Historical Contingency and Compensatory Evolution Constrain the Path of Evolution in a Genome Shuffling Experiment with *Saccharomyces cerevisiae*

Damien Biot-Pelletier

A Thesis

in

the Department

of

Biology

Presented in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy at Concordia University

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Damien Biot-Pelletier

Entitled: Historical Contingency and Compensatory Evolution Constrain the Path of Evolution in a Genome Shuffling Experiment with Saccharomyces cerevisiae

and submitted in partial fulfillment of the requirements for the degree of

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Wayne Brake — Chair

Christian Landry — External Examiner

Paul Joyce — External to Program

Christopher Brett — Examiner

Reginald Storms — Examiner

Vincent Martin — Thesis Supervisor

Approved by

Chair of Department or Graduate Program Director

2017

Dean of Faculty

ABSTRACT

Historical Contingency and Compensatory Evolution Constrain the Path of Evolutionary Engineering by Genome Shuffling in *Saccharomyces cerevisiae*

Damien Biot-Pelletier, Ph. D.

Concordia University, 2017

The research reported in this thesis builds on an evolutionary engineering experiment (Pinel 2011) that yielded strains of *Saccharomyces cerevisiae* tolerant to a lignocellulosic hydrolysate. A highly tolerant strain was later characterized by whole genome and transcriptome sequencing (Pinel 2015). The evolutionary trajectories of mutations identified by sequencing were probed by whole population amplicon sequencing, while their significance to the phenotype was assessed by genotyping of additional mutants. Results of this work suggested that our survey of mutations selected during evolutionary engineering was partial. I therefore hypothesized that a complete survey of mutational diversity by whole population genome sequencing would further refine our understanding of lignocellulosic hydrolysate tolerance in *S. cerevisiae*. I further conjectured that extending this survey to several time points would reveal some of the fundamental evolutionary mechanisms that shape the outcomes of genome shuffling experiments. In parallel, I hypothesized that phenotypic testing of reverse engineered point mutants would identify mutations responsible for lignocellulosic hydrolysate tolerance in our strains of *S. cerevisiae*. My data revealed that a stong founder effect and prevalent genetic hitchhiking during genome shuffling lead to the domination of compensatory patterns during evolution. Bias introduced by historical contingency lead to the selection of few genuinely beneficial mutations. In the specific context of lignocellulosic hydrolysate tolerance, mutations in genes *NRG1* and *GSH1*, conferring tolerance to acetic acid, oxidative, and potentially other stresses most prominently enhanced the phenotype.

Acknowledgements

I want to thank Dr. Vincent Martin for welcoming me into his laboratory. I bet neither him or myself could have guessed the kind of journey this Ph.D. grew into, when I naïvely walked into his office for the first time. I think we both agree that it was a good journey. He challenged me with ambitious projects, which were tremendous learning opportunities. I am grateful for the limitless scientific freedom and the unwavering support he granted me for all these years. My time in the Martin lab was as much a rich scientific learning opportunity as it was a formidable character building experience. I grew in all aspects of my personality thanks to this project, and I owe it to Dr. Martin. Merci beaucoup.

This thesis builds on prior work from Dr. Dominic Pinel and David Colatriano. Their blood, sweat and tears layed the foundations for my own research, and for that, I am in their debt.

I wish to thank my committtee members, Drs. Christopher Brett and Reginald Storms, who provided key ideas to help develop the project further. I thank them for their advice, and for challenging my claims and my ideas during committee meetings. I enjoyed each of our conversations.

Lab mates, past and present, deserve more than a few accolades. I owe one to every single one of them. They have taught me, forced me to question my work, provided advice, asked me thought provoking questions and helped when I needed it. Many thanks.

My sincere thanks to Dr. Hung Lee (University of Guelph) and to Kevin Shiell (Collège communautaire du Nouveau-Brunswick) for helping me secure a supply of

spent sulphite liquor. I thank Tembec and AV Cell for producing and graciously providing spent sulphite liquor for use in my research. I would like to thank Dr. Mathieu Bourgey of the Canadian Centre for Computational Genomics for advice in the analysis of NGS data. I also thank Dr. Vladimir Rheinharz of the McGill University Department of Computer Science for advice on the critical assessment of multiple linear regression results. Although I did not pursue the work presented in Annex I, it was a great learning experience. It was initiated thanks to a stimulating conversation with Dr. Dylan Fraser, whom I warmly thank for his time. I am grateful for a conversation with Dr. Lea Popovic of the Concordia Department of Mathematics, who nourrished my quantitative inclinations and helped me critically assess my work.

The long hours, the weekend-long experiments, the doubts and the mood swings of the Ph.D., no one endured them like my wife Olivia. She held the fort while I pursued my studies, and never stopped supporting me. I am spoiled and don't always realize it. My two sons, Alexandre and Félix, deserve my gratefulness: it can't be easy to have a father who is mostly absent working, and always preoccupied when home.

I owe quite a few nods to my parents. My mother taught me pride and perseverance. I wouldn't have made it without those virtues. My father is at the origin of my scientific vocation. I thank him for the intellectual awakening, and for the invaluable humanistic education he bequeathed upon me. He raised the rationalist, so that study and hard work could train the scientist.

Finally, this project would not have been possible without the support of the Fonds Québécois de la recherche sur la Nature et les Technologies, the NSERC Bioconversion Network, and BiofuelNet Canada.

List of figures

List of tables

List of abbreviations

ALE	adaptive laboratory evolution
CAGE	conjugative assembly genome engineering
CRISPR	clustered regularly interspaced short palindromic repeats
DSB	double-stranded break
EMS	ethyl methylsulphonate
GRAS	generally regarded as safe
GS	genome shuffling
GSR	general stress response
gTME	global transcription machinery engineering
HDR	homology-directed repair
HMF	hydroxymethylfurfural
HWSSL	hardwood spent sulphite liquor
LTEE	long-term experimental evolution
MAGE	multiplex automated genome engineering
MRSS	mean residual sum of squares
NGS	next-generation sequencing
NTG	N-methyl-N'-nitro-N-nitrosoguanidine
ORF	open reading frame
PAM	protospacer adjacent motif
PCR	polymerase chain reaction
PEG	polyethylene glycol
PTS	phosphotransferase system
qPCR	quantitative PCR
RFLP	restriction fragment length polymorphism
ROS	reactive oxygen species
RT-PCR	reverse transcriptase PCR
SNP	single nucleotide polymorphism
STRE	stress response element
SWSSL	softwood spent sulphite liquor
TRMR	trackable multiplex recombineering
VHG	very high gravity
YNB	yeast nitrogen base
YPD	yeast peptone dextrose

1. Introduction and Literature Review

with excerpts from :

Biot-Pelletier D, Martin VJJ (2014). Evolutionary engineering by genome shuffling. Appl Microbiol Biotechnol. 98(9):3877-87.

1.1 Thesis hypotheses, goals and objectives

Previous results from a genome shuffling experiment generated strains with increased tolerance to a lignocellulosic hydrolysate [8]. Genomic and phenotypic data provided insight into the genetic basis of this phenotype, and suggested that additional mutations could be uncovered [9]. Expanding on these findings, the following goals were formulated, leading to associated hypotheses and specific objectives.

Goal 1

Previous results suggested that our survey of mutations selected by genome shuffling was partial. I therefore decided to conduct an exhaustive survey of those mutations by performing whole population genome sequencing on seven pools from six time points of the evolutionary engineering experiment. I obtained high quality sequencing data from these pools and performed metagenomic analysis to uncover those mutations.

Goal 2

The metagenomic dataset extracted in Goal 1 provided information on the frequency of mutant alleles at each of the sampled time points. Allele frequency time series gave a snapshot of the evolutionary trajectory of each mutation in the pool. My second goals was thus to analyze and compare these evolutionary trajectories, with the aim of identifying patterns and regularities. Evidence for several evolutionary phenomena were extracted from the data, for example genetic hitchhiking, convergent evolution and compensatory evolution. I hypothesized that a better understanding of the evolutionary dynamics shaping the outcomes of genome shuffling would inform the better design of future experiments.

Goal 3

Mutations detected in Goal 1 and patterns identified in Goal 2 inform our understanding of the genetic basis of tolerance to SSL in *Saccharomyces cerevisiae*. The last goal of my doctoral research was to obtain phenotypic evidence to complement the genomic data and identify mutations responsible for SSL tolerance in genome shuffled mutants. Phenotypic scoring and targeted genotyping of random segregants of an SSL-tolerant strain was used to build a linear model predicting the contribution of individual mutations to the SSL-tolerance phenotype. Phenotypic scoring of single point mutants was further performed to obtain direct data on the contribution of individual mutant alleles to the SSL tolerance phenotype.

1.2 General introduction

The axioms of evolutionary biology were born from the empirical observations of Darwin and other 19th century naturalists. Yet, the ensuing debate between saltationists and gradualists illustrates the shift to essentially theoretical considerations that followed the publication of *The Origin of Species* [1]. These elaborate theoretical controversies could only be resolved by an appropriate understanding of the laws of inheritance, which the rediscovery of Mendel's work and the experiments of T.H. Morgan and Nilsson-Ehle could solely provide [2]. More recently, the advent of molecular population genetics sparked similar controversies over the neutral theory of molecular evolution. Once again, academic strife abated when the limits of empirical tools available to test theory were recognized, and when technological progress allowed the collection of new evidence. Both examples stem from the fact that the sister disciplines of evolutionary biology and population genetics embrace complex and diverse forces that shape large populations over multiple generations. The fundamental molecular and environmental mechanisms that underlie evolution act on a scale and a level of complexity that was well outside the experimental reach of 19th and early 20th century researchers.

Fortunately, it was not out of intellectual reach, and we owe much of our modern understanding of evolution to the models and predictions of pioneering statistician-biologists, such as J.B.S. Haldane and Sewall Wright. Ronald Fisher developed many of the tools of modern statistics by attempting to address questions of population genetics. Hence, key concepts in evolution, such as genetic drift, gene flow or mutation, to provide just a few examples, received their mathematical development from study of the implications of the Hardy-Weinberg principle, and especially the violation of its assumptions.

Early corroboration of theoretical predictions came from the observation of natural populations. Statistically significant agreement between patterns in the distribution of traits in the environment and the predictions of evolutionary models were the empirical evidence on which scientific progress relied. In this way, the morphology of fruit flies and ladybug beetles in nature were among the evidence that enabled Dobzhansky and his students to elaborate the modern synthesis of evolutionary biology. In the infancy of molecular biology, when access to protein electrophoresis, and later macromolecular sequences added details and refined knowledge on the genetic determinants of phenotypic traits, the tape of evolution could still not be replayed otherwise than in thought. Molecular methods enabled the differentiation of alleles at the sequence level, with their frequencies, spatial distribution and linkage estimated with renewed precision, but conditions, still not controlled, were dictated by the whims of nature [2].

Microorganisms, with their large populations, fast doubling rates and cheap nutritional requirements are amenable to controlled experiments on the dynamics of evolution [3]. By serial passage in defined media, or by prolonged incubation in chemostats, evolutionary response to precisely controlled environments can be followed in real time. At first essentially phenotypic, the tracking of experimental evolution now benefits from remarkable progress in sequencing technology [4]. Affordable and high throughput sequencing means that the genomes of large numbers of individuals and even whole populations can be sequenced over many experimental time points. The result is the highly sensitive detection of mutant alleles, and the precise measurement of their frequency over time. This has provided novel insight into the dynamics of evolution.

Adaptive evolution experiments can be used over a wide range of organisms and conditions, and over large numbers of generations, but still rely on the chance

occurrence of beneficial mutations. For example, in a long term evolution experiment, citrate utilization in *Escherichia coli* evolved after 30,000 generations, and occurred in only one out of 12 replicates [5]. Engineering methods aimed at either accelerating this process or studying phenomena for which natural diversity is limited or biased have been applied and complement the insights of experimental evolution [6]. Genome shuffling is one such method, which has been widely adopted for the engineering of complex traits in industrial microorganisms [7]. It has been successfully used in the Martin lab to generate strains of *Saccharomyces cerevisiae* that are tolerant to spent sulphite liquor (SSL), a byproduct of the acid bisulphite pulping process and prospective feedstock for the production of biochemicals and biofuels [8]. Strains generated by genome shuffling have been used in our laboratory to study the genetic determinants of lignocellulosic hydrolysate tolerance in *S. cerevisiae* [9]. In this thesis, I study the dynamics of this previous genome shuffling experiment. Using whole population genome sequencing, I surveyed the mutational diversity selected by the experiment, and tracked mutant allele frequency over the course of the evolution. The evolutionary trajectories of mutant alleles, coupled to a systematic dissection of mutant phenotypes further refined our understanding of lignocellulosic hydrolysate tolerance in yeast. Knowledge of the phenotype of interest and a detailed description of molecular evolution allowed us to identify some of the driving evolutionary forces that shaped the outcomes of our genome shuffling experiment. In particular, I posit that the data presented in this thesis make a strong case for the influence of historical contingency on the outcomes of evolutionary engineering experiments.

In this introductory chapter, I first provide a review of experimental evolution studies, describing how they have contributed to our understanding of evolutionary

biology. Next, I illustrate the emerging contribution of evolutionary engineering to our understanding of evolution. I then focus on genome shuffling, and review its application to the enhancement of industrially desirable traits in microbes. The specific genome shuffling (GS) experiment on which this thesis is based aimed at selecting for tolerance to lignocellulosic hydrolysates. Inhibition of production microbes by lignocellulosic hydrolysates is explained in particular as it relates to the physiology of *S. cerevisiae*. Special attention is allotted to our knowledge of the stress response in this yeast. I end the chapter with a description of the GS experiment on which this thesis is based, outlining key outcomes and summarizing lessons about lignocellulosic hydrolysate tolerance learned from the study of GS mutants.

