The dynamics and molecular basis of adaptive evolution in nutrient-limited environments

By

Jungeui Hong

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Biology

New York University

January, 2015

_____

Dr. David Gresham

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

DEDICATION

TO. *S.S.N.*

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. David Gresham, for his role in inspiring this project, as well as his guidance and support to fulfilling my research aims successfully. I am also indebted to committee members, Drs. Justin Blau, Kris Gunsalus, Mark Siegal and Christopher Mason for their direction and comments during my Ph.D. work and annual committee meetings. This dissertation could not have come to fruition without help and feedback from my talented lab mates: Niki, Sam, Nathan, Benjy, Naomi, Darach, Stephanie and all undergraduates and high school students who have devoted their time and efforts to the lab.

I am also everlastingly grateful to my wife, Seungji, for her patience and caring me and my boy, Joshua. Also, many thanks to my families in Korea for their supports in many ways so that I could make a big progression in my life in this different situation.

ABSTRACT

Understanding the dynamics and mechanisms of adaptive evolution is a central question in evolutionary biology. However, realizing this goal remains challenging due to the difficulty of observing adaptive evolution in real time and deducing the its molecular basis. Long-term Experimental Evolution (LTEE) using microbes and chemostats provides a means of overcoming to these limitations to address these central questions. I studied the evolution of genetic networks in *Saccharomyces cerevisiae* (budding yeast) populations propagated for more than 200 generations in different nitrogen-limiting conditions using chemostats. I find that rapid adaptive evolution in nitrogen-poor environments is dominated by the *de novo* generation and selection of copy number variants (CNVs), a large fraction of which contain genes encoding specific nitrogen transporters. The large fitness increases associated with these alleles limits the genetic heterogeneity of adapting populations even in environments with multiple nitrogen sources. Complete identification of acquired point mutations, in individual lineages and entire populations, identified heterogeneity at the level of genetic loci but common themes at the level of functional modules, including genes controlling phosphatidylinositol-3-phosphate metabolism and vacuole biogenesis. Adaptive strategies shared with other nutrient-limited environments point to selection of genetic variation in the TORC1 and Ras/PKA signaling pathways as a general mechanism underlying improved growth in nutrient-limited environments. By studying the fitness of individual alleles, and

their combination, as well as the evolutionary history of the evolving population, I find that the order in which adaptive mutations are acquired is constrained by epistasis. I observed the repeated selection of non-synonymous mutations in the zinc finger DNA binding domain of the GATA transcription factor, *GAT1*, an activator of the nitrogen catabolite repression (NCR) regulon. The functional effects of GAT1 mutations are exerted both directly, and indirectly by rewiring of incoherent feed-forward loops comprising multiple GATA transcription factors and their common NCR regulon targets. This suggests that under strong selection the evolution of gene expression is highly repeatable and that rewiring transcriptional networks can lead to both direct and indirect effects. Studies using LTEE are potentially applicable to understanding pathogenic strategies adopted by viruses, microbes and even human cancer cells. For example, recurrent mutations in the DNA binding domain of GAT1 is reminiscent of recurrent missense mutations in the DNA binding domain of TP53 found in a variety of tumors.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER1: INTRODUCTION

Natural selection has been one of the most influential ideas in biology since Charles Darwin brought his insight to the field in his seminal work *On the Origin of Species* [1] 150 years ago. Darwin's theory of evolution provides an explanation for the incredible biodiversity of taxa he observed in the natural world. Adaptive evolution is driven by natural selection in which a population becomes more suited to a particular environment by selecting variations that confer a reproductive advantage. Selection and heritable (genetic) variations are two major determinants of adaptive evolution. Since Darwin's idea, the concept of adaptive evolution has been applied to a variety of fields ranging from microbiology [2], ecology [3] and impacts virtually every aspect of biology.

Understanding the dynamics and mechanisms of adaptive evolution is a long-standing question in the field. Indeed, the Darwinian evolutionary framework has contributed to predictions concerning the evolutionary dynamics and outcomes. In adaptive evolution, mutation increases genetic diversity while selection favors only fitter genotypes thereby decreasing diversity. If selection is strong and the mutation supply rate is low, the fittest allele will fix in the population (i.e. be present in every individual) before the next mutation occurs, making the dynamics of evolution relatively simple (sequential hard sweep) [4,5]. By contrast, if multiple mutations that are near-equally beneficial can be introduced together, they can compete with each other in a process, which results in soft sweep [6].

However, a consensus about which evolutionary dynamics dominate in adapting populations has not been reached and our understanding remains mostly theoretical. One of the fundamental bottlenecks is that evolution is rarely observable in real-time as Darwin already remarked in his original work in 1859. The forces and processes underlying adaptive evolution are necessarily inferred from extant organisms making it hard to observe the evolutionary dynamics in real-time and to distinguish neutral from adaptive alleles. Therefore, most studies of adaptive evolution rely on comparative studies with living organisms and/or theoretical prediction by necessity.

The speed, causes, and dynamics of adaptive evolution are determined in part by the molecular basis of adaptive evolution. However, for much of the twentieth century evolutionary biology diverged from molecular biology in scientific culture. Recently, the reductionist approach of molecular biology has emerged as an essential complement to addressing various issues that remain unresolved in evolutionary biology [7]. The incorporation of the evolutionary theory into modern genetics and molecular biology shows a great success in isolating and characterizing many aspects of genetic architecture that underlies adaptive evolution in natural populations [8]. However, understanding the molecular basis of adaptive evolution at a systems level still remains a challenge.

An additional challenge to studying adaptive evolution is understanding and controlling multiple factors that determine the evolutionary dynamics. The types and strength of selective pressure is dependent on environmental conditions a

population is experiencing. A major source of genetic variations at the molecular level are spontaneous mutations that randomly arise from intrinsic DNA replication errors, spontaneous lesions and transposable genetic elements [9]. Among additional complicating factors are epistatic interaction between adaptive alleles [10-12], genetic drift [13-15] and random meiotic recombination [16-18]. All these factors contribute to the rate or mode of adaptive paths, and ultimately hinder fixation by the fittest clone. However, the molecular details underlying these factors and experimental approaches to monitor them in real time during evolution are far from complete.

Experimental evolution using microbes provide one means of overcoming these limitations in evolutionary biology [19-25]. Using microbes with short doubling times and archiving population samples allows maintenance of a 'fossil record' of the evolving population and observations of the evolutionary dynamics in the lab within a reasonable timeframe. In addition, well-controlled and replicated experimental set-ups enable us to rule out or minimize the effects of stochastic determinants such as drift and recombination and mainly focus on the interplay of selection and mutation. The effect of drift can be minimized using a constant culturing system with a large population size. Using asexually reproducing microbes rules out meiotic recombination that randomizes the order of fixed mutations in each lineage during the course of evolution. The types and strength of selection can be modulated by chemical treatment or nutrient limitation with a well-defined composition. Various genetic and molecular tools are readily available

to identify genomic variations and their phenotypic consequences in the relevant environmental condition. Finally, microbial populations themselves are of great importance in evolutionary biology not only as a pathogen to humans but also for their ecological roles in nature.

Experimental evolution using microbes has a relatively long history. A classical example of experimental evolution is a work done by Novick and colleagues in 1961 showing that *Escherichia coli* evolved to repetitively select amplification alleles of the lac operon under continuous lactose limiting media [26]. In later studies, other groups found examples of a 'mutator phenotype' while investigating the role of mutation rate in bacterial experimental evolutions [27,28]. The group of Richard Lenski is one of the founders of the 'Long-Term Experimental Evolution (LTEE)' [29-31]. They evolved twelve independent populations of bacteria for over 60,000 generations (literally more than 26 years as of the year of 2014) using a daily serial dilution method in a nutrient rich media. They have reported many fundamental aspects of adaptive evolution in bacteria based on analysis of LTEEs at the phenotypic and genotypic levels. The most striking recent result was that one *E. coli* population evolved to use citric acid as a carbon source in an aerobic condition [32]. *Saccharomyces cerevisiae* (budding yeast) is a eukaryotic microbe that is ideally suited to experimental evolutions owing to the available molecular genetic tools and its well-characterized genome features. For instance, Brown et al. found that amplification of HXT6/7 genes which encodes high affinity glucose transporters led to an increased rate of glucose uptake resulting in increased fitness

under glucose-limited conditions [33]. Despite progress in understanding adaptive evolution, these early studies are mostly based on analysis of specific target loci of interest and were technically limited in the scale of genotyping.

Recently, LTEEs combined with high-throughput analysis of genomic variations have provided a more comprehensive list of adaptive alleles and their dynamics [34-37]. Despite the variation in selective regimes and strength and models used, most results from different LTEEs are beginning to provide a consistent view of adaptive evolution. One emerging consensus from LTEEs studies is that evolutionary trajectories are constrained by clonal interference and epistasis under strong selective pressures [34,37-40]. The comparative ease of whole genome analysis means that LTEE is now able to answer more complex questions, i.e., the role of different types of genetic variation, the causes and consequences of antagonistic pleiotropy (the phenomenon where one variation has opposite effects on fitness depending on conditions), the distribution of fitness effects, how pathways and networks comprising multiple genes evolve, how interactions between genes influence the dynamics and outcome of adaptive evolution, whether evolution is historically contingent or convergent, and how adaptive strategies learned from microbial LTEEs be applied to other systems. My dissertation aims to address how functional modules or pathways such as gene regulatory networks (regulons) comprising multiple loci evolve under strong selective pressure.

In my research, I used budding yeast as a model system, nitrogen limitation as a source of selective pressure and chemostats as an experimental means of

maintaining evolving populations in constant environments over long time periods. Budding yeast has a number of properties that make it ideally suited to the study of molecular basis of adaptive evolution in LTEEs. First, yeast genetics provides a powerful tool for isolating and combining different adaptive alleles and studying their functional effects. In addition, genomic features and cell growth regulating pathways and processes in yeast are well characterized (**Figure 1.1A**).

For example, the molecular mechanisms underlying transcriptional control of nitrogen utilization in yeast via the so called nitrogen catabolite repression (NCR) regulon have been extensively studied [41-43] (**Figure 1.1B**). Limiting nitrogen concentrations in a culture medium imposes a very strong selective pressure since nitrogen is one of the essential nutrients. Under such a nitrogen limited or poor condition, transcriptional expression of a set of functionally related genes for nitrogen uptake and catabolism is regulated by four GATA factors: two activator (GAT1 and GLN3) and two repressors (DAL80 and GZF3). The well-characterized regulation of nitrogen utilization provides an ideal model system for studying the molecular basis of adaptive evolutions.

A chemostat was used as the culturing device in this work in order to generate a stable and continuous selective pressure. Its technical details and relative usefulness in LTEEs compared to the serial dilution in batch cultures will be discussed in the following **Section 1**. Next, I adopted next-generation sequencing (NGS) for high-throughput genome analysis in this study (see **Section 2**), which allowed me to identify the full spectrum of adaptive alleles and their dynamics.

A B

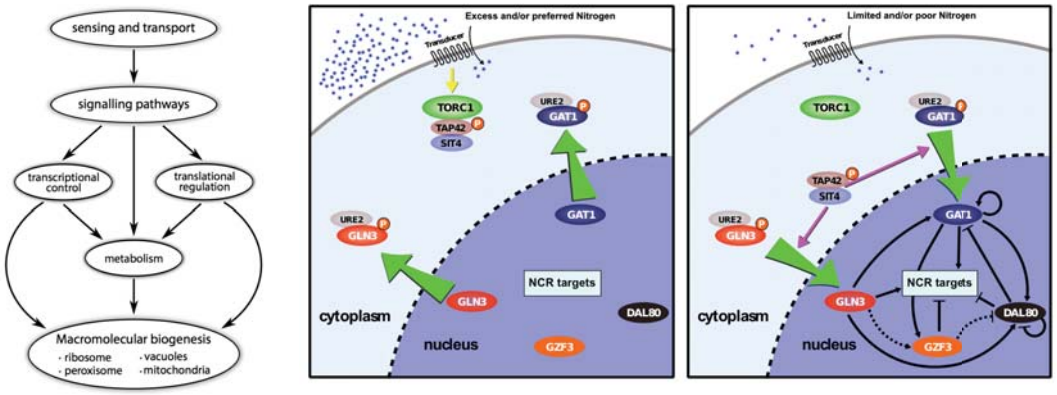

Figure 1.1. A. Scheme of central nutrient utilization pathway. B. Nitrogen Catabolite Repression (NCR) in yeast

My thesis project represents a successful application of LTEE to understanding the dynamics and molecular basis of adaptive evolution under constant nutrient-limited environments. One important long-term perspective from my studies is that LTEE is informative for understanding adaptive evolution of viruses, microbes and even cancer cells in the area of human health care [44-46]. Experimental studies of viral and bacterial infection suggest that adaptability of pathogens to their host is more complex than predictions made by classical theories and that their evolutionary outcomes are more stochastic (unpredictable). Theories of cancer evolution have been strongly influenced by evolutionary thought [47] but require more empirical evidence from experimental evolution approaches and high-throughput genomic screening methods. This dissertation serves as a starting point for applying concepts of the evolution of genetic networks to understanding the adaptive

strategies that tumor cells use to proliferate and metastasize as discussed in Chapter 2.

Before introducing the main results, two important technical backgrounds will be briefly reviewed here: (1) Chemostats and (2) Next-generation sequencing. Following three main chapters will cover an expanded version of research papers to which I contributed as a first author during the PhD training.

1.1. CHEMOSTATS

> *This section is based on the review paper "**The functional basis of adaptive evolution in chemostats**" by David Gresham and Jungeui Hong, copyright © 2014 by the FEMBS Microbiology Review, all rights reserved (Gresham and Hong, 2014). I performed computational simulation for inferring the dynamics of cell growth and nutrient concentration in a chemostat and generated all figures.*

Chemostats are devices for culturing microbes in a liquid medium of fixed volume at a constant growth rate by modulating constant in- and out-flows of the culture (see **Figure 1.2**). This method of culturing was first introduced by Jacques Monod [48], and Leo Szilard and Aaron Novick [49,50] independently in 1950. In chemostats, the growing population is in a steady-state where the growth rate is equal to the rate at which the culture is diluted.

The key technical advantages of the chemostats for studying experimental evolution are followings: (1) a steady-state of the cell culture can be maintained for a long time period, (2) the growth rate of cells is determined by experimentally controlled dilution rate, (3) the types, strength and consistency of selective pressure

experienced by the organism and population size is under precise experimental control, and (4) as a chemostat environment is usually new for the organism, fitness increases in evolving lineages and populations are typically large providing better statistical power for dissecting multi-locus alleles and their epistatic interactions.

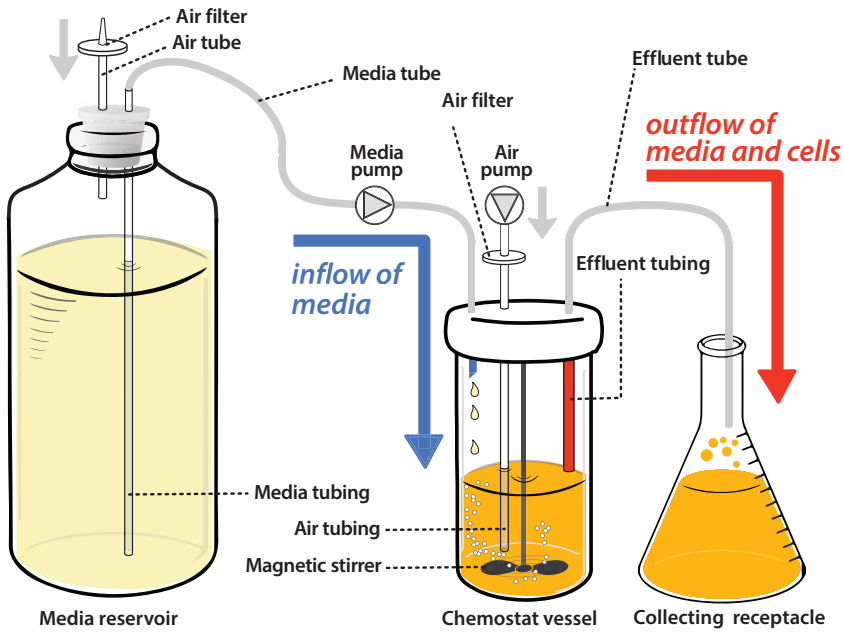

Figure 1.2. Design of a chemostat. Typically, a chemostat comprises a culture vessel in which the population grows under continuous agitation and aeration. New media flows into the vessel at a defined rate. At the same rate, culture containing cells and medium is removed from the chemostat. The flow of media and culture is maintained using a pumping apparatus and holding the chemostat vessel under positive pressure by means of a constant air flow.

The dynamics of cell growth in a chemostat is described using a hyperbolic function of the growth rate and two coupled differential equations:

$$\mu=\mu_{max}\cdot s/(K_s + s) \quad \text{(Eq 1)}$$

$$\frac{dx}{dt} = \mu_{\max} \frac{s}{K_s + s} x - Dx \quad \text{(Eq 2)}$$

$$\frac{ds}{dt} = DR - Ds - \frac{x}{Y} \mu_{\max} \frac{s}{K_s + s} \quad \text{(Eq 3)}$$

where '***x***' is the number of cells,
'$\boldsymbol{\mu}$' is the growth rate of the cells,
'$\boldsymbol{\mu_{max}}$' is the maximal growth rate of the cells,
'***s***' is the residual concentration of the limiting nutrient,
'***R***' is the concentration of the limiting nutrient in the feed medium
'$\boldsymbol{K_s}$' is the substrate concentration at half-maximal μ,
'***D***' is the culture dilution rate,
*'**Y**'* is the number of cells produced per mole of the limiting nutrient

Eq 1 is the relationship between growth rate (μ) and limiting nutrient concentration (s) and inferred based on the empirical measurements of *E. coli* growth rates in different nutrient concentrations [49,51]. *dx/dt* (Eq 2) and *ds/dt* (Eq 3) represent the temporal dynamics of the number of cells and the nutrient concentration change, respectively. In a steady state in the chemostat (see **Figure 1.3**), Eq 2 and Eq 3 are equal to zero (i.e. *dx/dt = 0* and *ds/dt = 0*), where growth rate (μ) is sub-maximal and exponential (i.e. constant per unit time) and equal to the culture dilution rate (D). Thus, the doubling time (i.e. the generation time) of the exponentially growing population is simply *ln(2)/D*. A variety of steady-state conditions can be established by simply varying the dilution rate (D) in a chemostat until the dilution

rate is greater than the maximal growth rate (μ_{max}) of the cells. If D > $\mu_{max,}$ cells in a chemostat will be washed out by the fresh medium. The three growth parameters – μ_{max}, K_s and Y – are intrinsic properties of the cell and therefore potentially modified by mutation and selection of experimental evolution in a chemostat.

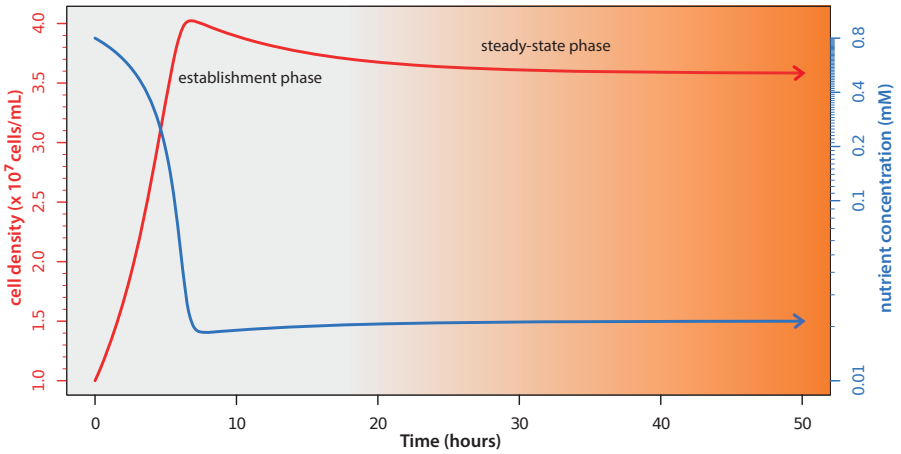

Figure 1.3. Establishment of a steady-state in the chemostat. Following inoculation and initiation of culture dilution the chemostat is characterized by a period during which the population increases and nutrient abundance declines. Eventually, a steady-state is established in which the cell population remains high and the concentration of the limiting nutrient remains low. The steady-state is predicted by the fundamental equations of the chemostat and depends on the parameter values used in the simulation. In this simulation, μ_{max} = 0.4 hr^{-1}, K_s = 0.05 mM, Y = 4.6 x 10^7 cells/mmole, R = 0.8 mM, and D = 0.12 hr^{-1}. The simulation was initialized with x = 1 x 10^7 cells/mL and s = 0.8 mM.

The continuous microbial culturing in a chemostat differs in many ways from batch culture growth. In a batch culture, a small number of cells are inoculated into a

fresh medium and then undergo a physiological and metabolic adjustment (‘lag phase’) and, following exponential cell growth (‘log phase’) and the cessation of cell growth and initiation of a quiescent state (‘stationary phase’). Experimental evolutions using serial dilution of batch cultures result in repeated population bottlenecks, a cycle of dramatic changes in physiological parameters such as pO2 (partial pressure of oxygen) and pH and nutrient concentration, and accumulation of cellular wastes. However, after a short period of initial batch-like growth, population growth in a chemostat is near constant and media composition is consistent over a long period of time owing to the continuous addition of fresh medium and removal of equal volume of the culture (see **Figure 1.3**). In a typical LTEE study using a chemostat, a single essential nutrient such as carbon, nitrogen, phosphorus or sulfur is limited while all other nutrients are present in excess. Thus, despite the increased experimental complexity of chemostats, their use greatly simplifies the selection for the purposes of experimental evolution.

1.2. NEXT-GENERATION SEQUENCING (NGS)

Isolation of causative (or driver) mutations in early LTEE studies was limited due to the difficulties of sequencing entire genomes. The advent of high-throughput sequencing methods enabled the use of chemostats for the study of evolution by mutation and selection [52]. For example, using unbiased whole genome tiling DNA microarray or next generation sequencings, recent LTEE studies have characterized the full spectrum of mutations and their functional relevance to the

evolutionary dynamics in *E. coli* or *S. cerevisiae* selected in nutrient-limited environments [34-39,53-56].

Unlike the 1st generation sequencing technique such as Sanger sequencing or microarray-based genotyping methods, next-generation sequencing (NGS) techniques including Roche 454, Illumina/Solexa Genome analyzer and Applied Biosystems (ABI) SOLiD generate high-throughput sequencing data in a cost- and labor-effective way. For example, the HiSeq 2500 and MiSeq, the latest model from Illumina as of 2014, allows researchers to obtain up to few thousands fold coverage for the yeast genome in a single lane. Different samples can be multiplexed using unique sample index in a single lane leading to further cost reduction.

The main goal of NGS in LTEEs is to comprehensively identify acquired genomic variants including single nucleotide polymorphisms (SNPs), small (< 1 Kb) insertions and deletions (INDELs), and copy number variations (CNVs) that include local amplification or deletion of long genomic regions and gain or loss of whole chromosomes. A general computational pipeline used in my thesis for identifying these different classes of variants is shown in **Figure 1.4**.

One important consideration in NGS based analysis is whether we sequence a clonal DNA sample or DNA prepared from the entire population. In microbes, there is no need to amplify an entire genome obtained from one single cell because it is very easy and quick to propagate one single cell to get a large clonal population of cells. Alleles identified from clonal sequencing of haploids should

have either 1 or 0 for their frequencies in the read alignment map except ones that may occur in a CNV. Clonal sequencing of asexual haploid strains that are isolated from evolving populations provides the list of mutations acquired in one single lineage. Assuming that no random meiotic recombination occur in the asexually reproducing cells, all the mutations in that lineage must have fixed in a sequential order, which can be estimated from allele frequencies (AFs) at the population level: higher AF mutations occurred earlier and lower AF ones later or more recently.

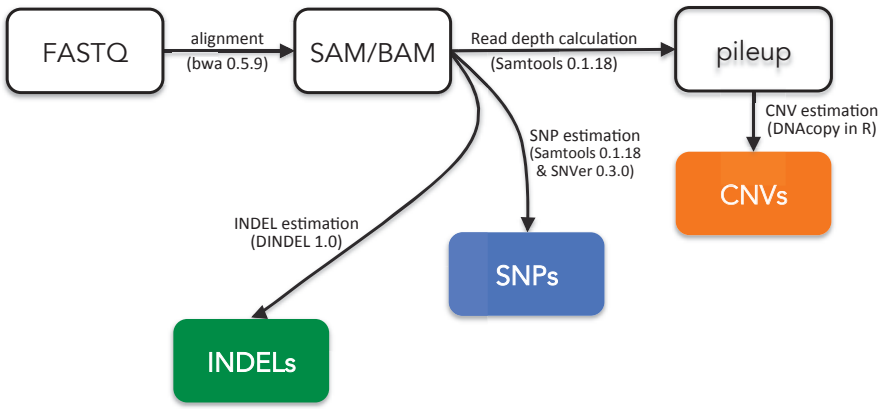

Figure 1.4. A typical bioinformatic workflow for analyzing NGS data in the fastq format. Three types of mutations are of major interest in this pipeline. Both data from HiSeq and MiSeq are processed using the same pipeline. Every analysis is run using the High-Performance Computing (HPC) in the Unix system (see all commands used in Appendix 1).

However, population level whole genome sequencing is challenging since an evolving population is a mixture of multiple different genotypes (AFs ranges from 0 to 1 as a continuous variable). Minor frequency mutations in a population (less

than 10 %) are hard to distinguish from random sequencing errors. Thus, read depth (coverage) is critical for the precise estimation of AFs in the population sequencing (**Figure 1.5**). One useful variation of population-level NGS is to target specific loci of interest and sequence them in order to get better resolution of AFs. The Illumina Miseq is well-suited for this as it generates enough read coverage for tens or hundreds of targeted loci in a cheap and rapid way.

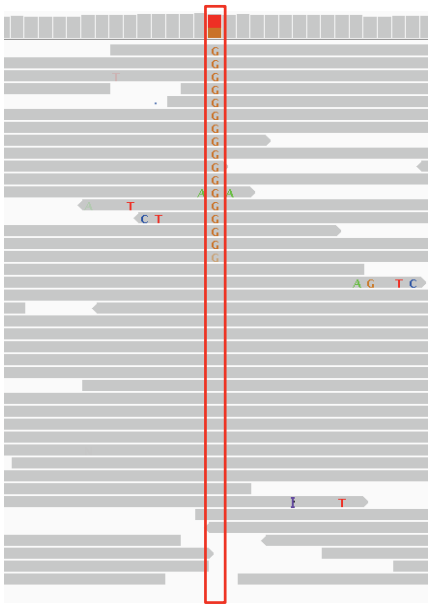

Figure 1.5. Allele frequency (AF) estimation from a population level sequencing data. AF is the proportion of the number of alternative (mutated) alleles among the number of all alleles from the alignment map. The detection limit of significant AFs is dependent on the read coverage.

An additional technical consideration of NGS is how to handle PCR bias originating from loci with extreme base compositions. The most typical Illumina sequencing library preparation (TruSeq®) includes a PCR amplification step for enriching properly ligated molecules to sequencing adapters. It has been suggested that high GC % region can be a source of PCR amplification bias in the final read coverage [57]. PCR free, enzyme based library preparation protocols such as

Illumina Nextera® kit can be used as an alternative although it is more expensive and only applicable when the amount of starting materials is very large. It is not clear how such bias affects the final result for AFs estimation of SNPs in DNA-seq or differential gene expression analysis in RNA-seq data. Currently, standard bioinformatics pipelines identifying PCR duplicates on the basis of the sequence identity. However, depending on the complexity of the sequenced material, the identified molecule can be generated by chance. Distinguishing unique molecules from PCR duplicates is not standardized yet.

