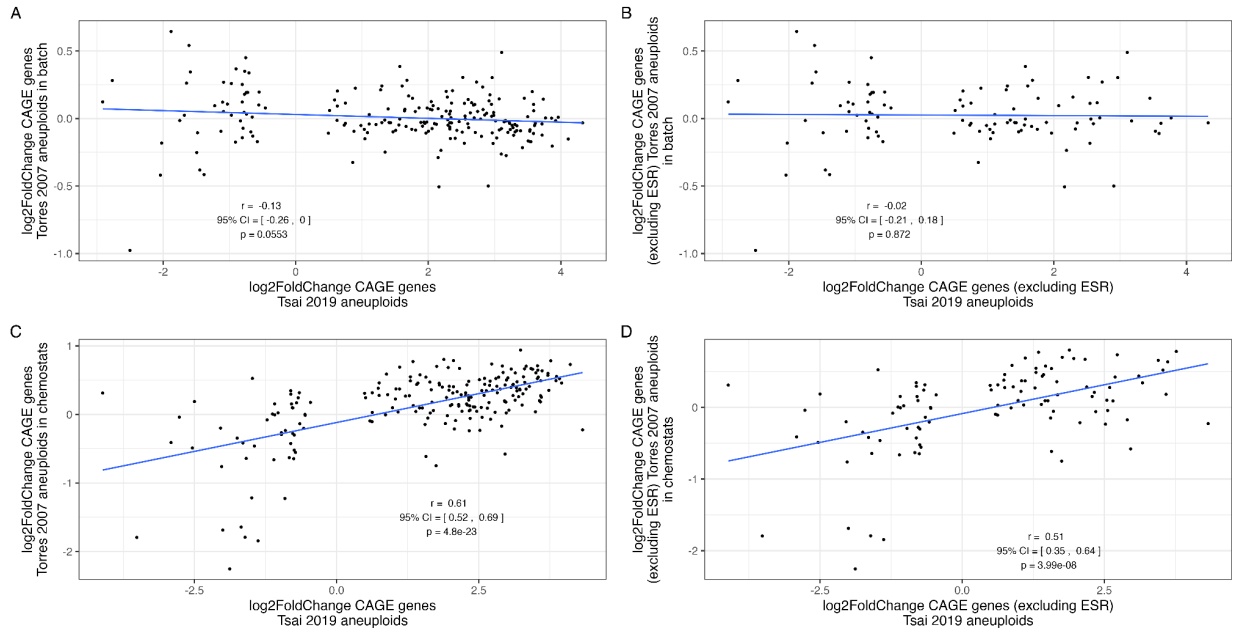

Figure 4.S10 Pearson correlation between Torres 2007 aneuploids and Tsai 2019 aneuploids for CAGE genes. The data from Torres is the mean for all aneuploid strains measured.

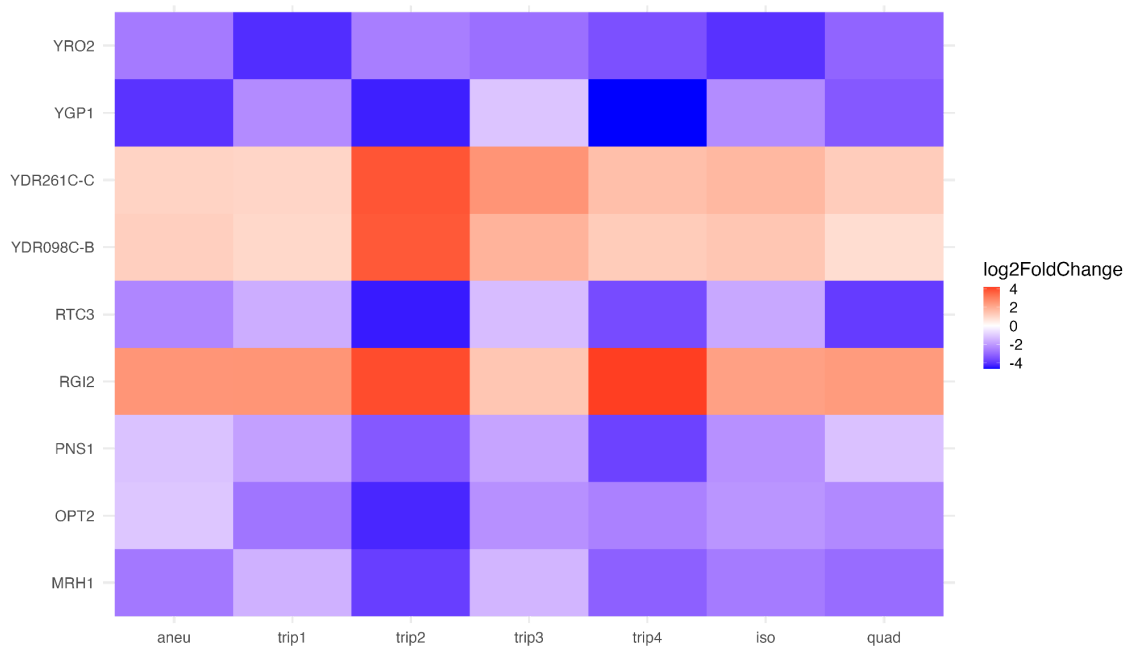

Figure 4.S11 Genes with significantly different mRNA expression from the euploid in all strains that are not on chromosome XI. Genes with positive log2FoldChange have higher expression in the CNV strain than the euploid strain.

Table 4.S1. Strain characteristics. More information about SNPs/indels including reference sequence and mutant sequence can be found in Lauer et al. 2018 S10 Table https://doi.org/10.1371/journal.pbio.3000069.s027.

Strain name	Gresham Lab Name	Clone ID in Lauer et al. 2018	Generation Isolated	SNPs/indels
eu	DGY1657	NA	NA	NA
aneu	DGY1728	gln_01_c1	150	YNL284C-B missense variant; YPL232W (SSO1) disruptive inframe deletion
trip1	DGY1734	gln_02_c3	250	YHL002W (HSE1) missense variant; Chr XIV:96555 non-coding variant; Chr XIV:96603 non-coding variant
trip2	DGY1747	gln_08_c2	150	YMR129W (POM152) missense variant; Chr V:431779 non-coding variant; Chr XII:915075 non-coding variant
trip3	DGY1751	gln_09_c3	250	YOL103W-A missense variant; YNR031C (SSK2) stop gained
trip4	DGY1736	gln_03_c2	250	YJR152W (DAL5) stop lost & splice region variant & conservative inframe deletion; Chr V:55180 non-coding variant; Chr X:524178 non-coding variant; Chr X:745685 non-coding variant
iso	DGY1744	gln_07_c1	250	YMR171C (EAR1) missense variant; YJL128C (PBS2) missense variant; Chr XV:594618 non-coding variant
quad	DGY1740	gln_05_c1	150	YOL077C (BRX1) missense variant; YNL338W frameshift_variant

Table 4.S2. Hermes mutagenesis library characteristics for uniquely identified insertion sites.

Sample	Total sites	Minimum reads per position	Maximum reads per position	Mean reads per position	Median reads per position
eu_1	172384	1	4761	20.09	8
eu_2	136167	1	2966	14.56	5
aneu	301220	1	26598	22.45	4
trip1	95152	1	2722	15.80	4
trip2	85327	1	2071	10.82	3
trip3	122326	1	8531	23.86	6
trip4	329624	1	10567	18.73	5
iso	126562	1	6620	23.58	6
quad	221218	1	8455	17.87	4

Table 4.S3 T-test for Log2FoldChange of gene expression for genes on chromosome 11. The column log2(copy number) indicates the expected log2 fold change based on the copy number of the amplified genes. The subset are either genes on chromosome XI that are amplified (i.e., part of the CNV) or genes on chromosome XI that are not amplified (i.e., not part of the CNV). The t-test tested if log2FoldChange for the group of genes was significantly different than zero (i.e., no change from euploid). Values are rounded to the nearest thousandth. 95% confidence intervals are shown.

strain	log2(copy number)	Mean log2 FC	group 1	group2	n	statistic	p	df	Con low	Conf high	H_a	subset
aneu	1	0.748	1	null model	334	22.596	0	326	0.683	0.813	two.sided	amplified genes
trip1	1.584	1.827	1	null model	22	15.493	0	19	1.58	2.074	two.sided	amplified genes
trip2	1.584	1.57	1	null model	46	12.71	0	41	1.32	1.819	two.sided	amplified genes
trip3	1.584	0.737	1	null model	65	11.526	0	62	0.609	0.865	two.sided	amplified genes
trip4	1.584	1.502	1	null model	79	23.635	0	76	1.375	1.628	two.sided	amplified genes
iso	1.584	1.395	1	null model	79	21.478	0	76	1.266	1.525	two.sided	amplified genes
quad	2	1.383	1	null model	18	10.056	0	15	1.09	1.676	two.sided	amplified genes
trip1		0.095	1	null model	293	4.583	0	292	0.054	0.136	two.sided	not amplified genes
trip2		0.003	1	null model	271	0.084	0.934	270	-0.072	0.078	two.sided	not amplified genes
trip3		-0.013	1	null model	253	-0.721	0.472	250	-0.048	0.022	two.sided	not amplified genes
trip4		-0.036	1	null model	239	-1.441	0.151	237	-0.085	0.013	two.sided	not amplified genes
iso		0.057	1	null model	238	1.721	0.087	236	-0.008	0.122	two.sided	not amplified genes
quad		-0.041	1	null model	297	-1.745	0.082	295	-0.086	0.005	two.sided	not amplified genes

Chapter 5: Conclusion

CNVs are a complex class of mutations, with important roles in evolution, and multifarious functional effects. In chapter two, I contributed to an investigation of the dynamics with which *GAP1* CNVs arise in populations of yeast in glutamine-limited chemostats. In chapter three, I used the population level dynamics observed in chapter two to infer the formation rate and selection coefficient associated with *GAP1* CNVs in glutamine-limited chemostats. In chapter four, I investigated the effect of diverse *GAP1* CNVs on strain fitness, genetic interactions, and mRNA expression.

5.1 Summary and Perspectives

5.1.1 Many competing *GAP1* CNVs contribute to rapid and repeatable adaptation

Using a fluorescent CNV reporter, we found that *GAP1* CNVs arise early in evolution in glutamine-limited chemostats and sweep through the population to rise to high frequencies, and this behavior is highly repeatable between replicate experimental populations. To determine if this behavior was due to a single lineage sweeping, or through many *GAP1* CNVs concurrently rising in frequency in a soft sweep, I combined the CNV reporter with a barcode lineage tracking library. I evolved these strains in glutamine-limited chemostats and found that in early generations there is extensive clonal interference, with hundreds to thousands of competing *GAP1* CNV lineages contributing to the rapid increase in frequency of *GAP1* CNVs in the population. However, very few CNVs ever rise to high frequency, and by the end of the experiment only 20-30 lineages remained. By analyzing which barcodes were found in the CNV subpopulation and when they arose, I determined that both standing variation and *de novo* variants contribute to these evolutionary dynamics. This study gave insights into the dynamics of CNVs in rapid adaptive evolution.

5.1.2 Simulation-based inference reveals *GAP1* CNVs have high rate and large effects

To estimate the rate at which CNVs are introduced and their fitness effects from the observed population evolutionary dynamics, I used likelihood-free simulation-based inference approaches. I compared the performance of two methods: Neural Posterior Estimation (NPE) and Approximate Bayesian Computation with Sequential Monte Carlo (ABC-SMC) with two evolutionary models, the Wright-Fisher model and the chemostat model. I found that NPE has several advantages over ABC-SMC, including more accurate estimation and reduced computational costs. I used NPE to estimate the *GAP1* CNV formation rate and effective selection coefficient from the *GAP1* CNV dynamics, and found they form at a rate of $10^{-4.7}$ - 10^{-4} CNVs per cell division, with selection coefficients of 0.04 - 0.1 per generation. I experimentally validated these estimates using barcode lineage tracking based estimation of CNV selection coefficients and pairwise fitness assays of CNV containing clones. This work demonstrated the utility of neural network based inference methods for estimation of evolutionary parameters from empirical data.

5.1.3 Diverse *GAP1* CNVs have common and strain specific effects

I investigated seven *GAP1* CNV strains with various structures which we had previously isolated from evolution experiments in glutamine-limited chemostats. I found that while CNV strains were as fit or more fit than the ancestral euploid strain in glutamine-limited chemostats, most were less fit in rich media. To investigate how CNVs impacted mutational tolerance, I performed transposon mutagenesis. I found that amplification of essential genes conferred new mutational tolerance, and CNVs result in novel genetic interactions. Several CNV strains had genetic interactions with genes involved in translation and mitochondrial function. However, I also observed strain specific genetic interactions. To better understand the functional effects of CNVs, I profiled the transcriptome of each CNV strain, and observed that while amplification results in increased gene expression relative to the euploid strain, some strains exhibited

dosage compensation, that is, the increase in mRNA was less than what would be expected based on the copy number. I did not observe previously described transcriptomic signatures of aneuploidy, instead, I observed that CNV strains tend to downregulate genes involved in cellular respiration, nucleoside biosynthetic processes, and small molecule metabolism, and upregulate genes involved in transposition, nucleic acid metabolic processes, and siderophore transport, though to different degrees in each strain. This study revealed the ways in which CNVs affect the mutational landscape and transcriptome.

It has been suggested that aneuploidy is a transient solution to strong and abrupt selective pressures, and that aneuploids will revert back to the euploid number of chromosomes after sufficient time has elapsed as to allow more specific mutations with fewer associated tradeoffs to arise (Yona et al. 2012). As my work has shown, different CNV structures may have different tradeoffs than aneuploidy. Would the tolerance of a particular genetic background to aneuploidy versus other CNV structures modulate this type of dynamic of CNV as a transient solution? Furthermore, cells can go from aneuploid to euploid (or vice versa) in a relatively simple manner, while more complex types of CNV may be less likely to revert to the ancestral state without causing further mutation or less likely to revert at all. How do all of these factors impact evolutionary trajectories? One might imagine that some genetic backgrounds might be more likely to tolerate certain CNVs. As I have shown, these CNVs then change the landscape of mutational tolerance and genetic interactions. This means that much of an evolutionary trajectories might be contingent on the type of CNV that arises in response to selection. If we can at some point predict which genetic backgrounds can tolerate which types of CNVs, this might give us insight into future evolutionary paths as well.

5.2 Future directions

5.2.1 Evolutionary dynamics of CNVs

It will be exciting to learn if the insights gleaned from this work, which focused on a single locus in a single genetic background, are broadly applicable. Performing similar evolution experiments in different genetic backgrounds, at different loci, and in different environmental conditions, and using simulation-based inference to infer the underlying rates and effects of CNVs, will provide further insight into what the general and condition specific aspects of CNV evolutionary dynamics.

The evolution experiments I have performed have been on relatively short timescales (less than 300 generations). It would be interesting to see what occurs if these experiments were carried out for longer periods, especially given the observed stochastic nature of the CNV dynamics in the later periods of the experimental evolution, and the reduction in the number of CNV lineages. While the labor of maintaining a chemostat evolution experiment and the probability of contamination are the primary reasons previous experiments have ended, we froze down whole population samples from those experiments. Populations of interest, such as the two populations that were barcode sequenced at several time points, could be used to start several replicate new evolution experiments. From these, several questions could be answered, including: does a single CNV eventually fix in the population, and if so, is it the same CNV lineage in each population? Are CNVs maintained over long time scales or are they eventually replaced by other variants that are not associated with costs? Do amplified genes accumulate further mutations that modify their function?

Finally, it would be very interesting to do further experiments exploring how the fitness effects and mutation rates of *GAP1* CNVs and other beneficial mutations interact with each other to shape evolutionary trajectories. Our experiments and inference procedures determined that in our evolution experiments *GAP1* CNV have high selection coefficients. The strength of

selection in chemostats can be modulated by increasing or decreasing the dilution rate, which in turn increases or decreases the steady-state concentration of the limiting nutrient (Gresham and Hong 2015). Theoretically, this may also modulate the fitness effect of *GAP1* CNVs. Similarly, the population size can be systematically varied by varying the concentration of the limiting nutrient in the feed media or by varying the volume in the growth vessel (Gresham and Hong 2015). This would allow investigation into how the extent of clonal interference between different CNV lineages, and between CNVs and other beneficial mutations affects evolutionary dynamics.