1.3 Experimental evolution

Experimental evolution consists in the propagation of populations of living organisms under defined and reproducible conditions across multiple generations. This typically entails simultaneous replication and the maintenance of control populations. Central to experimental evolution is the possibility to freeze the evolving population, or isolates for detailed genetic analysis [10 ch.1]. Several different kinds of experiments can be described as experimental evolution. Field experiments on the introduction of invasive or intentionally introduced novel species may fit the definition, but this thesis is mostly interested in strictly defined and controlled laboratory experiments. Most specifically, experiments performed with microbes are my focus, although examples from multicellular organisms exist [11].

Even within these boundaries, several different methodologies have been used to perform experimental evolution. Different modes of selection can be considered, such as truncation and culling, but the most widely reported approach is adaptive laboratory evolution (ALE)[10 ch.2]. ALE consists in the incubation of an initially isogenic microbial population into a novel, non-native environment over hundreds of generations. Examples of selective pressures applied in ALE experiments include nutrient limitation [12], physicochemical stress [13], or the conservation of a physiological function such as the ability to sporulate, [14] or social motilily [15]. Novel conditions may be applied in a continuous [16] or discontinuous manner [12]. Most adaptive evolution experiments are performed in a top-down fashion, whereby conditions are set by the experimenter to measure the evolutionary response, and compare the results to evolutionary theory. For example, adaptation to non-optimal temperature can be performed by prolonged culturing of *E. coli* at 42^{o}C. Resulting mutants are then characterized for their gain in fitness at novel temperatures. An alternate, or bottom up approach, involves comparing the fitness of well-defined mutants in competition assays testing different conditions.

Experimental evolution presents several advantages over the retrospective study of historical diversity, which make it a precious tool for the study of evolutionary phenomena. Those advantages can be summarized in two words: control and replication. Unlike natural evolution, the environmental conditions of laboratory experiments can be set and narrowly controlled. The evolutionary starting point is known and can be preserved, while samples can be regularly taken from the experiment and frozen for further characterization. The fitness gains in evolved mutants can thus be measured by head-to-head competition with the ancestor. More importantly, a detailed and almost real-time picture of evolution can be obtained from such experiments. This is

a significant advantage over the study of naturally occurring diversity, which relies on fossils and living samples from the present. The dynamic picture of evolutionary history provided by ALE allows the easier detection of genetic linkage and correlations. Because the ancestor is known and well characterized, phenotypic and genotypic differences with evolved mutants can be measured, enabling the identification of the genetic basis of adaption. Parallel ALE experiments in the ancestral conditions can be performed, enabling the segregation of genuine adaptive changes from the effects of genetic drift. In ALE, replication allows the study of the diversity of possible evolutionary solutions. Replication is also a means to assess parallelism, or the reproducibility of evolution. A well-characterized starting point further means that the effect of historical contingency can be better identified. The open-ended nature of evolution has the advantage that it can give unexpected results, linking genes to new functions, and hinting at novel pathways and molecular mechanisms. Finally, experimental evolution benefits from some of the advantages of microbes, which have large population sizes, short generation times and affordable nutritional requirements.

Experimental evolution is in no way a new practice: the first reported example of a controlled evolution experiment dates from the 1880s and is attributed to W.H. Dallinger, who gradually adapted strains of protozoa to elevated temperatures. At the beginning of his experiment, the protozoans thrived best around 16^{o}C, but after seven years of adaptation, he reported that his microbes could withstand 70^{o}C, earning the praise of Charles Darwin himself [10,17]. In a pioneering study following the invention of the chemostat, Novick and Szilard introduced one of the first modern and controlled examples of experimental evolution, providing evidence for clonal replacement in populations of bacteria growing under limiting nutrient conditions [18]. Thereafter,

experimental evolution has either been used to study the population genetics forces that guide evolution, or on the contrary focused on the specific mechanisms of adaptation leading to novel traits [10 ch.13]. Means to access the genetic architecture (i.e. the genetic basis of a phenotype) underlying selected traits was required to bridge the gap between both objectives. Early examples of such attempts used metabolic flux theory and experimental evolution to probe the mechanisms leading to selective neutrality [19,20]. However, as discussed below, the advent of affordable high throughput sequencing technologies has rapidly multiplied the number of studies which attempt to uncover the genetic architecture of selected traits to answer questions of evolutionary biology [21].

It is not my intent to provide an exhaustive survey of all experimental evolution studies and their contributions to evolutionary biology. In the rest of this section, I instead review some of the major issues of modern evolutionary biology for which experimental evolution has been successfully applied.

A review of experimental evolution studies, even partial, cannot be complete without mentioning the *Escherichia coli* long-term experimental evolution (LTEE) project. Started in 1988, this on-going experiment consists in twelve replicate lines of bacteria evolving in glucose-restricted minimal medium [22]. Lines from this experiment are propagated as batch cultures, passed daily by 1:100 dilution into fresh medium. As the first draft of this thesis is being written, the experiment has reached generation 65,829 [22]. Core findings from this experiment have been published in a series of thirteen articles all bearing the title *Long-term experimental evolution in Escherichia coli* [23–35], but several other studies about the LTEE have been published. The early phases of the experiment showed that the rate of adaption is initially rapid, with an increase in population fitness of 30% after the first 2000 generations [23]. Adaptation was subsequently demonstrated to

slow down, with the gain in fitness reaching 50% after 10,000 generations, suggesting a possible decline in the number of available beneficial mutations and a general pattern of diminishing returns for each new mutation [36]. Moreover, adaptation showed a high level of parallelism, suggesting that phenotypic evolution follows a predictable pattern.

Genotypic parallelism was later demonstrated for this same experiment. Among recurrent adaptive traits was the loss of ribose catabolism, rendered useless and burdensome by constant cultivation on a single carbon source [37]. Interestingly, this trait was repeatedly evolved following the same molecular mechanism involving disruption of sequences upstream of the *rbs* operon by an insertion element. Analysis of gene expression profiles at 20,000 generations detected a general alteration of the guanosine tetraphosphate and c-AMP receptor protein regulons. This observation led to candidate gene sequencing, which notably identified repeated mutations in gene *spoT*, and further demonstrated parallelism and the predictable character of evolution [38].

Loss of function, as illustrated by abolished ribose catabolism in the LTEE, is a recurrent observation in evolution experiments. This further argues for the deterministic nature of evolution in large populations. For example, adaptation of *E. coli* to constant glucose-limitation in otherwise benign conditions was shown to lead to the loss of stress response pathways, illustrated by the rapid proliferation of mutations that abolish the function of stress regulator RpoS, both in experimental and natural settings [16]. Similarly, culturing of *Bacillus subtilis* in excess glucose was shown to repeatedly lead to loss of sporulation ability [39]. More recently, Kvitek and Sherlock have applied whole population whole genome sequencing to generalize this loss of function pattern to evolution in constant environments [40]. Their review of selected mutations showed that over half of the mutations accumulated in replicate lines of yeast adapted to constant environments

led to a loss of function in the glucose signaling, Ras/cAMP/PKA and HOG pathways. The resulting loss of sensitivity to the environment was shown to be reproducible, underlining parallelism as well as a trade-off pattern in evolution. Arguably, the controlled environment of the laboratory is particularly conducive to adaptive loss of function. In the absence of environmental fluctuations, the fitness costs associated with stress response, nutrient sensing and a highly responsive metabolism become disadvantageous. Yet, this adaptive path represents a trade-off: environment-insensitive mutants isolated by Kvitek and Sherlock displayed loss of viability under starvation. Similarly, *rpoS*$^{-}$ isolates of *E. coli* are at a disadvantage during prolonged starvation [16]. Twenty percent of *B. subtilis* genes involved in sporulation are strongly expressed in non-inducing conditions, illustrating the strong fitness cost associated with this ability [14] and the scale of the apparent trade-off between growth rate and sporulation ability [39].

Phenotypic trade-offs pose a fundamental question about the mechanisms of evolution. In order to “spend on one side”, does nature necessarily need to “save on the other side”, as Darwin initially put it [1]? An alternative explanation, at least for some apparent trade-offs, would be that the loss of fitness in some traits associated with gains in other traits results from the accumulation of nearly neutral mutations by genetic drift. This dilemma between antagonistic pleiotropy (i.e., a mutation can have multiple, opposing effects on fitness) and mutation accumulation is a core question on the origins of genetic diversity. In other words, is natural diversity the result of natural selection, or did it arise because of drift, as predicted by the neutral theory? Experimental evolution has shed light on this debate. Decay of unused catabolic functions during the *E. coli* long-term evolution experiment was mentioned earlier. The mutation accumulation

hypothesis predicts random loss of function happening at a constant rate throughout evolution. On the contrary, the pattern of catabolic decay during the LTEE reproducibly happens during the early phases of rapid adaptation, and this behaviour was not affected by the spontaneous appearance of mutator phenotypes [41]. Yet another observation of the LTEE supports antagonistic pleiotropy. In the experiment, *E. coli* is adapted to glucose, a substrate uptaken via a phosphotransferase system (PTS). Diversity of fitness in adapted clones was measured on a variety of carbon sources, both PTS dependent and independent. The distribution of fitness was narrow on substrates imported via a PTS, while much greater diversity was observed on a PTS independent carbon source [25,26]. Offshoots of the *E. coli* long-term experiment were designed to specifically address the trade-off dilemma. The LTEE is conducted at 37^{o}C and pH 7.2, and an isolate from 2000 generations of adaptation was used as the founding ancestor for a series of branching adaptation experiments. Replicate lines of adaptive evolution to different environments, spanning temperatures of 20^{o}C to 41.5^{o}C and pH 5.4 to 8.0 were launched and maintained by serial passage for 2000 additional generations. Quantifiable fitness gains over the ancestor were observed in those niche conditions [42,43]. Competition experiments between niche-adapted clones in the other conditions revealed various degrees of trade-offs [43,44]. Trade-off was the general pattern observed in this series of studies, but it was not systematically observed [45], arguing that antagonistic pleiotropy may not be a general mechanism of adaptive specialization.

Specialist phenotypes are the textbook example of trade-off effects and antagonistic pleiotropy. The experimental evolution literature contains several interesting cases of adaptive specialization. I will mention two related examples, which to some extent illustrate artifactual disadvantages of adaptive laboratory evolution. Using the

same conditions as the LTEE, Treves and coworkers repeatedly observed the formation of two niche subgroups with differing carbon source preferences. The first group of bacteria feeds preferentially on glucose, fermenting it to acetate, which is used by the second group as its favourite carbon source [46,47]. While this observation is interesting for evolutionary ecology, it should serve as a caution to experimentalists interested in studying the causes of diversity. In this case, diversity is structurally maintained by the largely artifactual and chronic accumulation of waste products in serially transferred batch cultures. Parallel adaptive evolution in *Pseudomonas fluorescens* with and without stirring highlights another potential limitation of experimental evolution. Stirring of microbial cultures is meant to standardize aeration and the distribution of nutrients in culture vessels. It is part of the environmental control essential to meaningful experimental evolution. However, it leads to a loss of spatial organization, as illustrated by the specialization of *P. fluorescens* in unstirred adaptive evolution experiments. In still cultures, this bacterium rapidly and reproducibly separates into three groups, with spatial preference for the broth, the bottom of culture vessels, or the air-broth interface [48].

We have just seen how the conditions of evolution experiments can influence, sometimes in unexpected manners, the formation of ecological niches. I now complete my general discussion of experimental evolution with a presentation of some of the other limitations of this methodology. A central criticism addressed to laboratory evolution is that it ostensibly lacks realism. Indeed, it is performed in the idealized and tightly controlled environment of the lab, sacrificing the multidimensionality of nature. In particular, we have seen the effects that constant conditions can have on experimental evolution. However, in all fairness, the *ceteribus paribus* criticism may be addressed to all of experimental science. Microbial culture, with its ample nutrients and ideal

temperatures, can be considered as too simple and benign to reflect real life conditions, but on the contrary, there are instances where it may be too stressful. Indeed, typical batch culture, especially under a serial passage regimen, leads to cyclic starvation and to the accumulation of waste products. Since it requires large populations of rapidly evolving organisms, the scope of characters that can be tested using this approach is limited to those found in microbes and other life forms with short generation times. Despite this fact, it still represents an important time commitment, which many researchers may not be able to afford. The time scale of laboratory evolution experiments is too short for the study of macroevolution, limiting discussion to microevolutionary phenomena. A key aspect to consider is that the laboratory is a specific environment in itself, which requires adaption from natural organisms. Common laboratory strains of yeast and bacteria are in fact domesticated organisms. Therefore, organisms that go directly from nature to experimental evolution may be adapting simultaneously to the lab and the specific conditions tested by the experiment, which may confound downstream analyses. Inversely, lab strains may not reflect the behaviour of wild or industrial strains in the same conditions [10 ch.2,22].

1.3a Effect of sexual recombination on evolution

In the previous section, I have discussed the illustration of fundamental questions of evolutionary biology through experimental evolution. The influence of sex on evolution is of special interest to this thesis, because of the evolutionary engineering methods under study. There are relatively few reports of evolution experiments that incorporate sexual recombination, because clonal propagation is the preferred mode of microbial reproduction. Facultative sexual reproduction cycles in unicellular eukaryotes allow the

design of evolution experiments that study the impact of sex on evolutionary dynamics. Similar experiments in multicellular organisms exist, but they are of lesser relevance to this essentially microbiological work [11]. In this section, I mainly review three studies and the insight they provide into the influence of sex on the rate of adaption, on clonal interference, and on the effectiveness of purifying selection.

A distinguishing feature of microbial sex is its facultative nature. Facultative sex both represents a challenge and an experimental opportunity. Sexual reproduction is thought to confer a selective advantage in microbes, because it is maintained in wild populations [10 ch.16]. However, it is expendable: without the appropriate selective pressure, this ability is frequently lost to mutation accumulation. Examples of this "use it or lose it" character are documented in *Cryptococcus neoformans* [49], *Saccharomyces cerevisiae* [50], and *Chlamydomonas reinhardtii* [51]. Facultative sexual reproduction is an interesting tool for the experimenter, because it allows the comparative study of evolution with and without sex.