During my thesis work, I established experimental and computational methods for clonal, whole genome population and targeted amplicon deep sequencing. In addition, I presented a new cost-effective sequencing adapter design that enables identification of true positive PCR duplicates and multiplexing multiple sequencing libraries for the Illumina sequencing platforms (see Chapter 3). This technical innovation identifies genomic variants that were acquired during LTEEs with increased precision and sensitivity.

CHAPTER2. Molecular Specificity, Convergence and Constraint Shape Adaptive Evolution in Nutrient-Poor Environments

*This chapter is based on the research paper "**Molecular Specificity, Convergence and Constraint Shape Adaptive Evolution in Nutrient-Poor Environments**" by Jungeui Hong and David Gresham, published in PLoS genetics 2014.*

2.1. ABSTRACT

One of the central goals of evolutionary biology is to explain and predict the molecular basis of adaptive evolution. We studied the evolution of genetic networks in *Saccharomyces cerevisiae* (budding yeast) populations propagated for more than 200 generations in different nitrogen-limiting conditions. We find that rapid adaptive evolution in nitrogen-poor environments is dominated by the *de novo* generation and selection of copy number variants (CNVs), a large fraction of which contain genes encoding specific nitrogen transporters including *PUT4*, *DUR3* and *DAL4*. The large fitness increases associated with these alleles limits the genetic heterogeneity of adapting populations even in environments with multiple nitrogen sources. Complete identification of acquired point mutations, in individual lineages and entire populations, identified heterogeneity at the level of genetic loci but common themes at the level of functional modules, including genes controlling phosphatidylinositol-3-phosphate metabolism and vacuole biogenesis. Adaptive strategies shared with other nutrient-limited environments point to selection of genetic variation in the TORC1 and Ras/PKA signaling pathways as a general

mechanism underlying improved growth in nutrient-limited environments. Within a single population we observed the repeated independent selection of a multi-locus genotype, comprised of the functionally related genes *GAT1*, *MEP2* and *LST4*. By studying the fitness of individual alleles, and their combination, as well as the evolutionary history of the evolving population, we find that the order in which these mutations are acquired is constrained by epistasis. The identification of repeatedly selected variation at functionally related loci that interact epistatically suggests that gene network polymorphisms (GNPs) may be a frequent outcome of adaptive evolution. Our results provide insight into the mechanistic basis by which cells adapt to nutrient-limited environments and suggest that knowledge of the selective environment and the regulatory mechanisms important for growth and survival in that environment greatly increase the predictability of adaptive evolution.

2.2. INTRODUCTION

Increasingly, the fields of evolutionary and molecular biology are fusing in a research program that has been termed the "functional synthesis" [7]. The power of this approach is exemplified by the molecular reconstruction of ancestral proteins enabling the study of the functional properties [58] and evolutionary histories [59] of individual genes. By contrast, the evolution of pathways and networks comprising multiple genes has thus far been less amenable to functional studies. This is due in part to the difficulty of inferring and engineering ancestral states of

genetic networks. An alternative approach to the study of genetic network evolution is the study of long-term natural selection in laboratories. Experimental evolution using microbes has a number of useful features including the ability to monitor evolution in real time and to measure fitness in the relevant environmental condition [19] that makes it ideally suited to the study of gene network evolution.

Uniquely among experimental methods of long-term selection, continuous culturing using chemostats [48,50] enables establishment of a precise and invariant selective pressure in which cell growth is continuously constrained by the rate of provision of a growth limiting nutrient. In contrast to evolution experiments using serial dilution [19,30,60], in which cells undergo repeated cycles of feast and famine, the unchanging nutrient-poor environment of a chemostat reduces fitness to a single component – continuous growth in a nutrient-poor environment – facilitating testing and interpretation of the functional basis of beneficial mutations. Moreover, in chemostats, large population sizes can be maintained (in excess of a billion cells) during the long-term selection thereby minimizing the effects of genetic drift and population bottlenecks.

Despite recent progress in our understanding of the molecular basis of adaptive evolution in chemostats [34,61-65] many questions remain. Does selection target particular loci and preferentially utilize distinct types of alleles? What is the functional basis of adaptation and are there mechanistic relationships between beneficial mutations? Does increased environmental complexity result in increased heterogeneity within a population? To what extent does epistasis constrain adaptive

landscapes? Here, we describe the results of experimental evolution of the budding yeast, *Saccharomyces cerevisiae*, in different nitrogen-limited chemostat environments. Variation in nitrogen availability is frequently encountered in natural ecologies and use of this selection enables comparison with previous adaptive evolution studies in other nutrient-limited environments using chemostats [34,61,65].

Importantly, for the goal of understanding genetic network evolution the molecular mechanisms underlying nitrogen utilization in budding yeast have been extensively studied [41], which facilitates interpretation of the functional effects of adaptive mutations. In nitrogen-limited chemostats, the steady-state nitrogen concentration in the culture is extremely low and cells grow continuously in a nitrogen-poor environment. Under these conditions, expression of a set of coordinately regulated genes, the nitrogen catabolite repression (NCR) regulon, is activated by the GATA transcription factors, GLN3 and GAT1 [66]. NCR genes encode a number of transporter and catabolic enzymes for import and assimilation of diverse nitrogen sources, the expression of which is repressed during growth in a nitrogen-rich environment by the negative regulators GZF3 and DAL80 [66].

Despite the greatly simplified and invariant selective conditions of a chemostat, we find evidence for at least three distinct adaptive strategies in nitrogen-limited chemostats that operate with different levels of environmental specificity. Consistent with earlier studies in other nutrient limitations [33,34,61], comparative analysis among the different nitrogen-limited conditions revealed selection for copy

number variant (CNV) alleles that result in increased abundance of transporters specific for the molecular form of nitrogen provided in each environment. We show that these alleles are also selected when multiple nitrogen sources are simultaneously present in the environment and that their inordinate fitness effects likely limit the accumulation of genetic diversity, even in environments with increased environmental complexity. Novel alleles at some loci are recurrently selected in different nitrogen-limited environments, including *VAC14* and genes with related functions, pointing to a role for remodeling of phosphatidylinositol-3-phosphate production and vacuole biogenesis in adaptation to nitrogen-limitation. By integrating our results with previous studies we find that variation in a subset of loci is selected in both nitrogen-limited chemostats and glucose-limited chemostats providing evidence for a general adaptive strategy in nutrient poor environments through remodeling of the TORC1 and Ras/PKA pathways.

We also report a striking example of clonal interference in which independent lineages, defined by mutations in three functionally related loci, *GAT1*, *MEP2* and *LST4* co-evolve in a single population undergoing adaptive evolution in an ammonium-limited chemostat. By studying the individual and interactive effects of these alleles as well as reconstruction of lineage dynamics, we demonstrate that the order of mutations is constrained by epistatic interactions. We propose that this three-locus genotype comprising functionally related gene products represents a gene network polymorphism (GNP), which may be a more frequent outcome of adaptive evolution than previously appreciated.

2.3. RESULTS

To study adaptation in nitrogen-limited environments we founded populations with a haploid *Saccharomyces cerevisiae* strain isogenic to the reference genome (S288c) in different nitrogen-limited chemostats. A normalized concentration of 800μM nitrogen was used in all feed media making the molecular form of nitrogen the only variable in each environment. A single population in each different nitrogen-limited environment was maintained in continuous exponential growth (D = 0.12 culture volumes/hr; $t_{doubling}$ = 5.8 hours) for 250 generations.

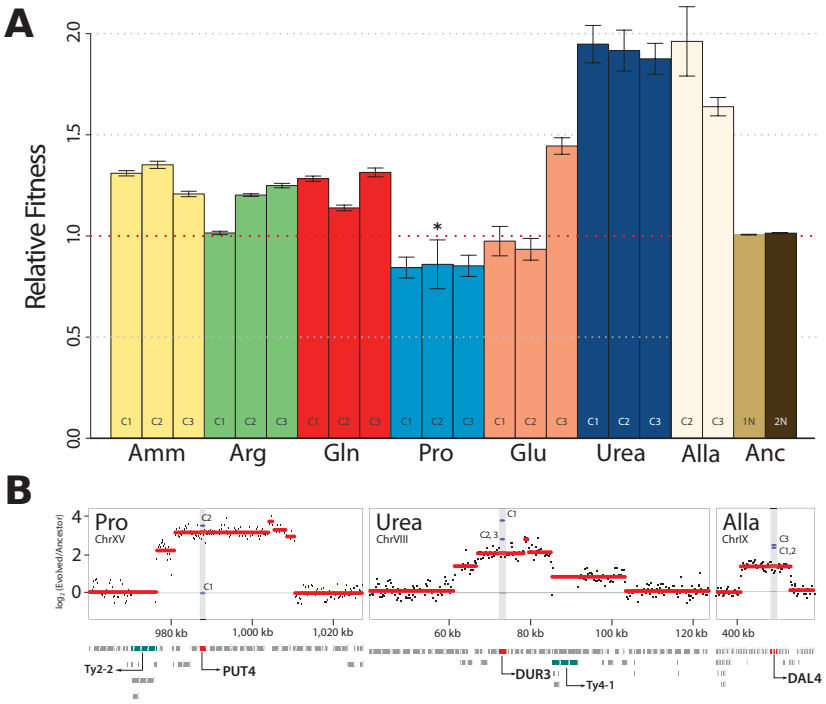

Figure 2.1. Increased fitness in nutrient-limited environments is associated with amplification of specific permease genes. (A) Fitness increases for clones recovered from each selection are typically >10%. Haploid (1N) and diploid (2N) ancestral strains were also tested in ammonium-limited chemostats but did not show fitness differences. (Amm :

ammonium, Arg : arginine, Gln : glutamine, Pro : proline, Glu : glutamate, Urea : urea, Alla : allantoin, Anc : ancestor). (B) DNA copy number was estimated using aCGH. Each black point represents a measurement from a unique probe on the microarray from analysis of population DNA samples. We detected CNVs containing genes with clear connections to nitrogen import at high frequencies in populations (red lines) and clones (blue lines). Retrotransposon (Ty) sequences were frequently found at the boundary regions of CNVs.

2.3.1 Adapted clones have dramatically increased fitness

Initially, we studied populations evolving in seven different nitrogen-limited environments. To identify phenotypically distinct clones within each adapted population of ~10^{10} cells following 250 generations of selection we performed batch culture growth rate assays on an unbiased sample of 94 clones from each population and selected three individuals that exhibited growth characteristics distinct from each other and the ancestral strain for further characterization (s**ee** methods). We determined the relative fitness of each clone in the appropriate nitrogen-limited chemostat environment and typically observed large increases in fitness (>10%) (**Figure 2.1A**). This is consistent with mutation and selection rapidly moving strains towards a fitness optimum. It is clear that the ancestral genotype differs in its distance to the fitness optimum with respect to different nitrogen limited environments: fitness increases in clones selected from ammonium-, arginine- and glutamine-limited chemostats are around 25% whereas fitness increases in clones evolved in urea- and allantoin-limited chemostats exceed 80%. In general, individuals from the same population had similar fitness. A

minority of clones did not show increased fitness using this assay for reasons that are not clear, but may be indicative of frequency-dependent selection. The majority of evolved clones were unaltered in their ability to grow in nitrogen-rich conditions or showed decreased fitness (typically less than 4%) (**Figure2.2**). Thus, mutations selected in the nitrogen-poor environments are uniquely beneficial in nitrogen-poor environments and exhibit antagonistic pleiotropy in nitrogen-rich environments.

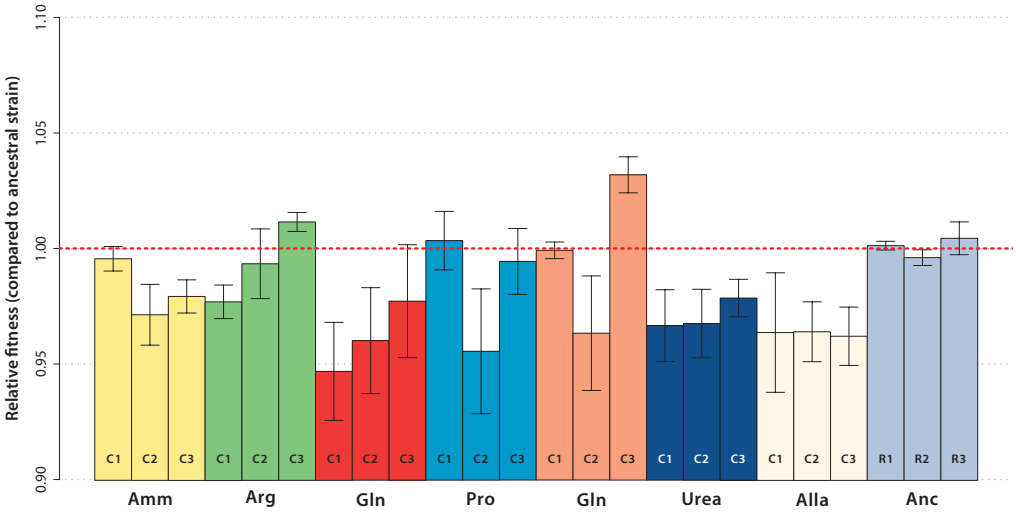

Figure 2.2. Evidence of antagonistic pleiotropy in evolved lineages. Each mutant recovered from evolved populations was competed against a common fluorescently-labeled ancestral strain in batch cultures supplied with 5 g/L ammonium sulfate. Evolved clones exhibited fitness decreases of up to 4% in nitrogen-rich environments.

2.3.2. Selection for amplification of specific transporter genes

To identify mutations associated with increased fitness we first analyzed the genomes of selected clones, and entire populations, using array comparative genomic hybridization (aCGH). We observed multiple copy number variants (CNVs), including duplicated and deleted genomic regions, typically greater than

~10kb, in individual clones and entire populations (**Figure 2.3**). Previously, we reported identification of amplification alleles that include the *GAP1* locus in clones adapted to glutamine- or glutamate-limitation [62]. A subset of CNVs present in other nitrogen-limited environments include compelling candidates that are likely to underlie selection of the amplified allele. These include a CNV containing the allantoin permease (*DAL4*) in allantoin-limited conditions, a CNV including the urea permease (*DUR3*) in urea-limited conditions and a CNV including the proline permease (*PUT4*) in proline-limited conditions (**Figure 2.1B**). Our ability to detect these CNV alleles in population samples using aCGH (**Figure 2.1B**) indicates that they are at high frequency following 250 generations of selection. Consistent with previous studies [61,67], CNVs are frequently proximal to retrotransposon sequences (**Figure 2.1B**), which may increase their spontaneous rate of generation. Previously, we, and others, have identified the repeated selection of copy number variants (CNVs) at the *HXT6/7* [33,61] and *SUL1* [61] locus in yeast strains selected from glucose- and sulfur-limited chemostats respectively. In *E. coli* evolved in lactulose-limiting conditions the lac operon, which includes the lactose permease (*lacY*), is frequently amplified [68]. Collectively, these findings make clear that in diverse nutrient-limiting conditions, increased production of specific nutrient transporters is a rapid route to increased fitness. The spontaneous rate at which amplification CNVs are generated appears to depend on context [69] ; however, estimates of gene amplification rates suggest that they are on the order of nucleotide substitution rates [70]. Selection for spontaneously generated

amplification alleles appears to be an expedient means of increasing production of specific nutrient transporters and these alleles are strongly selected in nutrient-poor conditions.

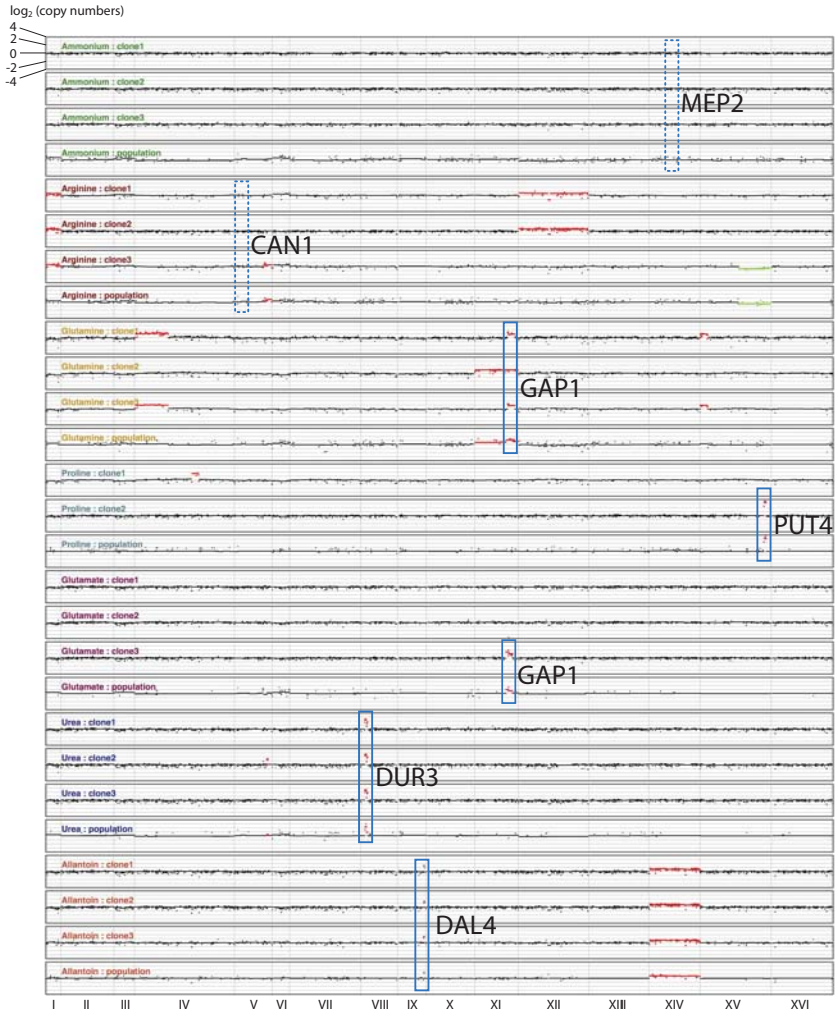

Figure 2.3. Complete aCGH results of all analyzed clones and populations that have undergone adaptive evolution in individual nitrogen sources. Most populations have acquired CNVs that include transporters of the specific nitrogen source except in the case of ammonium and arginine-limitation. For visualization, amplified or deleted regions with a minimum length of 10 kb and a $\log_2$ ratio > |0.5| are indicated by red (amplification) or green (deletion).

It is notable that we did not detect amplification alleles containing the known high affinity ammonium transporter gene, *MEP2,* in the ammonium-limited population or the arginine transporter, *CAN1,* in the arginine-limited population (**Figure 2.3**). It remains to be determined if amplification of *MEP2* or *CAN1* is beneficial in ammonium- or arginine-limited conditions or if these amplification alleles are deleterious for functional or genetic reasons. Moreover, we cannot exclude the possibility that amplification alleles were present at an earlier stage in these populations but were subsequently out-competed.

2.3.3. Aneuploidy and whole genome duplication may contribute to adaptive evolution

We observed additional copy number variants and entire chromosomal aneuploidies that include genes without obvious connections to growth in nitrogen-limited conditions (**Figure 2.3**). We identified 7 aneuploid clones among the 18 analyzed clones (~ 40%). The recurrent observation of aneuploidy in adaptive evolution studies [61,67] and as a mechanism of genetic suppression [71] suggests that they are likely to be adaptive, although the mechanistic basis for the selective advantage of aneuploidies remains to be determined.

We quantified the DNA content of all clones, using flow cytometry, and found that in populations adapted to allantoin- and urea-limitation a high frequency of cells had a 2N DNA content. These individuals are still of a haploid mating type (MATa) as demonstrated by successful mating with MATα cells. The resulting

triploid cells underwent sporulation, but typically yielded poor spore viability (< 10%) consistent with massive unbalanced chromosome content in the meiotic products of triploids. The maintenance of a MATa mating type in diploid cells recovered from chemostat selections indicates that they are the result of failed cytokinesis and not due to spontaneous mating type switching and subsequent mating. We did not detect a fitness advantage in the chemostat that is attributable to the diploid state *per se* (**Figure 2.1A**) consistent with previous studies [72]. Although the high frequency of diploid cells is consistent with selection, the lack of a detectable fitness effect in a wild type diploid cell suggests that selection for diploidization may require the prior acquisition of at least one mutation that is advantageous when increased in copy number as a result of a whole genome duplication.

2.3.4. mRNA expression levels are correlated with increased copy number at multiple scales

To study the functional basis of adaptation we performed genome-wide transcriptional profiling of evolved clones in the same chemostat environment as they had been selected. Divergence in the transcriptome between clones adapted to different nitrogen environments was qualitatively similar to that seen between clones adapted to glucose- and phosphorous-limited environments [61] (**Figure 2.4**). Some of the transcriptional variation in clones adapted to nitrogen-limited environments is a direct result of altered copy number due to CNVs as we detected

a small but significant positive correlation between DNA copy number and mRNA abundance (**Figure 2.5A**). In general, mRNAs corresponding to transporter genes found within CNVs were increased in abundance, consistent with increased DNA copy number resulting in increased transporter abundance (**Figure 2.5A**), providing further evidence that these genes drive selection of the CNV.

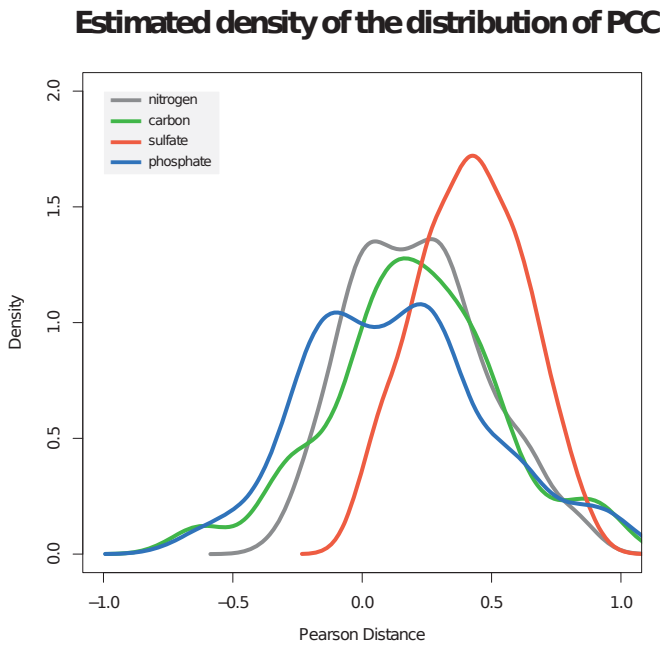

Figure 2.4. Comparison of transcriptional divergence between clones using the distribution of pair-wise Pearson correlation coefficients as in [61]**.** Transcriptional divergence among clones adapted to nitrogen limitation is similar to that found for glucose- and phosphate-limited selections. Clones adapted to sulfur-limitation show far greater convergence of transcriptional states.

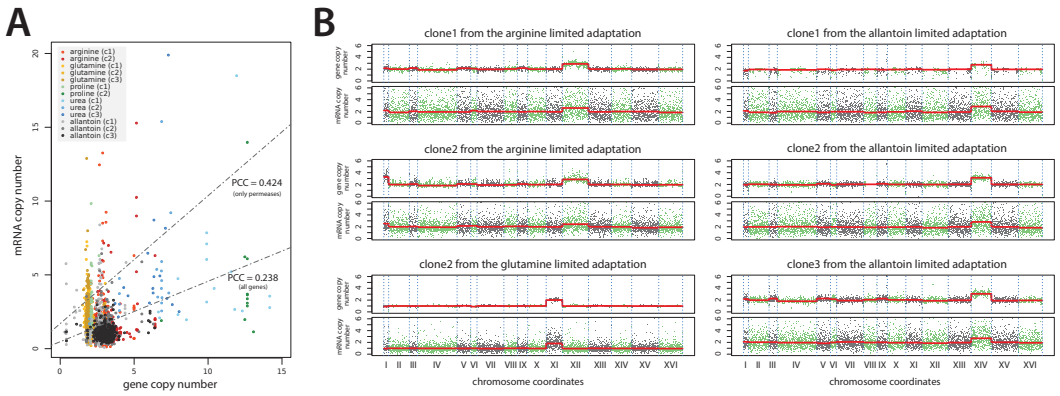

Figure 2.5. DNA copy number correlates with mRNA abundance. (A) CNVs result in increased gene expression. Nitrogen transporter genes located in CNVs tend to increase in expression with increased copy number. (B) All aneuploids identified showed increased mRNA expression of most genes in amplified chromosomes.

As previously observed [73], DNA copy number in disomic or trisomic chromosomes of aneuploid cells is proportional to mRNA abundance level (**Figure 2.5B**). In some cases this may explain the selection for a specific aneuploidy. For example, a clone recovered from the glutamine-limitation adaptation contains an additional entire copy of chromosome XI, which contains *GAP1* [62]. However, other chromosomal aneuploidies do not have an obvious connection to nutrient transport making it unclear how, or why, the large-scale increase in expression of genes along duplicated chromosomes of adapted clones contributes to fitness.

2.3.5. Defining the spectrum of point mutations associated with adaptation

To identify all mutations acquired during the selection experiments we performed whole genome sequencing of 18 clones from the seven populations (see methods).

We found an average of 4 SNPs per clone that together represent a broad range of classes (**Figure 2.6A**). The average number of SNPs is higher than expected (~1.0) based on the measured spontaneous nucleotide substitution rate [74] but is consistent with the average number of acquired SNPs (~3.3) reported for equivalent selections in glucose- or phosphorous-limited environments [24,64,75]. Whether this reflects an increased mutation rate under conditions of stress, as reported for *E. coli* [76], or heterogeneity in the number of mitotic events a particular lineage undergoes in a chemostat, remains to be determined. We detected a marginal but statistically significant bias towards SNPs in coding regions: 60/72 SNPs (83%) were found in coding regions, while 72% of yeast genome is coding (exact binomial test, p-value=0.035). Although the majority of base changes in coding regions were non-synonymous (52/72; 72%) this is not significantly different than the expected frequency (79%) of non-synonymous mutations [65] (exact binomial test, p-value=0.1912). We also identified 8 indels (7 deletions and 1 insertion) of one or two base pairs. The average number of indels per clone (~0.44) is higher than that expected on the basis of the known spontaneous rate of indel events (~0.06) [74]. All CNVs detected using aCGH were also identified on the basis of sequence read depth. Furthermore, we detected additional deleted genomic segments of several hundred base pairs suggesting that whole genome sequencing has superior sensitivity to aCGH for CNV detection [75]. In lineages that had undergone diploidization we detected both homozygous and heterozygous point mutations, which allowed us to distinguish mutations that occurred prior to, and

after, diploidization, respectively. In sum, comprehensive genome characterization indicates that in individual clones evolving in nitrogen-limited environments, multiple mutations are acquired in a short period of time that range from single nucleotide substitutions to complete duplication of the genome (**Figure 2.6B**).