5.2.2 Estimating additional parameters underlying evolutionary dynamics

The models that I used in chapter three were relatively simple models, with only two classes of mutations (*GAP1* CNV and other beneficial mutations), and only two parameters were inferred (*GAP1* CNV formation rate and selection coefficient). Additionally, the selection coefficients used for each class of mutations were a single effective coefficient. This is a simplification and does not capture the complexity of the evolving population. Future work should expand the model to include additional parameters to represent different types of CNVs that might have different rates and effects (e.g., aneuploidy, small tandem duplications, complex CNV), and represent fitness effects with a distribution instead of a single effective selection coefficient. These parameters, as well as the rates and effects of other beneficial mutations, could be simultaneously inferred. This would give greater insight into the parameters underlying dynamics, as well as the effect on the relationship between different classes of mutation on evolutionary dynamics.

5.2.3 The basis of CNV (in)tolerance

In order to ascertain if the fitness defects in the three least fit strains in chapter 4 (aneu, trip2, and quad) are due to the CNVs per se or due to the interaction between the CNVs and some other variant elsewhere in the genome, the CNV strain could be mated to an euploid

strain, and sporulated. The resulting tetrads could then be genotyped for CNV, and phenotype fitness in YPGal and sensitivity to CCCP. If the traits of decreased growth rate in YPGal and increased sensitivity to CCCP segregate with the CNV, then that would suggest that those traits are associated with the CNV itself. If those traits are only sometimes associated with the CNV in the tetrads, that would suggest that they are due to a genetic interaction between the CNV and another variant that is segregating independently.

Recent work (Hose et al. 2020) points to a relationship between mitochondrial state, proteostatic state, and aneuploid fitness. It would be interesting to know if this relationship holds for the CNV strains studied here. In chapter four, I performed fitness assays in rich media with galactose, and the strains that were least fit in galactose were also the most sensitive to CCCP, a drug which interferes with mitochondrial function. Mitochondrial state could be further investigated by staining with a marker such as MitoTracker and using microscopy to examine mitochondrial morphology. Galactose is simultaneously fermented and respired during exponential growth. It would be interesting to see if CNV strains exhibit similar sensitivity to CCCP when growing with glucose as the carbon source, which is fermented during exponential growth. Furthermore, it would be interesting to know if the CNV strains are also particularly sensitive to drugs which cause proteostatic stress.

5.2.4 Transposon-mutagenesis as a way to explore many questions

Transposon-mutagenesis is a powerful tool to study mutational tolerance and genetic interactions. As discussed in chapter 4, it overcomes many of the limitations of synthetic genetic arrays, which have been used extensively to study genetic interactions. It also has advantages to CRISPR based methods to ask similar questions: it does not require synthesis of complex and expensive oligo libraries, it can create heterozygous mutations when there are multiple copies of the same gene (e.g., in diploids or CNV containing strains), and it does not require high transformation efficiency (Noorani, Bradley, and de la Rosa 2020). This means it can be

used to study a wide variety of questions in many different strain backgrounds with relative ease. In addition to the already discussed applications in investigations of the effects of CNVs, transposon mutagenesis could be used to investigate genetic interactions with other mutations, combined with reporters to discover regulators of different processes such as transcription, and used in other environments to investigate gene by environment interactions.

However, the current protocol for transposon mutagenesis does have some limitations. First, it requires galactose for induction, which greatly reduces the number of conditions in which the experiment can be performed. To overcome this, the transposase could be placed under another inducible promoter. The estradiol-inducible ZEV system, which uses an artificial transcription factor to rapidly and specifically activate transcription (McIsaac et al. 2013), would be a good candidate for this, as the artificial transcription factor could be integrated on to the same plasmid as the transposon. Second, the current protocol requires propagation in the induction environment (YPGal) for several days. This makes it difficult to distinguish between mutational tolerance and the fitness effect of mutations, since different insertion mutants will be competing with each other. To overcome this, the protocol could be altered so that induction occurs in soft agar. This involves suspending cells in low-percentage agarose medium, which allows colonies to form separately, avoiding competition between mutants (Panasyuk et al. 2004). This also makes the protocol less laborious to perform.

I believe that an optimized transposon mutagenesis protocol would be a very effective way to introduce students, particularly undergraduates, to experimental techniques in yeast and the experience of doing research. There is sufficient flexibility in experimental design (e.g., the strain used, the environment studied) that an undergraduate and their mentor could design an experiment together to answer a question that the student is interested in, giving them ownership over the project. However, the protocol itself is relatively simple, and introduces a variety of different techniques, including basic microbiology techniques including culturing and

transformation, DNA extraction and PCR for library preparation, and analysis of sequencing data. The timing of many steps of the protocol are flexible, and there are several places it can be paused and resumed, but the entire process if conducted un-interrupted takes only a couple of weeks, making it suitable for an undergraduate who is simultaneously taking classes, or one who is working in the lab full time but for only a short while (e.g., in the summer undergraduate research program). I believe that this technique could be a powerful tool to both involve students in research and answer interesting questions.

References

Adam, D., N. Dimitrijevic, and M. Schartl. 1993. “Tumor Suppression in Xiphophorus by an Accidentally Acquired Promoter.” *Science* 259 (5096): 816–19.

Adler, Marlen, Mehreen Anjum, Otto G. Berg, Dan I. Andersson, and Linus Sandegren. 2014. “High Fitness Costs and Instability of Gene Duplications Reduce Rates of Evolution of New Genes by Duplication-Divergence Mechanisms.” *Molecular Biology and Evolution* 31 (6): 1526–35.

Aggeli, Dimitra, Yuping Li, and Gavin Sherlock. n.d. “Changes in the Distribution of Fitness Effects and Adaptive Mutational Spectra Following a Single First Step towards Adaptation.” https://doi.org/10.1101/2020.06.12.148833.

Aguilera, Andrés, and Hélène Gaillard. 2014. “Transcription and Recombination: When RNA Meets DNA.” *Cold Spring Harbor Perspectives in Biology* 6 (8): a016543–a016543.

Aigner, Johanna, Sergi Villatoro, Raquel Rabionet, Jaume Roquer, Jordi Jiménez-Conde, Eulàlia Martí, and Xavier Estivill. 2013. “A Common 56-Kilobase Deletion in a Primate-Specific Segmental Duplication Creates a Novel Butyrophilin-like Protein.” *BMC Genetics* 14 (July): 61.

Airoldi, Edoardo M., Darach Miller, Rodoniki Athanasiadou, Nathan Brandt, Farah Abdul-Rahman, Benjamin Neymotin, Tatsu Hashimoto, Tayebeh Bahmani, and David Gresham. 2016. “Steady-State and Dynamic Gene Expression Programs in Saccharomyces Cerevisiae in Response to Variation in Environmental Nitrogen.” *Molecular Biology of the Cell* 27 (8): 1383–96.

Alsing, Justin, Tom Charnock, Stephen Feeney, and Benjamin Wandelt. 2019. “Fast Likelihood-Free Cosmology with Neural Density Estimators and Active Learning.” *Monthly Notices of the Royal Astronomical Society*. https://doi.org/10.1093/mnras/stz1960.

Anders, Kirk R., Julie R. Kudrna, Kirstie E. Keller, Breanna Kinghorn, Elizabeth M. Miller, Daniel Pauw, Anders T. Peck, Christopher E. Shellooe, and Isaac J. T. Strong. 2009. “A Strategy for Constructing Aneuploid Yeast Strains by Transient Nondisjunction of a Target Chromosome.” *BMC Genetics* 10 (July): 36.

Anderson, P., and J. Roth. 1981. “Spontaneous Tandem Genetic Duplications in Salmonella Typhimurium Arise by Unequal Recombination between rRNA (rrn) Cistrons.” *Proceedings of the National Academy of Sciences of the United States of America* 78 (5): 3113–17.

Anderson, R. P., and J. R. Roth. 1977. “Tandem Genetic Duplications in Phage and Bacteria.” *Annual Review of Microbiology* 31 (1): 473–505.

Andersson, D. I. 2015. “Improving Predictions of the Risk of Resistance Development against New and Old Antibiotics.” *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 21 (10): 894–98.

Arguello, J. Roman, Ying Chen, Shuang Yang, Wen Wang, and Manyuan Long. 2006. “Origination of an X-Linked Testes Chimeric Gene by Illegitimate Recombination in Drosophila.” *PLoS Genetics* 2 (5): e77.

Arita, Yuko, Griffin Kim, Zhijian Li, Helena Friesen, Gina Turco, Rebecca Y. Wang, Dale Climie, et al. 2021. “A Genome-Scale Yeast Library with Inducible Expression of Individual Genes.” *Molecular Systems Biology* 17 (6): e10207.

Arlt, Martin F., Jennifer G. Mulle, Valerie M. Schaibley, Ryan L. Ragland, Sandra G. Durkin, Stephen T. Warren, and Thomas W. Glover. 2009. “Replication Stress Induces Genome-Wide Copy Number Changes in Human Cells That Resemble Polymorphic and Pathogenic Variants.” *American Journal of Human Genetics* 84 (3): 339–50.

Arlt, Martin F., Sountharia Rajendran, Shanda R. Birkeland, Thomas E. Wilson, and Thomas W.

Glover. 2012. “De Novo CNV Formation in Mouse Embryonic Stem Cells Occurs in the Absence of Xrcc4-Dependent Nonhomologous End Joining.” *PLoS Genetics* 8 (9): e1002981.
Ascencio, Diana, Guillaume Diss, Isabelle Gagnon-Arsenault, Alexandre K. Dubé, Alexander DeLuna, and Christian R. Landry. 2021. “Expression Attenuation as a Mechanism of Robustness against Gene Duplication.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (6). https://doi.org/10.1073/pnas.2014345118.
“Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data.” n.d. Accessed August 28, 2020. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
Bailey, Susan F., and Thomas Bataillon. 2016. “Can the Experimental Evolution Programme Help Us Elucidate the Genetic Basis of Adaptation in Nature?” *Molecular Ecology* 25 (1): 203–18.
Bank, Claudia, Ryan T. Hietpas, Alex Wong, Daniel N. Bolon, and Jeffrey D. Jensen. 2014. “A Bayesian MCMC Approach to Assess the Complete Distribution of Fitness Effects of New Mutations: Uncovering the Potential for Adaptive Walks in Challenging Environments.” *Genetics*. https://doi.org/10.1534/genetics.113.156190.
Barra, V., and D. Fachinetti. 2018. “The Dark Side of Centromeres: Types, Causes and Consequences of Structural Abnormalities Implicating Centromeric DNA.” *Nature Communications* 9 (1): 4340.
Barreiro, Luis B., Guillaume Laval, Hélène Quach, Etienne Patin, and Lluís Quintana-Murci. 2008. “Natural Selection Has Driven Population Differentiation in Modern Humans.” *Nature Genetics* 40: 340–45.
Barrick, Jeffrey E., Mark R. Kauth, Christopher C. Strelioff, and Richard E. Lenski. 2010. “Escherichia Coli rpoB Mutants Have Increased Evolvability in Proportion to Their Fitness Defects.” *Molecular Biology and Evolution* 27 (6): 1338–47.
Baryshnikova, Anastasia, and Brenda Andrews. 2012. “Neighboring-Gene Effect: A Genetic Uncertainty Principle.” *Nature Methods*.
Beach, Rebecca R., Chiara Ricci-Tam, Christopher M. Brennan, Christine A. Moomau, Pei-Hsin Hsu, Bo Hua, Rebecca E. Silberman, Michael Springer, and Angelika Amon. 2017. “Aneuploidy Causes Non-Genetic Individuality.” *Cell* 169 (2): 229–42.e21.
Beaumont, Mark A. 2010. “Approximate Bayesian Computation in Evolution and Ecology,” November. https://doi.org/10.1146/annurev-ecolsys-102209-144621.
Beaumont, Mark A., Wenyang Zhang, and David J. Balding. 2002. “Approximate Bayesian Computation in Population Genetics.” *Genetics* 162 (4): 2025–35.
Bell, Michael A. 1987. “Interacting Evolutionary Constraints in Pelvic Reduction of Threespine Sticklebacks, Gasterosteus Aculeatus (Pisces, Gasterosteidae).” *Biological Journal of the Linnean Society. Linnean Society of London* 31 (4): 347–82.
Ben-David, Uri, and Angelika Amon. 2020. “Context Is Everything: Aneuploidy in Cancer.” *Nature Reviews. Genetics* 21 (1): 44–62.
Bennett, Albert F., and Richard E. Lenski. 2007. “An Experimental Test of Evolutionary Trade-Offs during Temperature Adaptation.” *Proceedings of the National Academy of Sciences of the United States of America* 104 Suppl 1 (May): 8649–54.
Ben-Shitrit, Taly, Nir Yosef, Keren Shemesh, Roded Sharan, Eytan Ruppin, and Martin Kupiec. 2012. “Systematic Identification of Gene Annotation Errors in the Widely Used Yeast Mutation Collections.” *Nature Methods* 9 (4): 373–78.
Bermudez-Santana, Clara, Camille Attolini, Toralf Kirsten, Jan Engelhardt, Sonja J. Prohaska, Stephan Steigele, Peter F. Stadler, et al. 2010. “Genomic Organization of Eukaryotic tRNAs.” *BMC Genomics* 11 (1): 270–270.
Birchler, James A., and Reiner A. Veitia. 2012. “Gene Balance Hypothesis: Connecting Issues

of Dosage Sensitivity across Biological Disciplines." *Proceedings of the National Academy of Sciences of the United States of America* 109 (37): 14746–53.
Black, Joshua C., Elnaz Atabakhsh, Jaegil Kim, Kelly M. Biette, Capucine Van Rechem, Brendon Ladd, Paul D. Burrowes, et al. 2015. "Hypoxia Drives Transient Site-Specific Copy Gain and Drug-Resistant Gene Expression." *Genes & Development* 29 (10): 1018–31.
Black, Joshua C., Hailei Zhang, Jaegil Kim, Gad Getz, and Johnathan R. Whetstine. 2016. "Regulation of Transient Site-Specific Copy Gain by MicroRNA." *The Journal of Biological Chemistry* 291 (10): 4862–71.
Blanquart, François, and Thomas Bataillon. 2016. "Epistasis and the Structure of Fitness Landscapes: Are Experimental Fitness Landscapes Compatible with Fisher's Geometric Model?" *Genetics*. https://doi.org/10.1534/genetics.115.182691.
Blount, Zachary D., Jeffrey E. Barrick, Carla J. Davidson, and Richard E. Lenski. 2012. "Genomic Analysis of a Key Innovation in an Experimental Escherichia Coli Population." *Nature* 489 (7417): 513–18.
Blount, Zachary D., Christina Z. Borland, and Richard E. Lenski. 2008. "Historical Contingency and the Evolution of a Key Innovation in an Experimental Population of Escherichia Coli." *Proceedings of the National Academy of Sciences of the United States of America* 105 (23): 7899–7906.
Blount, Zachary D., Richard E. Lenski, and Jonathan B. Losos. 2018. "Contingency and Determinism in Evolution: Replaying Life's Tape." *Science*. https://doi.org/10.1126/science.aam5979.
Blum, Michael G. B., and Olivier François. 2010. "Non-Linear Regression Models for Approximate Bayesian Computation." *Statistics and Computing*. https://doi.org/10.1007/s11222-009-9116-0.
Blundell, Jamie R., Katja Schwartz, Danielle Francois, Daniel S. Fisher, Gavin Sherlock, and Sasha F. Levy. 2019. "The Dynamics of Adaptive Genetic Diversity during the Early Stages of Clonal Evolution." *Nature Ecology & Evolution* 3 (2): 293–301.
Bonney, Megan E., Hisao Moriya, and Angelika Amon. 2015. "Aneuploid Proliferation Defects in Yeast Are Not Driven by Copy Number Changes of a Few Dosage-Sensitive Genes." *Genes & Development*. https://doi.org/10.1101/gad.261743.115.
Brauer, Matthew J., Curtis Huttenhower, Edoardo M. Airoldi, Rachel Rosenstein, John C. Matese, David Gresham, Viktor M. Boer, Olga G. Troyanskaya, and David Botstein. 2008. "Coordination of Growth Rate, Cell Cycle, Stress Response, and Metabolic Activity in Yeast." *Molecular Biology of the Cell* 19 (1): 352–67.
Brewer, Bonita J., Celia Payen, Sara C. Di Rienzi, Megan M. Higgins, Giang Ong, Maitreya J. Dunham, and M. K. Raghuraman. 2015. "Origin-Dependent Inverted-Repeat Amplification: Tests of a Model for Inverted DNA Amplification." *PLoS Genetics* 11 (12): e1005699–e1005699.
Brewer, Bonita J., Celia Payen, M. K. Raghuraman, and Maitreya J. Dunham. 2011. "Origin-Dependent Inverted-Repeat Amplification: A Replication-Based Model for Generating Palindromic Amplicons." *PLoS Genetics* 7 (3): e1002016–e1002016.
Brewster, Jeffrey D. 2003. "A Simple Micro-Growth Assay for Enumerating Bacteria." *Journal of Microbiological Methods*. https://doi.org/10.1016/s0167-7012(02)00226-9.
Bridges, C. B. 1936. "THE BAR 'GENE' A DUPLICATION." *Science* 83 (2148): 210–11.
Brooks, Aaron N., Amanda L. Hughes, Sandra Clauder-Münster, Leslie A. Mitchell, Jef D. Boeke, and Lars M. Steinmetz. 2022. "Transcriptional Neighborhoods Regulate Transcript Isoform Lengths and Expression Levels." *Science*. https://doi.org/10.1126/science.abg0162.
Brown, C. J., K. M. Todd, and R. F. Rosenzweig. 1998a. "Multiple Duplications of Yeast Hexose Transport Genes in Response to Selection in a Glucose-Limited Environment." *Molecular Biology and Evolution* 15 (8): 931–42.