In *S. cerevisiae* and *C. reinhardtii*, meiosis occurs in response to prolonged starvation and results in the formation of spores. In those species, sex thus confers a direct survival advantage. It means that environmental changes are necessary to induce meiosis and enable sexual recombination. This creates a problem for the study of evolution in constant environments, but makes the execution of appropriate control experiments challenging. In yeast, this challenge is overcome by disruption of genes involved in the regulation of meiosis. Deletion of *IME1* abolishes meiosis completely: *imeΔ* cells simply wait through nitrogen starvation without sporulating [52]. An alternative is the deletion of genes *SPO11* and *SPO13*, which results in sporulation without meiosis

and recombination [53,54]. These mutants can be used as controls for the specific effects of sex, because they can undergo exactly the same treatments as experimental populations without undergoing sexual recombination.

Sex requires complex molecular machinery, notably for cell fusion and recombination of DNA. Maintenance of these systems, and the considerable commitments in time and energy involved with sex represent important fitness costs. Understanding the selective advantages that lead to the prevalence and maintenance of this complex mode of reproduction in nature is thus a central problem of evolutionary biology. The classical explanation for sex, termed the Fisher-Muller principle, states that recombination acts by reducing negative linkage disequilibrium at loci that are under selection. In other words, sex helps separating beneficial alleles from secondary deleterious mutations to which they may be linked: it shuffles alleles and diminishes their dependence on genetic background. This idea was tested by experimental evolution in *S. cerevisiae* by Goddard and coworkers. Yeast was cultured in chemostats using benign, glucose-limited conditions, and compared to harsh conditions of elevated temperature and increased osmolarity. Periods of asexual reproduction were regularly interrupted by episodes of induced sexual recombination. Asexual evolution of *spo11Δ spo13Δ* mutants was compared to that of wildtype cells. After 300 generations of mitotic propagation and five or ten cycles of sporulation, sexual yeast displayed a clear advantage in high salt and elevated temperature, but none in benign medium [55]. These observations are compatible with a faster rate of adaptation imparted by sexual recombination.

Reduced rates of adaptation in asexuals indicate negative linkage disequilibrium, but different mechanisms may explain how sex relieves this effect. One hypothesis is that sex accelerates evolution by reducing negative epistasis. This “ruby in the rubbish” model states that it is the direct uncoupling of beneficial alleles from their accompanying mutational load that constitutes the main advantage of sexual recombination. Another hypothesis states that sex reduces clonal interference. In asexual populations, beneficial mutations arise independently and compete against each other. This means that the fixation of one beneficial allele occurs at the expense of other ones. This clonal interference leads to patterns of incremental adaptation, in which beneficial mutations appear and accumulate in a single “winning” lineage. In sexual populations, two beneficial mutations do not need to sequentially arise by chance in a common lineage to both attain fixation. They may appear concurrently without necessarily competing: recombination facilitates their reunion into the same individuals. Epistasis and clonal interference are not predicted to react in the same way to population size. Epistasis is unaffected by population size, because the probability of recombination between alleles at two loci is the same regardless of the number of individuals. On the contrary, clonal interference is dependent on population size. Larger populations have a higher probability of generating beneficial mutations, meaning that they will accumulate more of them on average. Based on these predictions, Colegrave *et al.* evolved initially isogenic lines of *C. reinhardtii* with an intervening episode of sexual recombination and varying effective population sizes. Fitness increases were positively correlated with population size, indicating more accumulation of beneficial alleles in larger populations of these sexual microbes [56]. More recently, comparative whole population whole genome sequencing of evolution experiments in sexual and asexual lines of *S. cerevisiae*

described the effect of sex on molecular evolution [57]. This study showed that sexual and asexual lines produced comparable numbers of mutations, but that the sexuals accumulated a much smaller, mostly non-synonymous subset of those. Cohorts of mutations with correlated evolutionary trajectories were observed in the asexuals, but not in the sexuals. Together, these observations indicate pervasive hitchhiking in asexuals, and efficient purifying selection in sexuals. Patterns of clonal interference were observed in the asexual populations, in which cohorts of mutations were transiently selected then were outcompeted and declined to extinction. This study is the first reported investigation of the effects of sex on molecular evolution. Its results convincingly show that sex speeds adaptation both by reducing clonal interference and efficiently sorting beneficial mutations from hitchhikers.

1.3b Genomics of experimental evolution

Early investigations of experimental evolution, and most of the studies I have reviewed until now, relied on the tools of classical genetics and on phenotypic measurements. This technological constraint restricts access to patterns of molecular evolution and limits understanding of the genetic architecture of evolved traits. Rapid technological progress since the turn of the century has enabled researchers with ever increasing power and resolution in their investigations of evolutionary genomics and molecular evolution. Experimental evolution entered the molecular era with a study of adaptation to glucose limitation. Brown *et al.* hypothesized that amplification of glucose transporters was selected by evolution of yeast in glucose-restricted medium. Using Northern and Southern blotting and restriction mapping, they could demonstrate an increase in gene copy numbers of glucose transporters [58]. Albeit successful, this

approach was biased towards a very specific hypothesis that relied on prior knowledge of yeast physiology. Soon after, microarrays enabled the first genomic studies of experimental evolution. This agnostic approach confirmed previous findings on the amplification of hexose transporters [59], but could illustrate parallelism in the evolution of transcriptomes and large scale genome rearrangements [60]. Later generations of microarrays enabled the first investigations of single nucleotide changes in experimentally evolved mutants [61]. Sensitive microarrays were notably used to investigate the genetic basis of adaption of yeast to fluctuating glucose-galactose environments, identifying strong parallelism in the selection of mutant *gal80* alleles [62].

The development of massively parallel DNA sequencing technologies now allows for affordable and deep investigations of molecular evolution. Highly multiplexed candidate gene sequencing, high-resolution transcriptome profiling, whole genome sequencing and whole population whole genome sequencing (i.e. metagenomics) are all within reach of the average academic lab. In previous sections, I already have alluded to some of these recent genomics studies, and notably their contributions to our understanding of trade-off effects and the mechanisms by which sex speeds adaptation. Here, I propose to specifically review how high-resolution genomics have contributed to refine our understanding of evolutionary forces.

The *E. coli* long-term evolution experiment provided the material for the first in-depth investigation of ALE by whole genome sequencing of individual adapted clones [63] and whole population samples [64]. NGS investigation of this already well-studied experiment confirmed several previous findings. For example, it was known that the rate of adaptation was initially high, and later slowed down. On the contrary, the rate of genome-level change was remarkably constant, seemingly confirming predictions of the

neutral theory. However, stark overrepresentation of non-synonymous mutations, high genomic parallelism between replicate lines, temporal persistence of mutations, and demonstrated fitness effects of almost all detected mutations argued against a dominant effect for drift [63]. Whole population sequencing of the LTEE also provided evidence for clonal interference, mutation fixation, mutator phenotypes, and cross-feeding specialization [64]. A follow-up study to these pioneering investigations was recently published [65]. Tenaillon and coworkers sequenced the genomes of 264 clones from the 12 populations of the LTEE at 11 time points spanning 50,000 generations. Their data refine understanding of the experiment while confirming several key observations of previous studies. In contrast with the arguably less comprehensive 2009 studies, the data provide evidence for genetic hitchhiking of neutral mutations on beneficial alleles. However, it confirms that the early phases of evolution were dominated by beneficial mutations, which remained preponderant but declined in proportion as evolution progressed. Repeated selection of mutator phenotypes was also confirmed. Parallelism was prevalent, observable at the gene level. Overall, the authors argue that the LTEE is a strong demonstration in favour of a selectionist view of molecular evolution. As the main evidence for this position, the authors invoke the sustained fixation of a large number of beneficial mutations throughout the experiment.

Phenomena revealed by NGS of LTEE samples have been observed in similar experiments. Genetic hitchhiking, clonal interference and parallelism appear as molecular hallmarks of experimental evolution. For example, adaptation of 40 populations of *S. cerevisiae* to rich medium, revealed patterns of clonal interference and genetic hitchhiking similar to those observed in *E. coli* during the LTEE [66]. Review of the experimental evolution literature provides a rather compelling case for parallel evolution.

For example, Tenaillon *et al.* evolved 115 populations of *E. coli* to increased temperature, and sequenced a single isolate from each of these lines. The result showed high degrees of convergence at the gene, operon and functional complex levels [67]. Similar instances of evolutionary parallelism are reported, sometimes down to the nucleotide level, by Herron and Doebeli in clones of *E. coli* evolved in the presence of competing carbon sources [68]. Parallelism was also demonstrated in short-term evolution of *Burkholderia cenocepacia* for adherence and a wrinkly colony phenotype. Whole genome sequencing revealed reproducible clustering of mutations at the *wsp* locus [69]. Parallelism is thus well documented in asexual populations evolving from *de novo* mutations. Yeast undergoing frequent sexual recombination and evolved from standing genetic variation similarly displays high levels of parallelism [70].

Numerous deep sequencing studies on experimentally evolved isolates have provided examples of epistasis (i.e. dependence of the fitness effect of individual alleles on the genetic background). For example, sequencing of an *E. coli* mutant evolved for 1,2-propanediol utilization led to the observation that two mutations were required to elicit the desired phenotype [71]. Examples of reciprocal sign epistasis, in which a mutation produces opposite effects on fitness in different backgrounds, have also been revealed by WGS. Short-term adaptation of *E. coli* to lactate medium produces a mutation in gene *kdtA*, which leads to amino acid auxotrophy in the wildtype that is relieved in evolved mutants carrying six other mutations [72]. During adaptation of yeast to limiting glucose, beneficial mutations in genes *MTH1* and *HXT6*/*HXT7* are repeatedly selected, but Kvitek and Sherlock have shown that the presence of both mutations causes a reduction in fitness, preventing the selection of double mutants [73]. Similar observations were made in highly replicated evolution experiments of *E. coli*, in which

adaptation to elevated temperature selects for mutations in the RNA polymerase complex or in the termination factor *rho*, but not both [67]. Together, these examples demonstrate the influence that epistasis may have on evolution. Selection of one allele in a positive epistatic pair depends on the appearance and selection of the other. Reciprocal sign epistasis, on the contrary, has the potential to force evolution towards specific local fitness maxima, depending on the contingent and random order in which mutations appear and are selected.

1.3c Experimental study of compensatory evolution

Compensatory evolution is proposed to have played a major role in the evolutionary engineering experiment studied in this thesis. In this sub-section, I introduce this concept, and report major theoretical and experimental studies of this phenomenon. I discuss how compensatory evolution can be exploited in basic molecular biology research, drug target discovery, and biological engineering.

Compensatory evolution describes the process by which secondary mutations are selected to alleviate the fitness cost of other alleles. Therefore, this phenomenon is tightly linked to the concepts of epistasis and genetic interaction. The theoretical framework for compensatory evolution was established by Kimura [74] and further developed by Weinreich and Chao [75]. Models elaborated by these authors showed that fitness valleys could be crossed via the sequential accumulation of epistatic but individually deleterious or neutral alleles. These models further showed that the fixation of compensatory pairs was facilitated by large population sizes and linkage between the interacting loci [75].

Experimental investigation of compensatory evolution was conducted in viruses [76–79], bacteria [80–82], yeast [83,84] and worms [85,86]. Major conclusions from those studies are summarized below. In populations affected by a crippling mutational load, compensation is both rapid and pervasive [80,82,83,85,86]. A broad range of genetic defects can be compensated, but with unequal likelihood [82,83]. One of the main predictors of compensatory evolution is the strength of the fitness defects caused by deleterious alleles [80,83]. The path of compensatory evolution appears to depend highly on the selective environment [81,83,84]. Furthermore, environmental changes readily reveal the highly pleiotropic effect of compensatory mutations [83,84]. Both the dependence of compensatory evolution on the strength of fitness defects and the pleiotropic effects of compensating mutations are in line with Fisher's geometric model of adaptation. This model predicts that both the frequency of adaptive mutations and the tolerance of evolution to antagonistic pleiotropy increase with distance from the fitness optimum [87]. Historical contingency constrains the path of compensatory evolution, as shown by the divergent evolutionary solutions selected for the same defects in different yeast strains [84]. Finally, compensatory evolution is proposed to play a major role in evolutionary divergence. Indeed, in a context in which the rate of neutral and deleterious mutation is higher than that of beneficial mutation [88], widespread compensation is expected. Compensatory evolution does not typically restore the global expression pattern of the non-mutant ancestor, further potentiating divergence [83].

Compensatory evolution has potential as a tool for both basic and applied research. For example, it has been used to study the response of bacteria to the fitness cost of antibiotic resistance [81]. Similarly, the frequent role of gene loss in tumorigenesis identifies compensatory evolution as an attractive tool for the study of evolutionary

dynamics in cancer development and progression [83,89,90]. Because compensatory evolution converges at the functional network level rather than at the molecular level [79,83], it has been exploited to identify novel functional interactions and potential new drug targets for genetic diseases [84]. I finally propose that a finer knowledge of compensatory evolution may help genetic engineers alleviate the burdens caused by their interventions in the genomes of industrial organisms.

1.4 Evolutionary engineering

So far, I have discussed experimental evolution from the point of view of evolutionary biology. In that context, the precise phenotypes under selection are a secondary consideration: the focus is on forces and mechanisms that shape the path of evolution. However, it was realized early that Darwinian evolution could be exploited by the experimentalist to produce specific outcomes [91]. This approach of using evolution to produce a phenotype of interest is termed directed evolution or evolutionary engineering, with the former term mostly applied to single-molecule approaches. In this section, I review evolutionary engineering as an experimental method and field of study. I begin with a methodological definition of evolutionary engineering, presenting major experimental methods and approaches and illustrating them with key studies. I then provide a detailed review of genome shuffling, which is the specific evolutionary engineering approach studied in this thesis. I finish this section by placing evolutionary engineering in the wider context of synthetic biology, arguing that beyond specific design and engineering objectives, it represents an opportunity to investigate basic questions of biology.