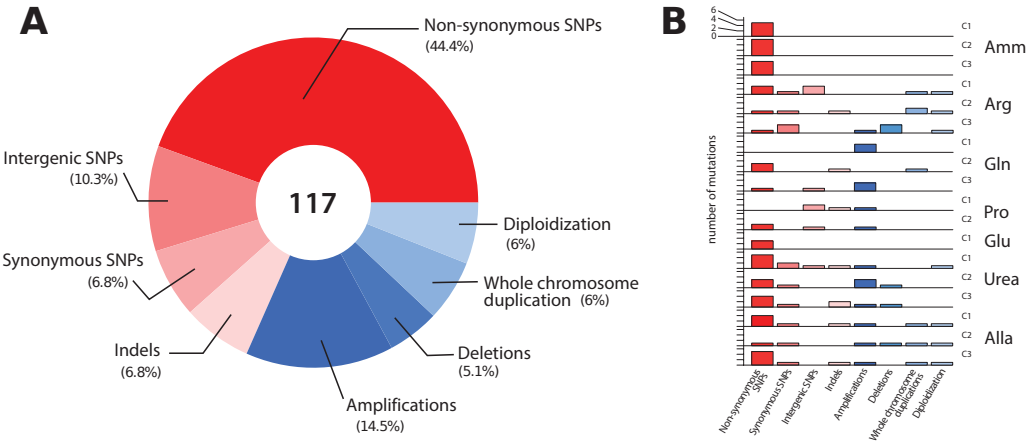

Figure 2.6. Overview of the classes of mutations identified in lineages adapted to nitrogen-limited conditions. (A) In total, 117 mutational events were identified in 18 sequenced clones resulting in sequence (red) and structural (blue) variation. (B) The number and type of mutations acquired in each individual clones genotyped using aCGH and whole genome sequencing. Non-synonymous SNPs and CNVs are found in most clones.

Whereas sequencing of clonal isolates provides information on individual lineages, deep sequencing of entire populations provides a means of assessing the genetic diversity in a population at a particular time point in the evolutionary history of the population [77]. We sought to identify all alleles that had risen to appreciable frequencies following 250 generations of selection using whole genome sequencing of entire populations. We identified fixed and non-fixed alleles and estimated their

frequencies on the basis of sequence read counts (**Figure 2.7**). Despite sequence read depths in excess of 300-fold, we detected few additional mutations in populations that were not identified in clones. Populations typically contained less than 10 SNPs at frequencies > 5% (**Table 2.1**). A single exception was identified; in the population adapted to allantoin-limitation we found 486 mutations, which is likely the result of mutator phenotype due to loss of function in the mismatch repair gene, *MSH2*, which we estimate to have a frequency of ~6% in the population.

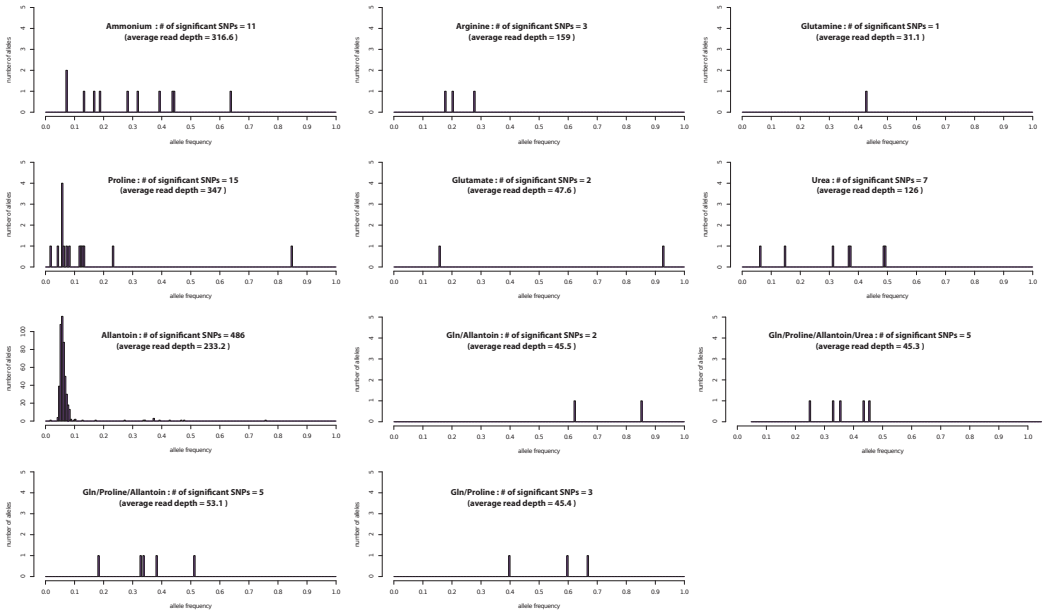

Figure 2.7. Allele frequencies distributions for each population based on whole genome sequencing. We estimated allele frequencies for all SNPs that were present at greater than ~ 5% using deep sequencing read counts in 11 different nitrogen-limited populations.

Selective environment (800 µM nitrogen)	Number of SNPs (> 5% frequency)
Ammonium (400 µM)	10
Arginine (200 µM)	3
Glutamine (400 µM)	1
Proline (800 µM)	11
Glutamate (800 µM)	2
Urea (800 µM)	7
Allantoin (200 µM)	486
Gln/Alla (200/100 µM)	2
Gln/Pro/Alla/Urea (100/200/50/100 µM)	5
Gln/Pro/Alla (133/166/67 µM)	6
Gln/Pro (200/400 µM)	4

Table 2.1. Genetic complexity of adapting populations. A small number of point mutations rose to appreciable frequencies in each population with the exception of the allantoin-limited population, which contains ~ 500 SNPs most of which have frequencies less than 10%. This population also contains a mutant *MSH2* gene, suggesting the existence of a low frequency mutator phenotypes [78,79]. Nitrogen concentrations were normalized between environments by adjusting the concentration of each compound according to its molecular composition.

2.3.6. Increased environmental complexity does not result in increased genetic diversity

We were surprised by the low genetic diversity in populations adapted to individual nitrogen sources (see **Table 2.1**) especially since previous analyses of *E. coli* populations evolving in glucose-limited chemostats have suggested the presence of multiple ecotypes [54,80]. We hypothesized that the low genetic diversity within populations may be a related to the presence of a single nitrogen source in the environment. To study the effect of increasing the complexity of environments on genetic variation in adapting populations, we performed additional long-term

selection experiments using mixtures of 2-4 different nitrogen sources. Following the same period of selection we did not detect increased genetic complexity, as assessed by population deep sequencing, in these selections compared with populations adapted to a single nitrogen source (**Table 2.1**). We performed aCGH on clones and populations evolved in the presence of mixed nitrogen sources and detected CNVs that include transporter genes specific to individual nitrogen sources present in each environment (**Figure 2.8**). However, we did not detect any lineages containing multiple CNVs that would improve transport of more than one of the available nitrogen sources in an environment, suggesting that lineages underwent specialization in the mixed environments. The highest frequency CNVs in populations adapted to mixed nitrogen sources transport non-preferred nitrogen sources (proline, allantoin and urea) (**Figure 2.8**), which also tend to be associated with the greatest individual fitness increases (**Figure 2.1A**). Collectively, our observations in single and mixed nitrogen-limited environments are consistent with a highly skewed distribution of fitness effects in which CNV alleles that include transporter genes have large fitness effects and therefore a high probability of sweeping to fixation. The large effect sizes of these CNV alleles limits genetic diversity even in environments of increased complexity.

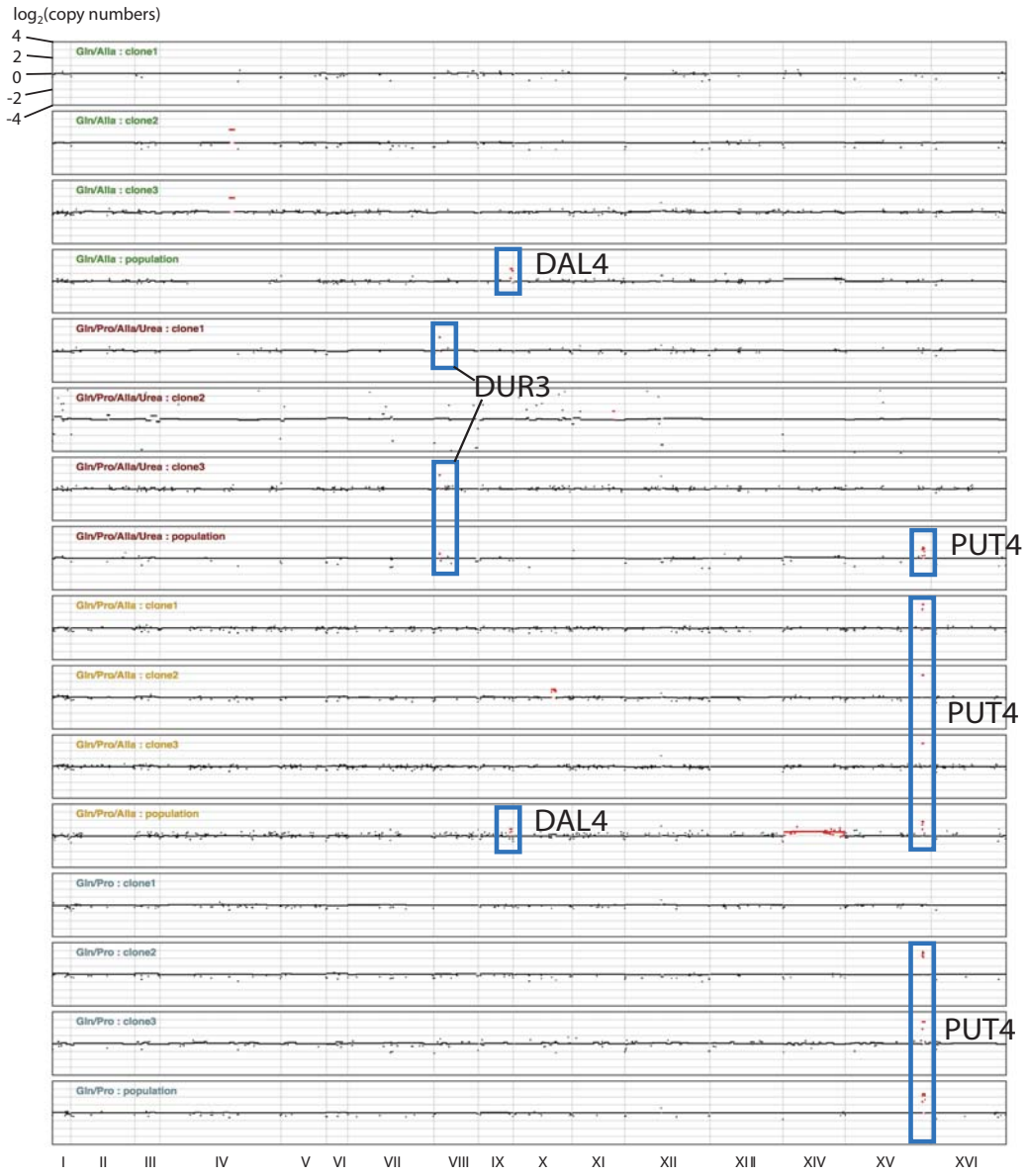

Figure 2.8. CNVs are frequently selected in the presence of mixed nitrogen sources. Complete aCGH results for all populations and clones evolved in mixed nitrogen source environments. CNVs that include transporters for non-preferred nitrogen sources (urea, allantoin and proline) are preferentially selected when multiple nitrogen sources are present.

2.3.7. Identification of specific and convergent targets of selection

High throughput sequencing of clones and populations revealed that genetic variation at a number of loci was repeatedly selected in different nitrogen-limited selections (**Figure 2.9A**). In addition to amplification of permease genes in conditions in which they increase import rates of nitrogen-containing compounds, we find that inactivating alleles are selected in conditions in which their function provides no benefit or may be deleterious. As we previously reported, this is the case for *GAP1*, which is amplified in glutamine- and glutamate-limited conditions and deleted when the nitrogen source is not an amino acid such as allantoin and urea [62] (**Figure 2.9A**). Similarly, amplification alleles containing *PUT4*, which encodes a proline permease, are selected in environments in which proline is a nitrogen source, but an inactivating mutation in *PUT4* was found in the arginine-limited environment. We hypothesize that loss of function mutations in these genes are selected as the NCR-derepressing conditions of a nitrogen-limited chemostat result in their high expression, which is futile in the absence of the substrate(s) they transport.

We identified six loci that acquired point mutations in multiple nitrogen-limitation selections. The most striking of these was *VAC14,* which is mutant in 8 of the 11 different selective environments. Sequence variants in *VAC14* are predominantly loss of function mutations and in two populations we found multiple independent *VAC14* alleles (**Figure 2.9A**). *VAC14* encodes a scaffold component of the protein complex regulating interconversion of phosphatidylinositide-3-phosphate (PI3P)

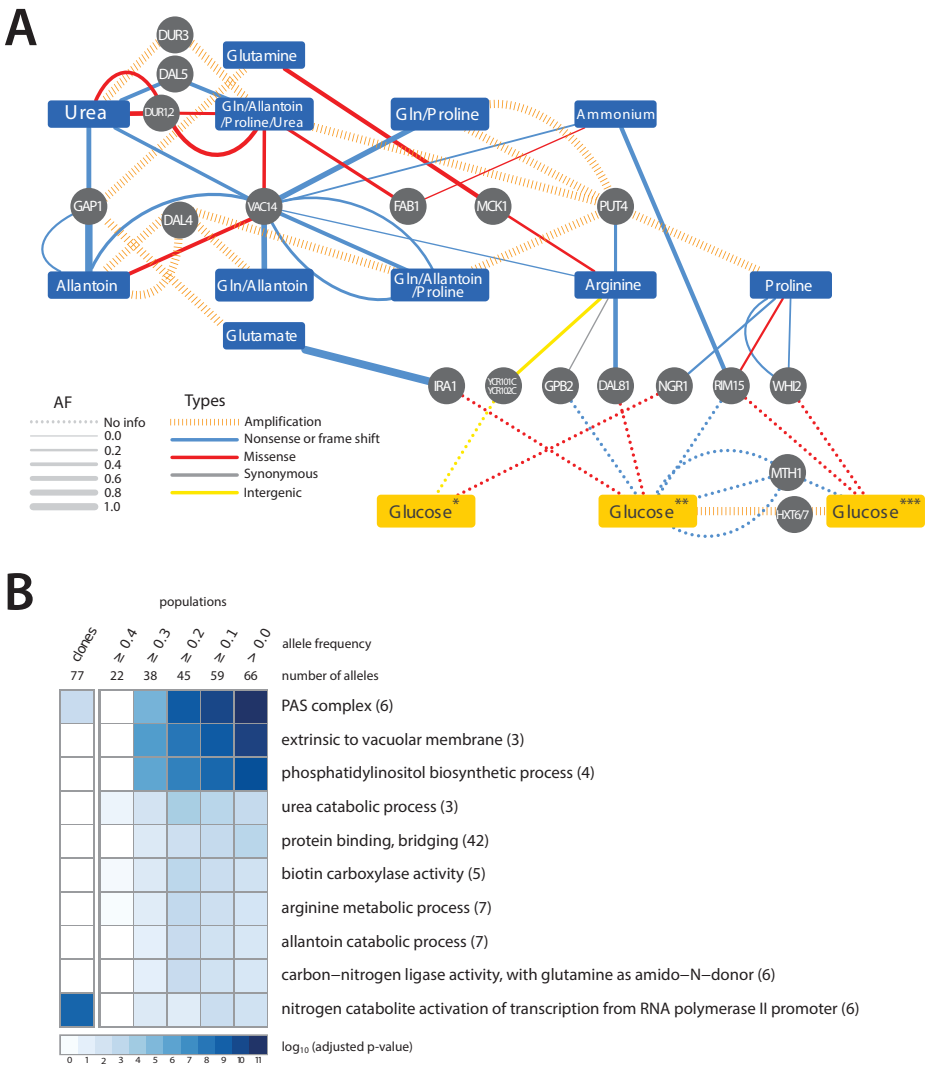

Figure 2.9. Adaptive mutations occur in functionally related loci (A) A small number of loci are mutated in multiple nitrogen-limited environments. Some loci found to be mutated in nitrogen-limiting conditions have also been reported as associated with adaptive evolution in glucose-limited environments (*Wenger, J. et al [65], **Kvitek, D.J. et al [64], ***Gresham, D. et al [61]). The color of edges represents the type of allele and the width of the edge represents the frequency of the allele in the population. (B) GO term enrichment analysis of mutated loci within clones and populations, analyzed at different allele frequency thresholds, identified in nitrogen-limited environments shows enrichment for specific cellular functions.

to phosphatidylinositide-3,5-bisphosphate ($PI(3,5)P_2$) [81]. Interestingly, an additional repeatedly mutated locus, *FAB1*, encodes the 1-phosphatidylinositol-3-phosphate 5-kinase that functionally interacts with VAC14. When all mutations identified in clones and populations are considered, there is a clear enrichment for molecular functions related to phosphatidylinositol biosynthetic processes and the related processes of autophagosome and vacuole biogenesis (**Figure 2.9B**) indicating that they are a convergent target of selection across nitrogen-poor environments. Functional enrichment analysis of mutations in populations and among clones also identified several additional molecular processes related to nitrogen metabolism (**Figure 2.9B**). Thus, the molecular basis of adaptive evolution in nitrogen-limited environments exhibits convergence at both the level of individual genes, and at the level of modules, defined by functionally related genes.

It is possible that some adaptive alleles recovered in our experiments are not specifically related to nitrogen utilization, but underlie adaptation to the requirement of continuous growth in nutrient-limited conditions. To identify such loci we compared the loci associated with adaptive evolution in nitrogen-limited environments with those identified in previous studies of adaptation to glucose-, phosphate- and sulfur-limited environments [34,61,64,65] (**Figure 2.9A**). Several loci mutated in both glucose- and nitrogen-limited chemostats encode components of signaling pathways that regulate cell growth in response to the nutritional state of the environment. At least two of these genes (*RIM15* and *WHI2*) regulate entry

into a quiescent (G_0) state. Loss of the ability to enter G_0 may be beneficial in the chemostat, as even transient entry into G_0 will prolong the cell division cycle leading to cells being outcompeted. Selection for this class of mutations may be analogous to the recurrent loss of function mutations found in the stress response sigma factor, *rpoS*, in experimental evolution of *E. coli* in chemostats [82]. No mutated loci were shared with phosphate and sulfur-limited selections.

2.3.8. Identification of a recurrently selected three-locus genotype comprising functionally related genes

The population adapted to ammonium-limitation was the only population in which we did not detect evidence of CNVs in either clones or the entire population (**Figure 2.6B**). However, clones from this population displayed the greatest divergence in nitrogen catabolite repression (NCR) gene expression among all clones analyzed (**Figure 2.10A** and **Figure 2.11**) and had large fitness increases (**Figure 2.1A**) suggesting that they had undergone significant adaptive evolution.

We found that these two clones, and a third that was not analyzed for gene expression, contain mutations in the DNA binding domain of the zinc finger transcription factor *GAT1* (**Figure 2.10B**), which encodes a positive regulator of NCR expression [41]. A subset of NCR genes is increased in expression in these clones including those encoding the high affinity (*MEP2*) and low affinity (*MEP1* and *MEP3*) ammonium permease genes (**Figure 2.10C**). Interestingly, several

NCR transcripts are also decreased in expression suggesting that the *GAT1* mutations may have differential effects on its transcriptional targets.

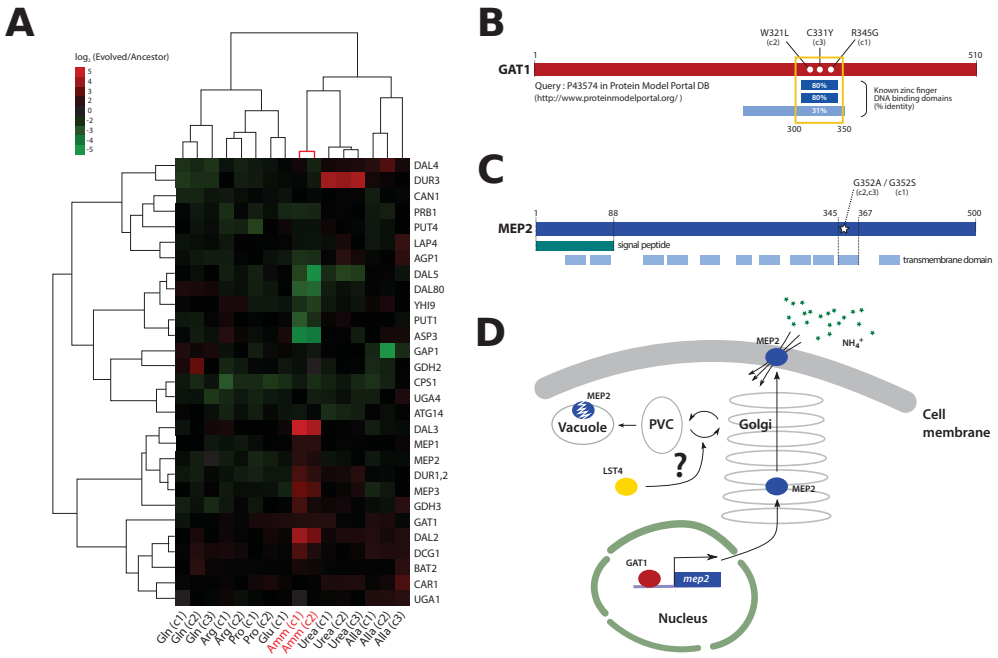

Figure 2.10. Functional effects of adaptive mutations in a gene network polymorphism. (A) NCR genes are altered in expression in clones recovered from ammonium-limited conditions. Only genes having at least one observation with $\log_2$ ratio > |1.5| were included (29 / 38 NCR genes [83]). Genes and samples are hierarchically clustered using centered correlation and complete linkage. (B) Three independently acquired *GAT1* mutations found in a single ammonium-limitation adapted population are clustered in the zinc finger DNA binding domain of the encoded protein. The wild type GAT1 protein sequence was queried using the Protein Model Portal database [84]. (C) Two different point mutations in MEP2 found in ammonium-limitation adapted clones change the identical codon within a putative trans-membrane domain. Domain information was obtained from SGD database (http://www.yeastgenome.org/). (D) GAT1 and LST4 likely regulate the production and delivery of MEP2 to the plasma membrane at the transcriptional and post-translational level, respectively.

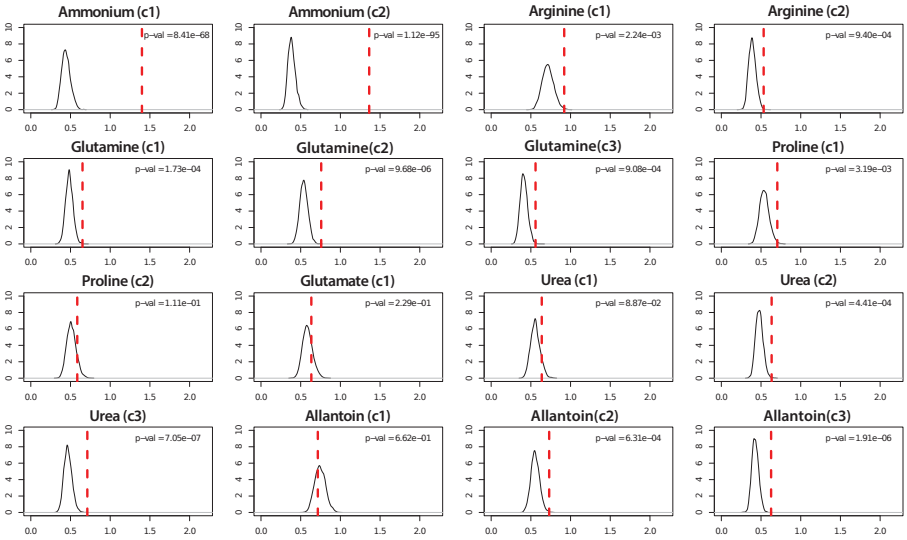

Figure 2.11. Significance analysis of NCR expression divergence in adapted clones. In most adaptations, NCR genes were significantly altered in expression. The statistical significance of NCR expression divergence (p-value) was calculated by 1) generating a null distribution by obtaining the mean absolute $\log_2$ gene expression ratio of 1,000 randomly chosen sets of 38 genes (without replacement) among all yeast ORFs on the microarray and then 2) computing the probability of obtaining an average absolute $\log_2$ gene expression ratio (indicated by a dotted red line) for the 38 measured NCR genes in the corresponding clone equal to or greater than that value. The greatest divergence in NCR expression is found among clones adapted to ammonium-limitation.

We found that these two clones, and a third that was not analyzed for gene expression, contain mutations in the DNA binding domain of the zinc finger transcription factor *GAT1* (**Figure 2.10B**), which encodes a positive regulator of NCR expression [41]. A subset of NCR genes is increased in expression in these clones including those encoding the high affinity (*MEP2*) and low affinity (*MEP1* and *MEP3*) ammonium permease genes (**Figure 2.10C**). Interestingly, several

NCR transcripts are also decreased in expression suggesting that the *GAT1* mutations may have differential effects on its transcriptional targets.

In addition to mutations in *GAT1*, we found that the three clones from the ammonium-limitation selection contained one of two different mutations in the identical codon of a predicted transmembrane domain of the high affinity ammonium transporter *MEP2*, a transcriptional target of GAT1 [85] (**Figure 2.10C**). Furthermore, two of these clones contained mutations in *LST4*, which encodes a protein required for efficient sorting of permeases from the Golgi to plasma membrane [86]. The three genes, *GAT1*, *MEP2* and *LST4* that comprise this recurrently selected multi-locus genotype encode functionally related gene products (**Figure 2.10D**) consistent with adaptive evolution proceeding via the sequential accumulation of variation in genetic networks within lineages.

2.3.9. Population dynamics of the three-locus genotype

We aimed to determine the temporal dynamics with which the mutations in *GAT1*, *MEP2* and *LST4* occurred and were selected. Population sequencing of the ammonium-limitation adapted population after 250 generations of selection identified 10 SNPs with detectable allele frequencies (> 5%). Allele frequencies in the population are informative about the order in which mutations were acquired in each asexually reproducing lineage; however, the timing of mutational events cannot be deduced on the basis of allele frequencies. To reconstruct the evolutionary history of the lineages we determined allele frequencies throughout

the evolution experiments using Sanger sequencing [61] (see methods). The resulting trajectories (**Figure 2.12A**) show that within a single population the same two locus genotype (*gat1*, *mep2*) was independently generated and selected three times (lineages A1, B1, and B3) and the three locus genotype (*gat1, mep2, lst4*) was generated at least twice (lineages A1 and B3). Interestingly, in both lineages, mutations in *GAT1* and *LST4* occurred in rapid succession and subsequently increased in frequency (i.e. lineage A0 and lineage B3 in **Figure 2.12A**), which is suggestive of a synergistic interaction between *LST4* and *GAT1*. Although we detect dramatic changes in allele frequencies during the selection no individual genotype swept to complete fixation (i.e. a "hard sweep"). Rather, competition (i.e. clonal interference) between lineages bearing different alleles in the identical multi-locus genotype resulted in alternating "soft sweeps".

2.3.10. Epistasis constrains the order of mutational events

As functionally related genes are enriched for genetic interactions [87], we hypothesized that epistatic interactions might exist between *GAT1*, *MEP2* and, *LST4*. To test this hypothesis we constructed strains containing the eight possible combinations of the *gat1-2*, *lst4-2* and *mep2-2* alleles identified in clone 3 (methods). The mutations in *MEP2* and *GAT1* are individually beneficial; however, the mutation in *LST4* does not confer a selective advantage on its own (**Figure 2.12B**). The double mutation genotypes comprised of either *mep2-2* and *lst4-4* or *gat1-2* and *lst4-2* are more fit than expected by summation of their individual

fitness effect providing evidence for positive epistasis. However, we found that the combined effect of the *gat1*-2/*lst4*-2/*mep2*-2 alleles does not result in significantly increased fitness compared with the *gat1-2*/*lst4-2* or *mep2-2*/*lst4-2* double mutant genotypes consistent with negative epistasis.