———. 1998b. “Multiple Duplications of Yeast Hexose Transport Genes in Response to Selection in a Glucose-Limited Environment.” *Molecular Biology and Evolution* 15 (8): 931–42.
Brügger, Kim, Peter Redder, Qunxin She, Fabrice Confalonieri, Yvan Zivanovic, and Roger A. Garrett. 2002. “Mobile Elements in Archaeal Genomes.” *FEMS Microbiology Letters* 206 (2): 131–41.
Burke, Molly K., Joseph P. Dunham, Parvin Shahrestani, Kevin R. Thornton, Michael R. Rose, and Anthony D. Long. 2010. “Genome-Wide Analysis of a Long-Term Evolution Experiment with Drosophila.” *Nature* 467 (7315): 587–90.
Bussotti, Giovanni, Evi Gouzelou, Mariana Cortes Boite, Ihcen Kherachi, and Gerald F. Spath. 2018. “Leishmania Genome Dynamics during Environmental Adaptation Reveals Strain-Specific Differences in Gene Copy Number Variation, Karyotype Instability, and Telomeric Amplification.” *mBio*, September. http://dx.doi.org/.
Cairns, J., and P. L. Foster. 1991. “Adaptive Reversion of a Frameshift Mutation in Escherichia Coli.” *Genetics* 128 (4): 695–701.
Camougrand, Nadine, Angela Grelaud-Coq, Esther Marza, Muriel Priault, Jean-Jacques Bessoule, and Stéphen Manon. 2003. “The Product of the UTH1 Gene, Required for Bax-Induced Cell Death in Yeast, Is Involved in the Response to Rapamycin.” *Molecular Microbiology* 47 (2): 495–506.
Camougrand, Nadine, Ingrid Kiššová, Gisèle Velours, and Stéphen Manon. 2004. “Uth1p: A Yeast Mitochondrial Protein at the Crossroads of Stress, Degradation and Cell Death.” *FEMS Yeast Research* 5 (2): 133–40.
Cardoso-Moreira, Margarida, J. Roman Arguello, and Andrew G. Clark. 2012. “Mutation Spectrum of Drosophila CNVs Revealed by Breakpoint Sequencing.” *Genome Biology* 13 (12): R119–R119.
Carr, Martin, Douda Bensasson, and Casey M. Bergman. 2012. “Evolutionary Genomics of Transposable Elements in Saccharomyces Cerevisiae.” *PloS One* 7 (11): e50978.
Carvalho, Claudia M. B., and James R. Lupski. 2016. “Mechanisms Underlying Structural Variant Formation in Genomic Disorders.” *Nature Reviews. Genetics* 17 (4): 224–38.
Casola, Claudio, and Esther Betrán. 2017. “The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses?” *Genome Biology and Evolution* 9 (6): 1351–73.
Caudal, Elodie, Anne Friedrich, Arthur Jallet, Marion Garin, Jing Hou, and Joseph Schacherer. 2021. “Population-Level Survey of Loss-of-Function Mutations Revealed That Background Dependent Fitness Genes Are Rare and Functionally Related in Yeast.” *bioRxiv*. https://doi.org/10.1101/2021.08.25.457624.
Chain, Frédéric J. J., Jullien M. Flynn, James K. Bull, and Melania E. Cristescu. 2019. “Accelerated Rates of Large-Scale Mutations in the Presence of Copper and Nickel.” *Genome Research* 29 (1): 64–73.
Chan, Yingguang Frank, Melissa E. Marks, Felicity C. Jones, Guadalupe Villarreal Jr, Michael D. Shapiro, Shannon D. Brady, Audrey M. Southwick, et al. 2010. “Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer.” *Science* 327 (5963): 302–5.
Chen, Guangbo, William D. Bradford, Chris W. Seidel, and Rong Li. 2012. “Hsp90 Stress Potentiates Rapid Cellular Adaptation through Induction of Aneuploidy.” *Nature* 482 (7384): 246–50.
Chen, Lu, Weichen Zhou, Cheng Zhang, James R. Lupski, Li Jin, and Feng Zhang. 2015. “CNV Instability Associated with DNA Replication Dynamics: Evidence for Replicative Mechanisms in CNV Mutagenesis.” *Human Molecular Genetics* 24 (6): 1574–83.
Chevin, Luis-Miguel. 2011. “On Measuring Selection in Experimental Evolution.” *Biology Letters*.

https://doi.org/10.1098/rsbl.2010.0580.
Clop, A., O. Vidal, and M. Amills. 2012. “Copy Number Variation in the Genomes of Domestic Animals.” *Animal Genetics* 43 (5): 503–17.
Cohen, S., and D. Segal. 2009. “Extrachromosomal Circular DNA in Eukaryotes: Possible Involvement in the Plasticity of Tandem Repeats.” *Cytogenetic and Genome Research* 124 (3-4): 327–38.
Colizzi, Enrico Sandro, and Paulien Hogeweg. 2019. “Transcriptional Mutagenesis Prevents Ribosomal DNA Deterioration: The Role of Duplications and Deletions.” *Genome Biology and Evolution* 11 (11): 3207–17.
Conant, Gavin C., and Kenneth H. Wolfe. 2008. “Turning a Hobby into a Job: How Duplicated Genes Find New Functions.” *Nature Reviews. Genetics* 9 (12): 938–50.
Copley, Shelley D. 2012. “Toward a Systems Biology Perspective on Enzyme Evolution.” *The Journal of Biological Chemistry* 287 (1): 3–10.
Costanzo, Michael, Jing Hou, Vincent Messier, Justin Nelson, Mahfuzur Rahman, Benjamin VanderSluis, Wen Wang, et al. 2021. “Environmental Robustness of the Global Yeast Genetic Interaction Network.” *Science* 372 (6542). https://doi.org/10.1126/science.abf8424.
Cowell, Annie N., Eva S. Istvan, Amanda K. Lukens, Maria G. Gomez-Lorenzo, Manu Vanaerschot, Tomoyo Sakata-Kato, Erika L. Flannery, et al. 2018. “Mapping the Malaria Parasite Druggable Genome by Using in Vitro Evolution and Chemogenomics.” *Science* 359 (6372): 191–99.
Cranmer, Kyle, Johann Brehmer, and Gilles Louppe. 2020. “The Frontier of Simulation-Based Inference.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (48): 30055–62.
Crow, James Franklin, and Motoo Kimura. 1970. *An Introduction to Population Genetics Theory*. Burgess International Group.
Csilléry, Katalin, Olivier François, and Michael G. B. Blum. 2012. “Abc: An R Package for Approximate Bayesian Computation (ABC).” *Methods in Ecology and Evolution*. https://doi.org/10.1111/j.2041-210x.2011.00179.x.
Cunha, Fernanda Marques da, Nicole Quesada Torelli, and Alicia J. Kowaltowski. 2015. “Mitochondrial Retrograde Signaling: Triggers, Pathways, and Outcomes.” *Oxidative Medicine and Cellular Longevity* 2015 (October): 482582.
Darwin, Charles. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*.
Dean, Antony M. 2005. “Protecting Haploid Polymorphisms in Temporally Variable Environments.” *Genetics*. https://doi.org/10.1534/genetics.104.036053.
DeBolt, Seth. 2010. “Copy Number Variation Shapes Genome Diversity in Arabidopsis over Immediate Family Generational Scales.” *Genome Biology and Evolution* 2 (July): 441–53.
Dephoure, Noah, Sunyoung Hwang, Ciara O’Sullivan, Stacie E. Dodgson, Steven P. Gygi, Angelika Amon, and Eduardo M. Torres. 2014. “Quantitative Proteomic Analysis Reveals Posttranslational Responses to Aneuploidy in Yeast.” *eLife* 3 (July): e03023.
Desai, Michael M., and Daniel S. Fisher. 2011. “The Balance between Mutators and Nonmutators in Asexual Populations.” *Genetics* 188 (4): 997–1014.
Destruelle, M., H. Holzer, and D. J. Klionsky. 1994. “Identification and Characterization of a Novel Yeast Gene: The YGP1 Gene Product Is a Highly Glycosylated Secreted Protein That Is Synthesized in Response to Nutrient Limitation.” *Molecular and Cellular Biology* 14 (4): 2740–54.
Dhami, Manpreet K., Thomas Hartwig, and Tadashi Fukami. 2016. “Genetic Basis of Priority Effects: Insights from Nectar Yeast.” *Proceedings. Biological Sciences / The Royal Society* 283 (1840). https://doi.org/10.1098/rspb.2016.1455.
Di Rienzi, Sara C., David Collingwood, M. K. Raghuraman, and Bonita J. Brewer. 2009. “Fragile

Genomic Sites Are Associated with Origins of Replication." *Genome Biology and Evolution* 1: 350–63.
Dodgson, Stacie E., Sharon Kim, Michael Costanzo, Anastasia Baryshnikova, Darcy L. Morse, Chris A. Kaiser, Charles Boone, and Angelika Amon. 2016. "Chromosome-Specific and Global Effects of Aneuploidy in Saccharomyces Cerevisiae." *Genetics* 202 (4): 1395–1409.
Dolatabadian, Aria, Dhwani Apurva Patel, David Edwards, and Jacqueline Batley. 2017. "Copy Number Variation and Disease Resistance in Plants." *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 130 (12): 2479–90.
Dombek, Kenneth M., Nataly Kacherovsky, and Elton T. Young. 2004. "The Reg1-Interacting Proteins, Bmh1, Bmh2, Ssb1, and Ssb2, Have Roles in Maintaining Glucose Repression in Saccharomyces Cerevisiae." *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.m400433200.
Domingo, Júlia, Pablo Baeza-Centurion, and Ben Lehner. 2019. "The Causes and Consequences of Genetic Interactions (Epistasis)." *Annual Review of Genomics and Human Genetics* 20 (August): 433–60.
Domitrovic, Tatiana, Guennadi Kozlov, João Claudio Gonçalves Freire, Claudio Akio Masuda, Marcius da Silva Almeida, Mónica Montero-Lomeli, Georgia Correa Atella, Edna Matta-Camacho, Kalle Gehring, and Eleonora Kurtenbach. 2010. "Structural and Functional Study of Yer067w, a New Protein Involved in Yeast Metabolism Control and Drug Resistance." *PLoS ONE*. https://doi.org/10.1371/journal.pone.0011163.
Dorsey, M., C. Peterson, K. Bray, and C. E. Paquin. 1992. "Spontaneous Amplification of the ADH4 Gene in Saccharomyces Cerevisiae." *Genetics* 132 (4): 943–50.
Douglas, Alison C., Andrew M. Smith, Sara Sharifpoor, Zhun Yan, Tanja Durbic, Lawrence E. Heisler, Anna Y. Lee, et al. 2012. "Functional Analysis with a Barcoder Yeast Gene Overexpression System." *G3* 2 (10): 1279–89.
Dulmage, Keely A., Cynthia L. Darnell, Angie Vreugdenhil, and Amy K. Schmid. 2018. "Copy Number Variation Is Associated with Gene Expression Change in Archaea." *Microbial Genomics*, August. https://doi.org/10.1099/mgen.0.000210.
Dunham, Maitreya J., Hassan Badrane, Tracy Ferea, Julian Adams, Patrick O. Brown, Frank Rosenzweig, and David Botstein. 2002. "Characteristic Genome Rearrangements in Experimental Evolution of Saccharomyces Cerevisiae." *Proceedings of the National Academy of Sciences of the United States of America* 99 (25): 16144–49.
Durkan, Conor, Artur Bekasov, Iain Murray, and George Papamakarios. 2019. "Neural Spline Flows." *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1906.04032.
Durkin, Sandra G., Ryan L. Ragland, Martin F. Arlt, Jennifer G. Mulle, Stephen T. Warren, and Thomas W. Glover. 2008. "Replication Stress Induces Tumor-like Microdeletions in FHIT/FRA3B." *Proceedings of the National Academy of Sciences of the United States of America* 105 (1): 246–51.
Eichler, E. E. 2001. "Recent Duplication, Domain Accretion and the Dynamic Mutation of the Human Genome." *Trends in Genetics: TIG* 17 (11): 661–69.
Elde, Nels C., Stephanie J. Child, Michael T. Eickbush, Jacob O. Kitzman, Kelsey S. Rogers, Jay Shendure, Adam P. Geballe, and Harmit S. Malik. 2012. "Poxviruses Deploy Genomic Accordions to Adapt Rapidly against Host Antiviral Defenses." *Cell* 150 (4): 831–41.
Ellis, B., P. Haaland, F. Hahne, N. L. Meur, N. Gopalakrishnan, and Spidlen J And Jiang. 2016. *flowCore: flowCore: Basic Structures for Flow Cytometry Data*.
Ewing, Adam D., Tracy J. Ballinger, Dent Earl, Broad Institute Genome Sequencing and Analysis Program and Platform, Christopher C. Harris, Li Ding, Richard K. Wilson, and David Haussler. 2013. "Retrotransposition of Gene Transcripts Leads to Structural Variation in Mammalian Genomes." *Genome Biology* 14 (3): R22.
Farslow, James C., Kendra J. Lipinski, Lucille B. Packard, Mark L. Edgley, Jon Taylor, Stephane