1.4a Definitions and experimental methods

We have defined evolutionary engineering as the deliberate use of Darwinian evolution for the realization of defined outcomes in biological systems. This approach stems directly from our incomplete understanding of biological systems, which precludes purely rational approaches. It requires four central components: i) a characteristic, either qualitative or quantitative, which can be scored; ii) a seed or canvas system used as an evolutionary starting point; iii) means to obtain diversity in the seed; and iv) a way to isolate fitter variants within that diversity. The precise nature of each these components will influence the kind of experiment to conduct. I choose to differentiate evolutionary engineering methods on the basis of three defining characteristics, which I call scale, continuity, and bias. Scale refers to the relative size of the biological unit undergoing evolution. Evolutionary engineering can be performed at the level of single biomolecules or genes; at the level of “genetic circuits” such a metabolic or regulatory pathways; or at the organismal level. Evolutionary engineering methods are either continuous or discontinuous. Classical discontinuous methods separate the generation of diversity, screening or selection, variant propagation and potential recombination into discreet steps, which typically require direct manipulations by researchers. Continuous methods rather proceed with minimal outside intervention, and the separation between evolutionary events appears seamless. A more fundamental and defining characteristic of evolutionary engineering methods is bias. By bias I refer to the relative number of assumptions on which an evolutionary engineering experiment is built. Therefore, I roughly categorize methods as semi-rational or agnostic. To some extent, bias is related to scale. For example, restricting the evolutionary search to a single regulatory pathway defines the scale of the experiment but implicitly ignores the effect of other cellular

subsystems. At the other end of the spectrum, the sequence space available to the evolution of whole organisms necessarily reduces bias, but we will see that methodological choices are never completely agnostic. Below, I divide exposition of evolutionary engineering based on scale, discussing the implications of continuity and bias.

Evolutionary engineering of single molecules

Evolutionary engineering of single biomolecules, both RNA and proteins, is a vast and rich field, which has demonstrated time and again a capacity to produce novel functions and capabilities [92]. The diversity of fine-tuned fluorescent proteins available to biologists is a Nobel prize-winning testimony of the power and impact of these approaches [93–95]. Other notable examples include a recombinase apt at excising the HIV virus [96], a P450 evolved into a propane hydroxylase [97], and an increase of more than 40^{o}C in thermostability of lipase A [98]. A central challenge for evolutionary protein engineers has been the inability to exhaustively sample the immense combinatorial sequence space available to even the simplest seed molecules. In this context, stepwise movement along the fitness landscape by accumulation of single mutations has been considered the most rewarding strategy [92]. This approach is agnostic to the extent that diversity is typically introduced at random by error-prone PCR. However the stepwise method is blind to epistasis involving individually neutral or deleterious mutations, and its ratchet-like behaviour can confine it to local fitness maxima. This asexual mode of propagation and mutation accumulation is vulnerable to a kind of artificial clonal interference, which can slow down and confine evolution, as we have seen in previous sections. *In vitro* recombination (or DNA shuffling) strategies to circumvent this

shortcoming have been devised [99–102]. Semi-rational approaches have been proposed to address the unsampled sequence space issue. Using prior knowledge of molecule structure and function, targeted evolution can increase the probability of identifying beneficial mutations. For example, evolution has been restricted to protein domains [97,103], or active sites and binding pockets [104–106]. Recombination of natural molecules is another way to reduce complexity by introducing an evolutionary bias [99,107,108].

More recently, systems for continuous *in vivo* mutagenesis and recombination of protein variants were reported. The first of these systems is phage-assisted continuous evolution (PACE). In this system, *E. coli* cells expressing an essential phage protein are continuously fed and removed from a pool of phages lacking that critical gene. Phages replicate faster than *E. coli*, and bacteria remain only transiently in the virus pool, such that evolution remains restricted to virions. The system must be designed so that selection links biomolecule evolution to the expression of the essential phage protein in infected bacteria. Therefore, evolution is directly tied to phage propagation. With this system, the inventors of PACE were capable of evolving T7 RNA polymerase to recognize the T3 promoter [109]. Another method, called heritable recombination (HR), takes advantage of the high homologous recombination rates induced in *S. cerevisiae* by double-stranded breaks to effect both recombination and mutagenesis. Mating and sporulation of yeast further facilitates recombination. This method addresses the transformation bottleneck, which limits sequence space exploration. Indeed, transformation is performed only once for the generation of initial diversity: further variation is induced by recombination and is propagated *in vivo* [110].

Fine-tuning of genetic circuits by evolutionary engineering

The poster child method for circuit evolution is multiplex automated genome engineering (MAGE). MAGE relies on the potentiation of oligo-mediated recombination in *E. coli* by the β ssDNA-binding protein of phage λ-red. Repeated transformation of oligonucleotide libraries, preferentially in an automated manner, leads to high-efficiency allelic replacement [111]. This efficiency enables the simultaneous modification of several targets, facilitating the evolution of entire genetic circuits, rather than single genes. Selection markers have improved the method by selecting for recombination-competent cells [112]. Bacterial conjugation was combined with MAGE to expedite recombination [113]. A similar oligo-mediated method has been reported in yeast, but with lower efficiency [114]. One of the most spectacular applications of MAGE was the genome-wide replacement of TAG stop codons in *E. coli* [115].

Evolutionary engineering of whole organisms

The evolutionary engineering of whole organisms implies complex traits involving intricate genetic networks. Even the smallest bacterial genomes represent enormous sequence spaces that evolutionary engineers can only scarcely sample. Unsurprisingly, targeted approaches have been devised at this scale to reduce complexity. For example, global transcription machinery engineering (gTME) is based on the assumption that mutagenesis of global regulators can be applied to simultaneously modulate numerous genes relevant to a phenotype of interest. With this approach, error-prone PCR of the σ^{70} transcription factor was enough to generate an *E. coli* strain

tolerant to sodium dodecyl sulfate and ethanol [116]. Variant libraries of TATA-binding proteins Spt15p and Taf25p were similarly applied to increase ethanol tolerance in *S. cerevisiae* [117].

Trackable multiplex recombineering (TRMR) is an agnostic variation on the MAGE method. In TRMR, genome-wide and oligo-mediated insertion of barcoded sequences followed by microarray analysis of selected mutants are used to identify loci relevant to the phenotype of interest [118]. In a hybrid approach, TRMR was used to identify targets, which were further engineered by MAGE, yielding *E. coli* mutants with enhanced tolerance to acetic acid, low pH and cellulosic hydrolysates [119].

Classical strain improvement methods are mostly agnostic [120]. They are based on the high throughput screening of mutants generated by random mutagenesis or adaptive laboratory evolution. Best mutants identified in these programs are then used in further rounds of mutagenesis and screening until the desired trait is achieved. These strain improvement programs are labour intensive and time consuming, typically leading to incremental improvements of 10% per year, as single mutants are selected and sequentially mutagenized [121]. Genome shuffling (GS) is a laboratory evolution method that addresses the limitations of classical strain improvement programmes (SIPs). Pioneered in the early 2000s [122,123], GS consists in the combinatorial evolution of complex phenotypes in whole organisms by genome-scale and recursive recombination of mutants. The work presented in this thesis is based on a genome shuffling experiment. This method is therefore reviewed in detail in the next section.

1.4b Evolutionary engineering by genome shuffling

Genome shuffling is a discontinuous and agnostic approach for the evolutionary engineering of whole organisms. Genetic diversity is introduced in a starting population of interest, and recursively recombined to rapidly generate new and potentially beneficial combinations of mutations. Intervening screening or selection steps may be applied at different points in the process to isolate improved mutants, which can be further recombined. This process can be performed repeatedly and stopped whenever the output is deemed satisfactory. Each time a mutant is isolated, it may be submitted to characterization.

The emphasis of this section is on the experimental design of GS experiments. Examination of the GS literature reveals that 14 years after the first genome shuffling reports [122], the number of studies in the field has exploded, with the last few years providing the largest publication harvests (Table 1.1). However, limiting the discussion of GS to examples from the published scientific literature would give an excessively narrow

Table 1.1 Published genome shuffling studies at the time this work was initiated

Recombination method	Type of phenotype	Product or agent	Source of diversity	Species	Study
			Prokaryotes		
Protoplast fusion	Production	ABE	NTG+UV+microwave	*Clostridium acetobutylicum*	Gao et al. (2012)
		Antimicrobial lipopeptide	NTG+UV+ion implantation	*Bacillus amyloliquefaciens*	Zhao et al. (2012)
		Avilamycin	Gamma rays	*Streptomyces viridochromogenes*	Lv et al. (2012)
		Ayamycin	EMS+UV	*Nocardia sp. ALAA 2000*	Gendy and El-Bondkly (2011)
		Epothilones	UV	*Sorangium cellulosum*	Gong et al. (2007)
		ε-poly-lysine	NTG+UV	*Streptomyces padanus, griseofuscus, graminearus, hygroscopicus, albulus*	Li et al. (2013)
		Hydroxycitric acid	NTG	*Streptomyces sp.U121*	Hida et al. (2007)
		Lactic acid	Nitrous acid+interspecies cross	*Bacillus amyloliquefaciens, Lactobacillus delbueckii*	John et al. (2008)
		Lipase	UV+DES	*Acinetobacter johnsonii*	Wang et al. (2012a)
		Natamycin	UV+5-BU	*Streptomyces gilvosporeus*	Luo et al. (2012)
		Rapamycin	UV	*Streptomyces hygroscopicus*	Chen et al. (2009)
		Spinosad	NTG+UV	*Saccharopolyspora spinosa*	Jin et al. (2009)
		Tylosin	NTG	*Streptomyces fradiae*	Zhang et al. (2002)
		Vitamin B12	NTG+UV	*Propionibacterium shermanii*	Zhang et al. (2010)
		Succinic acid	NTG+UV	*Actinobacillus succinogenes*	Zheng et al. (2013b)
		Daptomycin	NTG+UV	*Streptomyces roseosporus*	Yu et al. (2014)
	Resistance+production	1,3-propanediol	NTG	*Clostridium diolis*	Otte et al. (2009)
		Acid+lactic acid	NTG+UV	*Lactobacillus rhamnosus*	Wang et al. (2007)
		Acid+lactic acid	UV+DES	*Sporolactobacillus inulinus*	Zheng et al. (2010)
		Acid+propionic acid	UV+DES	*Proprionibacterium acidipropionici*	Guan et al. (2012)
		ε-poly-lysine+glucose	NTG+UV	*Streptomyces graminearus*	Li et al. (2012)
		Lactic acid	NTG+UV	*Lactobacillus rhamnosus*	Yu et al. (2008)
		Pristinamycin	UV	*Streptomyces pristnaespiralis*	Xu et al. (2008)
		Streptomycin+doramectin	NTG+UV	*Streptomyces avermitilis*	Zhang et al. (2013)
		Temprature+glutamic acid	DES+UV	*Corynebacterium glutamicum*	Zheng et al. (2012)
	Resistance	Acid	Adaptation + NTG	*Lactobacillus*	Patnaik et al. (2002)
	Resistance+degradation	PCP	NTG	*Sphingobium chlorophenolicum*	Dai and Copley (2004)
	Degradation	TNT	NTG+UV	*Stenotrophomonas maltophilia*	Lee et al. (2009)
	Antagonism	*S. scabies+P.infestans*	Interspecies/Interstrain cross	*Streptomyces melanosporofaciens, hygroscopicus*	Clermont et al. (2011)
			Eukaryotes		
Protoplast fusion	Production	Cellulase	EMS+UV	*Penicillium decumbens*	Cheng et al. (2009)
		Nuclease P1	Gamma rays	*Penicillium citrinum*	Wang et al. (2013)
		Taxol	UV	*Nodulisporium sylviforme*	Zhao et al. (2008)
		Xylanase	UV/NTG+NTG/EtBr	*Aspergillus sp. NRCF5*	El-Bondkly et al. (2012)
		Novel compounds	UV+NTG	*Tubercularia sp. TF5*	Wang et al. (2010)
		Ethanol	UV	*Scheffersomyces stipitis*	Shi et al. (2014)
	Resistance	NaCl	EMS	*Zygosaccharomyces rouxii*	Cao et al. (2010)
		Salt	EMS	*Hansenula anomala*	Cao et al. (2012)
	Resistance+production	Lactic acid+multistress+ethanol	UV	*Candida krusei*	Wei et al. (2008)
		Heat+ethanol	UV	*Saccharomyces cerevisiae*	Shi et al. (2009)
	Fermentation	Ethanol+glucose+xylose	Interstrain cross	*Saccharomyces cerevisiae*	Jingping et al. (2012)
	Reaction	Transglycosylation	Gamma rays	*Aspergillus niger*	Li et al. (2013)
	Resistance+production	Oxidative stress+protein	EMS	*Saccharomyces cerevisiae*	Li et al. (2011)
Sexual	Production	Ethanol	Not specified/Natural mutations	*Saccharomyces cerevisiae*	Tao et al. (2012)
			Genetic engineering+UV+EMS	*Saccharomyces cerevisiae*	Wang et al. (2012b)
			EMS	*Saccharomyces cerevisiae*	Liu et al. (2011)
	Resistance+production	Acetic acid+ethanol	UV	*Saccharomyces cerevisiae*	Zheng et al. (2011b)
		Ethanol	EMS	*Saccharomyces cerevisiae*	Hou (2010)
		Multistress+ethanol	Screened several strains	*Saccharomyces cerevisiae*	Zheng et al. (2011a)
			MBC	*Saccharomyces cerevisiae*	Zheng et al. (2013a)
	Resistance	Pulping effluent	UV	*Saccharomyces cerevisiae*	Pinel et al. (2011)
			UV	*Scheffersomyces stipitis*	Bajwa et al. (2010)
	Fermentation	Xylose	EMS	*Saccharomyces cerevisiae*	Demeke et al. (2013)
Whole genome transformation	Production	Ethanol	Whole genome transformation	*Scheffersomyces stipitis, Saccharomyces cerevisiae*	Zhang and Geng (2012)

picture of the potential of evolutionary engineering by GS. Patents and patent applications for GS give a much broader view of the method, and are a testimony to the scientific and technical possibilities of this technology (see [121] for a recent patent application). To provide a forward-looking vision of this technology, this section also covers projected yet never applied approaches to GS experiments.