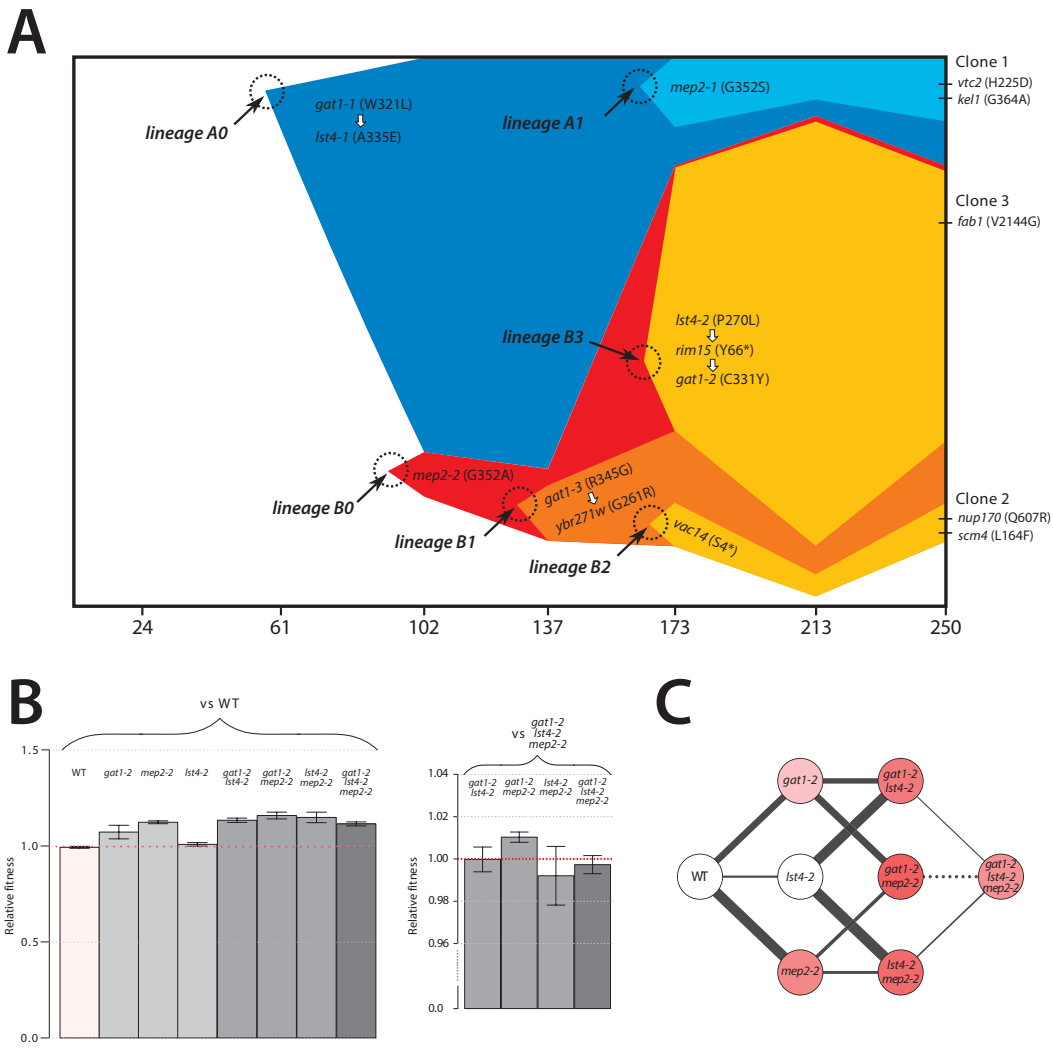

Figure 2.12. Recurrent selection and evolutionary dynamics of a GNP. (A) Estimated genotype dynamics during adaptive evolution. The time of introduction of each new mutation (dotted circles) is estimated on the basis of detecting an allele frequency of at least 5% in the population. Some

mutations were clustered based on their similarity in the dynamics. The temporal order of mutations that occurred in rapid succession (white arrows) was determined on the basis of their allele frequencies in the final evolved population estimated using deep sequencing data (**Figure 2.7**). (B) Fitness estimates of 8 backcrossed strains, representing all possible combinations of alleles that comprise the GNP, from clone 3 isolated from the ammonium-limitation selection were determined by direct competition with either the ancestral or the *gat1-2*/*lst4-2*/*mep2-2* genotypes. Error bars are 95% CI of the regression coefficient. (C) Fitness landscape reconstruction based on the fitness estimates for the 8 genotypes. The selection coefficient values of each strain are represented as color intensity. The width of each edge is proportional to the difference in fitness between two genotypes that edges connect. A solid line indicates a favored path whereas a dashed line indicates a disfavored path. Selection favors thicker, solid lines in the evolutionary trajectory.

To more accurately compare fitness effects of different genotypes we directly competed double mutant genotypes directly with the *gat1*-2/*lst4*-2/*mep2*-2 genotype. Consistent with our initial observations we find that the *gat1*-2/*lst4*-2/*mep2*-2 triple mutant genotype is not significantly fitter than the *gat1*-2/*lst4*-2 or *lst4*-2/*mep2*-2 double mutant genotypes and is in fact significantly less fit than the *gat1*-2/*mep2*-2 genotype. Thus, an *LST4* mutation is beneficial only in the background of an individual mutation in *GAT1* or *MEP2* whereas it is detrimental in the background of the *GAT1*/*MEP2* double mutant (**Figure 2.12**C). This sign epistatic interaction is consistent with the order of mutation acquisition in the three lineages in the population: an *LST4* mutation is observed after the occurrence of a *GAT1* mutation (lineage A0) or a *MEP2* mutation (lineage B3), but not in the lineage that contains a mutation in both *GAT1* and *MEP2* (lineage B1).

2.4. DISCUSSION

A major motivation for Novick and Szilard's introduction of the chemostat was the study of spontaneous mutations and evolution [50]. Seminal studies by Paquin and Adams in the 1980s established the use of budding yeast in experimental evolution studies in chemostats [88,89]. The advent of genome-scale methods for comprehensive identification of changes in gene expression [90], structural genomic variation [67] and DNA sequence [55] provided insight into the molecular basis of adaptive evolution in chemostats. For many years, experimental evolution using chemostats and budding yeast have primarily been performed using glucose as the growth limiting substrate. More recently, we reported a survey of adaptive evolution of budding yeast in glucose-, phosphate- and sulfur-limited environments [61]. Comparison among these selections revealed that the number of adaptive strategies differs as a function of the selective pressure and thus the details of the selective regime dictate the "repeatability" of evolution. Here, we have built on our recent report of adaptation in nitrogen-limited chemostats [62] to yield a comprehensive survey of adaptive strategies in environments that are limited for different sources of nitrogen. Our new study allows us to draw several general conclusions about the mechanistic bases of adaptive evolution in nutrient-poor environments and provide new insight in the complexity and dynamics of adaptive evolution.

2.4.1. Alleles that specifically increase the transport kinetics of the compound containing the growth-limiting nutrient are recurrently selected

In a chemostat, the rate of cell growth is constrained by the concentration of a single nutrient that is essential for growth [91]. Thus, there is intense selective pressure for adaptive strategies that improve the import or metabolism of the growth-limiting nutrient. In our study, we initially provided a single source of nitrogen at a growth-limiting concentration. We observed massively increased fitness of in selected lineages following 250 generations of selection when fitness was assessed in the same environment as that in which the selection was performed. In the majority of cases, analysis of individual lineages identified CNVs that include a transporter gene that specifically transports the molecular form of nitrogen provided in the environment. Thus, in addition to the amplification of the *GAP1* locus in glutamine- and glutamate-limited conditions [62], we find *DUR3* amplification alleles in urea-limited environments, *DAL4* amplification alleles in allantoin-limited environments and *PUT4* amplification alleles in proline-limited environments. The fact that these CNVs are detected in DNA samples of entire populations indicates that they are at high frequency in these populations, most likely as a result of selection. Transcriptome analysis indicates that these alleles result in increased gene expression, which likely results in increased protein production. Our new results are consistent with previous studies in budding yeast that have identified amplification of the *HXT6/7* locus in populations adapted to glucose-limitation [33,34,61] and amplification of the *SUL1* locus, encoding the

high affinity sulfur-permease, in populations adapted to sulfur limitation [61]. The large fitness increases attributable to these specific CNV alleles means that they dominate the evolutionary dynamics of adapting populations thereby limiting the genetic diversity in nutrient-limited environments. CNV alleles have been reported to underlie increased fitness in a diversity of selective environments and organisms, including humans, suggesting that they are a class of genetic variation that are of general importance for adaptive evolution.

Increased fitness associated with nutrient transporter amplification is specific to nutrient-poor environments. Using competitive growth rate assays in nitrogen-rich environment we find that evolved clones tend to have decreased fitness. Similar fitness trade-offs in carbon-rich environments have been reported for lineages adapted to glucose-limited chemostats [65]. Amplified transporter alleles may be an underlying source of this antagonistic pleiotropy. Previously, we have shown that inactivating mutations in *GAP1* are selected in chemostats containing limiting concentrations of non-amino acid nitrogen sources [62]. In the current study we identified a *PUT4* inactivating mutation in a lineage evolved under arginine limitation (**Figure 2.9A**). In environments in which the limiting nutrient is present in a predominant molecular form, loss of some transporter genes may be beneficial either through reduction in the energetic cost of their unnecessary production or as a result of a function that is deleterious in the particular environment. Future work will be required to rigorously test the hypothesis that CNV alleles are a molecular basis of antagonistic pleiotropy.

2.4.2. A hierarchy of generalist strategies underlies adaptive evolution in nutrient-poor environments

In addition to selection of specific transporter amplification alleles in different nitrogen-limited environments, we find evidence for convergent routes to increased fitness across different nitrogen-limited environments. The most striking evidence comes from the multiple inactivating and non-synonymous mutations that we identified in *VAC14*. We found at least one, and as many as three, independent alleles within the 2.6kb coding region of *VAC14* in eight of the eleven populations that we studied (**Figure 2.9A**). *VAC14* encodes a scaffold component of the protein complex regulating inter-conversion of phosphatidylinositide-3-phosphate (PI3P) to phosphatidylinositide-3,5-bisphosphate ($PI(3,5)P_2$) [81]. In addition, we found mutations in *FAB1*, which encodes a PI3P 5-kinase and *VAC7,* a regulator of FAB1, in different nitrogen-limited populations, albeit, much less frequently than *VAC14* mutations. Control of $PI(3,5)P_2$ levels by VAC14, VAC7 and FAB1 is important for several cellular processes including protein trafficking and maintenance of vacuole size and acidity [92,93]. Loss of function of VAC14 results in decreased $PI(3,5)P_2$ levels leading to enlarged vacuoles due to defective vacuolar fission [94]. Enlarged vacuoles may be beneficial in nitrogen-limited conditions as vacuoles function as a reserve for nitrogen stores as well as being the compartment for recycling of cytosolic proteins through autophagy [95]. Non-synonymous mutations in the *VAC7* and *FAB1* may have similar consequences on $PI(3,5)P_2$ levels and vacuole biogenesis as VAC14 loss of function mutations. Although

identifying the precise mechanistic basis by which mutations in these functionally related genes contribute to increased fitness in nitrogen-limited environments requires additional study, their selection in different nitrogen-limited environments, and their absence in the mutational spectra identified in other nutrient-limited conditions reported to date, suggests that novel alleles at these loci underlie a generalist strategy specific to nitrogen-limited conditions.

By integration of our results with previous studies in other nutrient-limited environments, we find evidence for adaptive strategies involving remodeling of the TORC1 and Ras/PKA signaling pathways that may be general to nutrient limitation. These signaling pathways control cellular growth rate in response to nutrient availability by regulating diverse cellular processes [96,97]. In particular, mutations in the regulator of cell cycle exit and entry into G_0, *RIM15* are found in different glucose- and nitrogen-limitation selections (**Figure 2.12A**). RIM15 is known to have an important role in integrating signals from multiple nutrient responsive signaling pathways including TORC1 and Ras/PKA [98,99]. A reduced capacity to enter a G_0 state could be beneficial in a variety of nutrient-limitations in chemostats. Consistent with this hypothesis, additional genes that are mutant in both nitrogen- and glucose-limited chemostats include *WHI2*, a negative regulator of G_1 cyclin expression, *IRA1* and *GPB2*, both of which are negative regulators of the Ras/PKA pathway, and *NGR1*, an RNA-binding protein involved in regulation of cell growth control. Selection for this class of mutations in different nutrient limitations is consistent with the argument that recurrent selection for loss of *rpoS*

in *E. coli* populations evolved in glucose-, nitrogen- [100] and phosphorous-limited [53] chemostats underlies a tradeoff between the cellular response to nutrient starvation and maintenance of stress resistance.

2.4.3. Selected variation accumulates in genetic networks under epistatic constraints

Although transporter amplifications dominate the majority of our adaptive evolution experiments, we did not identify transporter amplification alleles in two of our populations (ammonium and arginine limitation); the population that underwent adaptive evolution in an ammonium-limited environment was the only population in which we did not identify any CNVs or large-scale chromosomal events. Nutrient transport is still a primary target of selection in this population as we found two independently acquired non-synonymous SNPs that result in amino acid substitutions at the same amino acid residue in MEP2 (G352A and G352S). The mutated site is in a predicted trans-membrane domain (**Figure 2.10C**) making it likely that these mutations alter the affinity of MEP2 for ammonium either directly or indirectly. Fitness tests of one of a strain containing one of these mutations (G352A) show that this variant confers a fitness increase exceeding 10% (see *mep2-2* in **Figure 2.12B**). Interestingly, we find evidence that independently generated alleles containing this precise variant may have been selected in natural yeast populations. Although our ancestral strain, which is isogenic to S288c,

encodes a glycine at residue 352 in MEP2, this site is polymorphic among *S. cerevisiae* strains with 19/26 strains in the SGD database (http://www.yeastgenome .org) encoding an alanine at residue 352. Moreover, the reference genomes of *Saccharomyces sensu stricto* species, including *S. uvarum, S. mikatae,* and *S. paradoxus*, all contain an alanine at residue 352 in MEP2 homologues. It is interesting to note that a recent study reported recurrent selection of *MEP2* fusion alleles when a hybrid *S. cerevisiae*/*S. uvarum* strain was evolved in ammonium-limited chemostats [101]. *S. cerevisiae* and *S. uvarum* differ at 17 residues in the MEP2 protein, one of which is the 352nd amino acid. Consistent with the importance of the 352A allele under conditions of ammonium-limitation, all independently selected *S. cerevisiae*/*S. uvarum MEP2* fusion alleles retained the carboxy terminus-encoding portion of the *S. uvarum MEP2* allele, which codes for an alanine at codon 352. Collectively, these observations suggest that the selection that we imposed in the laboratory bears some resemblance to selection experienced by yeast cells in the natural world with a strikingly convergent response to selection at the molecular level.

The population adapted to ammonium-limitation provides evidence that accumulation of variation in functionally related genes underlies adaptive evolution in nutrient-limited environments. Two lineages within the population that contain mutations in *MEP2* also contained mutations in *GAT1*, which encodes a transcriptional activator of *MEP2* (in addition to other NCR genes) as well as mutations in *LST4*, which encodes a protein that functions in protein sorting to

plasma membranes [102]. Analysis of the dynamics with which these mutations were selected demonstrates that their sequential acquisition underlies clonal interference dynamics in this population. Clonal interference due to multiple independent mutations at the same locus has been documented in a variety of experimental evolution studies (e.g. [103]). Our current results show that competing lineages in the same population can accumulate mutations at multiple, common loci as has been observed in *E. coli* [77]. Interestingly, unlike the recurrently selected three locus genotype identified in [77] comprising variants in *spoT*, *rbs* and *nadR*, which encode functionally unrelated gene products the three loci that define the recurrently selected genotype identified in our study, *GAT1*, *MEP2* and *LST4,* comprise a functionally related gene network (**Figure 2.10D**).

The order in which mutations at these three loci are acquired appears to be constrained by epistatic interactions. By studying all possible allelic combinations at these three loci we determined that the *lst4-2* allele exhibits positive epistasis with the *mep2-2* and *gat1-2* alleles individually. However, the two locus *gat1-2*/*mep2-2* genotype is more fit than the three locus *gat1-2*/*mep2-2/lst4-2* genotype (**Figure 2.12C**). This negative epistatic interaction is consistent with the observation that an *LST4* mutation occurs in the background of a *GAT1* mutation (lineage A0) or a *MEP2* mutation (lineage B3), but does not occur in the lineage in which both a *GAT1* and *MEP2* mutation has already occurred (lineages B1 and B2) (**Figure 2.12A**). It is also interesting to note that the double mutant genotypes (*gat1-2*/*lst4-2* and *lst4-2*/*mep2-2*) and the triple mutant genotype (*gat1-2*/*lst4-*

2/*mep2-2*) do not differ significantly in their fitness (**Figure 2.12C**), suggesting that they will coexist in an evolving population. Consistent with this expectation, the lineages A0 and A1, which differ only at LST4 and the lineages B1 and B3, which differ at LST4 and two additional loci, co-exist that for around 100 generations (**Figure 2.12A**).

Increasingly, resolution of the multigenic basis of quantitative trait variation to nucleotide variants demonstrates that allelic variants in functionally related genes underlies adaptive evolution [104,105]. As the multi locus genotype that we have identified is 1) comprised of functionally related gene products that 2) interact epistatically with one another, we propose that it comprises a gene network polymorphism (GNP) similar to that reported for the galactose-utilization regulon segregating in diverged *Saccharomyces kudriavzevii* populations [106]. Given a sufficiently large population size, we show that nearly identical GNPs can be recurrently generated and selected within a population resulting in "soft sweeps" in which the GNPs are maintained at intermediate frequencies. The rapid generation of a GNP in a particular niche may lead to balanced unlinked GNPs (buGNPs) segregating in the larger population as observed in the *Saccharomyces kudriavzevii* population [106].

2.5. CONCLUSION

Our study provides new insight into the functional basis of adaptive evolution in nutrient-limited environments. Consistent with the low concentration of a single growth-limiting substrate representing the dominant selective pressure in a chemostat we find evidence for strong selection of alleles that enhance transport of the specific molecular form of the limiting nutrient. In addition, we have identified a mechanism underlying adaptive evolution that appears to be shared among different nitrogen-limited environments, involving phospholipid metabolism and vacuole biogenesis, and a mechanism shared between nitrogen- and carbon-limited environments, entailing nutrient-responsive growth regulating pathways. The identification of a finite number of adaptive strategies in nutrient-limited environments suggests that adaptive evolution of large populations in nutrient-limited environments proceeds along a limited number of paths. Thus, the combination of precise knowledge of the selective environment experienced by a population of organisms and the molecular mechanisms that underlie growth and survival in that environment is likely to greatly enhance the predictability of adaptive evolution.

2.6. MATERIALS AND METHODS

2.6.1. Strains and media. For all adaptive evolution experiments we founded populations with a haploid derivative (FY4) of the S288c reference strain. For competition assays, we integrated constitutively expressed mCherry or mCitrine-

labeled constructs, marked with the kanMX4 cassette, at the HO locus using the high efficiency yeast transformation protocol [107]. All nitrogen-limiting media contained 800μM nitrogen regardless of the molecular form of the 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 [108].

2.6.2. Long-term selection. We founded populations with FY4 in 200 mL of nitrogen-limited media. Chemostat cultures were maintained using Sixfors fermentors (Infors) at 30°C, constantly stirred at 400 rpm in aerobic conditions and diluted at a rate of 0.12 hr^{-1} (population doubling time 5.8 hr). Each steady-state population of ~10^{10} cells was maintained in continuous mode for 250 generations (~ 2 months). A 2 mL population sample was obtained every 20 generations and archived at -80°C in 15% glycerol.

2.6.3. Isolation of clones. Following 250 generations of selection we randomly plated cells onto rich media (YPD), and selected an unbiased sample of 94 clones. We grew all clones from each population in 96 well plates containing the same nitrogen source as that used in the selection experiment and recorded optical densities at 600 nm every 0.5 hr over 24 hours using a 96-well Tecan plate reader. Each plate included the ancestral strain (FY4) and a blank well. We estimated the growth rate and the saturation density of all strains using the '*grofit*' package [109] in R and selected three clones from each population for further analysis.

2.6.4. Determination of cell ploidy. We determined the DNA content of evolved clones by staining with Sytox green and analyzing at least 10,000 cells using flow cytometry. FY4 and an isogenic diploid (FY4/FY5) were used for calibration. In addition, each evolved clone was mated with an isogenic strain (FY5) of the opposite mating type (MATα). The resulting strain was sporulated and at least 20 tetrads were dissected using a micromanipulator. Spore viability was determined after three days growth on YPD at 30°C.

2.6.5. Fitness estimates. Each mutant was competed in a chemostat against the ancestral strain (FY4) or a mutant bearing *gat1-2*, *mep2-2*, and *lst4-2* mutations, engineered to constitutively express either mCherry or mCitrine, in the same nitrogen-limited condition used in the selection experiment. We inoculated the unlabeled evolved clone and labeled reference strain in separate chemostat vessels and obtained steady-state cultures of 200 mL. We then mixed the evolved clone with the labeled reference strain to a final ratio of 1:5. We obtained 2 mL samples every 2-3 generations over a total of ~20 generations. Samples were stored at 4°C in phosphate buffered saline (PBS) containing 0.01% Tween 20. The relative ratio of the fluorescently labeled reference strain and the unlabeled evolved clone was measured by counting at least 100,000 cells from each sample using flow cytometry. We used linear regression of the log transformed (ln) ratio of evolved/reference strain abundance against time (in generations) to estimate the selection coefficient (*s*, the slope of the fit linear line) and associated standard error

(*s.e*) using the '*lm*' function in R. We calculated the 95% confidence interval of the regression coefficient in R. The relative fitness, normalized to wild type, is *1+s*. Competition assays in batch culture were performed using synthetic deficient (SD) media containing 5 g/L ammonium sulfate and were performed using analogous methods by first growing evolved and fluorescently-labeled ancestral strains in isolation to log phase and then mixing them at a 1:1 ratio. Cultures were maintained in log phase growth for 24 hours (less than 12 generations) and sampled 5-6 times. The relative abundance of the two strains and fitness coefficients were determined using the same flow cytometry and analytical methods used for chemostat competitions.

2.6.6. DNA microarrays. RNA samples were obtained from evolved clones grown in chemostats limited for the same nitrogen source in which they had been selected. In addition, we obtained RNA samples of the ancestral strain (FY4) grown in each of the nitrogen-limited conditions. Gene expression profiling was performed using Agilent 60-mer DNA microarrays as previously described [61,73]. We used a common reference for all expression analysis, obtained from a sample of the ancestral strain grown in an ammonium sulfate-limited chemostat growing at a dilution rate of 0.12 hr^{-1}. We identified gene expression variation specific to evolved clones by normalizing each mRNA abundance measurement with the expression level of that transcript in the ancestral strain grown in the same environment. Array Comparative Genomic Hybridization (aCGH) was performed

using Agilent 60mer DNA microarrays as previously described [61,73]. Genomic DNA (gDNA) from evolved clones and entire populations was prepared using the QIAGEN genomic DNA extraction kit, labeled with Cy3 and co-hybridized with Cy5-labeled DNA from the ancestral strain. The resulting $\log_2$ transformed ratio was segmented using the '*DNAcopy*' package [110] in R.

2.6.7. Library preparation for next-generation sequencing. We obtained gDNA from each evolved clone and the ancestral strain (FY4) from 10 mL overnight cultures using the QIAGEN genomic DNA extraction kit. For population samples, gDNA was extracted from 10 mL samples taken directly from the adapting population. 1 μg of gDNA sample was then sonicated in a Covaris AFA to obtain fragments of 300-500 bp. To blunt the ends of fragmented gDNA we incubated with PNK (10 Unit) and T4 DNA polymerase (12 unit) at 20°C for 30 min, and then purified using QIAGEN Min-Elute Columns. Adenosine overhangs were added to the blunted DNA using Exo(-) Klenow (15 Unit) incubated at 37°C for 20 minutes, followed by purification using QIAGEN Min-Elute Column and elution in 19 μL EB buffer. To multiplex genome sequencing we ligated one of six unique 120 bp adapters (BIOO) using Quick ligase at 23°C for 20 minutes. The ligated samples were purified, and adaptor dimers removed, using AMPure XP beads (Agencourt). The purified samples were loaded on a 2% agarose gel with TAE buffer, run at 100 V for 60 min and then stained with SYBR gold. We excised a region of the gel corresponding to 300 to 500 bp and then recovered DNA using a

QIAquick Gel Extraction kit. The ligated DNA was PCR amplified using adapter-specific primers and High-Fidelity DNA polymerase in 25 μL reaction volume for 12 cycles to minimize amplification. The concentrations of libraries were determined by qPCR using the Kapa SYBR qPCR Master mix kit and the PhiX library sample as a control. The final samples were diluted in 10 mM Tris-HCl, pH 8.0 and 0.05% Tween 20 and 2 nM of each DNA library was loaded onto a flow cell.

2.6.8. Sequencing data generation and preprocessing. DNA libraries were sequenced using either single end (36 bp and 77 bp) or paired end (2x100 bp or 2x50 bp) protocols on a Illumina HiSeq 2000. Standard metrics were used to assess data quality. We used the *Saccharomyces cerevisiae* S288C reference genome, obtained from the SGD database on Feb 03, 2011 to align reads using BWA 0.5.9 [111]. We trimmed bases with base quality less than 20 from the 3' end of each read. We removed reads with mapping quality less than 20. In addition, PCR duplicates were removed using Picard 1.57 (http://picard.sourceforge.net). We generated BAM files from all remaining reads using samtools 0.1.18 [112]. The average read depth of all sequenced strains is ~160 X.

2.6.9. SNP and indel identification in clonal samples. To identify SNPs we used samtool 0.1.18 and bcftools 0.1.17 with the Bayesian inference option. We determined an empirical quality score cutoff of 160 using bcftools. For paired end

sequencing data we excluded all anomalous read pairs. As clonal individuals are haploid we required SNP alleles to have call frequencies close to 1.0. In duplicated genomic regions or diploidized clones, which may contain heterozygous SNPs, we lowered this requirement to a call frequency near 0.5. In addition, we excluded all SNP calls that were also identified in the ancestral strain. To identify small insertions and deletions (indels) we used the DINDEL package [113]. We first generated candidate variants from BAM files using DINDEL, and then realigned each of them to the reference sequence in order to minimize false positive calls that are frequent in repetitive regions. Indels detected by DINDEL package are therefore defined as those that are shorter than the sequence read length (50bp or 100bp depending on sequencing mode).

2.6.10. Identifying SNP alleles in heterogeneous population. We developed a heuristic threshold to identify low frequency SNPs in population sequencing data. First, we used two different BQ cutoffs, of 20 and 30, to identify SNPs using SNVer [114]. By comparing different population sequencing data to each other and to the ancestor, we identified SNPs in populations as ones that (1) are not found in the sequencing data from the ancestor and (2) exist uniquely in sequencing data from one population using both the high (30) and low (20) BQ cutoff options. We empirically found that optimal p-value cutoff of SNP calls generated using SNVer was $1x10^{-8}$, and the minimum total number of read counts covering the SNP location should be 50% of the average read counts in each population sequencing

data. Using these heuristics we were able to detect SNPs with frequencies of at least 5% in population sequencing data. The allele frequency of each SNP in a population was determined by dividing the number of reads containing the alternative base by the total number of bases mapping to that position.