Flibotte, Donald G. Moerman, Vaishali Katju, and Ulfar Bergthorsson. 2015. “Rapid Increase in Frequency of Gene Copy-Number Variants during Experimental Evolution in Caenorhabditis Elegans.” *BMC Genomics* 16 (December): 1044.
Fendt, Sarah-Maria, and Uwe Sauer. 2010. “Transcriptional Regulation of Respiration in Yeast Metabolizing Differently Repressive Carbon Substrates.” *BMC Systems Biology* 4 (February): 12.
Ferenci, Thomas, and Ram Maharjan. 2015. “Mutational Heterogeneity: A Key Ingredient of Bet-Hedging and Evolutionary Divergence?” *BioEssays*. https://doi.org/10.1002/bies.201400153.
Feuk, Lars, Andrew R. Carson, and Stephen W. Scherer. 2006. “Structural Variation in the Human Genome.” *Nature Reviews. Genetics* 7 (2): 85–97.
Fisher, Kaitlin J., Sean W. Buskirk, Ryan C. Vignogna, Daniel A. Marad, and Gregory I. Lang. 2018. “Adaptive Genome Duplication Affects Patterns of Molecular Evolution in Saccharomyces Cerevisiae.” *PLoS Genetics* 14 (5): e1007396.
Fitzgerald, Devon M., P. J. Hastings, and Susan M. Rosenberg. 2017. “Stress-Induced Mutagenesis: Implications in Cancer and Drug Resistance.” *Annual Review of Cancer Biology* 1 (March): 119–40.
Fitzgerald, Devon M., and Susan M. Rosenberg. 2019. “What Is Mutation? A Chapter in the Series: How Microbes ‘jeopardize’ the Modern Synthesis.” *PLOS Genetics*. https://doi.org/10.1371/journal.pgen.1007995.
Flagel, Lex, Yaniv Brandvain, and Daniel R. Schrider. 2019. “The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference.” *Molecular Biology and Evolution* 36 (2): 220–38.
Foster, Patricia L. 2007. “Stress-Induced Mutagenesis in Bacteria.” *Critical Reviews in Biochemistry and Molecular Biology* 42 (5): 373–97.
Franke, Martin, Daniel M. Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schöpflin, Katerina Kraft, et al. 2016. “Formation of New Chromatin Domains Determines Pathogenicity of Genomic Duplications.” *Nature* 538 (7624): 265–69.
Freeling, Michael, Michael J. Scanlon, and John E. Fowler. 2015. “Fractionation and Subfunctionalization Following Genome Duplications: Mechanisms That Drive Gene Content and Their Consequences.” *Current Opinion in Genetics & Development* 35 (December): 110–18.
Frenkel, Evgeni M., Benjamin H. Good, and Michael M. Desai. 2014. “The Fates of Mutant Lineages and the Distribution of Fitness Effects of Beneficial Mutations in Laboratory Budding Yeast Populations.” *Genetics* 196 (4): 1217–26.
Frickel, Jens, Philine G. D. Feulner, Emre Karakoc, and Lutz Becks. 2018. “Population Size Changes and Selection Drive Patterns of Parallel Evolution in a Host–virus System.” *Nature Communications* 9 (1): 1–10.
Gaisne, M., A. M. Bécam, J. Verdière, and C. J. Herbert. 1999. “A ‘Natural’ Mutation in Saccharomyces Cerevisiae Strains Derived from S288c Affects the Complex Regulatory Gene HAP1 (CYP1).” *Current Genetics* 36 (4): 195–200.
Gale, Andrew N., Rima M. Sakhawala, Anton Levitan, Roded Sharan, Judith Berman, Winston Timp, and Kyle W. Cunningham. 2020. “Identification of Essential Genes and Fluconazole Susceptibility Genes in by Profiling Transposon Insertions.” *G3* 10 (10): 3859–70.
Galhardo, Rodrigo S., P. J. Hastings, and Susan M. Rosenberg. 2007. “Mutation as a Stress Response and the Regulation of Evolvability.” *Critical Reviews in Biochemistry and Molecular Biology* 42 (5): 399–435.
Gallet, Romain, Tim F. Cooper, Santiago F. Elena, and Thomas Lenormand. 2012. “Measuring Selection Coefficients below 10(-3): Method, Questions, and Prospects.” *Genetics* 190 (1): 175–86.

Gallone, Brigida, Jan Steensels, Troels Prahl, Leah Soriaga, Veerle Saels, Beatriz Herrera-Malaver, Adriaan Merlevede, et al. 2016. “Domestication and Divergence of Saccharomyces Cerevisiae Beer Yeasts.” *Cell* 166 (6): 1397–1410.e16.
Gamazon, Eric R., Dan L. Nicolae, and Nancy J. Cox. 2011. “A Study of CNVs as Trait-Associated Polymorphisms and as Expression Quantitative Trait Loci.” *PLoS Genetics* 7 (2): e1001292.
Gangadharan, Sunil, Loris Mularoni, Jennifer Fain-Thornton, Sarah J. Wheelan, and Nancy L. Craig. 2010. “DNA Transposon Hermes Inserts into DNA in Nucleosome-Free Regions in Vivo.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (51): 21966–72.
Gao, Yuxia, Huayao Zhao, Yin Jin, Xiaoyu Xu, and Guan-Zhu Han. 2017. “Extent and Evolution of Gene Duplication in DNA Viruses.” *Virus Research* 240 (August): 161–65.
Garland, Theodore, and Michael Robertson Rose. 2009. *Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments*.
Gasch, Audrey P., James Hose, Michael A. Newton, Maria Sardi, Mun Yong, and Zhishi Wang. 2016. “Further Support for Aneuploidy Tolerance in Wild Yeast and Effects of Dosage Compensation on Gene Copy-Number Evolution.” *eLife* 5 (March): e14409.
Gasch, Audrey P., Paul T. Spellman, Camilla M. Kao, Orna Carmel-Harel, Michael B. Eisen, Gisela Storz, David Botstein, and Patrick O. Brown. 2000. “Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes.” *Molecular Biology of the Cell*. https://doi.org/10.1091/mbc.11.12.4241.
Geiger, Tamar, Juergen Cox, and Matthias Mann. 2010. “Proteomic Changes Resulting from Gene Copy Number Variations in Cancer Cells.” *PLoS Genetics* 6 (9): e1001090–e1001090.
Gelbart, W. M., and A. Chovnick. 1979. “Spontaneous Unequal Exchange in the Rosy Region of Drosophila Melanogaster.” *Genetics* 92 (3): 849–59.
Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis, Third Edition*. CRC Press.
Gerrish, P. J., and R. E. Lenski. 1998. “The Fate of Competing Beneficial Mutations in an Asexual Population.” *Genetica* 102-103 (1-6): 127–44.
Gerstein, Aleeza C., Jasmine Ono, Dara S. Lo, Marcus L. Campbell, Anastasia Kuzmin, and Sarah P. Otto. 2015. “Too Much of a Good Thing: The Unique and Repeated Paths toward Copper Adaptation.” *Genetics* 199 (2): 555–71.
Giaever, Guri, and Corey Nislow. 2014. “The Yeast Deletion Collection: A Decade of Functional Genomics.” *Genetics* 197 (2): 451–65.
Gietz, R. Daniel, and Robert H. Schiestl. 2007a. “Frozen Competent Yeast Cells That Can Be Transformed with High Efficiency Using the LiAc/SS Carrier DNA/PEG Method.” *Nature Protocols* 2 (1): 1–4.
———. 2007b. “High-Efficiency Yeast Transformation Using the LiAc/SS Carrier DNA/PEG Method.” *Nature Protocols* 2: 31–34.
Gillespie, Daniel T. 2001. “Approximate Accelerated Stochastic Simulation of Chemically Reacting Systems.” *The Journal of Chemical Physics*. https://doi.org/10.1063/1.1378322.
Gillespie, John H. 1984. “MOLECULAR EVOLUTION OVER THE MUTATIONAL LANDSCAPE.” *Evolution; International Journal of Organic Evolution* 38 (5): 1116–29.
———. 1991. *The Causes of Molecular Evolution*. Oxford University Press.
Girirajan, Santhosh, Catarina D. Campbell, and Evan E. Eichler. 2011. “Human Copy Number Variation and Complex Genetic Disease.” *Annual Review of Genetics* 45 (August): 203–26.
Gomez, Kevin, Jason Bertram, and Joanna Masel. 2020. “Mutation Bias Can Shape Adaptation in Large Asexual Populations Experiencing Clonal Interference.” *Proceedings. Biological Sciences / The Royal Society* 287 (1937): 20201503.

Gonçalves, Pedro J., Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan Öcal, Giacomo Bassetto, Chaitanya Chintaluri, et al. 2020. “Training Deep Neural Density Estimators to Identify Mechanistic Models of Neural Dynamics.” *eLife* 9 (September). https://doi.org/10.7554/eLife.56261.
Good, Benjamin H., and Michael M. Desai. 2015. “The Impact of Macroscopic Epistasis on Long-Term Evolutionary Dynamics.” *Genetics* 199 (1): 177–90.
Good, Benjamin H., Michael J. McDonald, Jeffrey E. Barrick, Richard E. Lenski, and Michael M. Desai. 2017. “The Dynamics of Molecular Evolution over 60,000 Generations.” *Nature* 551 (7678): 45–50.
Graves, J. L., Jr, K. L. Hertweck, M. A. Phillips, M. V. Han, L. G. Cabral, T. T. Barter, L. F. Greer, M. K. Burke, L. D. Mueller, and M. R. Rose. 2017. “Genomics of Parallel Experimental Evolution in Drosophila.” *Molecular Biology and Evolution* 34 (4): 831–42.
Grech, Leanne, Daniel C. Jeffares, Christoph Y. Sadée, María Rodríguez-López, Danny A. Bitton, Mimoza Hoti, Carolina Biagosch, et al. 2019. “Fitness Landscape of the Fission Yeast Genome.” *Molecular Biology and Evolution* 36 (8): 1612–23.
Greenberg, David S., Marcel Nonnenmacher, and Jakob H. Macke. 2019. “Automatic Posterior Transformation for Likelihood-Free Inference.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1905.07488.
Greenblum, Sharon, Rogan Carr, and Elhanan Borenstein. 2015. “Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species.” *Cell* 160 (4): 583–94.
Green, Michael R., and Joseph Sambrook. 2016. “Precipitation of DNA with Ethanol.” *Cold Spring Harbor Protocols* 2016 (12). https://doi.org/10.1101/pdb.prot093377.
Grenson, M., C. Hou, and M. Crabeel. 1970. “Multiplicity of the Amino Acid Permeases in Saccharomyces Cerevisiae. IV. Evidence for a General Amino Acid Permease.” *Journal of Bacteriology* 103 (3): 770–77.
Gresham, David, Michael M. Desai, Cheryl M. Tucker, Harry T. Jenq, Dave A. Pai, Alexandra Ward, Christopher G. DeSevo, David Botstein, and Maitreya J. Dunham. 2008. “The Repertoire and Dynamics of Evolutionary Adaptations to Controlled Nutrient-Limited Environments in Yeast.” *PLoS Genetics* 4 (12): e1000303.
Gresham, David, and Maitreya J. Dunham. 2014. “The Enduring Utility of Continuous Culturing in Experimental Evolution.” *Genomics* 104 (6): 399–405.
Gresham, David, and Jungeui Hong. 2015. “The Functional Basis of Adaptive Evolution in Chemostats.” *FEMS Microbiology Reviews* 39 (1): 2–16.
Gresham, David, Renata Usaite, Susanne Manuela Germann, Michael Lisby, David Botstein, and Birgitte Regenberg. 2010. “Adaptation to Diverse Nitrogen-Limited Environments by Deletion or Extrachromosomal Element Formation of the GAP1 Locus.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (43): 18551–56.
Griesbeck, O., G. S. Baird, R. E. Campbell, D. A. Zacharias, and R. Y. Tsien. 2001. “Reducing the Environmental Sensitivity of Yellow Fluorescent Protein. Mechanism and Applications.” *The Journal of Biological Chemistry* 276 (31): 29188–94.
Gruber, Jonathan D., Kara Vogel, Gizem Kalay, and Patricia J. Wittkopp. 2012. “Contrasting Properties of Gene-Specific Regulatory, Coding, and Copy Number Mutations in Saccharomyces Cerevisiae: Frequency, Effects, and Dominance.” *PLoS Genetics* 8 (2): e1002497.
Guo, Yabin, Jung Min Park, Bowen Cui, Elizabeth Humes, Sunil Gangadharan, Stevephen Hung, Peter C. FitzGerald, et al. 2013. “Integration Profiling of Gene Function with Dense Maps of Transposon Integration.” *Genetics* 195 (2): 599–609.
Gu, Zhenglong, Lars M. Steinmetz, Xun Gu, Curt Scharfe, Ronald W. Davis, and Wen-Hsiung Li. 2003. “Role of Duplicate Genes in Genetic Robustness against Null Mutations.” *Nature* 421 (6918): 63–66.

Hallatschek, Oskar. 2011. “The Noisy Edge of Traveling Waves.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (5): 1783–87.
Hall, David W., Rod Mahmoudizad, Andrew W. Hurd, and Sarah B. Joseph. 2008. “Spontaneous Mutations in Diploid Saccharomyces Cerevisiae: Another Thousand Cell Generations.” *Genetics Research* 90 (3): 229–41.
Hansche, P. E. 1975. “Gene Duplication as a Mechanism of Genetic Adaptation in Saccharomyces Cerevisiae.” *Genetics* 79 (4): 661–74.
Harari, Yaniv, Yoav Ram, and Martin Kupiec. 2018. “Frequent Ploidy Changes in Growing Yeast Cultures.” *Current Genetics* 64 (5): 1001–4.
Harari, Yaniv, Yoav Ram, Nimrod Rappoport, Lilach Hadany, and Martin Kupiec. 2018. “Spontaneous Changes in Ploidy Are Common in Yeast.” *Current Biology: CB* 28 (6): 825–35.e4.
Harrison, Marie-Claire, Abigail L. LaBella, Chris Todd Hittinger, and Antonis Rokas. 2021. “The Evolution of the GALactose Utilization Pathway in Budding Yeasts.” *Trends in Genetics*. https://doi.org/10.1016/j.tig.2021.08.013.
Harrison, Xavier A., Lynda Donaldson, Maria Eugenia Correa-Cano, Julian Evans, David N. Fisher, Cecily E. D. Goodwin, Beth S. Robinson, David J. Hodgson, and Richard Inger. 2018. “A Brief Introduction to Mixed Effects Modelling and Multi-Model Inference in Ecology.” *PeerJ* 6 (May): e4794.
Hastings, P. J., H. J. Bull, J. R. Klump, and S. M. Rosenberg. 2000. “Adaptive Amplification: An Inducible Chromosomal Instability Mechanism.” *Cell* 103 (5): 723–31.
Hastings, P. J., Grzegorz Ira, James R. Lupski, A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, et al. 2009. “A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation.” *PLoS Genetics* 5 (1): e1000327–e1000327.
Hastings, P. J., James R. Lupski, Susan M. Rosenberg, and Grzegorz Ira. 2009. “Mechanisms of Change in Gene Copy Number.” *Nature Reviews. Genetics* 10 (8): 551–64.
Hegreness, Matthew, Noam Shoresh, Daniel Hartl, and Roy Kishony. 2006. “An Equivalence Principle for the Incorporation of Favorable Mutations in Asexual Populations.” *Science* 311 (5767): 1615–17.
Henrichsen, Charlotte N., Nicolas Vinckenbosch, Sebastian Zöllner, Evelyne Chaignat, Sylvain Pradervand, Frédéric Schütz, Manuel Ruedi, Henrik Kaessmann, and Alexandre Reymond. 2009. “Segmental Copy Number Variation Shapes Tissue Transcriptomes.” *Nature Genetics* 41 (4): 424–29.
Hermisson, Joachim, and Pleuni S. Pennings. 2005. “Soft Sweeps: Molecular Population Genetics of Adaptation from Standing Genetic Variation.” *Genetics* 169 (4): 2335–52.
———. 2017. “Soft Sweeps and beyond: Understanding the Patterns and Probabilities of Selection Footprints under Rapid Adaptation.” *Methods in Ecology and Evolution* 8 (6): 700–716.
Hoffman, Charles S., and Fred Winston. 1987. “A Ten-Minute DNA Preparation from Yeast Efficiently Releases Autonomous Plasmids for Transformaion of Escherichia Coli.” *Gene* 57 (2): 267–72.
Hong, Jungeui, and David Gresham. 2014a. “Molecular Specificity, Convergence and Constraint Shape Adaptive Evolution in Nutrient-Poor Environments.” *PLoS Genetics* 10 (1): e1004041.
———. 2014b. “Molecular Specificity, Convergence and Constraint Shape Adaptive Evolution in Nutrient-Poor Environments.” *PLoS Genetics* 10 (1): e1004041.
Hope, Elyse A., Clara J. Amorosi, Aaron W. Miller, Kolena Dang, Caiti Smukowski Heil, and Maitreya J. Dunham. 2017. “Experimental Evolution Reveals Favored Adaptive Routes to Cell Aggregation in Yeast.” *Genetics* 206 (2): 1153–67.
Horiuchi, T., S. Horiuchi, and A. Novick. 1963. “The Genetic Basis of Hyper-Synthesis of