Each subsection below deals with one of the steps of the GS process. The first subsection deals with the species and phenotypes that have been evolved by GS. The second section is dedicated to methods for acquiring diversity in GS experiments. The power of GS relies on the recursive recombination of this genetic diversity and methods for achieving this are discussed in the third subsection.

Organisms and phenotypes evolved

Enhancements in the production of small molecules are the main objectives of a significant proportion of published GS studies. In particular,GS often has been used to evolve *Streptomyces* species producing antibiotics and other molecules. Improvement in chemical productivity in a variety of other microbes is reported (Table 1.1). Applying GS to improve ethanol titers from *S. cerevisiae*, the workhorse of the fuel and beverage alcohol industries, is another common objective (Table 1.1). Improvement in the ability to ferment xylose [124] or co-ferment glucose and xylose [125] also was evolved by GS in recombinant *S. cerevisiae*. Since *S. cerevisiae* does not natively ferment xylose, those studies represent attempts at evolving a rationally engineered but sub-optimal xylose fermentation phenotype. They highlight the potential of GS at evolving the complex genetic changes that are often required to fine-tune engineered organisms. In related

studies, the pentose fermenting yeast *Scheffersomyces stipitis* was also evolved by GS, in one case in conjunction with *S. cerevisiae* [126,127]. Aside from its industrial relevance, the ease with which *S. cerevisiae* and other yeasts can undergo sexual recombination contributes to their popularity as GS organisms (discussed below). Improved production of organic acids is another common aim of GS (Table 1.1), as is the production of proteins and enzymes [128–130]. Similarly, *Aspergillus niger* was genome shuffled to enhance its capacity to perform transglycosylation reactions for the production of isomaltooligosaccharides [131]. While the capacity to perform a reaction rather than production titers was the primary aim of the study, the ultimate output remained increased production of industrially relevant molecules.

The second most important phenotype evolved by GS is resistance to physicochemical stresses induced for example by salt [132], acid [133], or toxic industrial byproducts [8]. Resistance phenotypes are often linked to production titers: increased resistance to a given compound produced by a microbe often leads to a concomitant increase in its production. In other cases, production conditions themselves are stressful, and improving an organism's resistance to those key stresses leads to elevated chemical production [134–139].

Production and resistance phenotypes account for the vast majority of GS studies. Other interesting phenotypes have been reported, and are worth mentioning. For example, the bacterium *Sphingobium chlorophenolicum*, known for its ability to mineralize pentachlorophenol (PCP), was genome shuffled for increased resistance to this toxic pesticide, leading to a parallel increase in degradation capacity [140]. A degradative phenotype was similarly evolved in *Stenotrophomonas maltophilia* for the bioremediation of trinitrotoluene [141]. In a study related to the antibiotics-producing properties of

Streptomyces melanosporofaciens, antagonism for potato pathogens was enhanced by screening for increased inhibition of bacterial growth [142]. Data suggested increased antibiotics production was not the primary cause of the improvement in bactericidal properties of *S. melanosporofaciens*.

Source of genetic diversity

The first step in any GS experiment is the creation or acquisition of a genetically diverse population to be used for breeding. This section reviews how this diversity can be obtained. Here, diversity is defined as genome-level sequence diversity. Accordingly, the amount of the diversity is defined as the number of unique genome sequences. Most studies artificially induce diversity in an otherwise homogeneous starting population. It is possible to exploit diversity that is naturally available, and the several methods for tapping natural genetic variations for genome shuffling will be discussed. Figure 1.1A schematizes the sources of diversity that can be used in GS studies.

How does the source and amount of genetic diversity in the starting population affect the success of GS? The amount of diversity in the initial population depends on its source: methods of mutagenesis that induce point mutations may generate pools as large as the target genome, while focused libraries (discussed below) will be smaller by definition. As the evolutionary engineering process progresses, the diversity profile of the evolved population changes accordingly. Recombination (discussed further below) generates new permutations, adding a layer of complexity to the diversity landscape while intervening selection steps weed out neutral and deleterious mutations. The latter can have profound effects on the evolutionary process. Stringent and frequent selection

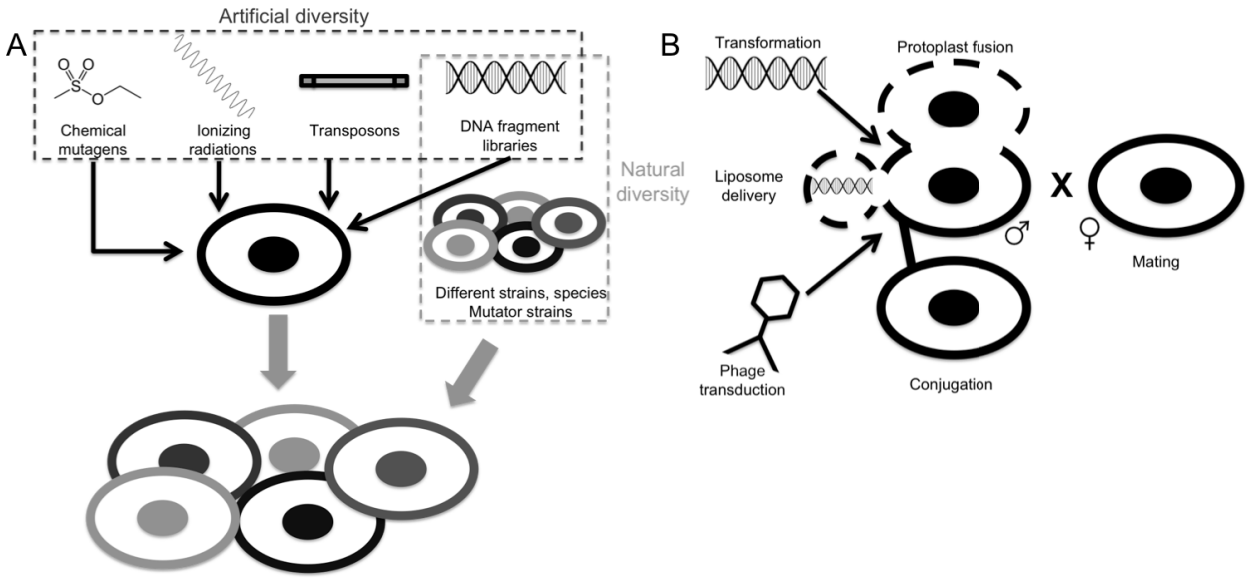


Figure 1.1 (A) Sources of diversity and (B) recombination methods for genome shuffling

may eliminate mutations that, if properly recombined with other mutations, could display beneficial epistatic interactions. It is therefore generally advisable to select permissively at the beginning of a GS experiment, increasing stringency as improved mutants are isolated. An excessively small pool may result from stringent selection, potentially leading to a hasty plateau in improvement. On the contrary, a selection pressure that is too permissive may slow down evolution by allowing neutral or deleterious mutations to clutter the pool.

Chemical or physical mutagens are most commonly used to induce genetic diversity in GS experiments (Table 1.1). It is generally assumed that random mutagenesis unbiasly covers the entire genome [120], in spite of evidence that nuance this assumption [143]. Frequently used chemical mutagens include nitrosoguanidine (NTG) and ethylmethylsulfonate (EMS), while ultraviolet (UV) light is widely used as an ionizing radiation to alter the genetic material of microbes. Mutagens are often used in combination. Although facultative, this is done to increase the diversity of induced mutations, as each mutagen has specific mechanisms of action that lead to different mutational bias [144–146] that vary between species [147].

Desired phenotypes can sometimes be found in nature, but not in the desired organism, while genes associated with a phenotype of interest may have natural homologs that can be exploited to accelerate the strain evolution process. Only a few examples report the use of natural genetic diversity as a starting point in evolving a strain by GS. In one example, nitrous acid mutagenesis was coupled to interspecies crosses to yield an organism with enhanced lactic acid production from starch [148]. An acid tolerant mutant of *Lactobacillus delbruecki* was crossed by protoplast fusion with

Bacillus amyloliquefaciens, a bacterium notable for its efficient starch utilization. In this example, the phenotypes sought were found in two distinct organisms, whose genomes were used as starting diversity. Natural diversity was similarly exploited in conjunction with mutagenesis to evolve higher production of ε-polylysine in five species of *Streptomyces*. In this study, the five species were separately evolved using UV and NTG mutagenesis as the source of diversity. Best isolates from all five species were subsequently submitted to interspecies hybridization, exploiting natural diversity to further improve productivity [149].

Clermont *et al.* used the diversity that naturally exists between two strains of *S. melanosporofaciens* and one strain of *S. hygroscopicus* and fused them to evolve an organism capable of controlling the proliferation of potato pathogens [142]. In another example, Zheng *et al.* compared 15 strains of *S. cerevisiae* for their resistance to multiple stresses and their ability to produce ethanol. Among those, two superior strains were identified and submitted to recursive mating to generate an enhanced hybrid [136]. In yet another example of exploiting natural diversity for GS, a strain of yeast was evolved to co-ferment glucose and xylose by transforming *S. cerevisiae* with a whole gDNA preparation from *S. stipitis* [127]. Isolates from this transformation where further shuffled by re-transforming with *S. cerevisiae* gDNA.

Genome-scale recombination

The choice of recombination method depends on several considerations. The organism will determine whether protoplast fusion, sexual recombination or other methods are feasible. For example, as will be discussed below, sexual recombination is only possible in organisms with characterized mating cycles. Other recombination

methods are tightly linked to the source of diversity. DNA fragment libraries, for example, cannot be delivered into cells by protoplast fusion. Protoplast fusion and sexual recombination account for nearly all published studies, while several other genome-scale recombination methods are suited for GS. Recombination methods discussed in this section are illustrated in Figure 1.1B.

Protoplast fusion is the most common recombination method in GS, and allows recombination between virtually any two or more cells. Protoplasts are cells stripped of their cell wall by digestion with lysosyme, zymolyase, or other cell wall-digesting enzymes. Fusion is promoted by submitting protoplasts to an electric pulse or by incubating them in the presence of PEG or surfactants that alter membrane fluidity. Recombination can then take place with genetic material from two or more cells enclosed within a single plasma membrane. Fusants are thereafter allowed to regenerate, and viable recombinants can be submitted to screening and selection. A common yet facultative step is protoplast inactivation. In this variation, protoplasts are rendered non-viable by exposure to UV light or heat. The only way for protoplasts to recover is to undergo fusion and recombination to repair fatal lesions. This approach prevents cells from the parent population from dominating the recombinant pool, as it results in failure of unfused protoplast to regenerate [150,151]. It may also induce further diversity via the action of an inactivating mutagen.

A potential advantage of protoplast fusion is that it enables poolwise recombination. In other words, any number of protoplasts can theoretically merge to yield a single progeny. This was demonstrated in *Streptomyces coelicolor* where a single round of protoplast fusion was sufficient to combine four different auxotrophic

markers into one cell, albeit with low efficiency [152]. This means that recombination between several mutants can occur at once, potentially accelerating the evolution process by creating more combinations and permutations. Protoplast fusion is less efficient in gram-negative than gram-positive bacteria. The periplasm and outer membrane of gram-negative bacteria harbour many important functions that are stripped away during protoplasting, making regeneration more challenging. In *E. coli*, the highest reported proportion of prototrophs from the fusion of two complementary auxotrophic populations is 0.7% [153]. A small number of GS studies in gram-negative bacteria has been published [140,141,154–156].

Sexual recombination is typically used when genome shuffling *S. cerevisiae* or other organisms with a sexual reproduction cycle. This approach takes advantage of the well characterized mating and sporulation cycles of yeast species to avoid some of the disadvantages of protoplast fusion. Using the natural ability of haploid yeast cells to fuse with one another circumvents the delicate task of generating protoplasts. The strategy takes advantage of the molecular meiotic machinery for recombination. Yet, sexual recombination has a number of disadvantages.

The most obvious is that it is limited to the subset of organisms with an easily manipulated sexual cycle. A second objection is that it only allows pairwise recombination, whereas other methods enable pool-wise recombination. In theory, GS based on sexual recombination could thus require more cycles than protoplast fusion to combine the same set of beneficial mutations into a single cell. However, using auxotrophic strains carrying four different auxotrophic markers, it was possible to generate 35% double auxotrophs after one round of mating [8], a proportion considerably

above what has been reported for protoplast fusion [122]. Other recombination methods can be envisioned in microbes. Mechanisms for horizontal gene transfer can be exploited to foster recombination between mutants. For example, in bacterial species with characterized fertility factors, conjugation may be an attractive way of effecting exchanges of genetic information. An illustration of the possibilities offered by horizontal gene transfer is conjugative assembly genome engineering (CAGE) [113]. CAGE allows the generation of precise chimeric bacterial genomes by decoupling genetic information from the conjugative machinery, and by using selection markers to precisely control the composition of the chimeric genome. This method has notably been used for the replacement of all TAG stop condons in *E. coli* [115]. Direct transformation may be exploited to deliver DNA libraries into cells. In their evolution of xylose fermentation in baker's yeast, Zhang and Geng [127] used existing transformation protocols to deliver entire gDNA preparations from *S. stipitis* and *S. cerevisiae* into baker's yeast. This approach has the advantage of being simple and straightforward. It is especially attractive for shuffling *S. cerevisiae* with related species because of the efficient homologous recombination capabilities of this organism.

1.4c Learning by doing: biological lessons from evolutionary engineering

Evolutionary engineering is used to change the properties of complex and incompletely understood biological systems. Aside from this applied objective, the evolved systems can be used to gain insight into their basic underlying biology. Recapitulation of the evolutionary process may provide additional information. In some cases, both the process and the outcomes can inform evolutionary biology in ways in which classical experimental evolution cannot. In this section, I review the contributions

of evolutionary engineering to basic biology. I first discuss studies that use evolutionary engineering to shed light on specific phenotypes with particular attention to examples from the literature on genome shuffling. The second subsection deals with real and proposed contributions of evolutionary engineering to the fundamental understanding of evolution.