2.6.11. Functional enrichment analysis. We collected all GO terms from '*GO.db*' and '*org.Sc.sgd.db*' packages in R, resulting in 6,366 ORFs assigned to 4,583 GO terms. We excluded any GO terms for which the number of assigned genes is less than 2 or more than 100. For a tested set of mutated genes we excluded ones without any GO annotation, incremented the count for each additional mutation identified in loci with multiple independent alleles and included both genes neighboring an intergenic SNPs. We then counted how many mutated loci are assigned to each term. We computed the p-value for each GO term using a one-tailed Fisher exact test. We used a Bonferroni correction to correct for multiple hypothesis testing.

2.6.12. Estimation of allele and genotype dynamics. We prepared gDNA from population samples taken at 7 intermediate time point in addition to the final generation (i.e. 24, 61, 102, 137, 173, 213, and 250 generations) using a rapid gDNA extraction protocol [115]. We amplified 200-500 bp length amplicons that contain the SNP at a central position. All amplicons were sequenced using Sanger sequencing and the resulting electropherogram analyzed using PeakPicker to

estimate allele frequencies as described [61,116]. Vectors of allele frequencies were clustered and averaged if the Pearson correlation coefficient of two mutations was greater than 0.97 and the difference in allele frequencies in the final generation (based on deep sequencing) was less than 4%. As allele frequency estimates from Sanger sequencing are less accurate than those obtained from deep sequencing data we excluded a small number of allele frequency estimates derived from Sanger sequencing that were inconsistent with our deep sequencing results.

2.6.13. Measurement of genetic interactions among alleles. We backcrossed clone 3, recovered from the ammonium-limited condition to the ancestral strain of opposite mating type (FY5; MATα), sporulated the hybrid diploid and dissected tetrads. All segregants were tested for mating type using halo assays [117]. We obtained more than one hundred backcrossed strains bearing different combinations of the 5 mutations acquired by clone 3. Genomic DNA for each strain was prepared using a rapid DNA extraction protocol [115]. Genotyping was performed using allele specific PCR. Eight strains identified by this process contained all possible combinations of the three mutations of interest – *gat1-2*, *mep2-2* and *lst4-2* – and the ancestral alleles of the two additional loci (*RIM15* and *FAB1*) that were not studied. Each strain was individually competed against the mCitrine-labeled reference strains as described.

2.6.14. Accession numbers. All DNA sequencing data are available from the NCBI Sequence Read Archive with accession number SRP032757. DNA microarray data are available through the NCBI Gene expression Omnibus with accession number GSE52787.

CHAPTER 3. Experimental evolution of a gene regulatory network

*This chapter is based on the research paper "**Experimental evolution of a gene regulatory network**" by Jungeui Hong and David Gresham (in preparation as of December, 2014)*

3.1. ABSTRACT

Understanding the molecular basis and dynamics of gene expression evolution has been of great interest in evolutionary biology. However, consensus from this field still remains elusive due to the lack of molecular tools of investigating genetic architecture underlying it at a systems level. To study the evolution of gene expression under conditions of strong selection, we performed experimental evolution of yeast cells growing in ammonium-limited chemostats. Following several hundred generations we found significant divergence of nitrogen responsive gene expression in lineages with increased fitness. We found repeated selection for non-synonymous mutations in the zinc finger DNA binding domain of the GATA transcription factor, *GAT1*, an activator of the nitrogen catabolite repression (NCR) regulon. The functional effects of GAT1 mutations are exerted both directly, and indirectly by rewiring of incoherent feed-forward loops comprising multiple GATA transcription factors and their common targets in the NCR regulon. We also find that evolving populations contain multiple GAT1 mutations at low frequencies (10^{-2}-10^{-3}) during the initial stages of the selection that fail to rise to appreciable frequencies due to clonal interference. Our study demonstrates that under strong

selection the evolution of gene expression is highly repeatable and that rewiring transcriptional networks can lead to both direct and indirect effects.

3.2. INTRODUCTION

The evolution of gene expression is a pervasive source of phenotypic diversity [118,119]. Most genetic variations causing such diversity result in either *cis*-regulatory or *trans*-regulatory changes. The relative importance of these two mechanisms is the source of a long-standing debate [120-123]. This is mainly because evolutionary biology is a retrospective science; the forces and processes underlying adaptive evolution are necessarily inferred from extant organisms making it hard to observe the evolutionary dynamics in real-time and to distinguish neutral from adaptive alleles.

Long-term experimental evolution (LTEE) provides a means of observing some of the inherent difficulties of evolutionary studies. In conjunction with recent advent of next-generation sequencing, many LTEE studies have successfully identified a comprehensive list of adaptive alleles and their dynamics. Copy number variants of nutrient specific transporter encoding genes are a highly convergent solution in various nutrient poor adaptive evolution [34,35,37,65]. Protein coding mutations are more frequent than *cis*-regulatory changes in LTEEs that have undergone near constant nutrient limitations in chemostats [34,37]. Such alterations are mostly missense, frame-shifting or stop codon mutations that are likely to confer severe pleiotropic effects compared to *cis*-regulatory changes [124-126]. However, studies

explaining the molecular basis of protein coding changes in adaptive evolution are still lacking.

We previously reported repeated selection within a single population evolving in ammonium-limited chemostats of independent missense mutations in the DNA binding domain of the transcription factor GAT1, an activator of nitrogen catabolite repression (NCR) genes for uptake and utilization of multiple nitrogen sources in *Saccharomyces cerevisiae* [37] (**Figure 3.1**). The expression of NCR genes is regulated by four GATA factors – two activators (GAT1 and GLN3) and two suppressors (DAL80 and GZF3) – that compete for the same binding sites in promoter regions (**Figure 1.1B**). NCR is an ideal system for studying the evolution of gene expression owing to the well-characterized properties of four regulators and the small number of direct targets (~ 40).

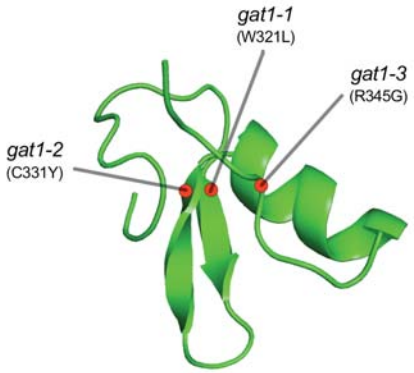

Figure 3.1. A model of 3D structure of predicted DNA binding domain of GAT1. 3D structure is based on all available GATA factor DNA binding domain structures in the 'modebase' database. Based on manual inspection, all amino acid changes from the GAT1 mutations appear to be deleterious to unknown degree to its protein functionality.

We aimed to determine whether selected mutations in GAT1 result in alterations in its regulatory activities and, if so, how rewiring of the regulatory network increases fitness in long-term nutrient limitations. Specifically, we sought to determine (1) Does rewiring confer pleiotropic effects? (2) How do *trans*-regulatory changes alter gene expression? (3) What is the dynamics of the evolution of gene expression? (4) Is regulatory rewiring convergent or contingent?

Here, we describe the first example of transcriptional factor evolution in real time and dissection of it functional effects. Missense mutations in GAT1 are under strong positive selection under conditions of ammonium-limitation and also show antagonistic pleiotropy in closely related selective regimes. The functional effects of *GAT1* mutations are exerted by rewiring feed-forward loops in the NCR transcriptional network. The evolutionary dynamics of the mutations suggest that rewiring of transcriptional regulator network is deterministic and predictable during the early stage of evolution but not in the later stage where stochastic outcomes dominate. Our findings may inform our understanding of gene expression evolution in pathogens and cancer cells [44-46].

3.3. RESULTS

3.3.1. GAT1 missense mutations exhibit antagonistic pleiotropy

We isolated three single GAT1 mutations – *gat1-1* (W321L), *gat1-2* (C331Y), and *gat1-3* (R345G) – using genetic backcrossing and allele specific PCR genotyping from three previously described lineages that evolved under ammonium-limitation

for 250 generations [37] (see **Figure 3.1**). Using competition assays (see methods), we determined fitness effects of the three GAT1 mutations and three original evolved mutants bearing additional 3 or 4 SNPs under various different nitrogen-limiting conditions in chemostats – ammonium, glutamine, proline and urea – in addition to YPD batch condition. We also compared them to the fitness effect of engineered GAT1 knockout mutation (**Figure 3.2**).

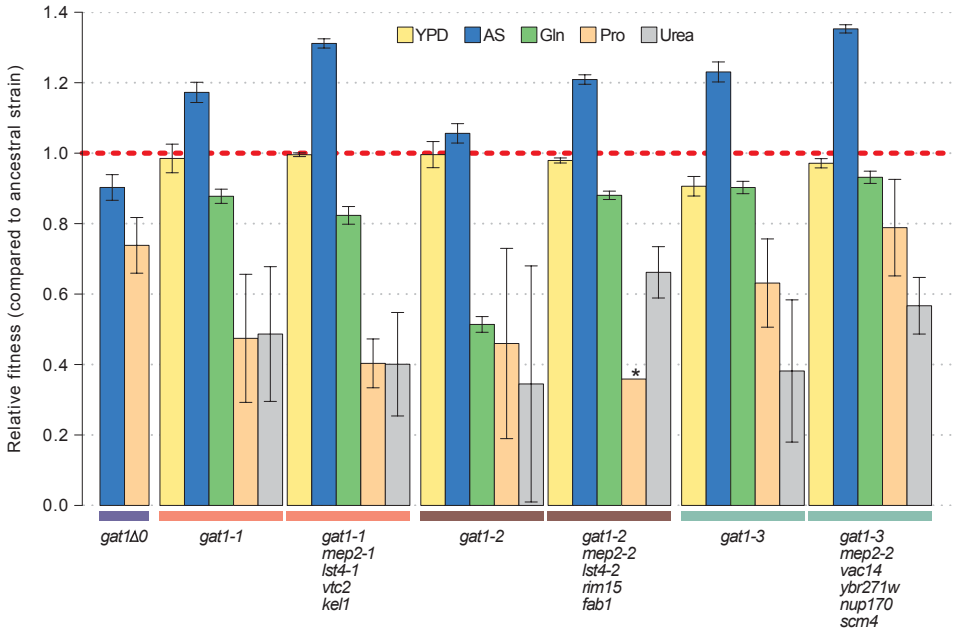

Figure 3.2. Antagonistic pleiotropy of GAT1 mutations. Relative fitness of the three original clones, the three *gat1* single loci and GAT1 knockout mutants compared to the ancestral WT strain in three different nitrogen-limited environments. GAT1 single mutations are not the fitness optimum and show severe antagonistic pleiotropy in fitness. They are only beneficial under constant limited concentration of ammonium nitrogen sources. Error bars represent 95% CI of linear regression analysis and (*) represents not statistically significant. (YPD : YPD batch culture, AS : Ammonium sulfate limited chemostat, Gln : Glutamine limited chemostat, Pro : Proline limited chemostat, Urea : Urea limited chemostat)

All strains bearing GAT1 missense mutations showed significantly increased fitness in ammonium-limited chemostats and dramatically decreased fitness in all other types of nitrogen limitations in chemostats. GAT1 mutations are nearly neutral or slightly detrimental in YPD batch media. Interestingly, the extent to which fitness decreases varies depending on the preference of yeast cell toward the nitrogen source being limited: marginal in preferred (glutamine or YPD) but dramatic in non-preferred sources (proline and urea) limitation in chemostats. By contrast, antagonistic pleiotropy was not observed in the GAT1 knockout strain under ammonium and proline limited chemostats. This suggests that missense mutations in the DNA binding domain of GAT1 recovered from LTEE have altered function and are not null mutations.

3.3.2. Selective alteration in gene regulation by GAT1 mutations

We aimed to determine how the transcriptional regulatory network activated by GAT1 is rewired due to adaptive missense mutations. To this end, we conducted RNA-seq for three individual GAT1 mutants. We compared expression profiles with those of lineages assayed using DNA microarrays in our previous report [37].

We find significant divergence of NCR gene expression in all tested strains (**Figure 3.3A**). Single GAT1 mutants showed up-regulation of genes encoding permeases for ammonium (*MEP1*, *MEP2*, and *MEP3*) as well as for other nitrogen sources such as urea, allantoin and GABA. The evolved lineages show more ‘fine-tuned’

gene expression pattern in which only ammonium permeases encoding genes are up regulated while other NCR targets are repressed.

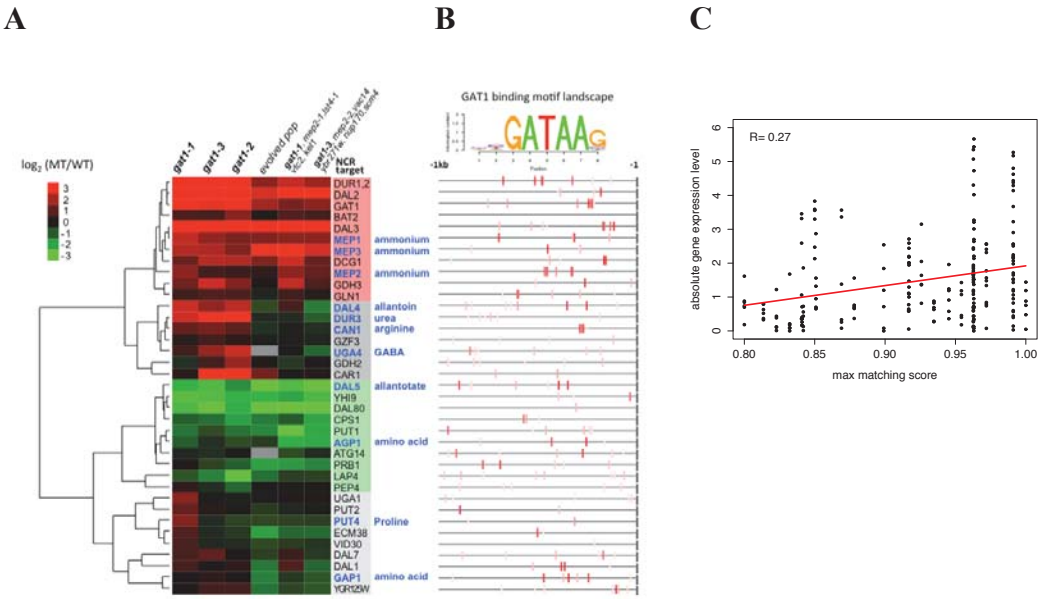

Figure 3.3. Correlation between gene expression level and binding landscape of NCR target genes. A, Clustering of gene expression profile of NCR target genes that are significantly up or down-regulated in GAT1 mutant background and binding motif analysis for target genes of Gat1p. 'Blue' colored genes in right side are permease encoding genes for different nitrogen sources. This clustering includes only 41 experimentally confirmed NCR target genes. Microarray data for the finally evolved population and two clones are adapted from our previous report (see ref [37]). B, GATA-factor binding landscapes for NCR target genes. Color intensity is proportional to the matching score against the consensus motif. C, Correlation between the maximum matching score of each motif to the consensus one and its corresponding absolute gene expression level in all measurements. R (Pearson correlation) is 0.27, weak but positive. X and Y-axes are absolute $\log_2$ transformed fold changes in gene expression compared to the ancestor strain and maximum matching score of each motif to the GAT1 binding consensus sequence.

Interestingly, we see a weak but positive correlation between the gene expression level and the maximum matching score of binding motifs for each gene in all

conditions (**Figure 3.3B and C**). For example, *DAL80* encoding a repressor for NCR targets is severely down regulated and has only one motif with very low matching score to the consensus motif, but *MEP2* is up regulated in all GAT1 single mutation backgrounds and has multiple high scoring motifs.

3.3.3. GAT1 mutations are recessive and hypomorphic

To measure changes in transcriptional activity of the GAT1 mutants, we engineered strains bearing GFP tagged promoter sequences of four transcriptional targets of GAT1 (*GAP1*, *MEP2, GZF3* and *DAL80*) in the background of the ancestor, the knockout, and the three GAT1 mutations (**Figure 3.4A**). GFP expression levels measured using this assay were highly comparable to RNAseq data and thereby a good proxy of transcriptional activation by GAT1 at each promoter (**Figure 3.4B**). Using this reporter assay, we compared the degree of transcriptional activation by different forms of GAT1 in selected promoter sequences under a variety of conditions.

We found significant differences of the transcriptional activities from *MEP2* and *DAL80* promoter constructs in nitrogen (both ammonium and proline)-limited conditions (**Figure 3.5A**). All adaptive GAT1 mutations resulted in suppression of *DAL80* expression and increased *MEP2* expression compared to the ancestor, while the GAT1 knockout mutation showed the opposite pattern. This result is consistent with the differential expression level of *MEP2* and *DAL80* genes in the RNAseq data and also supports the idea that the missense mutations in the DNA binding

domain of GAT1 are not null. We argue that this is the first example of hypomorphic mutations – partially impaired function in its regulation – in a transcriptional regulator from experimental evolutions that leads to rewiring of transcriptional regulatory network and confers antagonistic pleiotropy.

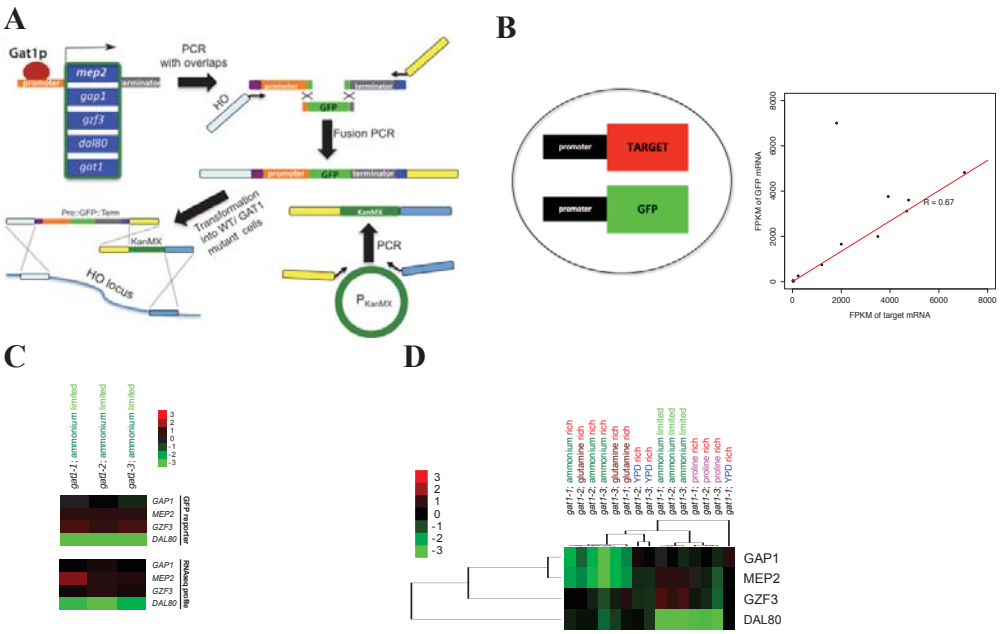

Figure 3.4. GFP reporter assays for selected target promoters of GAT1. A, Construction of GFP fused promoter sequences for *GAP1, MEP2*, *GZF3* and *DAL80* in WT/GAT1 mutant strains. GFP expression intensity measured in FACs is used as a proxy of gene expression profile of each target gene. B, Validation of GFP reporter constructs. We RNAseqed mRNA samples for three GFP constructs (GAP1, MEP2 and DAL80) as biological replicates for WT and three GAT1 mutant strains. Based on FPKM values, mRNA copy numbers of target genes and GFP bearing the same promoter region are highly correlated (R=0.67) in all sequenced samples (One outlier is due to low read coverage). C, GFP reporter assay is highly correlated with the RNAseq result conducted in ammonium-limited environment. Median GFP

intensity measured by FACs of each GAT1 mutant is divided by the one of WT, and then $\log_2$ transformed as a proxy of relative gene expression of each target. GFP expression intensity measured by FACs can be used as a proxy of mRNA expression level of four target genes. D, GFP-tagged promoter activity assays in batch cultures. Overnight cultures in YPD were transferred to SD batch media containing three different nitrogen sources (ammonium, proline, and glutamine), incubated for 1 hour at 30 °C and then measured for GFP expression using flow cytometry. X and Y-axes represent $\log_{10}$ transformed GFP intensity and frequency of cell counts, respectively. E, Comparison of *pDAL80* GFP reporter assays in 4 different nitrogen limited conditions in chemostats. Patterns of transcriptional activation of *pDAL80* by GAT1 are all identical regardless of types of nitrogen limitations.

We also tested whether these mutations are dominant or recessive using heterozygous diploid reporter constructs (*gat1*/*GAT1*) (**Figure 3.5B**). The transcriptional activities of the *DAL80* and *MEP2* promoters in the heterozygote backgrounds were identical to the haploid wide type GAT1 strain implying that the missense mutations of GAT1 are recessive. This means that the partially impaired functionality of the GAT1 alleles is complemented by the wild type GAT1 and excludes the possibility of dominant negative effects.

We found that the pattern of GAT1 activation of its targets, i.e. *DAL80,* is almost identical in all kinds of nitrogen 'limited' conditions in chemostats (ammonium, glutamine, proline and urea) and even non-preferred nitrogen source, i.e. proline, 'rich' conditions in batch cultures (**Figure 3.5D & E**). This result suggests that nitrogen utilization pathway transmits the same signals to downstream regulatory networks in response to nitrogen 'poor' conditions that are either limited

concentration of all types of nitrogen or rich concentration of non-preferred nitrogen sources.

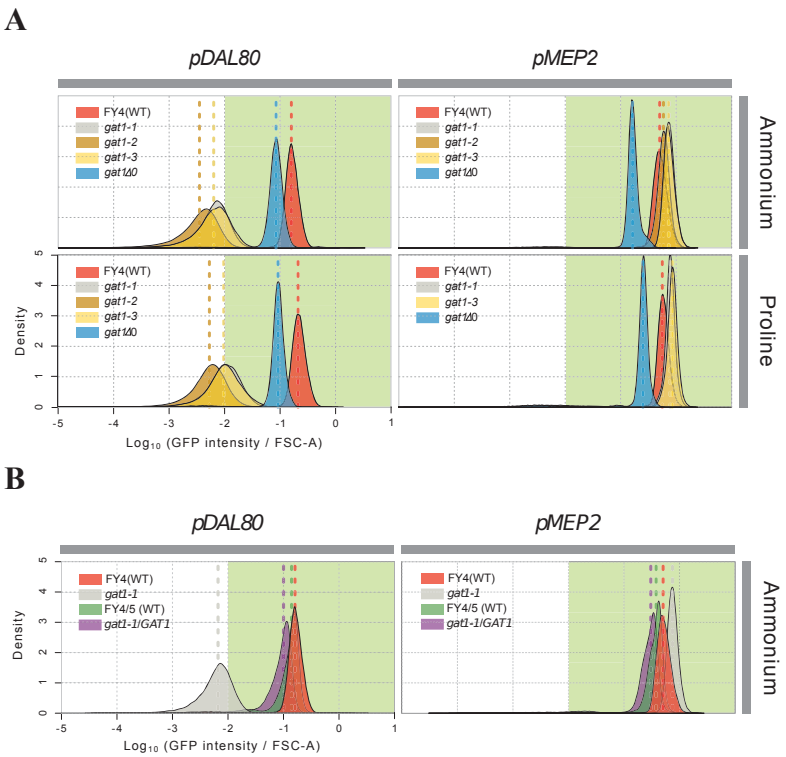

Figure 3.5. Transcriptional activation by different GAT1 mutations. A, Promoter binding activities for *MEP2* and *DAL80* in WT and MT GAT1 backgrounds under various nitrogen limited environments measured by GFP reporter constructs. X- and Y- axes are $\log_{10}$ transformed GFP intensity normalized by cell size (FSC-A) and cell counts measured by FACS, respectively. In either preferred (ammonium and glutamine) or non-preferred (proline and urea) nitrogen-limited chemostats, all strains showed up the exact same gene expression pattern for *DAL80*. Only in WT version of GAT1 strain, *DAL80* can be activated. B, Dominant / recessive test for GAT1 mutations using heterozygote diploid strains and FACS GFP reporter assay.

We suggest that the missense mutations in the DNA binding domain of GAT1 may lead to selective inactivation of its binding activities. Therefore, only a subset of its targets with 'strong' binding motifs such as *MEP2* is still highly activated while others with 'weak' motifs such as *DAL80* are not activated enough (**Figure 3.6A**). The decreased activation of *DAL80* by the adaptive *GAT1* mutations results in more elevated activation of *MEP2* indirectly since suppression of *MEP2* expression via DAL80 is decreased. However, knocking out *GAT1* did not induce such the same effect possibly because GLN3 still can activate *DAL80* expression perhaps due to the absence of competition with GAT1 for the same binding site (**Figure 3.6B**).

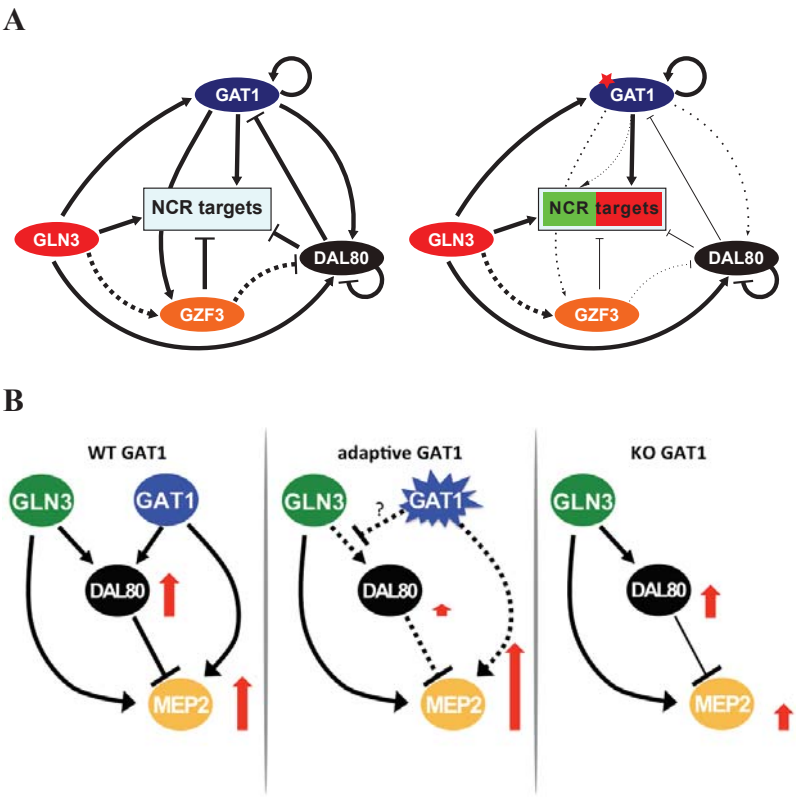

Figure 3.6. Models of rewiring of NCR regulon. A, Proposed model of rewired regulatory network in NCR regulon in the GAT1 mutants

background (red star). The WT model is based on experimentally verified published data. B, A simplified model of transcriptional regulation of GAT1 and DAL80 on the expression of MEP2. Solid and dotted lines represent known strong regulations and weak or putatively inactivated regulations, respectively. Red bars represent the expected level of gene expression.