Beta-Galactosidase." *Genetics* 48: 157–69.
Horiuchi, T., J. I. Tomizawa, and A. Novick. 1962. "Isolation and Properties of Bacteria Capable of High Rates of Beta-Galactosidase Synthesis." *Biochimica et Biophysica Acta* 55 (January): 152–63.
Hose, James, Leah E. Escalante, Katie J. Clowers, H. Auguste Dutcher, Deelegant Robinson, Venera Bouriakov, Joshua J. Coon, Evgenia Shishkova, and Audrey P. Gasch. 2020. "The Genetic Basis of Aneuploidy Tolerance in Wild Yeast." *eLife* 9 (January). https://doi.org/10.7554/eLife.52063.
Hose, James, Chris Mun Yong, Maria Sardi, Zhishi Wang, Michael A. Newton, and Audrey P. Gasch. 2015. "Dosage Compensation Can Buffer Copy-Number Variation in Wild Yeast." *eLife* 4 (May). https://doi.org/10.7554/eLife.05462.
Hughes, Austin L. 1994. "The Evolution of Functionally Novel Proteins after Gene Duplication." *Proceedings of the Royal Society of London B: Biological Sciences* 256.
Hughes, Julie M., Brian K. Lohman, Gail E. Deckert, Eric P. Nichols, Matt Settles, Zaid Abdo, and Eva M. Top. 2012. "The Role of Clonal Interference in the Evolutionary Dynamics of Plasmid-Host Adaptation." *mBio* 3: e00077–12.
Hughes, T. R., C. J. Roberts, H. Dai, A. R. Jones, M. R. Meyer, D. Slade, J. Burchard, et al. 2000. "Widespread Aneuploidy Revealed by DNA Microarray Expression Profiling." *Nature Genetics* 25 (3): 333–37.
Hull, Ryan M., Cristina Cruz, Carmen V. Jack, and Jonathan Houseley. 2017. "Environmental Change Drives Accelerated Adaptation through Stimulated Copy Number Variation." *PLoS Biology* 15 (6): e2001333.
Iafrate, A. John, Lars Feuk, Miguel N. Rivera, Marc L. Listewnik, Patricia K. Donahoe, Ying Qi, Stephen W. Scherer, and Charles Lee. 2004. "Detection of Large-Scale Variation in the Human Genome." *Nature Genetics* 36 (9): 949–51.
Iantorno, Stefano A., Caroline Durrant, Asis Khan, Mandy J. Sanders, Stephen M. Beverley, Wesley C. Warren, Matthew Berriman, David L. Sacks, James A. Cotton, and Michael E. Grigg. 2017. "Gene Expression in Leishmania Is Regulated Predominantly by Gene Dosage." *mBio* 8 (5). https://doi.org/10.1128/mBio.01393-17.
Imhof, M., and C. Schlotterer. 2001. "Fitness Effects of Advantageous Mutations in Evolving Escherichia Coli Populations." *Proceedings of the National Academy of Sciences of the United States of America* 98 (3): 1113–17.
Innan, Hideki, and Fyodor Kondrashov. 2010. "The Evolution of Gene Duplications: Classifying and Distinguishing between Models." *Nature Reviews. Genetics* 11 (2): 97–108.
Iskow, Rebecca C., Omer Gokcumen, Alexej Abyzov, Joanna Malukiewicz, Qihui Zhu, Ann T. Sukumar, Athma A. Pai, et al. 2012. "Regulatory Element Copy Number Differences Shape Primate Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 109: 12656–61.
Itsara, Andy, Gregory M. Cooper, Carl Baker, Santhosh Girirajan, Jun Li, Devin Absher, Ronald M. Krauss, et al. 2009. "Population Analysis of Large Copy Number Variants and Hotspots of Human Genetic Disease." *American Journal of Human Genetics* 84 (2): 148–61.
Izutsu, Minako, Jun Zhou, Yuzo Sugiyama, Osamu Nishimura, Tomoyuki Aizu, Atsushi Toyoda, Asao Fujiyama, Kiyokazu Agata, and Naoyuki Fuse. 2012. "Genome Features of 'Dark-Fly', a Drosophila Line Reared Long-Term in a Dark Environment." *PloS One* 7 (3): e33288.
Jack, Carmen V., Cristina Cruz, Ryan M. Hull, Markus A. Keller, Markus Ralser, and Jonathan Houseley. 2015. "Regulation of Ribosomal DNA Amplification by the TOR Pathway." *Proceedings of the National Academy of Sciences of the United States of America* 112 (31): 9674–79.
Jain, Neha, Petra Janning, and Heinz Neumann. 2021. "14-3-3 Protein Bmh1 Triggers Short-Range Compaction of Mitotic Chromosomes by Recruiting Sirtuin Deacetylase Hst2."

The Journal of Biological Chemistry 296 (January): 100078.
Jalvingh, Kirsten M., Peter L. Chang, Sergey V. Nuzhdin, and Bregje Wertheim. 2014. “Genomic Changes under Rapid Evolution: Selection for Parasitoid Resistance.” *Proceedings. Biological Sciences / The Royal Society* 281 (1779): 20132303.
Jennings, E., and M. Madigan. 2017. “astroABC : An Approximate Bayesian Computation Sequential Monte Carlo Sampler for Cosmological Parameter Estimation.” *Astronomy and Computing*. https://doi.org/10.1016/j.ascom.2017.01.001.
Joseph, Sarah B., and David W. Hall. 2004. “Spontaneous Mutations in Diploid Saccharomyces Cerevisiae.” *Genetics*. https://doi.org/10.1534/genetics.104.033761.
Kafri, Moshe, Eyal Metzl-Raz, Ghil Jona, and Naama Barkai. 2016. “The Cost of Protein Production.” *Cell Reports* 14 (1): 22–31.
Kang, Lin, Dau Dayal Aggarwal, Eugenia Rashkovetsky, Abraham B. Korol, and Pawel Michalak. 2016. “Rapid Genomic Changes in Drosophila Melanogaster Adapting to Desiccation Stress in an Experimental Evolution System.” *BMC Genomics* 17 (1). https://doi.org/10.1186/s12864-016-2556-y.
Kao, Katy C., and Gavin Sherlock. 2008. “Molecular Characterization of Clonal Interference during Adaptive Evolution in Asexual Populations of Saccharomyces Cerevisiae.” *Nature Genetics* 40 (12): 1499–1504.
Kaplan, Kenneth B., and Rong Li. 2012. “A Prescription for ‘stress’ – the Role of Hsp90 in Genome Stability and Cellular Adaptation.” *Trends in Cell Biology*. https://doi.org/10.1016/j.tcb.2012.08.006.
Kassen, Rees, and Thomas Bataillon. 2006. “Distribution of Fitness Effects among Beneficial Mutations before Selection in Experimental Populations of Bacteria.” *Nature Genetics* 38 (4): 484–88.
Kass, Robert E., and Adrian E. Raftery. 1995. “Bayes Factors.” *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1995.10476572.
Katju, Vaishali, and Ulfar Bergthorsson. 2013. “Copy-Number Changes in Evolution: Rates, Fitness Effects and Adaptive Significance.” *Frontiers in Genetics* 4 (December): 273.
Kawecki, Tadeusz J., Richard E. Lenski, Dieter Ebert, Brian Hollis, Isabelle Olivieri, and Michael C. Whitlock. 2012. “Experimental Evolution.” *Trends in Ecology & Evolution* 27 (10): 547–60.
Kezos, James N., Mark A. Phillips, Misty D. Thomas, Akamu J. Ewunkem, Grant A. Rutledge, Thomas T. Barter, Marta A. Santos, et al. 2019. “Genomics of Early Cardiac Dysfunction and Mortality in Obese.” *Physiological and Biochemical Zoology: PBZ* 92 (6): 591–611.
Khan, Aisha I., Duy M. Dinh, Dominique Schneider, Richard E. Lenski, and Tim F. Cooper. 2011. “Negative Epistasis between Beneficial Mutations in an Evolving Bacterial Population.” *Science* 332 (6034): 1193–96.
Klinger, Emmanuel, and Jan Hasenauer. 2017. “A Scheme for Adaptive Selection of Population Sizes in Approximate Bayesian Computation - Sequential Monte Carlo.” *Computational Methods in Systems Biology*. https://doi.org/10.1007/978-3-319-67471-1_8.
Klinger, Emmanuel, Dennis Rickert, and Jan Hasenauer. 2018. “pyABC: Distributed, Likelihood-Free Inference.” *Bioinformatics* 34 (20): 3591–93.
Kondrashov, Fyodor A. 2012. “Gene Duplication as a Mechanism of Genomic Adaptation to a Changing Environment.” *Proceedings. Biological Sciences / The Royal Society* 279 (1749): 5048–57.
Kondrashov, Fyodor A., and Alexey S. Kondrashov. 2010. “Measurements of Spontaneous Rates of Mutations in the Recent Past and the near Future.” *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2009.0286.
Korbel, Jan O., Alexander Eckehart Urban, Jason P. Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M. Kim, et al. 2007. “Paired-End Mapping Reveals Extensive

Structural Variation in the Human Genome." *Science* 318 (5849): 420–26.
Korotkevich, Gennady, Vladimir Sukhov, Nikolay Budin, Boris Shpak, Maxim N. Artyomov, and Alexey Sergushichev. n.d. "Fast Gene Set Enrichment Analysis." https://doi.org/10.1101/060012.
Koszul, Romain, Sandrine Caburet, Bernard Dujon, and Gilles Fischer. 2004. "Eucaryotic Genome Evolution through the Spontaneous Duplication of Large Chromosomal Segments." *The EMBO Journal* 23 (1): 234–43.
Kruschke, John K. 2014. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press.
Kryazhimskiy, Sergey, Daniel P. Rice, Elizabeth R. Jerison, and Michael M. Desai. 2014. "Microbial Evolution. Global Epistasis Makes Adaptation Predictable despite Sequence-Level Stochasticity." *Science* 344 (6191): 1519–22.
Kumar, Ravinder. 2017. "An Account of Fungal 14-3-3 Proteins." *European Journal of Cell Biology* 96 (2): 206–17.
Kvitek, Daniel J., and Gavin Sherlock. 2011. "Reciprocal Sign Epistasis between Frequently Experimentally Evolved Adaptive Mutations Causes a Rugged Fitness Landscape." *PLoS Genetics* 7 (4): e1002056–e1002056.
———. 2013. "Whole Genome, Whole Population Sequencing Reveals That Loss of Signaling Networks Is the Major Adaptive Strategy in a Constant Environment." *PLoS Genetics* 9 (11): e1003972.
Labib, Karim, Ben Hodgson, A. Admire, L. Shanks, N. Danzl, M. Wang, U. Weier, et al. 2007. "Replication Fork Barriers: Pausing for a Break or Stalling for Time?" *EMBO Reports* 8 (4): 346–53.
Lam, Kwan-Wood G., and Alec J. Jeffreys. 2006. "Processes of Copy-Number Change in Human DNA: The Dynamics of {alpha}-Globin Gene Deletion." *Proceedings of the National Academy of Sciences of the United States of America* 103 (24): 8921–27.
Lang, Gregory I., David Botstein, and Michael M. Desai. 2011. "Genetic Variation and the Fate of Beneficial Mutations in Asexual Populations." *Genetics* 188: 647–61.
Lang, Gregory I., Daniel P. Rice, Mark J. Hickman, Erica Sodergren, George M. Weinstock, David Botstein, and Michael M. Desai. 2013. "Pervasive Genetic Hitchhiking and Clonal Interference in Forty Evolving Yeast Populations." *Nature* 500: 571–74.
Langridge, J. 1969. "Mutations Conferring Quantitative and Qualitative Increases in β-Galactosidase Activity inEscherichia Coli." *Molecular & General Genetics: MGG* 105 (1): 74–83.
Larrimore, Katherine E., Natalia S. Barattin-Voynova, David W. Reid, and Davis T. W. Ng. 2020. "Aneuploidy-Induced Proteotoxic Stress Can Be Effectively Tolerated without Dosage Compensation, Genetic Mutations, or Stress Responses." *BMC Biology*. https://doi.org/10.1186/s12915-020-00852-x.
Lauer, Stephanie, Grace Avecilla, Pieter Spealman, Gunjan Sethia, Nathan Brandt, Sasha F. Levy, and David Gresham. 2018. "Single-Cell Copy Number Variant Detection Reveals the Dynamics and Diversity of Adaptation." *PLoS Biology* 16 (12): e3000069.
Lee, Jennifer A., Claudia M. B. Carvalho, and James R. Lupski. 2007. "A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders." *Cell* 131: 1235–47.
Lenski, Richard E., Michael R. Rose, Suzanne C. Simpson, and Scott C. Tadler. 1991. "Long-Term Experimental Evolution in Escherichia Coli. I. Adaptation and Divergence During 2,000 Generations." *The American Naturalist* 138: 1315–41.
Levitan, Anton, Andrew N. Gale, Emma K. Dallon, Darby W. Kozan, Kyle W. Cunningham, Roded Sharan, and Judith Berman. 2020. "Comparing the Utility of in Vivo Transposon Mutagenesis Approaches in Yeast Species to Infer Gene Essentiality." *Current Genetics* 66

(6): 1117–34.
Levy, Sasha F., Jamie R. Blundell, Sandeep Venkataram, Dmitri A. Petrov, Daniel S. Fisher, and Gavin Sherlock. 2015. “Quantitative Evolutionary Dynamics Using High-Resolution Lineage Tracking.” *Nature* 519 (7542): 181–86.
Lewis, E. B. 1978. “A Gene Complex Controlling Segmentation in Drosophila.” *Nature* 276 (5688): 565–70.
Li, Heng, and Richard Durbin. 2010. “Fast and Accurate Long-Read Alignment with Burrows–Wheeler Transform.” *Bioinformatics* 26 (5): 589–95.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp352.
Li, M., X. Fang, D. J. Baker, L. Guo, X. Gao, Z. Wei, S. Han, J. M. Van Deursen, and P. Zhang. 2010. “The ATM–p53 Pathway Suppresses Aneuploidy-Induced Tumorigenesis.” *Proceedings of the National Academy of Sciences* 107 (32): 14188–93.
Linder, Robert A., John P. Greco, Fabian Seidl, Takeshi Matsui, and Ian M. Ehrenreich. 2017. “The Stress-Inducible Peroxidase TSA2 Underlies a Conditionally Beneficial Chromosomal Duplication in Saccharomyces Cerevisiae.” *G3 Genes|Genomes|Genetics*. https://doi.org/10.1534/g3.117.300069.
Lipinski, Kendra J., James C. Farslow, Kelly A. Fitzpatrick, Michael Lynch, Vaishali Katju, and Ulfar Bergthorsson. 2011. “High Spontaneous Rate of Gene Duplication in Caenorhabditis Elegans.” *Current Biology: CB* 21 (4): 306–10.
Liu, Chang, Dewald van Dyk, Yue Li, Brenda Andrews, and Hai Rao. 2009. “A Genome-Wide Synthetic Dosage Lethality Screen Reveals Multiple Pathways That Require the Functioning of Ubiquitin-Binding Proteins Rad23 and Dsk2.” *BMC Biology* 7 (November): 75.
Liu, Haoxuan, and Jianzhi Zhang. 2019. “Yeast Spontaneous Mutation Rate and Spectrum Vary with Environment.” *Current Biology: CB* 29 (10): 1584–91.e3.
Liu, Zhengchang, Takayuki Sekito, Mário Spírek, Janet Thornton, and Ronald A. Butow. 2003. “Retrograde Signaling Is Regulated by the Dynamic Interaction between Rtg2p and Mks1p.” *Molecular Cell* 12 (2): 401–11.
Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.
Lueckmann, Jan-Matthis, Pedro J. Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H. Macke. 2017. “Flexible Statistical Inference for Mechanistic Models of Neural Dynamics.” In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 1289–99. Curran Associates, Inc.
Lupiáñez, Darío G., Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, et al. 2015. “Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions.” *Cell* 161 (5): 1012–25.
Lupiáñez, Darío G., Malte Spielmann, and Stefan Mundlos. 2016. “Breaking TADs: How Alterations of Chromatin Domains Result in Disease.” *Trends in Genetics: TIG* 32 (4): 225–37.
Lynch, Michael. 2010. “Rate, Molecular Spectrum, and Consequences of Human Mutation.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (3): 961–68.
Lynch, Michael, and John S. Conery. 2000. “The Evolutionary Fate and Consequences of Duplicate Genes.” *Science* 290 (5494).
Lynch, Michael, Way Sung, Krystalynne Morris, Nicole Coffey, Christian R. Landry, Erik B.