Evolutionary engineering and the genetic architecture of complex traits

A relatively small percentage of genome shuffling studies are followed by investigations of the genetic architecture of evolved strains. Yet, investigating the genetic changes in GS mutants can prove rewarding for future rational and semi-rational approaches, such as reverse metabolic engineering or MAGE. Linking phenotype to genotype can prove interesting from a basic science point of view, contributing to our understanding of the systems biology of complex traits. Relatively few examples of such investigations can be found in the literature.

Examination of changes in gene expression profiles of GS-evolved strains can help identify the causes of phenotypic improvements. Targeted qPCR is the most frequent method used in GS studies interested in investigating the genetic architecture of engineered strains. This method is not agnostic and may not exhaustively explore the genetic basis of traits engineering by genome shuffling. It has still been successfully used to identify mechanisms that contribute to selected phenotypes. For example, after GS of *S. cerevisiae* for increased performance in very high gravity (VHG) fermentation, Tao and coworkers monitored changes in expression of genes involved in trehalose metabolism [157]. Trehalose is a disaccharide tightly associated with the stress response in yeast, and analysis of cells revealed that the evolved strain accumulated more

trehalose. Activity of trehalose-producing enzymes was augmented, while trehalose degradation activity was decreased. Quantitative PCR (qPCR) of genes involved in trehalose metabolism revealed a constitutive expression pattern in the GS-evolved strain, whereas induced expression was observed in the parent. Pulse-field gel electrophoresis revealed chromosomal rearrangements hypothesized to cause the changes in gene expression.

In another example, qPCR was used to probe levels of surfactin synthetase (*srfA*) gene expression in *Bacillus amyloliquefaciens* evolved by GS [158]. The shuffled strain, which produced a greater than 10-fold increase in the surfactin titer as compared to its parent, contained around 15 times more *srfA* mRNA. Also using qPCR, Jin and coworkers [159] investigated gene expression variations in a previously identified GS mutant of *Streptomyces pristinaespiralis* with increased pristinamycin production [160], concentrating their investigation on known components of the pristinamycin biosynthesis pathway. The expression of two of these genes (*snbA, snaB*), involved in distinct sections of the synthesis process, declined during prolonged fermentation in the parent strain, while it was maintained in the genome shuffled mutant. A third gene involved in resistance to the antibiotic was expressed earlier during fermentation by the mutant than by the parent. Restriction fragment length polymorphisms (RFLP) analysis was used to visualize chromosomal alterations potentially involved in pristinamycin yield improvements. Cloning of fragments present in the mutants but not in the parent strain identified two novel genes hypothesized to play a role in pristinamycin synthesis by *S. pristinaespiralis*.

A yeast species with the potential for flavor enhancement of soy sauce, *Zygosaccharomyces rouxii,* was genome shuffled to yield a strain with increased

resistance to high salt concentrations [161]. In a follow-up study, the causes of this increased resistance were investigated [162]. The Hog1 mitogen-activated protein kinase is known to activate genes involved in glycerol synthesis in *S. cerevisiae*, and it was reasoned that the *Z. rouxii* homolog (*ZrHOG1*) was a likely hit in the shuffled strain. Sequence comparison of the parental and mutant *ZrHOG1* genes revealed two nucleotide substitutions in the open reading frame, resulting in a single amino acid change, and a single base change upstream of the start codon. While the amino acid change in *ZrHOG1* suggested no obvious changes in protein function, estimation of transcription levels by qPCR pointed to elevated activity in the shuffled mutant. Furthermore, expression of mutant *ZrHOG1* in *S. cerevisiae* led to increased glycerol content.

More recent studies have used this targeted qPCR approach to corroborate hypotheses on the genetic architecture of GS mutants. For example, genome shuffling of *S. cerevisiae* for increased glutathione content was found to lead to an increase in expression of *GSH1*, encoding a key enzyme in the glutathione biosynthetic pathway [163]. Similarly, GS mutants of *B. amyloliquefaciens* with increased fengycin production expectedly displayed elevated expression of the fengycin synthetase gene [164]. Confirmation of straightforward hypotheses about gene expression in GS mutants is certainly useful, but cannot reveal novel or unexpected mechanisms about selected phenotypes. More open-ended approaches increase the probability of identifying such genes or pathways. These approaches may provide a wider, system-level picture of the changes taking place in evolved strains. In an early example, a shuffling mutant of *Proprionibacterium shermanii* with improved vitamin B12 production was submitted to a cursory proteomics analysis by 2D-gel electrophoresis. Comparison of the gel profiles of

the parent and enhanced strains identified 38 proteins with altered levels, including several enzymes involved directly or indirectly in vitamin B12 synthesis [165]. Other studies have explored the proteomics of GS mutants using mass spectrometry. Expanding on previous work with *B. amyloliquefaciens*, Zhao and coworkers provided a survey of proteins with altered expression in a surfactin-overproducing mutant [166]. A similar study in *P. acidipropionici* revealed 24 differentially expressed proteins in propionic acid tolerant mutants. The expression of nine of the genes encoding these proteins was further checked by qPCR. Overexpression of four of these genes in *E. coli* conferred increased tolerance to propionic acid.

In recent years, the cost of whole genome sequencing has decreased considerably. This enables the detailed investigation of genetic and transcriptional changes in genome shuffled recombinants. For example, strains of *S. cerevisiae* evolved for increased stress resistance and ethanol titers in VHG fermentation were characterized by a combination of physicochemical and genetic methods that include karyotyping, qPCR, array-comparative genome hybridization (aCGH) and RNA sequencing (RNA-Seq) [167]. This report is unique among GS studies, in that it used a chemical mutagen, methyl benzimidazole-2-yl-carbamate (MBC) that mainly induces large-scale structural rearrangements of the genome rather than point mutations. Pulse-field gel electrophoresis showed the evolved strains displayed altered karyotypes compared to their parent. Quantitative PCR showed copy number variations throughout the genomes of the mutants. Microarray-based comparative genomic hybridization of the most productive shuffled mutant identified the largest copy number variations on chromosomes 8, 11 and 14. RNAseq analysis confirmed the presence of several differentially expressed genes from those chromosomes. Mitotic cell cycle, small

molecule metabolism and stress response were the main functional annotations among differentially expressed genes. For example, catalase and trehalose metabolism genes showed increased transcription. This observation correlated with increased catalase and trehalose titers in all mutants tested. Two other genes with suspected roles in stress resistance (*YFL052W* and *SKN7*) had increased transcription in the most productive mutant. Their effect on the stress resistance phenotype was confirmed by overexpression in the parent background. Together, these results showed a clear link between copy number variation, transcription level and the stress resistance phenotype.

Whole genome sequencing of GS mutants of *Pachysolens tannophilus* was recently reported, revealing 60 SNPs shared by three different strains [168]. This highlights an important challenge of whole genome sequencing approaches, which is to assess the contribution of each SNP to the phenotype of interest. Because of synergistic effects and hitchhiker mutations, studying SNPs in isolation may not reveal the importance of each mutation. In addition, the systematic study of epistasis in relatively small pools of mutations rapidly becomes unmanageable because of the combinatorial explosion. In section 1.5 and in chapters 2, 3 and 4 of this thesis, I described strategies used by alumni of the Martin lab and myself to address these challenges.

Mechanisms of evolution probed by evolutionary engineering

To date, directed evolution studies in proteins have best described the dynamics that shape evolutionary engineering and its outcomes. Single molecule evolutionary engineering studies have illustrated phenomena familiar to evolutionary biologists, and amply discussed in the experimental evolution literature (Section 1.1). Notably, studies on proteins were a privileged model to study epistasis, showing how it both potentiates

and constrains evolution. The role of antagonistic pleiotropy, or evolutionary trade-offs, was also illustrated in directed evolution studies on single molecules.

During the evolution of proteins, epistasis has repeatedly been related to stability [169–171]. For example, the evolution of a cytochrome P450 fatty acid hydroxylase into a propane monooxygenase has revealed that increases in turnover rate for the desired reaction were correlated with concomitant decreases in thermostability [97,172]. This trend leads to situations where stabilizing mutations are required to enable catalytically beneficial mutations. This has been convincingly demonstrated in β-lactamases evolved to confer resistance to cephalosporins. In those studies, the effect of active site mutations directly depended on the presence of a specific stabilizing mutation [173,174]. Other studies on β-lactamases have demonstrated the influence of epistasis on protein evolution. A 10^5-fold increase in tolerance to the antibiotic cefotaxime is conferred by a specific variant of the enzyme carrying five precise point mutations. There are 120 trajectories leading to the accumulation of these five mutations from the ancestral enzyme. Testing each trajetory, Weinreich and coworkers demonstrated that only 18 trajectories led to a gain in fitness at each step, therefore identifying the restricted number of adaptive evolutionary trajectories [175]. Evolutionary engineering of proteins has shown the potentiating effect of genetic drift on the path of evolution. Indeed, because of epistasis, mutations neutral in a given background may contribute to fitness in another [176,177]. A cautionary comment must be added to this discussion of evolutionary dynamics of protein engineering experiments. The context in which proteins are selected in the laboratory differs from the conditions encountered in nature. This may in part explain the generalized observation of stability-mediated epistasis. Indeed,

while stability is generally not explicitly selected in the laboratory, it is most probably an important determinant of fitness in naturally evolving organisms. Generalization of evolutionary mechanisms observed in the lab must therefore proceed with caution.

Many of the methods presented in section 1.2a are still in their infancy, but they may allow researchers to pose questions that cannot be answered by traditional experimental evolution. Evolution of some fundamental cellular properties, like genome size and codon usage cannot be explored with existing organisms at practical timescales. For example, the ability to replace degenerate *E. coli* codons by MAGE has allowed their reassignment to non-natural amino acids [115]. This technique enables studies on the response of evolution to expanded or restricted amino acid repertoires, and may inform our knowledge on the origins of the standard genetic code. Reports on protein engineering with non-natural amino acids [178,179] and reduced genetic codes [180,181] have already started to shed light on these questions. On a more modest scale, one of the objectives of this thesis is to shed light on the dynamics of evolutionary engineering, in the hope that it will aid the design of future experiments, and contribute to our understanding of evolution.

1.5 The engineering problem: tolerance of yeast to lignocellulosic hydrolysates

One of the central aims of synthetic biology is to enable a sustainable bioeconomy. Biotechnologists propose to replace non-renewable fossil resources by biological feedstock to produce fuels and chemicals. To this end, sustainable feedstocks and reliable biocatalysts are required. The established bioethanol industry and burgeoning biochemical enterprises rely heavily on corn and sugarcane as carbon sources for their microbial biocatalysts [182]. This practice has raised concerns related to

sustainability, economic viability, and food security [183,184]. Alternative feedstocks are therefore required. Lignocellulosic biomass contains some of the most abundant biomolecules found on Earth. Its main component, cellulose, is a β-(1,4) linked polymer of glucose. Given the abundance and composition of lignocellulosic biomass, it would make for a marvellous source of carbon in industrial applications, such as biofuel production, were it not for its recalcitrance to hydrolysis. Liberation of lignocellulosic sugars often requires harsh conditions that produce a wide range of side products. The resulting hydrolysates are complex and heterogeneous mixtures of sugars, nutrients and inhibitory substances [185,186]. The toxicity of target molecules like fuel ethanol, stressful fermentation conditions [187,188] and the presence of inhibitors in industrial feedstocks [189–191] lead to a challenging environment for biocatalysts. The ability to withstand the stresses imposed by lignocellulosic hydrolysates could thus greatly aid a sustainable bioeconomy.

The workhorse biocatalyst of the biofuel and biochemical industries is the domesticated yeast *Saccharomyces cerevisiae*. This is explained by its age-old use for the production of ethanol, its GRAS (generally recognized as safe) status, and its easy culturing and genetic manipulation. The work reported in this thesis builds on an evolutionary engineering experiment aimed at increasing the tolerance of *S. cerevisiae* to a specific lignocellulosic hydrolysate known as spent sulphite liquor (SSL). This section therefore provides a brief introduction to SSL and its chemical composition. A review of prior knowledge on the response of yeast to the main stresses encountered in SSL is then provided. The final part of this section summarizes the evolutionary engineering experiments on which this thesis is based.

1.5a. Spent sulphite liquor: a model lignocellulosic hydrolysate

The classical framework for lignocellulose bioconversion consists of three steps: pretreatment, hydrolysis and fermentation [192]. Pre-treatment consists of physical and chemical transformation of lignocellulosic biomass to facilitate access of cellulose-degrading enzymes to their substrate. The second step is hydrolysis and requires the participation of cellulases. It liberates sugars for step three, fermentation, which is conducted by microorganisms such as yeast. While this model is adopted by the emerging biofuel industry, pulp and paper has been involved with lignocellulosic biomass for a long time. Pulping consists of the liberation of fibrous material by removal of the lignin matrix and as such, it is a well-established form of pre-treatment [193]. Harsh conditions of temperature and pH during pulping lead to the degradation of hemicelluloses, which consist of short polymers of acetylated pentose and hexose sugars. Pulping thus yields two products: pulp, which contains the fibrous cellulose material, and a liquor that contains delignifying chemicals and products of lignin and hemicellulose degradation. Liquor can be evaporated and the resulting solids burned for energy. An alternative is to use the sugars liberated by the hydrolysis of hemicellulose as substrate for fermentation.