3.3.4. Convergent evolution of GAT1 mutations

To test whether the rewired GAT1 regulatory network is convergent or stochastic, we performed 'replay' experimental evolution in triplicates populations under the identical ammonium-limited chemostats as the original experiment. We found that evolution is highly convergent and parallel both at the phenotypic and genotypic levels. Following 250 generations we find significant increase in population level fitness along with deceleration in the rate of fitness improvement at the later stage of evolution as seen in the previous LTEE study [36] (**Figure 3.7A**). Using whole genome, whole population sequencing of the evolving populations, we also find recurrent selection for the same types of missense mutations in the DNA binding domain of GAT1 at the early stages of selection (100G). Unlike the original evolution, we found amplification alleles including *MEP2* with high frequency only at the finally evolved stage, 250G (**Figure 3.7B**). This suggests that GAT1 transcription factor evolution is a deterministic solution during the early stage of ammonium-limited adaptations.

However, we did not find any high frequency mutations at 50G although the population fitness had already increased more than 15% (**Figure 3.7A**). We hypothesized that multiple minor frequency mutations in genes that were targeted

at later generations such as *GAT1* might exist in the earliest generation. Therefore, we identified additional 'hidden' mutations that are less frequent (down to 1%) in the evolved populations using targeted amplicon sequencing (**Figure 3.7C**; see method). We find that evolving populations contain multiple *GAT1* mutations mostly in the same DNA binding domain at very low frequencies (10^{-2}-10^{-3}) during the initial stages of the selection that fail to rise to appreciable frequencies likely due to clonal interference. The dynamics of adaptive mutations clearly shows that *GAT1* alleles are the primary and initial target but are gradually outcompeted by other adaptive alleles such as *MEP2* amplifications. Interestingly, *DAL81*, *TPK3* and *WHI5* that are known as general targets also in other types of nutrient-limited adaptations acquired mutations in their coding regions only at the last step of replayed evolution in a more stochastic manner than *GAT1*.

We questioned whether the mutational hotspot in the DNA binding domain of GAT1 is under positive selection in natural populations. From sequencing data of 42 different wild isolates of yeast strains (19 of *S. cerevisiae*, 23 of *S. paradoxus*) [127], dN and dS values in each amino acid position of GAT1 were calculated (**Figure 3.7D**). We find that this domain is under strong purifying selection in natural environments. This result is highly consistent with the antagonistic pleiotropy of GAT1 mutations (**Figure 3.2**); missense mutations of a transcription factor governing nutrient utilization are hypomorphic and only beneficial in a specific type of constant selective pressures but not in other types of fluctuating environments like in the wild.

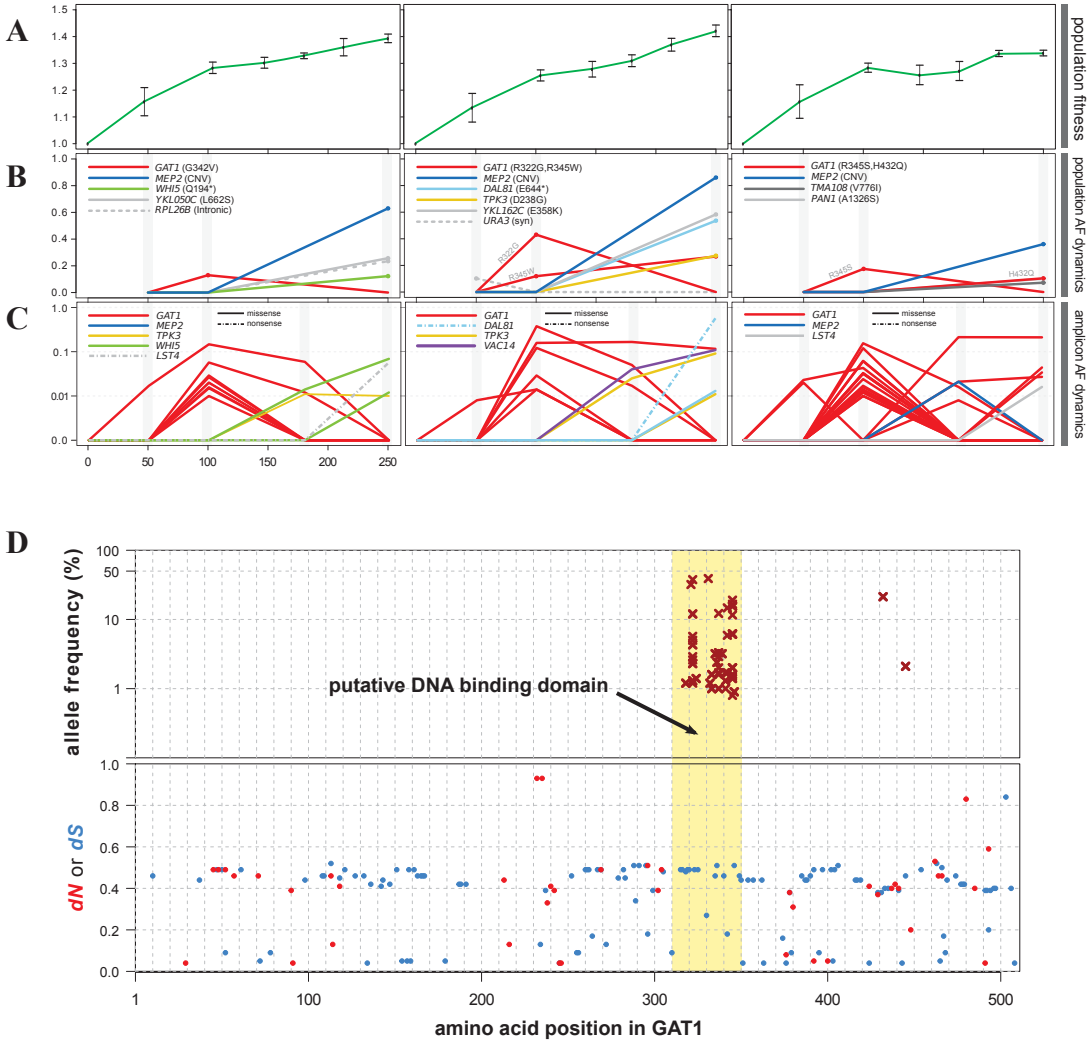

Figure 3.7. The DNA binding domain of GAT1 is target of positive selection. A, The dynamics of population level fitness. The rate of fitness improvement decelerates over time in all replicated evolutions in a very similar fashion. Error bar is 95% CI inferred using linear regression from competition assays. Six different generations were tested. **B, Allele dynamics in parallel evolutions**. 50, 100 and 250 generation samples were sequenced for identifying major frequency mutations (> ~10%) using whole genome population sequencings (Illumina HiSeq2500, 2x50 paired mode; average read depth was ~ 50X) during replayed adaptations. **C, Detection of minor frequency mutations of targeted loci from 'replayed' ammonium-limited adapted populations using amplicon sequencing.** 10 genes that are already found as adaptive targets with high allele frequency under multiple nitrogen limited environments were selectively amplified and sequenced using MiSeq

2x250 option. *GAT1* (red lines) is the primary target from the early generation but is gradually replaced with other alleles such as amplified loci including *MEP2* (blue lines) and protein truncating mutations in genes related with signaling pathway governing cell cycle and growth possibly due to clonal interference. **D, The DNA binding domain of GAT1 is under positive selection under constant ammonium-limitation while under purifying selection in the wild.** With the exception of two mutations (H432Q and S445T), all identified GAT1 mutations are enriched in the putative DNA binding domain in all experiments. *dN* and *dS* values for GAT1 of 42 different wild yeast strains at each amino acid position is also calculated by SNAP v1.1.1 **(**http://www.hiv.lanl.gov/content/sequence/SNAP /SNAP.html). Sequencing data for wild yeast strains was adapted from Bergström et al., 2014 Mol. Biol. Evol. There is no evidence of positive selection in wild but only purifying selection for the domain, implying that non-synonymous mutations are likely detrimental in dynamic environments.

3.4. DISCUSSIONS

3.4.1. Alterations in regulatory network are dominant under strong, constant selective pressure conditions

The main finding of this study is the repeated selection of missense mutations in the zinc-finger DNA binding domain of GAT1 in ammonium-limited environments. The DNA binding domain of a TF is critical for the binding specificity to its target motifs and thereby makes the evolution of specificity very slow [128]. However, there is evidence that conformational changes in DNA-binding domain of TFs during the evolution in the wild can provide flexibility and modularity in gene regulation [129,130]. Many, if not all, TFs in human can recognize multiple different binding sites [128]. We argue that multi-specificity of DNA-binding in wild-type GAT1 is disrupted or altered by missense mutations in

its DNA binding domain. Multiple binding modes of a TF may be unnecessary or even deleterious under very simple, strong selective pressure. The same claim can be applied to missense mutations in the *TP53* gene frequently found in majority of human cancers, which are known to disrupt flexibility in DNA binding activity *in vitro* [131]. Another example is CTCF, a poly-zinc finger transcription factor regulating oncogenes and tumor suppressor genes, that has repetitive somatic missense mutations in the DNA binding domain altering binding specificity to its target genes in tumors [132]. Our study describes the first experimentally derived example of DNA-binding specificity alteration (or evolution) in a transcription factor. The alteration is not simply 'loss of function' or 'inactivation' but rather hypomorphic given that no null, frame-shifting or truncating mutations were found in the binding domain of GAT1. The mutated GAT1 maintains its functionality but with new properties. We suggest that *trans*-regulatory changes may predominate over *cis*-regulatory changes under very strong and constant selective regimes such as cancer evolution.

3.4.2. How does gene expression evolution result in fitness increase?

The evolution of gene expression is an adaptive solution in nutrient limitations. However, it is not clear how such alteration results in fitness increase of an evolved cell. Amplification of nutrient specific transporters is a dominant outcome of selection in chemostats [37]. However, the altered gene expression in GAT1 mutants does not have the same functional effects. For example, *GAT1* mutations

induce up-regulation of *DUR3* and *DAL4* encoding a urea and an allantoin transporter but do not confer fitness benefit in urea and allantoin-limited chemostats (**Figure 3.2** and **Figure 3.3A**). On the contrary, the same antagonistic pleiotropy of the finally evolved clones is explained by the transporter amplification model given that they showed up-regulation of only ammonium specific transport encoding genes (*MEP1*, *MEP2* and *MEP3*) but suppression of others for allantoin, urea, arginine, GABA, allantotate and amino acids (**Figure 3.3A**). Here, we notice that two different non-synonymous mutations targeting the same 352nd amino acid residue in the transmembrane domain of MEP2 (*mep2-1* and *mep2-2;* **see ref 10 and Figure 3.2**) are common to original clones and contribute to additional fitness gain. Indeed, MEP2 can act as not only a high affinity ammonium transporter but also a post-transcriptional regulatory sensor that possibly controls PKA signaling transduction pathway [133] and thereby alters gene expression related with various down-stream cellular pathways [134]. We speculate that signaling pathways mediated by MEP2 may have additional effects on the fitness of *GAT1* mutations under non-ammonium limited conditions.

3.4.3. Temporal contribution of contingency and convergence in evolution

We find that there seems to be a mutual exclusivity between missense mutations in *GAT1* and MEP2 amplifications and/or loss of function mutations in other signaling pathway genes such as *DAL81*, *TPK3* and *WHI5* (**Figure 3.6B & C**). It is also interesting to see that GAT1 mutations are early adaptive alleles while others are

late ones. The similar dynamic pattern was also seen in a glucose-limited experimental evolution study where null mutations in MTH1, a negative regulator of glucose signaling pathway, and amplifications of HXT6/7, high affinity glucose transporters, are mutually exclusive due to reciprocal sign epistasis [34,56]. We hypothesize that signaling or regulatory alterations are immediate adaptive solutions but are outcompeted by transporter amplifications that harbor less metabolic costs and thereby are fitness optimum in nutrient limitations. It is less likely, but worth considering that mutations in those regulators revert once they acquire amplification alleles of transporters in the same lineage due to the reciprocal sign epistasis. This possibility has been raised in a protein evolution study [135] but never reported in the evolution of a functional module comprising multiple genes at a different level. We also cannot exclude the possibility that the transient dynamics of GAT1 mutations is due to subtle but substantial changes in nutrient concentration and/or physicochemical parameters such as pH and pO_2 within chemostats media as faster growing cells become dominant and consume more resources in the culture.

Our results raise an alternative hypothesis for the following central question: are the evolutionary trajectories historically contingent or convergent. Lenski's group has suggested that evolutionary outcomes were divergent in several independently replayed LTEEs that were conducted under an identical nutrient-limited condition in *E. coli* [32]. In contrast, there are countered reports that even if the evolutionary trajectories are many, the consequences are limited by selection: evolution is

convergent and parallel [37-39,136]. This discrepancy may result from the different culturing systems (batch or chemostat) or the different level at which each group defined convergence. Our study suggests that convergence and contingency may have temporally different contributions to evolution.

3.5. CONCLUSION

The DNA binding specificity change in the NCR transcriptional activator, GAT1 is an initial but transient 'driver' as an evolving population seeks the fitness optimum during the course of ammonium-limited adaptations. Since the binding domain is highly conserved in various distant yeast strains and under purifying selection in the wild, we suggest that the evolution of the NCR regulon is specific to the constant ammonium-limited environment. Our study demonstrates that the evolution of gene expression is highly repeatable under strong selection due to alteration in the binding specificity of a transcription factor resulting in rewiring transcriptional networks. We propose this result can be applied to understanding of the adaptive strategies that tumor cells use to proliferate. For example, the mutational "hotspot" in the DNA binding domain of the *GAT1* is reminiscent of a hotspot in the *TP53* gene found in a variety of tumors [137]: surprisingly, the molecular mechanism underlying the oncogenic activity of these missense mutations remains elusive.

3.6. MATERIALS AND METHODS

3.6.1. Strains and media. The ancestral strain used is a haploid derivative (FY4; MATa) of the S288c reference strain. For isolations of single GAT1 mutations, we backcrossed each original evolved mutant (see **Figure 3.1A**) to a opposite mating type haploid (FY5; MATα) strain and recovered tens of random spores bearing different combinations of original adaptive alleles. For genotyping individual segregants, we used allele-specific PCRs where two WT and MT type forward primers and one common reverse primer were prepared. We also knocked out the entire *GAT1* locus by replacing it with G418 marker using the high efficiency transformation protocol. We also engineered *GAT1* single loci strains such that they harbor GFP fused promoter sequences of four GAT1 NCR targets (*GAP1*, *MEP2*, *DAL80* and *GZF3*) into HO locus using homology based transformation methods. For this, we fused GFP with 1kb of 5' and 3' UTR regions of each target gene. We also tried to construct a strain bearing the *GAT1* promoter sequence but even 2kb long 5'UTR region didn't induce any GFP expression in WT and MT GAT1 backgrounds (data not shown). Finally, we also constructed diploid strains that are heterozygous in the *GAT1* locus by mating the GFP reporter constructs made in the WT and the three *GAT1* mutants backgrounds to the opposite mating type FY5 strain. All constructed strains were finally genotyped using Sanger sequencing. All medium conditions and recipes were identical with the ones in our previous study (see ref [37]): all nitrogen limited media contains 800 μM of nitrogen regardless of the molecular forms.

3.6.2. Competition fitness assays. For all detailed protocols and analysis steps, see method section in the ref [37]. In short, we used a mCitrine-labeled FY4 strain for all competition assays as the ancestor strain. We tested fitness effects of three original evolved mutants, three GAT1 single loci segregants and one engineered GAT1 knockout strain under 4 different nitrogen limited-media – two preferred sources (ammonium and glutamine) and two non-preferred ones (proline and urea) – in chemostats and YPD rich medium in batch cultures. Sampling was done at least 5 and up to 10 times for the precise estimation of relative fitness using linear regression analysis. We noticed that less frequent sampling resulted in much broader range of 95 % confidence interval but mostly showed good significant differences in fitness comparing to the ancestor. For the competition assay in YPD rich condition, we sampled twice per day and back-diluted the competing culture (1/200) into a fresh medium every night for 3 or 4 days at an incubator. For population-level fitness assay, we seeded mCitirine-labeled FY4 strain into chemostats vessels beside the vessels already containing evolving populations and let them reach to the same steady state for few days in the same dilution rate and then mixed the evolving population samples with the fluorescence labeled ancestor strain in ratio of 1:9. The competing cultures were independently sampled at multiple time points (less than 20 generations) for FACs. All analysis steps were the same as the clonal fitness assay method.

3.6.3. Directional RNA-seq. We seeded three GFP reporter constructs of *GAP1*, *MEP2* and *DAL80* for each of the WT ancestor and the three *GAT1* single loci

mutants as three biological replicates. In sum, we cultured 12 chemostats vessels (3 GFP reporters x 4 different backgrounds) and harvested a 10 mL of each culture at steady state using the vacuum filtration method, frozen them immediately using liquid nitrogen and kept at -80°C until use. RNA extraction was done using a phenol-chloroform method. For mRNA enrichment, we used poly-A selection and the final yield of selected RNA molecules was around 10 – 50 ng in total. For cDNA synthesis, we used Superscript III kit (Invitrogen) and dNTPs mixtures for the 1st strand synthesis and *E. coli* DNA ligase and polymerase I (Invitrogen) for 2nd strand synthesis with a mixture of dATP, dCTP, dGTP and dUTP. Then, we followed end-repair, A-tailing and adapter ligation based on the general Illumina library preparation protocol. All cleanups in-between each steps were done using AMPure XP beads (Beckman Coulter, Inc). For directional sequencing of 1st strand sequences only, we treated UNG (Thermo) and amplified the ligated molecules using Phusion high fidelity DNA polymerase (NEB) in 12 or 15 of PCR cycles. Adapter dimers were further removed by conducting AMPure XP beads selections twice. We checked the proper amplification of ligated molecules in the BioAnalyzer and then finally quantified library concentration using qPCR method with KAPA Library Quantification kits (KAPA Biosystems). All sequencings were done in the Illumina HiSeq2500 2x50 paired end fast-run mode. We used Tophat v2.0.11 to align sequencing reads to the *Saccharomyces cerevisiae* S288C reference genome, obtained from the SGD database on Feb 03, 2011. From the

RNA-seq count data, we ran 'edgeR' v3.8.5 [138] to determine differential expression (DE) level of all genes compared to the ancestor.

3.6.4. Binding motif analysis. We computationally verified all potential binding sites of GAT1 in the 37 NCR target genes. Position weight matrix (PWM) of the consensus motif ('GATAAG') of GAT1 was obtained from JASPAR DB using 'MotifDb' package in R and only 1kb of 5' UTR region of each target was scanned to find all potential binding sites with at least 80% of matching score to the consensus motif. The matching score was assigned to each motif as 'red' color intensity in the **Figure 3.3B**.

3.6.5. GFP reporter assay. All constructed reporter strains were seeded in the same chemostats conditions. Samples were taken at a steady state from the chemostat cultures that are run in the same dilution rate (0.12/h), sonicated and FACSed in a PBS buffer. We additionally tested the GFP activity in nitrogen batch cultures in AS-ammonium, glutamine, and Proline media and YPD batch culture. All GFP intensity values are normalized by the size factor (FSC-A) and log 10 transformed in the figures (**Figure 3.4D & E**). For diploid strains used for the dominance and recessive test (**Figure 3.3E**), additional size factor between haploid and diploid cells (the median size of diploid cells divided by the one of haploid cells) was also multiplied to the normalized GFP intensity values of diploid cells. This normalization removes all cell size effects to the GFP intensity.

3.6.6. Replayed LTEEs. All conditions for replaying LTEEs were identical as described in the previous study (see ref [37]). Dilution rate (0.12/h) was checked every one or two day over 250 generation (~ 2 months), and intermediate samples were archived at every 20 generations. We confirmed that cell density was consistent as ~ $3x10^7$ cells/mL in 200 mL of cultures across the entire culturing period.

3.6.7. 3D structure of GAT1 DNA binding domain. The ModBase model of yeast GAT1 (UniProt P43574) from an NMR structure (PDB 4GAT) complexed with Zn and DNA, with nearly identical fold to a high resolution crystal structure (PDB 2VUS), is the best model available. The 3D image was visualized by Polyview-3D (http://polyview.cchmc.org/polyview3d.html).

3.6.8. Whole genome population sequencing. We sequenced the entire population genomes at 50, 100 and 250 generations for all three LTEE cultures. All library preparation steps were identical as our previous study (see ref [37]). We used 2x50 paired end mode in HiSeq 2500 and all fastq files were processed using bwa, samtools, and SNVer to generate a list of mutations with significant allele frequencies at the population levels. The average read coverage in these sequencing data was less than 50 so the detection limit of SNPs was around 10%. We also identified CNVs normalized by the median read depth at the entire genome using the pileup data from alignment BAM files using samtools and R. We noticed a

significant CNV region that includes *MEP2* locus in all three population sequencing data only at the 250 generation. We estimated the allele frequency of the CNV at the population level by randomly selecting 96 clones and conducting qPCR to measure the copy number of *MEP2* loci for all clones. We found that most clones possess one or two copy of *MEP2* but interestingly some possesses even up to 8 or 9 copies (data not shown). All clones bearing more than 2 copies of *MEP2* locus were counted as mutants for the allele frequency estimation of the CNV.

3.6.9. Targeted amplicon sequencing. Due to the low read coverage of the population level sequencing, we repeated sequencings only for 12 selected target loci: *GAT1*, *MEP2*, *LST4*, *VAC14*, *RIM15*, *YBR271W*, *RPL26B*, *YKL050C*, *WHI5*, *DAL81*, *TPK3* and *YKL162C*. They are either target genes that posses any high frequency SNPs from the population level sequencing at any time point or repetitively selected target genes in many LTEEs of nutrient limitations in yeast. We amplified 12 loci with +/- 1kb of up and down stream sequences using 30 cycles of PCRs, randomly fragmented using sonication (~ 250 bp long) and then prepared DNA libraries for them using the same protocol we used for the population sequencing. We ran the Illumina MiSeq 2x250 option for these sheared amplicon libraries such that two paired reads are overlapped after alignment to the reference genome. Only the overlapped region was selected for the allele frequency estimation and thereby we were able to dramatically reduce the false positive SNP calls rate. For each SNP frequency estimation, the average read coverage was ~

50,000. Using SNVer, we collected minor frequency mutations down to ~ 1% that is highly strict and so significant as shown in the **Figure 3.4B**.

3.6.10. dN and dS test. We obtained 42 different *GAT1* sequences from natural yeast isolates that were sequenced in a previous study (see ref [127]). dN and dS represent the proportion of observed Non-synonymous substitutions among all potential Non-synonymous substitutions and the proportion of observed Synonymous substitutions among all potential Synonymous substitutions, respectively. Actual values for each amino acid position were calculated from http://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html.

CHAPTER 4. Estimation of the effects of PCR duplicates in next-generation sequencing data analysis using a sequencing adapter design for unique molecule identification

*This chapter is based on the research paper "**Estimation of the effects of PCR duplicates in next-generation sequencing data analysis using a sequencing adapter design for unique molecule identification**" by Jungeui Hong and David Gresham (in preparation as of December, 2014)*

4.1. ABSTRACT

One of the technical issues arising from preparing next-generation sequencing libraries is how PCR duplicates that are generated during the library amplification step should be handled and processed for downstream analysis of detecting rare variants. We present a new cost-effective sequencing adapter design that enables both removing true positive PCR duplicates and multiplexing multiple sequencing libraries for the Illumina HiSeq and MiSeq sequencing platforms. Conventional bioinformatics approaches remove PCR duplicates by choosing from multiple reads that align to precisely the same genomic coordinates. However, this approach cannot discriminate true positive PCR duplicates from false negative ones thereby resulting in decrease in read coverage and possible bias in the final copy number or frequency estimation. We introduce a simple custom sequencing adapter design where the sample multiplexing index is moved to the end of adapters directly ligated to the insert DNA and a random barcode is located at the site that usually

contains the multiplexing index. Using this new design, we performed multiple DNA-seq, Amplicon-seq and RNA-seq assays. We showed that removal of PCR duplicates by coordinate alignment alone results in dramatic data loss with impacts on estimation of allele frequency or gene expression profiling. We find that majority of duplicated molecules originate from random fragmentation especially for smaller size targeted sequencing methods such as Amplicon-seq or RNA-seq. We illustrate the cost-effectiveness and power of this new adapter design for minimizing false positive PCR duplicates calls while maintaining the possibility of library multiplexing without any additional cost for preparing Illumina sequencing libraries.

4.2. INTRODUCTION

The recent advent of next-generation sequencing technologies enables rapid and cost-effective identification of rare alleles from pooled population samples or a panel of multiple isogenic cell lines and expression profiling of entire transcriptomes. One technical issue in sequencing library preparation protocols that use PCR amplifications such as Illumina TruSeq is how to handle and minimize PCR duplicates [57,139,140]. PCR bias in library amplification occurs mainly due to unbalanced GC composition in the genome [57,139,141,142]. Actual PCR duplicate rates are typically orders of magnitude higher than expected and increase as much as 50% or higher depending on the number of PCR cycles used.

PCR-free library preparation protocols provide a straightforward alternative but are limited in their usage due to the high cost of reagents and kits and the requirement of greater amount of starting materials. Optimizing PCR setting and buffer systems can be a trivial solution to minimize PCR bias [57,140] but requires extensive manual calibrations depending on experimental settings and sample conditions. Due to these technical limitations, in general, PCR duplicates are removed after completing sequencing using bioinformatic tools such as samtools [112] and Picard (http://picard.sourceforge.net) that detect duplicates based on the coordinate information after aligning reads to a known reference genome and then simply discard them.

In PCR based sequencing method, it is also critical to understand how PCR duplicates affect the final quality of sequencing data analysis. However, systematic studies on this issue are lacking. PCR duplicates represent redundant information for certain DNA sequences, inflate perceived read depth and therefore require careful handling for proper downstream data analysis. One recent study based on the targeted sequencing technology showed that PCR bias should be minimized when detecting minor frequency alleles in heterogeneous populations such as tumor tissues [143]. In that study, the presence of duplicates can introduce more variability in estimating population heterogeneity and thereby cause false interpretation in the final call of adaptive alleles in cancer evolution.

The Illumina TruSeq sequencing adapter comprises of two unique single stranded oligonucleotides (P5 and P7) that generally possess one sample index for

multiplexing (**Figure 4.1**). Multiplexing samples is generally preferred in order to minimize the cost of sequencing multiple samples. This can be easily achieved by introducing a sample index or barcode into one of the sequencing adapter oligos (P7) used in the TruSeq library preparation protocol. The Illumina Genome Analyzer sequencing run typically includes an additional reading phase for the short sample index sequence as well as for the original target DNA.

Some variations are possible in the adapter design depending on different applications [144]. Dual sample indexing can be adopted to minimize false positive multiplexing call [145]. In this design, an additional sample index is added to the opposite (P5) adapter oligo in order to increase accuracies for detecting rare variants from pooled samples where cross sample contamination should be critically avoided. One practical disadvantage of this design is that this requires an additional sequencing phase for the additional index. Also, this sequencing should be run only when all lanes contain dual-indexed libraries. Instead of the sample index that is a known sequence, one can introduce a random barcode into the adapter oligo for removal of PCR duplicates [146]. The random barcode can be combined with the coordinate information of each read aligned to the reference genome, and used for discriminating real PCR duplicates from false positive ones.

We introduce a novel Illumina sequencing adapter design – "unique molecule identification (UMI)" – enabling both removal of PCR duplicates and library multiplexing. We combined both the sample index for de-multiplexing and the random barcode for removal of PCR duplicates in one sequencing adapter (**Figure**

4.1). Starting with the TruSeq adapter oligos, we moved the multiplexing index to the 3' end position of adapters that are directly ligated to the target DNA and introduced a random barcode at the sample index site. These customized adapters are far cheaper than commercial ones and can be directly used for the same TruSeq library preparation protocol without any alteration in the general sequencing pipeline and the number of sequencing reads.