Dopman, W. Joseph Dickinson, et al. 2008. “A Genome-Wide View of the Spectrum of Spontaneous Mutations in Yeast.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (27): 9272–77.
MacLean, R. Craig, and Angus Buckling. 2009. “The Distribution of Fitness Effects of Beneficial Mutations in Pseudomonas Aeruginosa.” *PLoS Genetics* 5 (3): e1000406.
Maddamsetti, R., R. E. Lenski, and J. E. Barrick. 2015. “Adaptation, Clonal Interference, and Frequency-Dependent Interactions in a Long-Term Evolution Experiment with Escherichia Coli.” *Genetics*. https://doi.org/10.1534/genetics.115.176677.
Makanae, Koji, Reiko Kintaka, Takashi Makino, Hiroaki Kitano, and Hisao Moriya. 2013. “Identification of Dosage-Sensitive Genes in Saccharomyces Cerevisiae Using the Genetic Tug-of-War Method.” *Genome Research* 23 (2): 300–311.
Mani, Ramamurthy, Robert P. St Onge, John L. Hartman 4th, Guri Giaever, and Frederick P. Roth. 2008. “Defining Genetic Interaction.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (9): 3461–66.
Mansisidor, Andres R., Temistocles Molinar, Priyanka Srivastava, Hannah Blitzblau, Hannah Klein, and Andreas Hochwagen. 2018. “Genomic Copy-Number Loss Is Rescued by Self-Limiting Production of DNA Circles.” *bioRxiv*. https://doi.org/10.1101/255471.
Marjoram, Paul, John Molitor, Vincent Plagnol, and Simon Tavare. 2003. “Markov Chain Monte Carlo without Likelihoods.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (26): 15324–28.
Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.journal* 17 (1): 10–12.
Mayo, Sonia, Sandra Monfort, Mónica Roselló, Carmen Orellana, Silvestre Oltra, Alfonso Caro-Llopis, and Francisco Martínez. 2017. “Chimeric Genes in Deletions and Duplications Associated with Intellectual Disability.” *International Journal of Genomics and Proteomics* 2017 (May): 4798474.
McIsaac, R. Scott, Benjamin L. Oakes, Xin Wang, Krysta A. Dummit, David Botstein, and Marcus B. Noyes. 2013. “Synthetic Gene Expression Perturbation Systems with Rapid, Tunable, Single-Gene Specificity in Yeast.” *Nucleic Acids Research* 41 (4): e57.
Merla, Giuseppe, Cédric Howald, Charlotte N. Henrichsen, Robert Lyle, Carine Wyss, Marie-Thérèse Zabot, Stylianos E. Antonarakis, and Alexandre Reymond. 2006. “Submicroscopic Deletion in Patients with Williams-Beuren Syndrome Influences Expression Levels of the Nonhemizygous Flanking Genes.” *American Journal of Human Genetics* 79 (2): 332–41.
Messer, Philipp W., and Dmitri A. Petrov. 2013. “Population Genomics of Rapid Adaptation by Soft Selective Sweeps.” *Trends in Ecology & Evolution* 28 (11): 659–69.
Michel, Agnès H., Riko Hatakeyama, Philipp Kimmig, Meret Arter, Matthias Peter, Joao Matos, Claudio De Virgilio, and Benoît Kornmann. 2017. “Functional Mapping of Yeast Genomes by Saturated Transposition.” *eLife* 6 (May). https://doi.org/10.7554/eLife.23570.
Miller, Aaron W., Corrie Befort, Emily O. Kerr, and Maitreya J. Dunham. 2013. “Design and Use of Multiplexed Chemostat Arrays.” *Journal of Visualized Experiments: JoVE*, no. 72 (February): e50262.
Mishra, Sweta, Capucine Van Rechem, Sangita Pal, Thomas L. Clarke, Damayanti Chakraborty, Sarah D. Mahan, Joshua C. Black, et al. 2018. “Cross-Talk between Lysine-Modifying Enzymes Controls Site-Specific DNA Amplifications.” *Cell* 175 (6): 1716.
Molina, Jessica, Paulina Carmona-Mora, Jacqueline Chrast, Paola M. Krall, César P. Canales, James R. Lupski, Alexandre Reymond, and Katherina Walz. 2008. “Abnormal Social Behaviors and Altered Gene Expression Rates in a Mouse Model for Potocki-Lupski Syndrome.” *Human Molecular Genetics* 17 (16): 2486–95.
Møller, Henrik D., Kaj S. Andersen, and Birgitte Regenberg. 2013. “A Model for Generating

Several Adaptive Phenotypes from a Single Genetic Event: Saccharomyces Cerevisiae GAP1 as a Potential Bet-Hedging Switch." *Communicative & Integrative Biology* 6 (3): e23933.
Møller, Henrik D., Lance Parsons, Tue S. Jørgensen, David Botstein, and Birgitte Regenberg. 2015. "Extrachromosomal Circular DNA Is Common in Yeast." *Proceedings of the National Academy of Sciences of the United States of America* 112 (24): E3114–22.
Montgomery, Stephen B., David L. Goode, Erika Kvikstad, Cornelis A. Albers, Zhengdong D. Zhang, Xinmeng Jasmine Mu, Guruprasad Ananda, et al. 2013. "The Origin, Evolution, and Functional Impact of Short Insertion–deletion Variants Identified in 179 Human Genomes." *Genome Research*. https://doi.org/10.1101/gr.148718.112.
Morgenthaler, Andrew B., Ryan K. Fritts, and Shelley D. Copley. 2022. "Amplicon Remodeling and Genomic Mutations Drive Population Dynamics after Segmental Amplification." *Molecular Biology and Evolution* 39 (1). https://doi.org/10.1093/molbev/msab289.
Morgenthaler, Andrew B., Wallis R. Kinney, Christopher C. Ebmeier, Corinne M. Walsh, Daniel J. Snyder, Vaughn S. Cooper, William M. Old, and Shelley D. Copley. 2019. "Mutations That Improve Efficiency of a Weak-Link Enzyme Are Rare Compared to Adaptive Mutations Elsewhere in the Genome." *eLife* 8: e53535.
Moriya, Hisao. 2015. "Quantitative Nature of Overexpression Experiments." *Molecular Biology of the Cell* 26 (22): 3932–39.
Mount, Harley O'connor, Nicole M. Revie, Robert T. Todd, Kaitlin Anstett, Cathy Collins, Michael Costanzo, Charles Boone, Nicole Robbins, Anna Selmecki, and Leah E. Cowen. 2018. "Global Analysis of Genetic Circuitry and Adaptive Mechanisms Enabling Resistance to the Azole Antifungal Drugs." *PLoS Genetics* 14 (4): e1007319.
Moura de Sousa, Jorge A., Paulo R. A. Campos, and Isabel Gordo. 2013. "An ABC Method for Estimating the Rate and Distribution of Effects of Beneficial Mutations." *Genome Biology and Evolution* 5 (5): 794–806.
Moxon, Richard, Chris Bayliss, and Derek Hood. 2006. "Bacterial Contingency Loci: The Role of Simple Sequence DNA Repeats in Bacterial Adaptation." *Annual Review of Genetics* 40: 307–33.
Muenzner, Julia, Pauline Trebulle, Federica Agostini, Christoph B. Messner, Martin Steger, Andrea Lehmann, Elodie Caudal, et al. n.d. "The Natural Diversity of the Yeast Proteome Reveals Chromosome-Wide Dosage Compensation in Aneuploids." https://doi.org/10.1101/2022.04.06.487392.
Muller, H. J. 1932. "Some Genetic Aspects of Sex." *The American Naturalist* 66 (703): 118–38.
Myhre, Simen, Ole-Christian Lingjærde, Bryan T. Hennessy, Miriam R. Aure, Mark S. Carey, Jan Alsner, Trine Tramm, et al. 2013. "Influence of DNA Copy Number and mRNA Levels on the Expression of Breast Cancer Related Proteins." *Molecular Oncology* 7 (3): 704–18.
Nair, Shalini, Becky Miller, Marion Barends, Anchalee Jaidee, Jigar Patel, Mayfong Mayxay, Paul Newton, François Nosten, Michael T. Ferdig, and Tim J. C. Anderson. 2008. "Adaptive Copy Number Evolution in Malaria Parasites." *PLoS Genetics* 4 (10): e1000243.
Natesuntorn, Waranya, Kotaro Iwami, Yuki Matsubara, Yu Sasano, Minetaka Sugiyama, Yoshinobu Kaneko, and Satoshi Harashima. 2015. "Genome-Wide Construction of a Series of Designed Segmental Aneuploids in Saccharomyces Cerevisiae." *Scientific Reports* 5 (July): 12510.
Neymotin, Benjamin, Rodoniki Athanasiadou, and David Gresham. 2014. "Determination of in Vivo RNA Kinetics Using RATE-Seq." *RNA* 20 (10): 1645–52.
Nguyen Ba, Alex N., Ivana Cvijović, José I. Rojas Echenique, Katherine R. Lawrence, Artur Rego-Costa, Xianan Liu, Sasha F. Levy, and Michael M. Desai. 2019. "High-Resolution Lineage Tracking Reveals Travelling Wave of Adaptation in Laboratory Yeast." *Nature* 575 (7783): 494–99.

Noorani, Imran, Allan Bradley, and Jorge de la Rosa. 2020. “CRISPR and Transposon in Vivo Screens for Cancer Drivers and Therapeutic Targets.” *Genome Biology* 21 (1): 204.
Ohno, Susumu. 1970. *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer Berlin Heidelberg.
Ohye, Tamae, Hidehito Inagaki, Mamoru Ozaki, Toshiro Ikeda, and Hiroki Kurahashi. 2014. “Signature of Backward Replication Slippage at the Copy Number Variation Junction.” *Journal of Human Genetics* 59 (5): 247–50.
Orr, H. Allen. 2003. “The Distribution of Fitness Effects among Beneficial Mutations.” *Genetics* 163 (4): 1519–26.
Ottaviani, Diego, Magdalena LeCain, and Denise Sheer. 2014. “The Role of Microhomology in Genomic Structural Variation.” *Trends in Genetics: TIG* 30 (3): 85–94.
Otto, Sarah P., and Troy Day. 2007. “A Biologist’s Guide to Mathematical Modeling in Ecology and Evolution.” https://doi.org/10.1515/9781400840915.
Panasyuk, Ganna, Ivan Nemazanyy, Valeriy Filonenko, and Alexander Zhyvoloup. 2004. “Large-Scale Yeast Transformation in Low-Percentage Agarose Medium.” *BioTechniques* 36 (1): 40–42, 44.
Papamakarios, George, and Iain Murray. 2016. “Fast \epsilon -Free Inference of Simulation Models with Bayesian Conditional Density Estimation.” In *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 1028–36. Curran Associates, Inc.
Papamakarios, George, Theo Pavlakou, and Iain Murray. 2017. “Masked Autoregressive Flow for Density Estimation.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1705.07057.
Papp, Balázs, Csaba Pál, and Laurence D. Hurst. 2003. “Dosage Sensitivity and the Evolution of Gene Families in Yeast.” *Nature* 424 (6945): 194–97.
Paulander, Wilhelm, Dan I. Andersson, and Sophie Maisnier-Patin. 2010. “Amplification of the Gene for Isoleucyl-tRNA Synthetase Facilitates Adaptation to the Fitness Cost of Mupirocin Resistance in Salmonella Enterica.” *Genetics* 185 (1): 305–12.
Pavani, Mattia, Paolo Bonaiuti, Elena Chiroli, Fridolin Gross, Federica Natali, Francesca Macaluso, Ádám Póti, et al. 2021. “Epistasis, Aneuploidy, and Functional Mutations Underlie Evolution of Resistance to Induced Microtubule Depolymerization.” *The EMBO Journal* 40 (22): e108225.
Pavelka, Norman, Giulia Rancati, Jin Zhu, William D. Bradford, Anita Saraf, Laurence Florens, Brian W. Sanderson, Gaye L. Hattem, and Rong Li. 2010. “Aneuploidy Confers Quantitative Proteome Changes and Phenotypic Variation in Budding Yeast.” *Nature* 468 (7321): 321–25.
Payen, Celia, Sara C. Di Rienzi, Giang T. Ong, Jamie L. Pogachar, Joseph C. Sanchez, Anna B. Sunshine, M. K. Raghuraman, Bonita J. Brewer, and Maitreya J. Dunham. 2014. “The Dynamics of Diverse Segmental Amplifications in Populations of Saccharomyces Cerevisiae Adapting to Strong Selection.” *G3* 4 (3): 399–409.
Payen, Celia, Romain Koszul, Bernard Dujon, Gilles Fischer, J. A. Bailey, E. E. Eichler, A. J. Sharp, et al. 2008. “Segmental Duplications Arise from Pol32-Dependent Repair of Broken Forks through Two Alternative Replication-Based Mechanisms.” *PLoS Genetics* 4 (9): e1000175–e1000175.
Payen, Celia, Anna B. Sunshine, Giang T. Ong, Jamie L. Pogachar, Wei Zhao, and Maitreya J. Dunham. 2016. “High-Throughput Identification of Adaptive Mutations in Experimentally Evolved Yeast Populations.” *PLoS Genetics* 12 (10): e1006339.
Peng, Zhen, Weichen Zhou, Wenqing Fu, Renqian Du, Li Jin, and Feng Zhang. 2015. “Correlation between Frequency of Non-Allelic Homologous Recombination and Homology Properties: Evidence from Homology-Mediated CNV Mutations in the Human Genome.” *Human Molecular Genetics* 24 (5): 1225–33.

Peter, Jackson, Matteo De Chiara, Anne Friedrich, Jia-Xing Yue, David Pflieger, Anders Bergström, Anastasie Sigwalt, et al. 2018. “Genome Evolution across 1,011 Saccharomyces Cerevisiae Isolates.” *Nature* 556 (7701): 339–44.
Pettersson, Mats E., Song Sun, Dan I. Andersson, and Otto G. Berg. 2009. “Evolution of New Gene Functions: Simulation and Analysis of the Amplification Model.” *Genetica* 135 (3): 309–24.
Phillips, Mark A., Grant A. Rutledge, James N. Kezos, Zachary S. Greenspan, Andrew Talbott, Sara Matty, Hamid Arain, Laurence D. Mueller, Michael R. Rose, and Parvin Shahrestani. 2018. “Effects of Evolutionary History on Genome Wide and Phenotypic Convergence in Drosophila Populations.” *BMC Genomics* 19 (1): 743.
Pös, Ondrej, Jan Radvanszky, Gergely Buglyó, Zuzana Pös, Diana Rusnakova, Bálint Nagy, and Tomas Szemes. 2021. “DNA Copy Number Variation: Main Characteristics, Evolutionary Significance, and Pathological Aspects.” *Biomedical Journal*. https://doi.org/10.1016/j.bj.2021.02.003.
Prangle, Dennis. 2017. “Adapting the ABC Distance Function.” *Bayesian Analysis*. https://doi.org/10.1214/16-ba1002.
Pränting, Maria, and Dan I. Andersson. 2011. “Escape from Growth Restriction in Small Colony Variants of Salmonella Typhimurium by Gene Amplification and Mutation.” *Molecular Microbiology* 79 (2): 305–15.
Press, Maximilian Oliver, Ashley N. Hall, Elizabeth A. Morton, and Christine Queitsch. 2019. “Substitutions Are Boring: Some Arguments about Parallel Mutations and High Mutation Rates.” *Trends in Genetics: TIG* 35 (4): 253–64.
Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. 1999. “Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites.” *Molecular Biology and Evolution* 16 (12): 1791–98.
Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42.
Ramirez, Oscar, Iñigo Olalde, Jonas Berglund, Belen Lorente-Galdos, Jessica Hernandez-Rodriguez, Javier Quilez, Matthew T. Webster, et al. 2014. “Analysis of Structural Diversity in Wolf-like Canids Reveals Post-Domestication Variants.” *BMC Genomics* 15 (1): 465–465.
Ram, Yoav, Eynat Dellus-Gur, Maayan Bibi, Kedar Karkare, Uri Obolski, Marcus W. Feldman, Tim F. Cooper, Judith Berman, and Lilach Hadany. 2019. “Predicting Microbial Growth in a Mixed Culture from Growth Curve Data.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (29): 14698–707.
Rancati, Giulia, Norman Pavelka, Brian Fleharty, Aaron Noll, Rhonda Trimble, Kendra Walton, Anoja Perera, Karen Staehling-Hampton, Chris W. Seidel, and Rong Li. 2008. “Aneuploidy Underlies Rapid Adaptive Evolution of Yeast Cells Deprived of a Conserved Cytokinesis Motor.” *Cell* 135 (5): 879–93.
Raynes, Y., and P. D. Sniegowski. 2014. “Experimental Evolution and the Dynamics of Genomic Mutation Rate Modifiers.” *Heredity* 113 (5): 375–80.
Reams, Andrew B., Eric Kofoid, Michael Savageau, and John R. Roth. 2010. “Duplication Frequency in a Population of Salmonella Enterica Rapidly Approaches Steady State with or without Recombination.” *Genetics* 184 (4): 1077–94.
Reams, Andrew B., and Ellen L. Neidle. 2004. “Gene Amplification Involves Site-Specific Short Homology-Independent Illegitimate Recombination in Acinetobacter Sp. Strain ADP1.” *Journal of Molecular Biology* 338 (4): 643–56.
Reams, Andrew B., and John R. Roth. 2015. “Mechanisms of Gene Duplication and Amplification.” *Cold Spring Harbor Perspectives in Biology* 7 (2): a016592.
Reinders, Joerg, René P. Zahedi, Nikolaus Pfanner, Chris Meisinger, and Albert Sickmann.