More than 90% of pulp worldwide is made using the alkaline Kraft process. Because this process degrades the majority of the carbohydrate raw material to hydroxyl and dicarboxylic acids, liquor derived from its application are unsuitable for fermentation [193,194]. About 6% of worldwide pulp is produced following the acid sulphite process. It is performed at high temperature (125^{o}C-145^{o}C) in acidic conditions (pH 1-2), which results in partial hydrolysis of hemicellulose and release of fermentable monomeric sugars [193]. The sugar-containing by-product of sulphite pulping is called spent sulphite

liquor (SSL). Conditions of acid sulphite pulping are similar to those of the dilute acid process, currently the most economical pre-treatment method used for lignocellulosic ethanol production [195]. Therefore, SSL and common lignocellulosic hydrolysates show several similarities in terms of chemical composition [196]. Aside from hexoses and pentoses, SSL contains other products of lignin and hemicellulose degradation, namely lignosulfonates, acetic acid, and low levels of furan aldehydes and sulphite. Up to a quarter of the carbohydrate is in the form of oligosaccharides [193]. The nature of the wood used for pulping has a strong effect on the exact composition of SSL. The main determinant is whether SSL is derived from softwood (SW) or hardwood (HW). The main differences between softwoods and hardwoods concern the proportion and composition of the hemicellulose. There is slightly more hemicellulose and less cellulose in softwoods. Hemicellulose from hardwood contains more pentose monomers, and is more highly acetylated [193]. This is reflected in the composition of HWSSL, which contains higher concentrations of pentoses and acetic acid than SWSSL. As a consequence, SWSSL is much less inhibitory than HWSSL, and is easier to ferment for *S. cerevisiae*. Valorisation of SWSSL by fermentation is a well-established practice, tracing back to the early 20th century [193]. However, 50% of the annual SSL output is derived from hardwoods, and its biological conversation remains a challenge.

SSL contains diverse inhibitors that slow down growth and fermentation kinetics [197]. They can be divided into four groups: sugar degradation products, lignin degradation products, compounds derived from extractives and heavy metal ions [198,199]. Sugar degradation products include the furan aldehydes furfural (derived from pentoses) and hydroxymethylfurfural (HMF) as well as levulinic acid (both derived from hexoses). Furfural is generally inhibitory [199], but some organisms like *S. stipitis* display high

tolerance [200]. HMF is less abundant, notably in low hexose HWSSL because it is readily converted to levulinic acid at low pH. The combination of furfural, HMF and levulinic acid with compounds derived from lignin degradation present synergistic toxicity, forming secondary products [201]. Among lignin degradation products, sulphonated oligomers of lignin called lignosulphonates are highly prevalent [201]. Various other phenolics found in SSL, especially the low molecular weight derivatives of lignin, are proposed to have higher toxicity than furfurals. It is proposed that these phenolics compromise the integrity of the cytoplasmic membrane. Microbes exposed to these compounds display reduced growth and sugar assimilation [199,202]. Syringaldehyde and vanillic acid, model phenolic compounds, have documented effects on yeast growth and fermentative ability [197,203]. Extractives include acidic resins, tanninic, terpenic and acetic acids. They are generally less toxic than lignin degradation products [198], but high amounts of hydrolysable tannins in some hardwoods lead to the formation of compounds, like gallic acid and pyrogallol, with antifungal properties [201,204,205]. Acetic acid is a well documented inhibitor of yeast growth [199,206,207]. Protonated forms of the acid are lipophilic, allowing them to diffuse into cells. Dissociation in the cytoplasm leads to anion accumulation [207]. Heavy metals from the corrosion of pulping equipment also may contribute to inhibition [202,208].

The yeast *S. cerevisiae* is an efficient fermenter of hexose sugars, and displays high native tolerance to inhibitors found in lignocellulosic biomass, explaining its early adoption for the fermentation of SWSSL [209]. The pentose-rich HWSSL represents an extra challenge, due to its higher toxicity and because brewer's yeast does not efficiently metabolize xylose [210]. Indeed, *S. cerevisiae* does not possess specific xylose transporters, and expresses xylose metabolizing enzymes at low levels [210]. Genetic engineering of *S. cerevisiae* to enable pentose fermentation has been attempted [211].

Pentose-fermenting yeast, such as, *Candida shehatae*, *Pachysolen tannophilus* and especially the high ethanol producer *S. stipitis* have been proposed as alternatives [212]. Pentose-fermenting yeasts are highly sensitive to HWSSL inhibitors [209,213]. Adaptive evolution [214,215] and genome shuffling [126,168,216] have been used to increase their tolerance. Similarly, in section 1.6 further below, I describe how genome shuffling was used in the Martin lab to specifically address the SSL-toxicity challenge in *S. cerevisiae*.

1.5b. Response of yeast to stress in lignocellulosic hydrolysates

SSL can be detoxified by treatment with activated charcoal, ion exchange chromatography, overliming with calcium hydroxide, or solvent extraction. These methods present effective means of removing inhibitors from SSL, but are either too costly, or lead to degradation of industrial equipment [193]. Biological detoxification by white-rot fungus, such as *Trametes versicolor*, has been reported [198]. *Paecilomyces variotti* also has been used to effectively remove acetic, gallic acid and pyrogallol from HWSSL [213]. Those organisms notably produce lignolytic enzymes such as laccase, manganese peroxidase and lignin peroxidase that help to remove pollutants from lignocellulosic hydrolysates [198]. On-going efforts, in our and other labs, to increase stress tolerance of fermenting microbes aim to address the reduced fermentation performance due to SSL toxicity to *S. cerevisiae*. Engineering of yeast for tolerance to SSL requires the knowledge of stress response. Below, I survey the present knowledge of the yeast general stress response. I discuss aspects of the response to weak organic acid, and to oxidative and osmotic stresses, which are suggested by the results of this thesis to be the main determinants of yeast survival in SSL.

General stress response

In response to various environmental conditions, notably oxidative, pH, heat and osmotic stresses, yeast activates a non-specific cellular response termed the general (or global) stress response (GSR) [217–219]. It involves the upregulation of approximately 200 genes involved in various functions [220,221]. Upregulation of these genes is directed by the pentanucleotide CCCCT promoter element, referred to as the stress responsive element (STRE). Activation of STRE-controlled genes depends on the zing finger transcriptional activators Msn2p and Msn4p [218,219,222] which are activated in response to a variety of stresses [217,223]. The existence of this general stress response system explains why adaptation to one stress often confers tolerance to other unrelated stresses [224]. However, the GSR is transient [220], as exemplified by the rapid degradation of Msn2p after the onset of stress response [225]. Among processes directed by the GSR is the biosynthesis of trehalose, a major stress protectant in *S. cerevisiae* [226]. Trehalose accumulation appears intimately linked to the general stress response, as it requires Msn2p and Msn4p [227]. Trehalose confers stability to the plasma membrane [228] and to enzymes [229]. It ensures the proper folding of proteins [230] and can be used as a carbon source during starvation [231]. It acts as a general protectant, involved in response to heat, toxic chemicals, and osmotic, oxidative and ethanol stresses [188]. Recent studies suggest that, with the notable exception of extreme desiccation, the protective effect of trehalose may be indirect rather than an immediate result of its physicochemical properties [232,233]. Elements of the GSR partially overlap with mechanisms of stress response to specific inhibitors, as described below.

Weak organic acid stress

Degradation of acetylated sugars during pulping leads to an accumulation of organic acids, foremost acetic acid, in SSL. The specific effects of organic acid stress are distinct from those of simple low pH conditions. Protons do not traverse the plasma membrane: their effect is confined to the extracellular domains of surface-exposed lipids and proteins [234]. On the contrary, at sufficiently low pH, weak organic acids diffuse freely through the plasma membrane in undissociated form [207], or may enter by facilitated diffusion through Fsp1p [235]. The neutral pH of the cytosol leads to organic acid dissociation, trapping the proton and carboxylate anion inside the cell. By carrying protons to the cytosol, organic acids may affect intracellular pH and deregulate cell homeostasis. In addition, as acidity increases, so does the hydrophobicity of organic acids, which affects lipid organization of membranes and increase their permeability [236,237]. This in turn leads to an increase in proton influx and to a loss of transmembrane potential [237]. In this respect, lipophilic acids do not cause changes in intracellular pH, while more hydrophilic ones like acetic acid do affect cytosolic acidity [238].

The organic acid stress response is inducible. Growth of unadapted yeast upon acid exposure is delayed. Once induced, adapted cells can be inoculated in fresh medium containing organic acids without further halting their growth. This adaptation is mediated by several mechanisms. A primary physiological response to organic acids is the pumping of protons to the extracellular environment by Pma1p and their sequestration into the vacuole by the V-ATPase [239]. Multidrug efflux pumps have been implicated in organic acid detoxification [236,240]. Another organic acid tolerance strategy is to reduce permeability by altering cell wall structure [241–243], internalizing diffusion channels [235], and changing plasma membrane composition [244]. Starvation and energy

limitation is observed in yeast stressed by organic acids, leading to the activation of the TOR pathway [245]. Chemogenomic screening using the yeast deletion mutant collection identifies vacuolar acidification, intracellular trafficking, and ergosterol biosynthesis as core functions of the general organic acid response [234].

Transcriptional responses are specific to each organic acid, but tend to overlap within similar cellular functions. The only gene common to all organic acids encodes the multidrug transporter Tpo3p [234]. A number of key regulons have been implicated in organic stress response. The GSR and its key regulators Msn2p/Msn4p are activated by the presence of weak organic acids [246]. Transcriptional regulators Pdr1p/Pdr3p, generally involved in drug resistance [247], are specifically activated by lipophilic acids [240,248]. The repressor Rim101p was first characterized as a regulator of the alkaline pH response [249] and was involved in the assembly of the cell wall [250]. However, evidence shows a broader role in the regulation of pH homeostasis [251] as shown by its activation by propionic acid stress [243]. Protein War1p is a nuclear protein activated by organic acid stress, possibly by direct binding of carboxylate anions [252]. Only one of its targets, the plasma membrane ABC transporter Pdr12p, has been identified as protective against organic acids [246].

Acetic acid is the most prevalent organic acid encountered in SSL. Cytosolic acidification is proposed to be its main physiological effect [238,253]. It reduces import of histidine, leucine, and glucose, possibly because of reduced ATP levels and direct inhibition of transporters by low pH. Exposure to acetic acid also induces programmed cell death [254]. An important mechanism to cope with acetic acid is the limitation of influx through internalization and vacuolar degradation of Fsp1p [235]. This is regulated by Hog1p, which associates with Fsp1p in the absence of acetic acid. Acid stress activates

Hog1p transiently, causing phosphorylation of Fsp1p. This triggers ubiquitination, internalization and degradation in the vacuole [255]. Several studies have pointed to regulation of carbohydrate metabolism and ribosome biogenesis as key processes involved in the specific response to acetic acid [234]. This notably echoes reports that acetic acid challenge leads to extensive degradation of rRNA in yeast [256]. A key regulator in the response to acetic acid is transcriptional regulator Haa1p [257] which also responds to lactic acid [241]. In the absence of stress, Haa1p is phosphorylated and actively exported from the nucleus via interactions with Msn5p. Dephosphorylation and rapid localization of Haa1p to the nucleus is proposed to be an important mechanism of acetic acid response [238,258]. The protective effect of this regulator and many of its targets against acetic acid has been demonstrated [236,259]. The most notable protective targets of Haa1p are *SAP30* (involved in the Rpd3L histone deacetylase complex) and *HRK1* (involved in regulation of membrane transporters) [257]. Other targets of Haa1p conferring tolerance to acetic acid include efflux antiporters Tpo2p and Tpo3p, and the cell wall associated Ygp1p, which is proposed to mediate remodelling of the cell envelope to prevent acetate re-entry [236].

Oxidative stress

The oxygen paradox describes the contradictory effects of oxygen on cell physiology. While it is essential to respiratory metabolism, derivatives of molecular oxygen, termed reactive oxygen species (ROS), can cause cellular damage, contribute to ageing and lead to cell death. The main ROS produced in the cell is the superoxide anion, the majority of which results from electron leakage of the mitochondrial electron transport chain. It is actively detoxified by superoxide dismutases, which converts it to

molecular oxygen and hydrogen peroxide. Both superoxide and hydrogen peroxide are relatively unreactive, but they are readily converted to the highly reactive hydroxyl radical [260]. Reaction of superoxide anions with nitric oxide radicals can lead to the formation of reactive nitrogen species. The highly reactive hydroxyl radical is the most damaging ROS, and reacts indiscriminately with most biomolecules [261]. Yeast possesses enzymatic and non-enzymatic mechanisms to cope with ROS normally produced by respiration. Oxidative stress is induced when ROS production exceeds the normal detoxification capacity of the cell [261]. A reduction of cytosolic pH may trigger oxidative stress response [187]. Damage caused by ROS affects nucleic acids [262,263], proteins [264] and lipids [265]. Oxidative damage to DNA has been implicated in mutagenesis, and in carcinogenesis in humans [266]. Elevated oxidative stress was demonstrated to induce interchromosomal recombination, suggesting an activation of DNA repair mechanisms [267,268]. ROS notably induce the formation of 8-hydroxyguanine, which leads to GC to TA base pair transversions if left unrepaired [269]. ROS damage also leads to the cross-linking and breakdown of proteins [270]. The main targets of ROS are unsaturated lipids, involved in a process of autocatalytic peroxidation [261]. Results of ROS attacks on lipids include membrane cross-linking and the formation of peroxyl radicals. Peroxyl radicals can further react with other cellular components to form lipid hydroperoxides, which are highly toxic to yeast via extensive membrane damage [271]. The breakdown of lipid hydroperoxide generates highly reactive aldehydes, which can cause damage to proteins by carbonylation [272].

Yeast synthesizes a diversity of antioxidant molecules. Non-enzymatic defences include glutathione, ubiquinol, D-erythroascorbic acid, flavohaemoglobin, metallothioneins, polyamines, trehalose and ergosterol [188]. Glutathione is a tripeptide of

glutamate, cysteine and glycine. In its reduced form, it is the main thiol found in yeast. Glutathione scavenges ROS by direct or enzyme-mediated oxidation of its thiol group to yield a dimeric disulphide form. Oxidized glutathione is actively recycled to the reduced form by Glr1p, a specific NADPH-dependent reductase [273]. Ubiquinol, the reduced form of coenzyme Q, is an important component of the mitochondrial respiratory chain. Soluble in lipid membranes, it is an important antioxidant, capable of inhibiting lipid peroxidation [274]. It is encountered in the membranes of several organelles [275]. D-erythroascorbic acid is highly similar to ascorbic acid (vitamin C). Much like its sister compound, it reacts readily with ROS and lipid peroxy radicals. Deletion and overexpression studies on genes involved in the synthesis of D-erythroascorbic acid have demonstrated its protective effects against oxidative stress in yeast [276].