Figure 4.1. Scheme of new adapter design: comparison between commercial TruSeq and our own adapter.

Using this new adapter design, we tested the effect of PCR duplicates on the results of three different sequencing assays: whole genome DNA sequencing, targeted loci amplicon sequencing, and RNA-seq. We present evidence that PCR duplicates rates are highly proportional to the number of PCR cycles and overestimated when

using only the coordinate based methods such as samtools [112] and Picard. We show that PCR duplicates result in miscalculation of frequencies of minor alleles (less than 10%) in heterogeneous populations and differentially expressed mRNA levels. For accurate sensitive detection of polymorphisms or differential gene expression, removal of true PCR duplicates is necessary.

4.3. RESULTS

4.3.1. Estimating 'true' PCR duplicate rates

We estimated the 'true' PCR duplicate rate in each sequencing library by cross-checking both the coordinate information of aligned paired reads and their unique 6mer random barcode (**Figure 4.3**). PCR duplicate rate based on only the coordinate information ranges from 20 to 40% when less than 10 PCR cycles were used, and were as high as 90% when 15 PCR cycles were used only for some RNA-seq libraries with very low amount of starting materials (less than ~10ng in total). Using unique molecule identification by random barcodes altogether, the rate decreased to less than 10 % for when less than 10 PCR cycles were used. This result suggests that the number of PCR cycle should be less than 10 and majority of PCR duplicates determined only by the coordinate information are false positives.

Interestingly, we found that the PCR duplicate rates in whole genome DNA sequencing data were very low - less than 5% - regardless of which detection methods were used (see the rectangular data points in the **Figure 4.2**). This is possibly because whole genome sequencing data generally has lower read coverage

(< 50X) and is thereby less likely to have redundant duplicates. However, the difference in the duplicate rates between two methods becomes more dramatic in the targeted amplicon sequencing or the RNA-seq data (see triangle or circle data points in the **Figure 4.2**). We suggest that sequencing libraries targeting narrower genomic regions tend to have more random fragmentation duplicates that do not originate from the PCR amplification step.

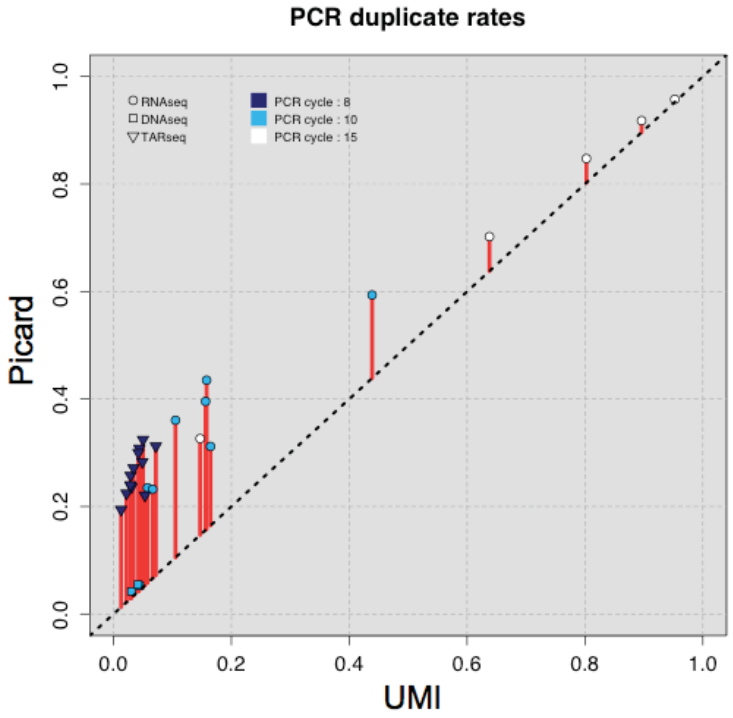

Figure 4.2. Comparison of PCR duplicates rates generated by the new adapter design and the conventional bioinformatics approaches.

4.3.2. Effects of PCR duplicates on detecting SNPs from heterogeneous populations

We conducted targeted amplicon sequencing and whole genome DNA sequencing to detect allele frequencies of SNPs from multiple different heterogeneous population samples. Using our new adapter design, we studied effects of PCR bias on detecting SNPs in population samples. We compared differences in the allele frequencies for SNPs identified by SNVer after removing PCR duplicates using UMI vs Picard tools (**Figure 4.3**). The AF difference in two methods increases as read depth decreases, implying that AF estimation using only Picard tools is misleading when analyzing low read depth sequencing data. The maximum AF difference is ~ 3%, which is significant difference in minor frequency allele estimation. Thus, our method enables more sensitive detection of minor alleles in heterogeneous populations.

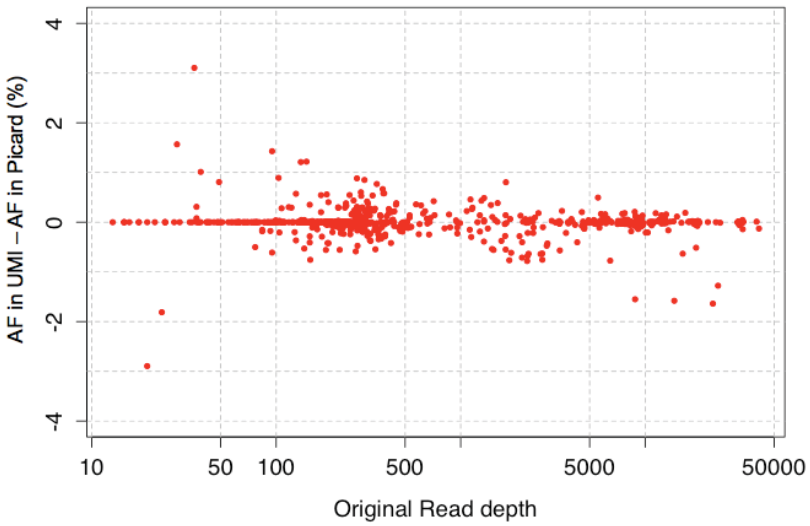

Figure 4.3. Differences in allele frequencies of SNPs using UMI or Picard tool vs Read depth.

4.3.3. Effects of PCR duplicates in RNA-seq data analysis

We analyzed 12 RNA-seq libraries that include one control and three different tested samples with 3 biological replicates. We compared differences in read counts for each gene using UMI and Picard tool (**Figure 4.4**). Interestingly, there was a clear gene size effect on the differences in the reads count values. Smaller size genes tend to have bigger loss of read counts when using only Picard, implying that most of duplicates detected and removed by Picard are not true PCR duplicates but just randomly generated duplicates with the same coordinate that are nonetheless unique.

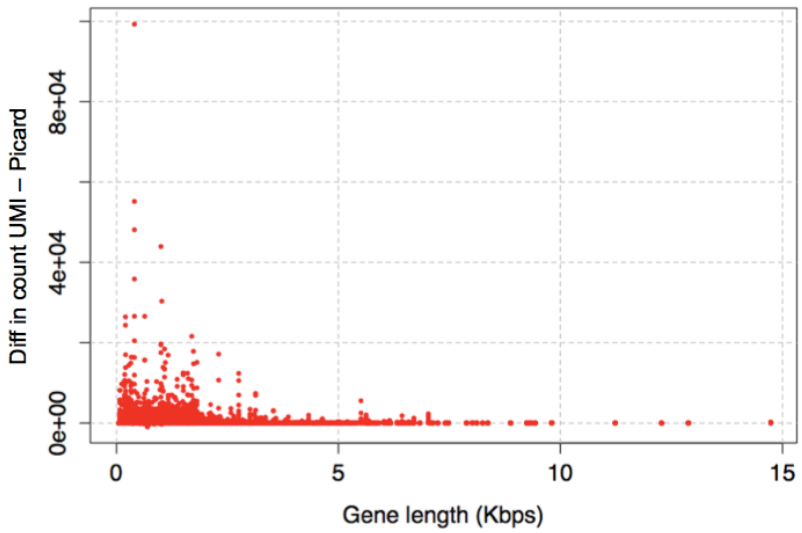

Figure 4.4. Differences in count values of each gene for RNAseq data between UMI and Picard.

4.4. DISCUSSION

We suggest that the majority of PCR duplicates identified based only on coordinate information are not real PCR duplicates but randomly generated unique duplicates; therefore, there will be massive loss of data when using the conventional coordinate based detection methods such as samtools or Picard. This should be considered when preparing amplicon sequencing or RNA-seq that targets much narrower genomic regions and therefore tend to have higher probability of generating unique molecules that appear to be duplicated fragments. There is an anti-correlation between the size of the sequencing region and the rate of PCR duplicates.

We argue that PCR duplicates should be removed using a combination of coordinate information and a random barcode especially in highly sensitive applications such as detection of minor frequency alleles (< 10%) from whole genome sequencing of heterogeneous population samples or targeted amplicon sequencing and RNA-seq data that are targeting small genomic regions.

This study illustrates that our new adapter design can easily distinguish PCR duplicates from random fragment duplicates. This design requires only one single index reading for the random barcode; therefore, there is no need for additional sequencing reagents and primers unlike dual indexing adapter design [145]. This is also highly cost effective since it is possible to make a ~ 500 ul of 20 μM adapter stock only for ~ $150 that can be used for constructing hundreds or thousands of libraries.

One technical disadvantage is that the first seven nucleotides in each read should be trimmed, lowering the final sequencing yield. This, however, can be easily compensated by the fact that massive amount of non-PCR duplicates can be saved using this design. Another concern is the low diversity issue for the first 7 nucleotides, which may impact sequencing cluster identification in the Illumina sequencers. We suggest an adapter design such that diversity of sample indices can be maximized when multiplexed as shown in the **Table 4.1**. In order to further increase sequence diversity at the 7^{th} 'T' nucleotide, one can vary the length of sample indices for multiplexing. Adding 5% PhiX control in the pooled library sample is also recommended.

4.5. MATERIALS AND METHODS

4.5.1. New sequencing adapter preparation Two modified adapter oligos (P5 and P7) were prepared (www.idtdna.com) so that the P5 oligo possesses a phosphorothioate bond between the 3' end T and its neighbor base and the 6-mer sample index plus one additional 'T', and P7 oligo possesses a 5' phosphate group, the 6-mer random barcode for tagging PCR duplicates and the complementary nucleotides against the sample index plus 'T' sequence in the P5 oligo (**Table 4.1**). The additional 'T' next to the sample index is required for proper priming site of the sequencing read primers (**Figure 4.1**). Two partially complementary oligos were annealed to form the final Y-shaped adapter as following: (1) Each individual oligonucleotide was re-suspended at the same molar concentration (20 μM) in the

annealing buffer (10 mM Tris, pH 7.5–8.0, 50 mM NaCl, 1 mM EDTA). (2) Equal volumes of both complementary oligos were mixed, placed in a standard heatblock at 95°C for 5 minutes, and then cooled to room temperature on the workbench for around 1 hour. (3) The annealed adapters were checked on a non-denaturing 5-6% PAGE gel. Around 90 % of bands should be found at around 300-400 bp due to the intrinsic property of the Y-shaped partial dsDNA. (4) The annealed adapters were kept at -20°C for long-term storage.

Adapter ID	Sample index	Adapter ID	Sample index
DGseq_1	**CGATGT**	**DGseq_13**	**GACTTA**
DGseq_2	**TGACCA**	**DGseq_14**	**AGTCTA**
DGseq_3	**ACAGTG**	**DGseq_15**	**TATCGA**
DGseq_4	**GATCAG**	**DGseq_16**	**TCTGAT**
DGseq_5	**CTCAGA**	**DGseq_17**	**GAACGT**
DGseq_6	**TAGCTT**	**DGseq_18**	**TGTACT**
DGseq_7	**GTGGCC**	**DGseq_19**	**TCGAAA**
DGseq_8	**ACTTGA**	**DGseq_20**	**ACAAGT**
DGseq_9	**GCCAAT**	**DGseq_21**	**TGAAAC**
DGseq_10	**CAGATC**	**DGseq_22**	**TCACAG**
DGseq_11	**AGTTCC**	**DGseq_23**	**TTACGC**
DGseq_12	**TTCGAG**	**DGseq_24**	**AAATGC**

Table 4.1. Sample indices used in new adapter design. Every pair of indices has a minimum of 3 hamming distance and guarantee based balance for the first 6 cycles if combined as the order in the table.

4.5.2. Library sequencing protocol

We tested the new sequencing adapters in whole genome population DNA-seq (3 libraries), targeted amplicon sequencing (12 libraries), and RNA-seq (12 libraries) for multiple different applications with *Saccharomyces cerevisiae*. While most library preparation steps were identical as the typical TruSeq protocol, some

variations were introduced: (1) all cleanups and DNA insert size selections were done using AMPure beads. (2) For, targeted loci amplicon sequencing, each amplicon was fragmentized before adapter ligation to further randomize DNA inserts. (3) For RNA-seq where the amount of starting materials was very little and limited, only 0.5 μM of the adapter was used for ligation. Otherwise, 20 μM was used and confirmed as a good standard concentration to make enough ligated molecules and minimize the unnecessary adapter dimer formation. (4) The number of PCR cycles varied from 8 to 15 depending on the amount of starting materials. (5) The necessary final concentration of libraries loaded onto a flow-cell was empirically determined to be higher than the standard requirement of 2nM. (6) 5 – 25 % of PhiX control was added in each library in order to minimize any possible negative effect of the low diversity issue in the first 7 sequencing cycles detecting 6-mer sample index plus one 'T' overhang in the 7^{th} position. Since multiple different sample indices were multiplexed in each lane, only the 7^{th} position was identical in all sequence reads. We empirically determined 5% of PhiX is sufficient to avoid the low diversity issue at the 7^{th} position. Multiplexed libraries were sequenced using either 2x50 bp paired end mode for DNA-seq and RNA-seq in the Illumina HiSeq 2500 or MiSeq 2x250 bp paired end for targeted amplicon-seq.

4.5.3. Data processing and analysis

De-multiplexing was done using a custom perl and Unix scripts. Only one mismatch in the sample index was allowed. The first 7 nucleotides including the

sample index and 'T' overhang in every read were trimmed for downstream analysis. For reads alignment, BWA -mem [111] option was used to the *Saccharomyces cerevisiae* S288C reference genome, obtained from the SGD database on Feb 03, 2011. PCR duplicate rates were calculated based on the SAM format alignment file using a custom perl script: first, all alignments possessing the same coordinate information were selected, and then only alignments with the unique 6mer barcode (perfect match) were determined as non-PCR duplicates (**Figure 4.5**). All poorly (mapping quality less than 10) and mis-aligned paired reads were removed in this analysis. Picard tools (http://picard.sourceforge.net) were also used for removal of PCR duplicates without considering the random barcode information. For SNP detection in population samples based on the DNA-seq or the targeted loci amplicon sequencing, SNVer [114] was used a minimum detection limit was set as ~ 1%). EdgeR [138] was used to determine differently expressed genes between control and treated samples from the RNA-seq data. All additional downstream data analysis was conducted using R.

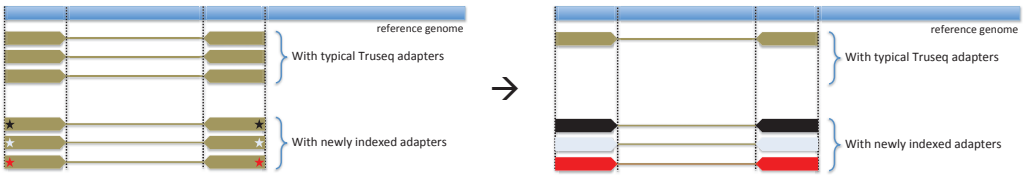

Figure 4.5. Illustration about how to remove PCR duplicates using our new adapter. Different colored stars represent unique barcode tags implemented in the new adapter design.

CHAPTER 5. CONCLUSION

5.1. SUMMAARY AND CONCLUSION

This dissertation aims to understand the molecular basis of adaptive alleles and their dynamics during the course of nutrient limited adaptations in *S. cerevisiae*. Experimental evolution using chemostats and asexual haploid yeast were used to achieve this goal because of the precise control of genetic and non-genetic factors that determine evolutionary outcomes and dynamics. Chemostat culturing maintains a large population and a constant nutrient – nitrogen in this study – 'poor' conditions over the entire experimental evolution. Thus, selection is strong and the mutation supply rate is very high, resulting in a very rapid adaptive evolution that is visible within a reasonable timeframe in the laboratory. Next-generation sequencing is a great addition to this type of study since it allows genome wide discovery of the full spectrum of adaptive alleles acquired during the course of experimental evolutions. In this thesis, I introduced several technical improvements to the area of experimental evolutions and successfully applied them to achieve my research goals as followings.

In Chapter 2, I examined the specific and convergent adaptive solutions at multiple molecular levels among different nitrogen limited selections in chemostats. The main questions I addressed were (1) what are the major selection targets that are common or specific to different nitrogen-limited adaptations? (2) At which level(s)

does selection act? (3) How are evolutionary dynamics and fitness landscapes constrained and shaped? It is clear that selection for improved nutrient transport capabilities underlies adaptive evolution in constant nutrient limitations. The most dominant selective alleles in chemostats are CNVs containing nutrient transporter genes specifically corresponding to the conditions in which they are selected. Amplification of such transporter genes results in increased production of mRNA molecules and consequently the proteins that they encode. Despite the possible increased metabolic burden, additional copies of nutrient transporters in a cell are beneficial in terms of nutrient transport capabilities under extreme nutrient poor environments. Interestingly, I found no adaptive alterations in enzymatic functions in metabolic genes although they are central to optimized growth in nutrient poor conditions. In addition, there is little evidence that translational regulation of nutrient transporters is targeted by selection. More general (common) solutions in chemostats regardless of types of nutrient limitations are loss of function mutations in regulatory genes responsible for initiation of a quiescent G_0 phase entry in response to nutrient starvations. Loss of such regulatory function may be advantageous specifically in chemostat selections since the environment requires continuous cell division. In addition, loss of function mutations in regulation of phosphatidylinositol-3,5-bisphosphate production with roles in protein trafficking and vacuole biogenesis were a highly repeatable solution across different nitrogen-limited chemostat conditions. The molecular mechanisms or relevance of these mutations to nitrogen utilization remain to be investigated. I also found one

interesting example of repeated selection of SNPs in functionally related loci comprising a transcriptional activator (GAT1), its direct target, an ammonium transporter (MEP2) and a post-translational regulator (LST4) from an ammonium-limited adaptation. Using yeast genetics, fitness assays and sequencing techniques, I verified that complex epistasis underlies such adaptive loci and also constrains the evolutionary dynamics. From these analyses in Chapter 1, I concluded that targets of selection converge at multiple levels from one single nucleotide to genes and functional modules and that the evolutionary dynamics are constrained by epistatic interactions and clonal interference.

In Chapter 3, I continued to examine the molecular details of the evolution of GAT1 found in the ammonium-limited environment from the study of Chapter 1 and understood the first completely characterized example of a transcription factor evolution. My main questions for this section were: (1) How do protein coding changes in a transcription factor affect its function under strong selective pressure? (2) What is the molecular mechanism underlying the evolution of gene expression? (3) Is the evolution of gene expression convergent, stochastic or historically contingent? My major finding was the repetitive selection of missense mutations in the zinc finger DNA binding domain of GAT1, a GATA transcriptional activator for multiple NCR genes including *MEP2*, a high affinity ammonium transporter-encoding gene, from ammonium-limited adaptations. Interestingly, all such mutations in the DNA binding domain appear to be 'partial' loss of function in

transcriptional regulation. I confirmed that missense mutations in GAT1, result in partial impairment in its DNA binding activity and thereby selective activation of its target genes in the NCR regulon consisting of a complex feed forward loop with multiple GATA transcription factors. Using next-generation sequencing, fitness assay and other molecular analyses, I determined these mutations are under strong positive selection under conditions of ammonium-limitation and also show antagonistic pleiotropy. Their functional effects are exerted by rewiring feed-forward loops in the transcriptional network. From replayed experimental evolutions in ammonium-limited environments, I also found that the outcomes of adaptive evolutions are highly convergent at the level of nucleotides, genes and functional loci at the early stage of evolution but seem to be more stochastic as additional mutations accumulate.

In Chapter 4, I offered one way of improving sensitivity in estimating allele frequencies and gene expression levels from the Illumina platform-based sequencing techniques. The main question was how PCR duplicates that are generated during the Illumina sequencing library preparation step affect the final quality of allele frequency estimation in DNA-seq or expression profiling in RNA-seq. This is a very important, but not yet resolved, question in the area of next-generation sequencings. I estimated how much portion of sequencing reads is really ‘true positive’ and ‘unique’ and whether available computational tools in the scientific community are reliable for the proper removal of PCR duplicates. My

new sequencing adapter design is very straightforward and cost-effective but highly sensitive for filtering out the true PCR duplicates. I concluded that the new adapter design enables increased sensitivity of detection of minor frequency alleles in a population sequencing data or subtle changes of gene expression in RNA-seq data.

5.2. Future directions and application

5.2.1. Many adaptive alleles are still missing

It will be important to continue to identify the pathways that are targets of selection in other types of nutrient limitations, their dynamics in real time and the functional basis of other types of adaptive alleles that remained unexamined in this dissertation. I mainly analyzed evolved population samples from the ammonium limitation where SNPs were the major source of adaptive alleles. However, massive structural variations such as large indels, aneuploidy and diploidization are pervasive in other types of nitrogen limitations such as allantoin and urea. Their molecular basis and roles in adaptive evolution should be further investigated in future studies. Additional population and clonal sequencing for early time point samples from those LTEE studies would provide a more general and complete view of evolutionary dynamics in nutrient limitations.

5.2.2. New insight about the evolutionary convergence and contingency

This dissertation serves as a great starting point to answer the long-standing evolutionary biology question about whether adaptive evolution is historically

contingent or convergent. There is controversy in the results of different LTEEs between Lenski's group suggesting that evolution is historically contingent [32] and others claiming that evolution is convergent and parallel [38,39,136]. Such discrepancy may originate from different experimental settings or different definitions about molecular 'convergence' or 'contingency'. I determined that the dynamics and outcomes of adaptive evolutions in chemostats are highly reproducible at the level of nucleotides or genes in early stage of adaptations by replaying experimental evolutions in parallel. However, there is a tendency for the outcomes at later generations to be more stochastic at the same level of nucleotides or genes. From these results, I suggest that convergence and contingency might have temporally different contributions to adaptive evolution, as shown in the evolutionary dynamics of *GAT1*. It will be interesting to investigate whether such temporally differential effects really exist even in other types of nutrient limitations. Longer culturing of yeast as the Lenski group's did with *E. coli* will be also useful to get a clearer picture regarding this question.

5.2.3. The role of epistasis requires further investigation

A possible mutual exclusivity seen in the dynamics between alterations in a transcriptional regulator, GAT1, and its target gene encoding a nutrient transporter, MEP2, is another interesting point to be further explored. Another similar example of this is shown in a glucose-limited adaptation study [34,56], where loss of function in *MTH1* and amplification of *HXT6/7* were shown to be mutually

exclusive due to reciprocal sign epistasis. In both of these studies, alterations in the regulators (GAT1 and MTH1) always precede the amplification of the nutrient transporters (MEP2 and HXT6/7) during the course of adaptive evolution. The early onset and rapid disappearance of adaptive alleles of such regulators suggest that they are a highly recurrent but transient solution in nutrient limited adaptation. One mechanism explaining such dynamics might be reciprocal sign epistasis between two different groups of alleles, followed by a reversion mutation on the alleles of the regulators. Selection of reversion mutations is less likely in general but has been shown to be a potentially important mechanism from a case study of antibiotic resistance evolution [135]. Moreover, it will be very interesting to investigate why alterations in regulators are identified early but amplifications of transporters are not detected until later in the evolution experiments.

5.2.4. Medical applications of experimental evolution

From a broader perspective, experimental approaches used in this study are applicable to understanding pathogenic strategies adopted by viruses, microbes and even cancer cells in the area of human health [44-46]. For example, cancer is also an evolutionary process of clonal somatic populations in human. Clonal evolution of tumor cells has been investigated for several decades in theory [47,147] mainly due to the lack of fundamental understanding of adaptive evolution and experimental means of high-throughput genotyping and/or genetic techniques. LTEE studies with microbes can provide novel insights for studying cancer by

circumventing such limitations. My dissertation showed that evolving yeast populations share many of the properties of tumor evolution including mutational heterogeneity and the presence of multiple competing lineages. Notably, repeated selection of missense mutations in the DNA binding domain of *GAT1* in this study is a reminiscent of mutational hotspots in the DNA binding domain of *TP53* frequently found in human cancer evolutions. The comprehensive description of the dynamics and molecular basis of adaptive evolution in this study may hold benefits for understanding tumorigenesis or evolution of drug resistance in cancer cells. The recent advent of single cell sequencing and genetic alteration tools such as CRISPR/CAS9 [148] will also be applicable to identify major adaptive alleles in cancer cell lines and their functional studies. The combination of population level and clonal sequencings for identifying the order of mutations in evolved mutants in this study would provide a means of distinguishing 'driver' mutations from 'passenger' mutations between primary and metastatic tumor tissues. Moreover, LTEE studies with microbes in chemostats can be a basis for designing experimental evolution of cancer cells in the lab in order to study their real time dynamics, which has not yet been tried in cancer research.

5.2.5. QTL studies

Finally, this study provides empirical evidence of the impact of epistasis on phenotypic variations mediated by quantitative trait loci (QTLs). It is known that QTLs are a major determinant for most heritable variations in natural populations

[149] but epistasis among such traits is rarely studied mainly due to the lack of systematic tools to gain sufficient statistical power. With the advent of recent high-throughput genomic tools used in this study, it is now possible to relate epistasis with phenotypic variations mediated by QTLs in lab evolved and natural populations. My study tested the role of epistasis in fitness from laboratory evolution under carefully defined growth limiting conditions using chemostats. The study of *GAT1* evolution and its possible epistatic interaction with amplification allele of *MEP2* might also be a good example of genetic basis underlying variation in gene expression, i.e., expression QTLs. Many phenotyping and genotyping assays available are still low-throughput or requires technical improvement for better resolution. However, the use of high-throughput fitness assays such as synthetic genetic array (SGA) technology [150] or microscopy based fitness assay [151,152] are likely to speed the study of these problems using experimental evolution studies in microbes.

5.3. CONCLUDING REMARK

This dissertation provided many fundamental insights about the functional effects and dynamics of adaptive variations that are selected under constant selective pressures. Experimental evolution in chemostats are powerful models for isolating the full spectrum of adaptive alleles and monitoring their mode of action in real-time in the lab. The recent advent of high-throughput next-generation sequencing and genome editing tools now enables more comprehensive molecular level studies

of adaptive alleles. High-throughput phenotyping tools are becoming more available for their functional studies. One future approach will be to characterize natural selection under fluctuating environmental conditions to find a connection to the evolution of natural populations in the wild. My study serves as a basis for understanding pathogenic strategies of virus or bacteria in animal hosts or the evolution of drug resistance in cancer cells.

REFERENCES

1. Darwin C (1859) On The Origin Of Species Or The Preservation Of Favoured Races In The Struggle For Life. John Murray. 1 pp.