2006. “Toward the Complete Yeast Mitochondrial Proteome: Multidimensional Separation Techniques for Mitochondrial Proteomics.” *Journal of Proteome Research* 5 (7): 1543–54.
Reinders, Jörg, Karina Wagner, Rene P. Zahedi, Diana Stojanovski, Beate Eyrich, Martin van der Laan, Peter Rehling, Albert Sickmann, Nikolaus Pfanner, and Chris Meisinger. 2007. “Profiling Phosphoproteins of Yeast Mitochondria Reveals a Role of Phosphorylation in Assembly of the ATP Synthase.” *Molecular & Cellular Proteomics: MCP* 6 (11): 1896–1906.
Remolina, Silvia C., Peter L. Chang, Jeff Leips, Sergey V. Nuzhdin, and Kimberly A. Hughes. 2012. “Genomic Basis of Aging and Life-History Evolution in Drosophila Melanogaster.” *Evolution; International Journal of Organic Evolution* 66 (11): 3390–3403.
Rezelj, Veronica V., Laura I. Levi, and Marco Vignuzzi. 2018. “The Defective Component of Viral Populations.” *Current Opinion in Virology* 33 (August): 74–80.
Rice, Alan M., and Aoife McLysaght. 2017a. “Dosage Sensitivity Is a Major Determinant of Human Copy Number Variant Pathogenicity.” *Nature Communications* 8 (February): 14366.
———. 2017b. “Dosage-Sensitive Genes in Evolution and Disease.” *BMC Biology* 15 (1): 78.
Rippey, Caitlin, Tom Walsh, Suleyman Gulsuner, Matt Brodsky, Alex S. Nord, Molly Gasperini, Sarah Pierce, et al. 2013. “Formation of Chimeric Genes by Copy-Number Variation as a Mutational Mechanism in Schizophrenia.” *American Journal of Human Genetics* 93 (4): 697–710.
Robinson, Deelegant, Michael Place, James Hose, Adam Jochem, and Audrey P. Gasch. 2021. “Natural Variation in the Consequences of Gene Overexpression and Its Implications for Evolutionary Trajectories.” *eLife* 10 (August). https://doi.org/10.7554/eLife.70564.
Rodrigo, Guillermo, and Mario A. Fares. 2018. “Intrinsic Adaptive Value and Early Fate of Gene Duplication Revealed by a Bottom-up Approach.” *eLife* 7 (January). https://doi.org/10.7554/eLife.29739.
Rokyta, Darin R., Craig J. Beisel, Paul Joyce, Martin T. Ferris, Christina L. Burch, and Holly A. Wichman. 2008. “Beneficial Fitness Effects Are Not Exponential for Two Viruses.” *Journal of Molecular Evolution* 67 (4): 368–76.
Rokyta, Darin R., Paul Joyce, S. Brian Caudle, and Holly A. Wichman. 2005. “An Empirical Test of the Mutational Landscape Model of Adaptation Using a Single-Stranded DNA Virus.” *Nature Genetics* 37 (4): 441–44.
Rosin, Dalia, Gil Hornung, Itay Tirosh, Ariel Gispan, and Naama Barkai. 2012. “Promoter Nucleosome Organization Shapes the Evolution of Gene Expression.” *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1002579.
Roth, John R., and D. I. Andersson. 2012. “Poxvirus Use a ‘Gene Accordion’ to Tune out Host Defenses.” *Cell*.
Rouzine, Igor M., Eric Brunet, and Claus O. Wilke. 2008. “The Traveling-Wave Approach to Asexual Evolution: Muller’s Ratchet and Speed of Adaptation.” *Theoretical Population Biology* 73 (1): 24–46.
Sakofsky, Cynthia J., Sandeep Ayyar, Angela K. Deem, Woo-Hyun Chung, Grzegorz Ira, and Anna Malkova. 2015. “Translesion Polymerases Drive Microhomology-Mediated Break-Induced Replication Leading to Complex Chromosomal Rearrangements.” *Molecular Cell* 60 (6): 860–72.
Schacherer, Joseph, Jacky de Montigny, Anne Welcker, Jean-Luc Souciet, and Serge Potier. 2005. “Duplication Processes in Saccharomyces Cerevisiae Haploid Strains.” *Nucleic Acids Research* 33 (19): 6319–26.
Schacherer, Joseph, Yves Tourrette, Serge Potier, Jean-Luc Souciet, and Jacky de Montigny. 2007. “Spontaneous Duplications in Diploid Saccharomyces Cerevisiae Cells.” *DNA Repair* 6 (10): 1441–52.
Schacherer, Joseph, Yves Tourrette, Jean-Luc Souciet, Serge Potier, and Jacky De Montigny. 2004. “Recovery of a Function Involving Gene Duplication by Retroposition in

Saccharomyces Cerevisiae." *Genome Research* 14 (7): 1291–97.
Schenk, Martijn F., Mark P. Zwart, Sungmin Hwang, Philip Ruelens, Edouard Severing, Joachim Krug, and J. Arjan G. M. de Visser. 2022. "Population Size Mediates the Contribution of High-Rate and Large-Benefit Mutations to Parallel Evolution." *Nature Ecology & Evolution*, March. https://doi.org/10.1038/s41559-022-01669-3.
Scherer, Stephen W., Charles Lee, Ewan Birney, David M. Altshuler, Evan E. Eichler, Nigel P. Carter, Matthew E. Hurles, and Lars Feuk. 2007. "Challenges and Standards in Integrating Surveys of Structural Variation." *Nature Genetics* 39 (7): S7–15.
Schrider, Daniel R., and Matthew W. Hahn. 2010. "Lower Linkage Disequilibrium at CNVs Is due to Both Recurrent Mutation and Transposing Duplications." *Molecular Biology and Evolution* 27 (1): 103–11.
Schrider, Daniel R., Fabio C. P. Navarro, Pedro A. F. Galante, Raphael B. Parmigiani, Anamaria A. Camargo, Matthew W. Hahn, and Sandro J. de Souza. 2013. "Gene Copy-Number Polymorphism Caused by Retrotransposition in Humans." *PLoS Genetics* 9 (1): e1003242.
Scopel, Eduardo F. C., James Hose, Douda Bensasson, and Audrey P. Gasch. 2021. "Genetic Variation in Aneuploidy Prevalence and Tolerance across Saccharomyces Cerevisiae Lineages." *Genetics* 217 (4). https://doi.org/10.1093/genetics/iyab015.
Sebat, Jonathan, B. Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Pär Lundin, Susanne Månér, et al. 2004. "Large-Scale Copy Number Polymorphism in the Human Genome." *Science* 305 (5683): 525–28.
Segal, Ella Shtifman, Vladimir Gritsenko, Anton Levitan, Bhawna Yadav, Naama Dror, Jacob L. Steenwyk, Yael Silberberg, et al. 2018. "Gene Essentiality Analyzed by In Vivo Transposon Mutagenesis and Machine Learning in a Stable Haploid Isolate of Candida Albicans." *mBio*. https://doi.org/10.1128/mbio.02048-18.
Sellis, Diamantis, Daniel J. Kvitek, Barbara Dunn, Gavin Sherlock, and Dmitri A. Petrov. 2016. "Heterozygote Advantage Is a Common Outcome of Adaptation in Saccharomyces Cerevisiae." *Genetics* 203 (3): 1401–13.
Selmecki, Anna, Anja Forche, and Judith Berman. 2006. "Aneuploidy and Isochromosome Formation in Drug-Resistant Candida Albicans." *Science* 313 (5785): 367–70.
Selmecki, Anna, Maryam Gerami-Nejad, Carsten Paulson, Anja Forche, and Judith Berman. 2008. "An Isochromosome Confers Drug Resistance in Vivo by Amplification of Two Genes, ERG11 and TAC1." *Molecular Microbiology* 68 (3): 624–41.
Selmecki, Anna M., Keely Dulmage, Leah E. Cowen, James B. Anderson, and Judith Berman. 2009. "Acquisition of Aneuploidy Provides Increased Fitness during the Evolution of Antifungal Drug Resistance." *PLoS Genetics* 5 (10): e1000705.
Shapira, Stuart K., and Victoria G. Finnerty. 1986. "The Use of Genetic Complementation in the Study of Eukaryotic Macromolecular Evolution: Rate of Spontaneous Gene Duplication at Two Loci ofDrosophila Melanogaster." *Journal of Molecular Evolution* 23 (2): 159–67.
Sharp, Nathaniel P., Linnea Sandell, Christopher G. James, and Sarah P. Otto. 2018. "The Genome-Wide Rate and Spectrum of Spontaneous Mutations Differ between Haploid and Diploid Yeast." *Proceedings of the National Academy of Sciences of the United States of America* 115 (22): E5046–55.
Sheltzer, Jason M., Heidi M. Blank, Sarah J. Pfau, Yoshie Tange, Benson M. George, Timothy J. Humpton, Ilana L. Brito, Yasushi Hiraoka, Osami Niwa, and Angelika Amon. 2011. "Aneuploidy Drives Genomic Instability in Yeast." *Science* 333 (6045): 1026–30.
Sheltzer, Jason M., Eduardo M. Torres, Maitreya J. Dunham, and Angelika Amon. 2012. "Transcriptional Consequences of Aneuploidy." *Proceedings of the National Academy of Sciences of the United States of America* 109 (31): 12644–49.
Shewaramani, Sonal, Thomas J. Finn, Sinead C. Leahy, Rees Kassen, Paul B. Rainey, and Christina D. Moon. 2017. "Anaerobically Grown Escherichia Coli Has an Enhanced

Mutation Rate and Distinct Mutational Spectra." *PLoS Genetics* 13 (1): e1006570.
Shlien, Adam, and David Malkin. 2009. "Copy Number Variations and Cancer." *Genome Medicine* 1 (6): 62–62.
Shor, Erika, Catherine A. Fox, and James R. Broach. 2013. "The Yeast Environmental Stress Response Regulates Mutagenesis Induced by Proteotoxic Stress." *PLoS Genetics* 9 (8): e1003680.
Siguier, Patricia, Edith Gourbeyre, and Mick Chandler. 2014. "Bacterial Insertion Sequences: Their Genomic Impact and Diversity." *FEMS Microbiology Reviews* 38 (5): 865–91.
Sisson, S. A., Y. Fan, and Mark M. Tanaka. 2007. "Sequential Monte Carlo without Likelihoods." *Proceedings of the National Academy of Sciences of the United States of America* 104 (6): 1760–65.
Skelly, Daniel A., Gennifer E. Merrihew, Michael Riffle, Caitlin F. Connelly, Emily O. Kerr, Marnie Johansson, Daniel Jaschob, et al. 2013. "Integrative Phenomics Reveals Insight into the Structure of Phenotypic Diversity in Budding Yeast." *Genome Research* 23 (9): 1496–1504.
Skourti-Stathaki, Konstantina, and Nicholas J. Proudfoot. 2014. "A Double-Edged Sword: R Loops as Threats to Genome Integrity and Powerful Regulators of Gene Expression." *Genes & Development* 28 (13): 1384–96.
Slack, Andrew, P. C. Thornton, Daniel B. Magner, Susan M. Rosenberg, and P. J. Hastings. 2006. "On the Mechanism of Gene Amplification Induced under Stress in Escherichia Coli." *PLoS Genetics* 2 (4): e48.
Slechta, E. Susan, Kim L. Bunny, Elisabeth Kugelberg, Eric Kofoid, Dan I. Andersson, and John R. Roth. 2003. "Adaptive Mutation: General Mutagenesis Is Not a Programmed Response to Stress but Results from Rare Coamplification of dinB with Lac." *Proceedings of the National Academy of Sciences of the United States of America* 100 (22): 12847–52.
Sniegowski, P. D., P. J. Gerrish, and R. E. Lenski. 1997. "Evolution of High Mutation Rates in Experimental Populations of E. Coli." *Nature* 387 (6634): 703–5.
Sonti, R. V., and J. R. Roth. 1989. "Role of Gene Duplications in the Adaptation of Salmonella Typhimurium to Growth on Limiting Carbon Sources." *Genetics* 123 (1): 19–28.
Sopko, Richelle, Dongqing Huang, Nicolle Preston, Gordon Chua, Balázs Papp, Kimberly Kafadar, Mike Snyder, et al. 2006. "Mapping Pathways and Phenotypes by Systematic Gene Overexpression." *Molecular Cell* 21 (3): 319–30.
Sousa, Jorge A. Moura de, Jorge A. Moura de Sousa, Paulo R. A. Campos, and Isabel Gordo. 2013. "An ABC Method for Estimating the Rate and Distribution of Effects of Beneficial Mutations." *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evt045.
Spielmann, Malte, Darío G. Lupiáñez, and Stefan Mundlos. 2018. "Structural Variation in the 3D Genome." *Nature Reviews. Genetics* 19 (7): 453–67.
Sprouffske, Kathleen, and Andreas Wagner. 2016. "Growthcurver: An R Package for Obtaining Interpretable Metrics from Microbial Growth Curves." *BMC Bioinformatics* 17 (April): 172.
Stanbrough, M., and B. Magasanik. 1995. "Transcriptional and Posttranslational Regulation of the General Amino Acid Permease of Saccharomyces Cerevisiae." *Journal of Bacteriology* 177 (1): 94–102.
———. 1996. "Two Transcription Factors, Gln3p and Nil1p, Use the Same GATAAG Sites to Activate the Expression of GAP1 of Saccharomyces Cerevisiae." *Journal of Bacteriology* 178 (8): 2465–68.
Stankiewicz, Paweł, and James R. Lupski. 2002. "Genome Architecture, Rearrangements and Genomic Disorders." *Trends in Genetics: TIG* 18 (2): 74–82.
Starlinger, P. 1977. "DNA Rearrangements in Procaryotes." *Annual Review of Genetics* 11: 103–26.
Steinrueck, Magdalena, and Călin C. Guet. 2017. "Complex Chromosomal Neighborhood Effects Determine the Adaptive Potential of a Gene under Selection." *eLife* 6 (July).