2. CIBA Foundation Symposium (2008) Antibiotic Resistance. John Wiley & Sons. 1 pp.

3. Grant PR (1999) Ecology and Evolution of Darwin's Finches. Princeton University Press. 1 pp.

4. Muller HJ (1932) Some genetic aspects of sex. American Naturalist: 118–138.

5. Gillespie JH (1984) Molecular Evolution Over the Mutational Landscape. Evolution 38: 1116. doi:10.2307/2408444.

6. Desai MM, Fisher DS, Murray AW (2007) The speed of evolution and maintenance of variation in asexual populations. Curr Biol 17: 385–394. doi:10.1016/j.cub.2007.01.072.

7. Dean AM, Thornton JW (2007) Mechanistic approaches to the study of evolution: the functional synthesis. Nature Reviews Genetics 8: 675–688. doi:10.1038/nrg2160.

8. Olson-Manning CF, Wagner MR, Mitchell-Olds T (2012) Adaptive evolution: evaluating empirical support for theoretical predictions. Nature Reviews Genetics 13: 867–877. doi:10.1038/nrg3322.

9. Griffiths AJF (2000) An introduction to genetic analysis. W.H. Freeman. 1 pp.

10. Carter AJR, Hermisson J, Hansen TF (2005) The role of epistatic gene interactions in the response to selection and the evolution of evolvability. Theor Popul Biol 68: 179–196. doi:10.1016/j.tpb.2005.05.002.

11. Weinreich DM, Watson RA, Chao L (2005) Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. Evolution 59: 1165–1174.

12. Phillips PC (2008) Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics 9: 855–867. doi:10.1038/nrg2452.

13. Futuyma DJ (1998) Evolutionary Biology. Sinauer Associates Incorporated. 1 pp.

14. Gillespie JH (2001) Is the population size of a species relevant to its evolution? Evolution 55: 2161–2169.

15. Masel J (2011) Genetic drift. Curr Biol 21: R837–R838. doi:10.1016/j.cub.2011.08.007.

16. Otto SP, Lenormand T (2002) Resolving the paradox of sex and recombination. Nature Reviews Genetics 3: 252–261. doi:10.1038/nrg761.

17. Coop G, Przeworski M (2007) An evolutionary view of human recombination. Nature Reviews Genetics 8: 23–34. doi:10.1038/nrg1947.

18. Webster MT, Hurst LD (2012) Direct and indirect consequences of meiotic recombination: implications for genome evolution. Trends Genet 28: 101–109. doi:10.1016/j.tig.2011.11.002.

19. Elena SF, Lenski RE (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nature Reviews Genetics 4: 457–469. doi:10.1038/nrg1088.

20. Zeyl C (2006) Experimental evolution with yeast. FEMS Yeast Research 6: 685–691. doi:10.1111/j.1567-1364.2006.00061.x.

21. Bennett AF, Hughes BS (2009) Microbial experimental evolution. Am J Physiol Regul Integr Comp Physiol 297: R17–R25. doi:10.1152/ajpregu.90562.2008.

22. Buckling A, Craig Maclean R, Brockhurst MA, Colegrave N (2009) The Beagle in a bottle. Nature 457: 824–829. doi:10.1038/nature07892.

23. Conrad TM, Lewis NE, Palsson BØ (2011) Microbial laboratory evolution in the era of genome-scale science. Molecular Systems Biology 7: 509. doi:10.1038/msb.2011.42.

24. Dettman JR, Rodrigue N, Melnyk AH, Wong A, Bailey SF, et al. (2012) Evolutionary insight from whole-genome sequencing of experimentally evolved microbes. Mol Ecol 21: 2058–2077. doi:10.1111/j.1365-294X.2012.05484.x.

25. Desai MM (2013) Statistical questions in experimental evolution. Journal of Statistical Mechanics: Theory and ….

26. Novick A, Horiuchi T (1961) Hyper-production of beta-galactosidase by Escherichia coli bacteria. Cold Spring Harb Symp Quant Biol 26: 239–245.

27. Gibson TC, Scheppe ML, Cox EC (1970) Fitness of an Escherichia coli

mutator gene. Science 169: 686–688.

28. Cox EC, Gibson TC (1974) Selection for high mutation rates in chemostats. Genetics 77: 169–184.

29. Lenski RE, Rose MR, Simpson SC, Tadler SC (1991) Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. American Naturalist: 1315–1341.

30. Lenski RE, Travisano M (1994) Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. Proceedings of the National Academy of Sciences 91: 6808–6814.

31. Gerrish P, Lenski RE (1997) Evolution of high mutation rates in experimental populations of E. coli. Nature.

32. Blount ZD, Borland CZ, Lenski RE (2008) Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. Proceedings of the National Academy of Sciences 105: 7899–7906. doi:10.1073/pnas.0803151105.

33. Brown CJ, Todd KM, Rosenzweig RF (1998) Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. Mol Biol Evol 15: 931–942.

34. Kao KC, Sherlock G (2008) Molecular characterization of clonal interference during adaptive evolution in asexual populations of Saccharomyces cerevisiae. Nature genetics 40: 1499–1504. doi:10.1038/ng.280.

35. Gresham D, Desai MM, Tucker CM, Jenq HT, Pai DA, et al. (2008) The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. PLoS Genet 4: e1000303. doi:10.1371/journal.pgen.1000303.

36. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, et al. (2009) Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature 461: 1243–1247. doi:doi:10.1038/nature08480.

37. Hong J, Gresham D (2014) Molecular specificity, convergence and constraint shape adaptive evolution in nutrient-poor environments. PLoS Genet 10: e1004041. doi:10.1371/journal.pgen.1004041.

38. Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, et al. (2013) Pervasive genetic hitchhiking and clonal interference in forty

evolving yeast populations. Nature 500: 571–574. doi:10.1038/nature12344.

39. Kryazhimskiy S, Rice DP, Jerison ER, Desai MM (2014) Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science 344: 1519–1522. doi:10.1126/science.1250939.

40. Wiser MJ, Ribeck N, Lenski RE (2013) Long-Term Dynamics of Adaptation in Asexual Populations. Science. doi:10.1126/science.1243357.

41. Magasanik B, Kaiser CA (2002) Nitrogen regulation in Saccharomyces cerevisiae. Gene 290: 1–18. doi:10.1016/S0378-1119(02)00558-9.

42. Beltran G, Novo M, Rozès N, Mas A, Guillamón JM (2004) Nitrogen catabolite repression in Saccharomyces cerevisiae during wine fermentations. FEMS Yeast Research 4: 625–632.

43. Boczko EM, Cooper TG, Gedeon T, Mischaikow K, Murdock DG, et al. (n.d.) Structure theorems and the dynamics of nitrogen catabolite repression in yeast. pnasorg.

44. Bedhomme S, Lafforgue G, Elena SF (2012) Multihost Experimental Evolution of a Plant RNA Virus Reveals Local Adaptation and Host-Specific Mutations. Mol Biol Evol 29: 1481–1492. doi:10.1093/molbev/msr314.

45. Ensminger AW (2013) Experimental Evolution of Pathogenesis: “Patient” Research. PLoS Pathog 9: e1003340. doi:10.1371/journal.ppat.1003340.

46. Sprouffske K, Merlo LMF, Gerrish PJ, Maley CC, Sniegowski PD (2012) Cancer in Light of Experimental Evolution. Curr Biol 22: R762–R771. doi:10.1016/j.cub.2012.06.065.

47. Nowell PC (1976) The clonal evolution of tumor cell populations. Science 194: 23–28.

48. Monod J (1950) La technique de culture continue. Théorie et applications. Ann Inst Pasteur 79: 390–410.

49. Novick A, Szilard L (1950) Experiments with the Chemostat on spontaneous mutations of bacteria. Proceedings of the National Academy of Sciences 36: 708–719.

50. Novick A, Szilard L (1950) Description of the chemostat. Science 112: 715–716.

51. Monod J (1949) The growth of bacterial cultures. Annu Rev Microbiol 3: 371–394.

52. Ziv N, Brandt NJ, Gresham D (2013) The use of chemostats in microbial systems biology. J Vis Exp. doi:10.3791/50168.

53. Wang L, Spira B, Zhou Z, Feng L, Maharjan RP, et al. (2010) Divergence involving global regulatory gene mutations in an Escherichia coli population evolving under phosphate limitation. Genome Biol Evol 2: 478–487. doi:10.1093/gbe/evq035.

54. Maharjan RP, Ferenci T, Reeves PR, Li Y, Liu B, et al. (2012) The multiplicity of divergence mechanisms in a single evolving population. Genome Biology 13: R41. doi:10.1186/gb-2012-13-6-r41.

55. Gresham D, Ruderfer DM, Pratt SC, Schacherer J, Dunham MJ, et al. (2006) Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. Science 311: 1932–1936. doi:10.1126/science.1123726.

56. Kvitek DJ, Sherlock G (2011) Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. PLoS Genet 7: e1002056. doi:10.1371/journal.pgen.1002056.

57. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology 12: R18. doi:10.1186/gb-2011-12-2-r18.

58. Thomson JM, Gaucher EA, Burgan MF, De Kee DW, Li T, et al. (2005) Resurrecting ancestral alcohol dehydrogenases from yeast. Nature genetics 37: 630–635. doi:10.1038/ng1553.

59. Bridgham JT, Ortlund EA, Thornton JW (2009) An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature 461: 515–519. doi:10.1038/nature08249.

60. Lang GI, Botstein D, Desai MM (2011) Genetic variation and the fate of beneficial mutations in asexual populations. Genetics 188: 647–661. doi:10.1534/genetics.111.128942.

61. Gresham D, Desai MM, Tucker CM, Jenq HT, Pai DA, et al. (2008) The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. PLoS Genet 4: e1000303. doi:10.1371/journal.pgen.1000303.t005.

62. Gresham D, Usaite R, Germann SM, Lisby M, Botstein D, et al. (2010) Adaptation to diverse nitrogen-limited environments by deletion or extrachromosomal element formation of the GAP1 locus. Proceedings of the National Academy of Sciences 107: 18551–18556. doi:10.1073/pnas.1014023107.

63. Ferenci T (2007) Bacterial physiology, regulation and mutational adaptation in a chemostat environment. Advances in microbial physiology 53: 169–315. doi:10.1016/S0065-2911(07)53003-1.

64. Kvitek DJ, Sherlock G (2011) Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. PLoS Genet 7: e1002056. doi:10.1371/journal.pgen.1002056.

65. Wenger JW, Piotrowski J, Nagarajan S, Chiotti K, Sherlock G, et al. (2011) Hunger artists: yeast adapted to carbon limitation show trade-offs under carbon sufficiency. PLoS Genet 7: e1002202. doi:10.1371/journal.pgen.1002202.

66. Cooper TG (2002) Transmitting the signal of excess nitrogen in Saccharomyces cerevisiae from the Tor proteins to the GATA factors: connecting the dots. FEMS Microbiology Reviews 26: 223–238. doi:10.1111/j.1574-6976.2002.tb00612.x.

67. Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, et al. (2002) Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America 99: 16144–16149. doi:10.1073/pnas.242624799.

68. Zhong S, Khodursky A, Dykhuizen DE, Dean AM (2004) Evolutionary genomics of ecological specialization. Proceedings of the National Academy of Sciences of the United States of America 101: 11719. doi:10.1073/pnas.0404397101.

69. Zhang H, Zeidler AFB, Song W, Puccia CM, Malc E, et al. (2013) Gene copy-number variation in haploid and diploid strains of the yeast Saccharomyces cerevisiae. Genetics 193: 785–801. doi:10.1534/genetics.112.146522.

70. Dorsey M, Peterson C, Bray K, Paquin CE (1992) Spontaneous amplification of the ADH4 gene in Saccharomyces cerevisiae. Genetics 132: 943–950.

71. Rancati G, Pavelka N, Fleharty B, Noll A, Trimble R, et al. (2008)

Aneuploidy underlies rapid adaptive evolution of yeast cells deprived of a conserved cytokinesis motor. Cell 135: 879–893. doi:10.1016/j.cell.2008.09.039.

72. Gerstein AC, Otto SP (2011) Cryptic fitness advantage: diploids invade haploid populations despite lacking any apparent advantage as measured by standard fitness assays. PloS one 6: e26599. doi:10.1371/journal.pone.0026599.

73. Torres EM, Sokolsky T, Tucker CM, Chan LY, Boselli M, et al. (2007) Effects of aneuploidy on cellular physiology and cell division in haploid yeast. Science 317: 916–924. doi:10.1126/science.1142210.

74. Lynch M, Sung W, Morris K, Coffey N, Landry CR, 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: 9272–9277. doi:10.1073/pnas.0803466105.

75. Araya CL, Payen C, Dunham MJ, Fields S (2010) Whole-genome sequencing of a laboratory-evolved yeast strain. BMC Genomics 11: 88. doi:10.1186/1471-2164-11-88.

76. Notley-McRobb L, Seeto S, Ferenci T (2003) The influence of cellular physiology on the initiation of mutational pathways in Escherichia coli populations. Proc Biol Sci 270: 843–848. doi:10.1098/rspb.2002.2295.

77. Herron MD, Doebeli M (2013) Parallel Evolutionary Dynamics of Adaptive Diversification in Escherichia coli. PLoS Biol 11: e1001490. doi:10.1371/journal.pbio.1001490.

78. Drotschmann K (1999) Mutator phenotypes of yeast strains heterozygous for mutations in the MSH2 gene. Proceedings of the National Academy of Sciences 96: 2970–2975. doi:10.1073/pnas.96.6.2970.

79. Barrick JE, Lenski RE (2009) Genome-wide mutational diversity in an evolving population of Escherichia coli. Cold Spring Harb Symp Quant Biol 74: 119–129. doi:10.1101/sqb.2009.74.018.

80. Maharjan R, Seeto S, Notley-McRobb L, Ferenci T (2006) Clonal adaptive radiation in a constant environment. Science 313: 514–517. doi:10.1126/science.1129865.

81. Jin N, Chow CY, Liu L, Zolov SN, Bronson R, et al. (2008) VAC14 nucleates a protein complex essential for the acute interconversion of PI3P and PI(3,5)P(2) in yeast and mouse. The EMBO Journal 27: 3221–3234.

doi:10.1038/emboj.2008.248.

82. Notley-McRobb L, King T, Ferenci T (2002) rpoS mutations and loss of general stress resistance in Escherichia coli populations as a consequence of conflict between competing stress responses. J Bacteriol 184: 806–811. doi:10.1128/JB.184.3.806-811.2002.

83. Godard P, Urrestarazu A, Vissers S, Kontos K, Bontempi G, et al. (2007) Effect of 21 different nitrogen sources on global gene expression in the yeast Saccharomyces cerevisiae. Molecular and Cellular Biology 27: 3065–3086. doi:10.1128/MCB.01084-06.

84. Arnold K, Kiefer F, Kopp J, Battey JND, Podvinec M, et al. (2009) The Protein Model Portal. J Struct Funct Genomics 10: 1–8. doi:10.1007/s10969-008-9048-5.

85. Scherens B, Feller A, Vierendeels F, Messenguy F, Dubois E (2006) Identification of direct and indirect targets of the Gln3 and Gat1 activators by transcriptional profiling in response to nitrogen availability in the short and long term. FEMS Yeast Research 6: 777–791. doi:10.1111/j.1567-1364.2006.00060.x.

86. Roberg KJ, Bickel S, Rowley N, Kaiser CA (1997) Control of amino acid permease sorting in the late secretory pathway of Saccharomyces cerevisiae by SEC13, LST4, LST7 and LST8. Genetics 147: 1569–1584.

87. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, et al. (2010) The Genetic Landscape of a Cell. Science 327: 425–431. doi:10.1126/science.1180823.

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

89. Paquin CE, Adams J (1983) Relative fitness can decrease in evolving asexual populations of S. cerevisiae. Nature 306: 368–370.

90. Ferea TL, Botstein D, Brown PO, Rosenzweig RF (1999) Systematic changes in gene expression patterns following adaptive evolution in yeast. Proceedings of the National Academy of Sciences 96: 9721–9726. doi:10.1073/pnas.96.17.9721.

91. Kubitschek HE (1970) Introduction to research with continuous cultures. Prentice Hall. 1 pp.

92. Rudge SA, Anderson DM, Emr SD (2004) Vacuole size control: regulation of PtdIns(3,5)P2 levels by the vacuole-associated Vac14-Fig4 complex, a PtdIns(3,5)P2-specific phosphatase. Molecular Biology of the Cell 15: 24–36. doi:10.1091/mbc.E03-05-0297.

93. Dove SK, McEwen RK, Mayes A, Hughes DC, Beggs JD, et al. (2002) Vac14 controls PtdIns (3, 5) P2 synthesis and Fab1-dependent protein trafficking to the multivesicular body. Curr Biol 12: 885–893. doi:10.1016/S0960-9822(02)00891-6.

94. Weisman LS (2003) Yeast vacuole inheritance and dynamics. Annu Rev Genet 37: 435–460. doi:10.1146/annurev.genet.37.050203.103207.

95. Li SC, Kane PM (2009) The yeast lysosome-like vacuole: endpoint and crossroads. Biochim Biophys Acta 1793: 650–663. doi:10.1016/j.bbamcr.2008.08.003.

96. Cardenas ME, Cutler NS, Lorenz MC, Di Como CJ, Heitman J (1999) The TOR signaling cascade regulates gene expression in response to nutrients. Genes Dev 13: 3271–3279. doi:10.1101/gad.13.24.3271.

97. Tamanoi F (2011) Ras signaling in yeast. Genes Cancer 2: 210–215. doi:10.1177/1947601911407322.

98. Cameroni E, Hulo N, Roosen J, Winderickx J, De Virgilio C (2004) The novel yeast PAS kinase Rim15 orchestrates G0-associated antioxidant defense mechanisms. Cell Cycle 3: 460–466. doi:10.4161/cc.3.4.791.

99. Swinnen E, Wanke V, Roosen J, Smets B, Dubouloz F, et al. (2006) Rim15 and the crossroads of nutrient signalling pathways in Saccharomyces cerevisiae. Cell Div 1: 3. doi:10.1186/1747-1028-1-3.

100. Ferenci T (2003) What is driving the acquisition of mutS and rpoS polymorphisms in Escherichia coli? Trends Microbiol 11: 457–461. doi:10.1016/j.tim.2003.08.003.

101. Dunn B, Paulish T, Stanbery A, Piotrowski J, Koniges G, et al. (2013) Recurrent Rearrangement during Adaptive Evolution in an Interspecific Yeast Hybrid Suggests a Model for Rapid Introgression. PLoS Genet 9: e1003366. doi:10.1371/journal.pgen.1003366.

102. Rubio-Texeira M, Kaiser CA (2006) Amino acids regulate retrieval of the yeast general amino acid permease from the vacuolar targeting pathway. Molecular Biology of the Cell 17: 3031–3050. doi:10.1091/mbc.E05-07-0669.

103. Lee M-C, Marx CJ (2013) Synchronous waves of failed soft sweeps in the laboratory: remarkably rampant clonal interference of alleles at a single locus. Genetics 193: 943–952. doi:10.1534/genetics.112.148502.

104. Gerke J, Lorenz K, Cohen B (2009) Genetic interactions between transcription factors cause natural variation in yeast. Science 323: 498–501. doi:10.1126/science.1166426.

105. Steiner CC, Weber JN, Hoekstra HE (2007) Adaptive variation in beach mice produced by two interacting pigmentation genes. PLoS Biol 5: e219. doi:10.1371/journal.pbio.0050219.g001.

106. Hittinger CT, Gonçalves P, Sampaio JP, Dover J, Johnston M, et al. (2010) Remarkably ancient balanced polymorphisms in a multi-locus gene network. Nature 464: 54–58. doi:10.1038/nature08791.

107. Gietz RD, Schiestl RH (2007) High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nature protocols 2: 31–34. doi:10.1038/nprot.2007.13.

108. Brauer MJ, Huttenhower C, Airoldi EM, Rosenstein R, Matese JC, et al. (2008) Coordination of Growth Rate, Cell Cycle, Stress Response, and Metabolic Activity in Yeast. Molecular Biology of the Cell 19: 352–367. doi:10.1091/mbc.E07-08-0779.

109. Kahm M, Hasenbrink G, Lichtenberg-Fraté H, Ludwig J, Kschischo M (2010) grofit: fitting biological growth curves with R. Journal of Statistical Software 33: 1–21.

110. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23: 657–663. doi:10.1093/bioinformatics/btl646.

111. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589–595. doi:10.1093/bioinformatics/btp698.

112. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. doi:10.1093/bioinformatics/btp352.

113. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, et al. (2011) Dindel: accurate indel calls from short-read data. Genome Res 21: 961–973. doi:10.1101/gr.112326.110.

114. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H (2011) SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Research 39: e132. doi:10.1093/nar/gkr599.

115. Lõoke M, Kristjuhan K, Kristjuhan A (2011) Extraction of genomic DNA from yeasts for PCR-based applications. BioTechniques 50: 325–328. doi:10.2144/000113672.

116. Ge B, Gurd S, Gaudin T, Dore C, Lepage P, et al. (2005) Survey of allelic expression using EST mining. Genome Res 15: 1584–1591. doi:10.1101/gr.4023805.

117. Julius D, Blair L, Brake A, Sprague G, Thorner J (1983) Yeast alpha factor is processed from a larger precursor polypeptide: the essential role of a membrane-bound dipeptidyl aminopeptidase. Cell 32: 839–852.

118. Jacob F, Monod J (1961) On the Regulation of Gene Activity. Cold Spring Harb Symp Quant Biol 26: 193–211. doi:10.1101/SQB.1961.026.01.024.

119. Jacob F (1977) Evolution and Tinkering. Science 196: 1161–1166.

120. Stern DL, Orgogozo V (2008) The loci of evolution: how predictable is genetic evolution? Evolution 62: 2177. doi:10.1111/j.1558-5646.2008.00450.x.

121. Soskine M, Tawfik DS (2010) Mutational effects and the evolution of new protein functions. Nature Reviews Genetics 11: 572–582. doi:10.1038/nrg2808.

122. Hoekstra HE, Coyne JA (2007) The locus of evolution: evo devo and the genetics of adaptation. Evolution 61: 1016. doi:10.1111/j.1558-5646.2007.00105.x.

123. Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. Nature 430: 85–88. doi:10.1038/nature02698.

124. Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nature Reviews Genetics 8: 206–216. doi:10.1038/nrg2063.

125. Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. 134: 25–36. doi:10.1016/j.cell.2008.06.030.

126. Ostman B, Hintze A, Adami C (2011) Impact of epistasis and pleiotropy on evolutionary adaptation. Proceedings of the Royal Society B: Biological

Sciences. doi:10.1098/rspb.2011.0870.

127. Liti G, Carter DM, Moses AM, Warringer J, Parts L, et al. (2009) Population genomics of domestic and wild yeasts. Nature 458: 337–341. doi:doi:10.1038/nature07743.

128. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, et al. (2013) DNA-Binding Specificities of Human Transcription Factors. Cell 152: 327–339. doi:10.1016/j.cell.2012.12.009.

129. Cheatle Jarvela AM, Brubaker L, Vedenko A, Gupta A, Armitage BA, et al. (2014) Modular Evolution of DNA-Binding Preference of a Tbrain Transcription Factor Provides a Mechanism for Modifying Gene Regulatory Networks. Mol Biol Evol. doi:10.1093/molbev/msu213.

130. Nakagawa S, Gisselbrecht SS, Rogers JM, Hartl DL, Bulyk ML (2013) DNA-binding specificity changes in the evolution of forkhead transcription factors. Proceedings of the National Academy of Sciences of the United States of America 110: 12349–12354. doi:10.1073/pnas.1310430110.

131. Chene P (1999) Mutations at position 277 modify the DNA-binding specificity of human p53 in vitro. Biochem Biophys Res Commun 263: 1–5. doi:10.1006/bbrc.1999.1294.

132. Filippova GN, Qi CF, Ulmer JE, Moore JM, Ward MD, et al. (2002) Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter its DNA-binding specificity. Cancer Res 62: 48–52.

133. Kriel J, Haesendonckx S, Rubio-Texeira M, Van Zeebroeck G, Thevelein JM (2011) From transporter to transceptor: signaling from transporters provokes re-evaluation of complex trafficking and regulatory controls: endocytic internalization and intracellular trafficking of nutrient transceptors may, at least in part, be governed by their signaling function. Bioessays 33: 870–879. doi:10.1002/bies.201100100.

134. Rutherford JC, Chua G, Hughes T, Cardenas ME, Heitman J (2008) A Mep2-dependent transcriptional profile links permease function to gene expression during pseudohyphal growth in Saccharomyces cerevisiae. Molecular Biology of the Cell 19: 3028–3039. doi:10.1091/mbc.E08-01-0033.

135. DePristo MA, Hartl DL, Weinreich DM (2007) Mutational reversions during adaptive protein evolution. Mol Biol Evol 24: 1608–1610. doi:10.1093/molbev/msm118.

136. Fong SS, Joyce AR, Palsson BØ (2005) Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res 15: 1365–1372. doi:10.1101/gr.3832305.

137. Freed-Pastor WA, Prives C (2012) Mutant p53: one name, many proteins. Genes Dev 26: 1268–1286. doi:10.1101/gad.190678.112.

138. Robinson MD, McCarthy DJ, Smyth GK (2009) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140. doi:10.1093/bioinformatics/btp616.

139. Hoeijmakers WAM, Bártfai R, Françoijs K-J, Stunnenberg HG (2011) Linear amplification for deep sequencing. Nature protocols 6: 1026–1036. doi:10.1038/nprot.2011.345.

140. Dabney J, Meyer M (2012) Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. BioTechniques 52: 87–. doi:10.2144/000113809.

141. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, et al. (2010) Target-enrichment strategies for next-generation sequencing. Nature Methods 7: 111–118. doi:10.1038/nmeth.1419.

142. Chen Y-C, Liu T, Yu C-H, Chiang T-Y, Hwang C-C (2013) Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. PloS one 8: e62856. doi:10.1371/journal.pone.0062856.

143. Smith EN, Jepsen K, Khosroheidari M, Rassenti LZ, D'Antonio M, et al. (2014) Biased estimates of clonal evolution and subclonal heterogeneity can arise from PCR duplicates in deep sequencing experiments. Genome Biology 15: 420. doi:10.1186/s13059-014-0420-4.

144. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, et al. (2014) Library construction for next-generation sequencing: overviews and challenges. BioTechniques 56: 61–4–66–68–passim. doi:10.2144/000114133.

145. Kircher M, Sawyer S, Meyer M (2011) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Research 40: e3–e3. doi:10.1093/nar/gkr771.

146. Kvitek DJ (2013) Whole Genome, Whole Population Sequencing Reveals That Loss of Signaling Networks Is the Major Adaptive Strategy in a

Constant Environment. PLoS Genet 9: e1003972. doi:10.1371/journal.pgen.1003972.s009.

147. Greaves M, Maley CC (2012) Clonal evolution in cancer. Nature 481: 306–313. doi:10.1038/nature10762.

148. Wang T, Wei JJ, Sabatini DM, Lander ES (2014) Genetic screens in human cells using the CRISPR-Cas9 system. Science 343: 80–84. doi:10.1126/science.1246981.

149. Barton NH, Keightley PD (2002) Understanding quantitative genetic variation. Nature Reviews Genetics 3: 11–21. doi:10.1038/nrg700.

150. Tong AHY, Boone C (2006) Synthetic genetic array analysis in Saccharomyces cerevisiae. Methods Mol Biol 313: 171–192.

151. Levy SF, Ziv N, Siegal ML (2012) Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant. PLoS Biol 10: e1001325. doi:10.1371/journal.pbio.1001325.

152. Ziv N, Siegal ML, Gresham D (2013) Genetic and Nongenetic Determinants of Cell Growth Variation Assessed by High-Throughput Microscopy. Mol Biol Evol. doi:10.1093/molbev/mst138.