https://doi.org/10.7554/eLife.25100.
Stingele, Silvia, Gabriele Stoehr, Karolina Peplowska, Jürgen Cox, Matthias Mann, and Zuzana Storchova. 2012. “Global Analysis of Genome, Transcriptome and Proteome Reveals the Response to Aneuploidy in Human Cells.” *Molecular Systems Biology* 8: 608.
Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. 2009. “The Cancer Genome.” *Nature* 458 (7239): 719–24.
Straus, D. S. 1975. “Selection for a Large Genetic Duplication in Salmonella Typhimurium.” *Genetics* 80 (2): 227–37.
Sturtevant, A. H. 1925. “The Effects of Unequal Crossing over at the Bar Locus in Drosophila.” *Genetics* 10 (2): 117–47.
Sui, Yang, Lei Qi, Jian-Kun Wu, Xue-Ping Wen, Xing-Xing Tang, Zhong-Jun Ma, Xue-Chang Wu, et al. 2020. “Genome-Wide Mapping of Spontaneous Genetic Alterations in Diploid Yeast Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (45): 28191–200.
Sung, Way, Matthew S. Ackerman, Jean-François Gout, Samuel F. Miller, Emily Williams, Patricia L. Foster, and Michael Lynch. 2015. “Asymmetric Context-Dependent Mutation Patterns Revealed through Mutation–Accumulation Experiments.” *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msv055.
Sunnåker, Mikael, Alberto Giovanni Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. 2013. “Approximate Bayesian Computation.” *PLoS Computational Biology* 9 (1): e1002803.
Sunshine, Anna B., Celia Payen, Giang T. Ong, Ivan Liachko, Kean Ming Tan, and Maitreya J. Dunham. 2015. “The Fitness Consequences of Aneuploidy Are Driven by Condition-Dependent Gene Effects.” *PLoS Biology* 13 (5): e1002155.
Sun, Song, Otto G. Berg, John R. Roth, and Dan I. Andersson. 2009. “Contribution of Gene Amplification to Evolution of Increased Antibiotic Resistance in Salmonella Typhimurium.” *Genetics* 182 (4): 1183–95.
Sun, Song, Rongqin Ke, Diarmaid Hughes, Mats Nilsson, and Dan I. Andersson. 2012. “Genome-Wide Detection of Spontaneous Chromosomal Rearrangements in Bacteria.” *PloS One* 7 (8): e42639.
Suzuki, Yo, Robert P. St Onge, Ramamurthy Mani, Oliver D. King, Adrian Heilbut, Vyacheslav M. Labunskyy, Weidong Chen, et al. 2011. “Knocking out Multigene Redundancies via Cycles of Sexual Assortment and Fluorescence Selection.” *Nature Methods* 8 (2): 159–64.
Tanaka, Mark M., Andrew R. Francis, Fabio Luciani, and S. A. Sisson. 2006. “Using Approximate Bayesian Computation to Estimate Tuberculosis Transmission Parameters From Genotype Data.” *Genetics*. https://doi.org/10.1534/genetics.106.055574.
Tang, Yun-Chi, and Angelika Amon. 2013. “Gene Copy-Number Alterations: A Cost-Benefit Analysis.” *Cell* 152 (3): 394–405.
Tavaré, Simon, David J. Balding, R. C. Griffiths, and Peter Donnelly. 1997. “Inferring Coalescence Times From DNA Sequence Data.” *Genetics*. https://doi.org/10.1093/genetics/145.2.505.
Tejero-Cantero, Alvaro, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro Gonçalves, David Greenberg, and Jakob Macke. 2020. “Sbi: A Toolkit for Simulation-Based Inference.” *Journal of Open Source Software*. https://doi.org/10.21105/joss.02505.
Terhorst, Allegra, Arzu Sandikci, Abigail Keller, Charles A. Whittaker, Maitreya J. Dunham, and Angelika Amon. 2020. “The Environmental Stress Response Causes Ribosome Loss in Aneuploid Yeast Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (29): 17031–40.
Thomas, Barbara J., and Rodney Rothstein. 1989. “Elevated Recombination Rates in

Transcriptionally Active DNA." *Cell* 56 (4): 619–30.
Todd, Robert T., and Anna Selmecki. 2020. "Expandable and Reversible Copy Number Amplification Drives Rapid Adaptation to Antifungal Drugs." *eLife* 9 (July). https://doi.org/10.7554/eLife.58349.
Todd, Robert T., Tyler D. Wikoff, Anja Forche, and Anna Selmecki. 2019. "Genome Plasticity in Candida Albicans Is Driven by Long Repeat Sequences." *eLife*. https://doi.org/10.7554/elife.45954.
Torada, Luis, Lucrezia Lorenzon, Alice Beddis, Ulas Isildak, Linda Pattini, Sara Mathieson, and Matteo Fumagalli. 2019. "ImaGene: A Convolutional Neural Network to Quantify Natural Selection from Genomic Data." *BMC Bioinformatics* 20 (Suppl 9): 337.
Torres, Eduardo M., Noah Dephoure, Amudha Panneerselvam, Cheryl M. Tucker, Charles A. Whittaker, Steven P. Gygi, Maitreya J. Dunham, and Angelika Amon. 2010. "Identification of Aneuploidy-Tolerating Mutations." *Cell* 143 (1): 71–83.
Torres, Eduardo M., Tanya Sokolsky, Cheryl M. Tucker, Leon Y. Chan, Monica Boselli, Maitreya J. Dunham, and Angelika Amon. 2007. "Effects of Aneuploidy on Cellular Physiology and Cell Division in Haploid Yeast." *Science* 317 (5840): 916–24.
Tsai, Hung-Ji, and Anjali Nelliat. 2019. "A Double-Edged Sword: Aneuploidy Is a Prevalent Strategy in Fungal Adaptation." *Genes*. https://doi.org/10.3390/genes10100787.
Tsai, Hung-Ji, Anjali R. Nelliat, Mohammad Ikbal Choudhury, Andrei Kucharavy, William D. Bradford, Malcolm E. Cook, Jisoo Kim, et al. 2019. "Hypo-Osmotic-like Stress Underlies General Cellular Defects of Aneuploidy." *Nature* 570 (7759): 117–21.
Turner, Daniel J., Marcos Miretti, Diana Rajan, Heike Fiegler, Nigel P. Carter, Martyn L. Blayney, Stephan Beck, and Matthew E. Hurles. 2008. "Germline Rates of de Novo Meiotic Deletions and Duplications Causing Several Genomic Disorders." *Nature Genetics* 40 (1): 90–95.
Turner, Kristen M., Viraj Deshpande, Doruk Beyter, Tomoyuki Koga, Jessica Rusert, Catherine Lee, Bin Li, et al. 2017. "Extrachromosomal Oncogene Amplification Drives Tumour Evolution and Genetic Heterogeneity." *Nature* 543 (7643): 122–25.
Turner, Thomas L., Andrew D. Stewart, Andrew T. Fields, William R. Rice, and Aaron M. Tarone. 2011. "Population-Based Resequencing of Experimentally Evolved Populations Reveals the Genetic Basis of Body Size Variation in Drosophila Melanogaster." *PLoS Genetics* 7 (3): e1001336.
Tutuncuoglu, Beril, and Nevan J. Krogan. 2019. "Mapping Genetic Interactions in Cancer: A Road to Rational Combination Therapies." *Genome Medicine*. https://doi.org/10.1186/s13073-019-0680-4.
Usakin, Lev A., Galina L. Kogan, Alla I. Kalmykova, and Vladimir A. Gvozdev. 2005. "An Alien Promoter Capture as a Primary Step of the Evolution of Testes-Expressed Repeats in the Drosophila Melanogaster Genome." *Molecular Biology and Evolution* 22 (7): 1555–60.
Van den Bergh, Bram, Toon Swings, Maarten Fauvart, and Jan Michiels. 2018. "Experimental Design, Population Dynamics, and Diversity in Microbial Experimental Evolution." *Microbiology and Molecular Biology Reviews: MMBR* 82 (3). https://doi.org/10.1128/MMBR.00008-18.
Veitia, Reiner A. 2004. "Gene Dosage Balance in Cellular Pathways: Implications for Dominance and Gene Duplicability." *Genetics*.
Veitia, Reiner A., Samuel Bottani, and James A. Birchler. 2008. "Cellular Reactions to Gene Dosage Imbalance: Genomic, Transcriptomic and Proteomic Effects." *Trends in Genetics: TIG* 24 (8): 390–97.
Venkataram, Sandeep, Barbara Dunn, Yuping Li, Atish Agarwala, Jessica Chang, Emily R. Ebel, Kerry Geiler-Samerotte, et al. 2016. "Development of a Comprehensive Genotype-to-Fitness Map of Adaptation-Driving Mutations in Yeast." *Cell* 166 (6): 1585–96.e22.

Venkataram, Sandeep, Ross Monasky, Shohreh H. Sikaroodi, Sergey Kryazhimskiy, and Betul Kacar. 2020. “Evolutionary Stalling and a Limit on the Power of Natural Selection to Improve a Cellular Module.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (31): 18582–90.
Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods* 17 (3): 261–72.
Wang, Yuexing, Guosheng Xiong, Jiang Hu, Liang Jiang, Hong Yu, Jie Xu, Yunxia Fang, et al. 2015. “Copy Number Variation at the GL7 Locus Contributes to Grain Size Diversity in Rice.” *Nature Genetics* 47 (8): 944–48.
Weinreich, Daniel M., Nigel F. Delaney, Mark A. Depristo, and Daniel L. Hartl. 2006. “Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins.” *Science* 312 (5770): 111–14.
Wei, Xinzhu, and Jianzhi Zhang. 2019. “Patterns and Mechanisms of Diminishing Returns from Beneficial Mutations.” *Molecular Biology and Evolution* 36 (5): 1008–21.
Werdyani, Salem, Yajun Yu, Georgia Skardasi, Jingxiong Xu, Konstantin Shestopaloff, Wei Xu, Elizabeth Dicks, et al. 2017. “Germline INDELs and CNVs in a Cohort of Colorectal Cancer Patients: Their Characteristics, Associations with Relapse-Free Survival Time, and Potential Time-Varying Effects on the Risk of Relapse.” *Cancer Medicine*. https://doi.org/10.1002/cam4.1074.
Whale, A. J., M. King, R. M. Hull, F. Krueger, and J. Houseley. 2021. “Stimulation of Adaptive Gene Amplification by Origin Firing under Replication Fork Constraint.” *bioRxiv*. https://www.biorxiv.org/content/10.1101/2021.03.04.433911v1.abstract.
Whoriskey, S. K., V. H. Nghiem, P. M. Leong, J. M. Masson, and J. H. Miller. 1987. “Genetic Rearrangements and Gene Amplification in Escherichia Coli: DNA Sequences at the Junctures of Amplified Gene Fusions.” *Genes & Development* 1 (3): 227–37.
Wiles, Amy M., Houjian Cai, Fred Naider, and Jeffrey M. Becker. 2006. “Nutrient Regulation of Oligopeptide Transport in Saccharomyces Cerevisiae.” *Microbiology* 152 (Pt 10): 3133–45.
Wilson, Thomas E., Martin F. Arlt, So Hae Park, Sountharia Rajendran, Michelle Paulsen, Mats Ljungman, and Thomas W. Glover. 2015. “Large Transcription Units Unify Copy Number Variants and Common Fragile Sites Arising under Replication Stress.” *Genome Research* 25 (2): 189–200.
Winzeler, E. A., D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, et al. 1999. “Functional Characterization of the S. Cerevisiae Genome by Gene Deletion and Parallel Analysis.” *Science* 285 (5429): 901–6.
Wright, Dominic, Henrik Boije, Jennifer R. S. Meadows, Bertrand Bed’hom, David Gourichon, Agathe Vieaud, Michèle Tixier-Boichard, et al. 2009. “Copy Number Variation in Intron 1 of SOX5 Causes the Pea-Comb Phenotype in Chickens.” *PLoS Genetics* 5 (6): e1000512–e1000512.
Yona, Avihu H., Yair S. Manor, Rebecca H. Herbst, Gal H. Romano, Amir Mitchell, Martin Kupiec, Yitzhak Pilpel, and Orna Dahan. 2012. “Chromosomal Duplication Is a Transient Evolutionary Solution to Stress.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (51): 21010–15.
Yu, G., F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang. 2010. “GOSemSim: An R Package for Measuring Semantic Similarity among GO Terms and Gene Products.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq064.
Yu, Guangchuang, Li-Gen Wang, Yanyan Han, and Qing-Yu He. 2012. “clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters.” *OMICS: A Journal of Integrative Biology*. https://doi.org/10.1089/omi.2011.0118.
Zarrei, Mehdi, Jeffrey R. MacDonald, Daniele Merico, and Stephen W. Scherer. 2015. “A Copy

Number Variation Map of the Human Genome." *Nature Reviews. Genetics* 16 (3): 172–83.
Zhang, Feng, Mehrdad Khajavi, Anne M. Connolly, Charles F. Towne, Sat Dev Batish, and James R. Lupski. 2009. "The DNA Replication FoSTeS/MMBIR Mechanism Can Generate Genomic, Genic and Exonic Complex Rearrangements in Humans." *Nature Genetics* 41 (7): 849–53.
Zhang, F., W. Gu, M. E. Hurles, and J. R. Lupski. 2009. "Copy Number Variation in Human Health, Disease, and Evolution." *Annual Review of Genomics and Human Genetics* 10: 451–81.
Zhang, Hengshan, Ane F. B. Zeidler, Wei Song, Christopher M. Puccia, Ewa Malc, Patricia W. Greenwell, Piotr A. Mieczkowski, Thomas D. Petes, and Juan Lucas Argueso. 2013. "Gene Copy-Number Variation in Haploid and Diploid Strains of the Yeast Saccharomyces Cerevisiae." *Genetics* 193 (3): 785–801.
Zhang, L., and A. Hach. 1999. "Molecular Mechanism of Heme Signaling in Yeast: The Transcriptional Activator Hap1 Serves as the Key Mediator." *Cellular and Molecular Life Sciences: CMLS* 56 (5-6): 415–26.
Zhao, Lu, Zhimin Liu, Sasha F. Levy, and Song Wu. 2017. "Bartender: A Fast and Accurate Clustering Algorithm to Count Barcode Reads." *Bioinformatics* , October. https://doi.org/10.1093/bioinformatics/btx655.
Zhou, Dan, Nitin Udpa, Merril Gersten, Deeann W. Visk, Ali Bashir, Jin Xue, Kelly A. Frazer, et al. 2011. "Experimental Selection of Hypoxia-Tolerant Drosophila Melanogaster." *Proceedings of the National Academy of Sciences of the United States of America* 108 (6): 2349–54.
Zhou, Kai, Abram Aertsen, and Chris W. Michiels. 2014. "The Role of Variable DNA Tandem Repeats in Bacterial Adaptation." *FEMS Microbiology Reviews* 38 (1): 119–41.
Zhu, Jin, Hung-Ji Tsai, Molly R. Gordon, and Rong Li. 2018. "Cellular Stress Associated with Aneuploidy." *Developmental Cell* 44 (4): 420–31.
Zhu, Yuan O., Gavin Sherlock, and Dmitri A. Petrov. 2016. "Whole Genome Analysis of 132 Clinical Saccharomyces Cerevisiae Strains Reveals Extensive Ploidy Variation." *G3* 6 (8): 2421–34.
Zhu, Yuan O., Mark L. Siegal, David W. Hall, and Dmitri A. Petrov. 2014. "Precise Estimates of Mutation Rate and Spectrum in Yeast." *Proceedings of the National Academy of Sciences of the United States of America* 111 (22): E2310–18.
Zichner, Thomas, David A. Garfield, Tobias Rausch, Adrian M. Stütz, Enrico Cannavó, Martina Braun, Eileen E. M. Furlong, and Jan O. Korbel. 2013. "Impact of Genomic Structural Variation in Drosophila Melanogaster Based on Population-Scale Sequencing." *Genome Research* 23 (3): 568–79.
Żmieńko, Agnieszka, Anna Samelak, Piotr Kozłowski, and Marek Figlerowicz. 2014. "Copy Number Polymorphism in Plant Genomes." *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 127 (1): 1–18.
Zuellig, Matthew P., and Andrea L. Sweigart. 2018. "Gene Duplicates Cause Hybrid Lethality between Sympatric Species of Mimulus." *PLoS Genetics* 14 (4): e1007130.