Trehalose has been implicated in the protection of cells against ROS. Its protective effects have been shown in the face of various oxidative stress-inducing conditions [277,278]. It may protect against protein carbonylation [277] and lipid peroxidation [278]. However, rapid degradation of trehalose is needed for normal metabolism to resume. Trehalose notably inhibits glutathione reductase, involved in reducing oxidative damage [279], and impaired trehalose degradation leads to oxidative stress sensitivity [280].

Enzymes protective against oxidative stress include superoxide dismutases, catalase, glutaredoxin, thioredoxin, cytochrome c peroxidase, glutathione peroxidase, glutathione reductase and peroxidases [188]. There are two superoxide dismutases in yeast: Sod1p located in the cytoplasm, and Sod2p located in the mitochondria [281]. As mentioned earlier, these enzymes catalyze the conversion of superoxide anions to molecular oxygen and hydrogen peroxide. Oxidative stress-inducing drugs and respiratory metabolism both induce the expression of superoxide dismutase in yeast [281].

Mutants lacking either enzymes display hypersensitivity to oxygen [282]. Yeast contains two catalase genes: a cytoplasmic (*CTT1*)[283] and a peroxisomal version (*CTA1*)[284]. These enzymes catalyze the conversion of hydrogen peroxide into water and oxygen. Accordingly, their removal leads to sensitivity to hydrogen peroxide [285]. Thioredoxins and glutaredoxins are sulfhydryl proteins that scavenge ROS by oxidation of the adjacent cysteines of a CXXC motif. Both are returned to the reduce state at the expense of NADPH, via the action of thioredoxing reductase or, in the case of glutaredoxins, direct reaction with glutathione. Thioredoxin and glutathione peroxidases are involved in the detoxification of bulky hydroperoxides, which do not react with catalase. In addition to thioredoxin and glutathione, peroxidases can have cytochrome c and ascorbate as reducing co-factors. Yeast glutathione peroxidases are peculiar in that they are attached to membranes via phospholipid anchors and are capable of reducing lipid hydroperoxides that are esterified to membranes [286]. Metallothionein are small proteins, encoded by genes *CUP1* and *CRS5*, responsible for the scavenging of toxic metal ions, especially copper ions [287]. They also have an antioxidant role. Their overexpression can notably suppress the effects of superoxide dismutase mutants [261]. They further have been shown to scavenge superoxide anions and hydroxyl radicals [261].

Transcription factors critical to the oxidative stress response are Yap1p, Skn7p, and possibly the osmotic stress regulator Hog1p. All of these regulators have been involved in the activation or expression of specific antioxidants [187,288,289]. In presence of oxidative stress, formation of specific disulphide bridge in Yap1p by a specialized peroxidase leads to a conformational change that triggers accumulation in the nucleus [290,291]. There, Yap1p notably activates genes involved in glutathione and thioredoxin metabolism [292–294]. Skn7p was originally identified in screens for mutants hypersensitive

to hydrogen peroxide [295]. It is a known regulator of catalase-encoding *CTT1* and superoxide dismutase-encoding *SOD1* [293,294]. The Hog1p mitogen-activated protein kinase (MAPK), primarily involved in osmotic stress, is known to regulate oxidative stress response genes, either directly of via interaction with transcriptional repressor Sko1p [296,297]. Furthermore, *hog1Δ* mutants display sensitivity to hydrogen peroxide [298]. Yet, Hog1p is not responsive to oxidative stress or non-osmotic stresses [296], suggesting an indirect role in the response to ROS.

Osmotic stress

Osmotic stress is defined as an imbalance between extra and intracellular osmolarities, which negatively impacts cell physiology [299]. It may arise in hypo or hyperosmotic environments, leading to water influx and efflux, respectively. The latter effect is expected in SSL, which displays high solute concentration. Mechanisms used by yeast to cope with osmotic stress are classified into two general categories: osmotolerance and osmoadaptation.

Osmotolerance describes the innate ability of a given yeast strain to withstand osmotic stress, and is conferred by general traits such as membrane structure [300], vacuole function [301] and residual trehalose levels [230]. Osmoadaptation refers to inducible responses to specific stimuli, which can be either chronic or acute. Chronic response is a signal transduction alteration that alters the level of specific proteins, while acute response is invoked upon sudden exposure to high external osmolarity [301]. The general physiological mode of adaptation to osmotic stress consists in the intracellular accumulation of solutes to balance osmolarity of the cytosol with that of the extracellular environment. These solutes, that do not otherwise alter cell physiology, are termed

compatible solutes [302]. Glycerol is the main compatible solute in *S. cerevisiae* during osmotic shock, and it is required for survival in hyperosmotic conditions [303]. Upon exposure to high osmotic potential, glycerol is quickly accumulated [303,304]. Increases in glycerol content lead to enhanced osmotic stress tolerance [303,305–308]. High osmolarity activates the high osmolarity glycerol pathway (HOG) [309]. The output of this pathway is the phosphorylation of Hog1p, a MAPK that stimulates the hyperproduction and hyperaccumulation of glycerol as compatible solute. Notable target genes of Hog1p are *GPD1* and *GPP2*, which encode for key enzymes of the glycerol biosynthesis pathway. Hog1p also regulates activity of the aquaglyceroporin Fsp1p to reduce glycerol efflux [310]. Two parallel sensory pathways, downstream of membrane sensors Sln1p and Sho1p, activate the Hog1p kinase. As seen earlier, the HOG pathway has a wider implication in the stress response. It is required for the activation of Msn2p and Msn4p, suggesting that Hog1p plays a role in the regulation of the general stress response via STREs [296]. Because osmotic stress is associated with changes in cell size and turgor pressure, reorganization of the cell wall and cytoskeleton are elements of the osmotic stress response [311,312].

1.6 Genome shuffling of *Saccharomyces cerevisiae* for increased tolerance to spent sulphite liquor

Previous members of the Martin lab used genome shuffling to increase the tolerance of *Saccharomyces cerevisiae* to hardwood spent sulphite liquor. From an initial pool of random UV mutants, recursive rounds of mating and selection identified recombinant mutants with increased resistance to SSL. Individual mutants were extensively characterized, measuring their tolerance to various inhibitors and assessing

their ability to grow and produce ethanol in presence of high concentrations of SSL. Among characterized mutants, a single isolate showing high inhibitor tolerance and ability to produce ethanol from SSL was identified [8]. This mutant, named R57, was further characterized by whole genome sequencing and RNAseq. These analyses identified single nucleotide differences and transcriptional changes between R57 and its wildtype parent. Frequency of mutations identified in R57 was determined by amplicon sequencing on frozen pools of mutants from several time points of the evolutionary engineering process. Additional endpoint mutants were genotyped at the R57 mutant loci [9]. This work is the foundation for the research presented in this thesis. In this last section of the introduction, I provide a detailed overview of this prior work. I explain the evolutionary engineering process, mutant strain characterization, and sequencing results, finishing with a presentation of RNAseq results.

1.6a Genome shuffling by recursive mating

Starting points for the evolutionary engineering of SSL-tolerant *S. cerevisiae* were prototrophic haploid derivatives of the common laboratory strain CEN.PK. Pools of random mutants were generated by UV irradiation for both mating types (*MATα* and *MATa*). Mutants with increased SSL-tolerance were selected using gradient plates. Those were agar plates with increasing concentrations of SSL from one end to another. Yeast growing at SSL concentrations above maximum wildtype levels was scraped off the gradient plate. The two pools of selected haploid mutants were mixed for mating, and resulting diploids were selected on SSL gradient plates as described above. This first pool of selected diploids was submitted to sporulation and mating, generating a first pool of shuffled mutants. Four additional cycles of selection, sporulation and mating

were performed. An aliquot of the propagated shuffled mutants was reserved for each of the five rounds of genome shuffling and kept at -80^{o}C. The genome shuffling experiment is described in detail in a previous publication [8] and is summarized in Chapter 2.

1.6b Characterization of genome shuffling mutants

Wildtype parental strains and mutants from various time points of the evolutionary engineering experiments were compared for their ability to survive and thrive in the presence of undiluted of SSL. This revealed an increase in average tolerance to SSL as the experiment progressed. While parental strains rapidly died in SSL, selected mutants from later rounds of genome shuffling survived exposure and even proliferated in undiluted SSL. Furthermore, three GS mutants from rounds 3 (R311) and 5 (R57 and R511) were able to thrive and produce ethanol during repeated serial passage in undiluted SSL. Resistance of these mutants to acetic acid, HMF, high salt concentrations, hydrogen peroxide, and high osmotic pressure was measured by spot assays. The mutants showed notable increases in tolerance to acetic acid, hydrogen peroxide and sorbitol-induced osmotic stress. The same mutants displayed moderate increases in tolerance to high salt, but almost none to hydroxymethylfurfural. Strain R57 displayed the highest overall tolerance to inhibitors and produced the highest amounts of ethanol from undiluted HWSSL.

1.6c Sequencing of SSL tolerant mutants

The genome of strain R57 was sequenced to identify mutations responsible for high SSL tolerance. Comparison with the CEN.PK reference genome [296] revealed 21 single nucleotide changes affecting 17 genes involved in various functions. These SNPs

are listed in Table 1.2. A SIFT score was assigned to each SNP to predict their effect on protein structure and function, and most substitutions were suggested to be neutral, with the notable exception of *ubp7-T2466A* and *art5-G454T*. Two independent mutations were mapped to genes *GDH1* and *MAL11*, suggesting repeated selection at critical loci.

Amplicon sequencing of the R57 mutant loci was performed on the pools of mutants from four time points of the evolutionary engineering experiment. The chosen time points corresponded to the initial selection of UV mutants, and rounds 1, 3 and 5. This revealed strong prevalence of mutant alleles *ubp7-T2466A* and *art5-G454T*, both involved in protein homeostasis. Several mutations displayed similar and seemingly correlated frequencies with *ubp7-T2466A*, suggesting a hitchhiking phenomenon. Additional isolates from round 5 were genotyped by amplicon sequencing at the R57 mutant loci. This analysis confirmed the strong prevalence of the *ubp7-T2466A* mutation, prompting the generation of a single mutant in the wildtype background. This mutant displayed increased tolerance, indicating that the *ubp7-T2466A* allele enhanced fitness of the parental strain in the presence of SSL.

1.6d Transcriptional changes in mutant strain R57

Transcriptional profiles of the wildtype and R57 strains revealed 149 differentially expressed genes, 131 of which were upregulated in the mutant. Of those, *NRG1* was the only mutated gene displaying increasing transcription. *NRG1* encodes a transcriptional regulator initially involved in carbohydrate metabolism. It was previously involved in response to acetic acid stress [217]. The largest group of upregulated genes were implicated in translation regulation, ribosome biogenesis, and monosaccharide

Table 1.2 SNPs identified by whole genome sequencing of strain R57

Mutation	Chr	Amino acid substitution	Genotype
ssa-C91A	I	(Q31K)	Hetero
nrg1-C137A	IV	(P46Q)	Homo
ste5-C512T	IV	(S171F)	Hetero
ste5-T2649C		Silent	Hetero
aro1-C1283T	IV	S428F	Hetero
aro1-C1284T		Silent	Hetero
dop1-A40T	IV	N14Y	Hetero
art5-C454A	VII	L152I	Hetero
pbp1-T(-191)A	VII	Non-coding	Hetero
mal11-C310T	VII	P104S	Hetero
mal11-T482A		M161K	Hetero
ubp7-T2466A	IX	N822K	Homo
gsh1-A(-73)T	X	Non-coding	Hetero
tof2-C2141T	XI	S714L	Hetero
ynl058c-A7G	XIV	K3E	Hetero
sgo1-C575A	XV	S192Y	Hetero
nop58-A(+25)T	XV	Non-coding	Hetero
gdh1-C47T	XV	S16F	Hetero
gdh1-T68G		F23C	Hetero
fit3-C(+43)T	XV	Non-coding	Hetero

metabolism, as previously observed under acetic acid stress [217]. Cell wall organization, metal ion-binding and membrane-associated genes figured prominently among upregulated transcripts in R57**.**

1.6e Summary of foundational work

Previous evolutionary engineering efforts in the Martin lab have successfully produced strains of *S. cerevisiae* tolerant to elevated concentrations of spent sulphite liquor. The resulting mutants show tolerance to a wide range of stresses, and can ferment sugars found in SSL to ethanol. Among these, a superior strain was identified and called R57. Twenty one SNPs was detected in this strain by whole genome sequencing. Targeted probing of evolutionary dynamics, genotyping of round 5 mutants by amplicon sequencing, and direct phenotypic testing of a single mutant identified key mutations in genes *UBP7* and *ART5*. This suggested protein homeostasis as a critical cellular process selected by genome shuffling. Repeated and independent selection of mutations in genes *MAL11* and *GDH1* pointed to these genes as important for the SSL-tolerance phenotype. Finally, upregulation of transcripts involved in ribosome biogenesis and carbohydrate metabolism, notably in mutated gene *NRG1*, argued for a determinant role for acetic acid stress response in resistance to SSL.

1.7 Conclusion

Experimental evolution and evolutionary engineering allow researchers to follow evolution in almost real time in the controlled environmental of the laboratory. They further enable the generation and identification of organisms with novel or enhanced traits that are often too complex to engineer rationally using current knowledge. Modern

high-throughput sequencing technologies permit the dissection of complex genetic architectures, and enable the tracking of evolutionary dynamics leading to phenotypes of interest. This sheds light on the molecular and cell biology of complex traits, and provides invaluable insight into fundamental mechanisms that shape evolution.

Spent sulphite liquor is a toxic byproduct of the pulp and paper industry. It serves as a model lignocellulosic hydrolysate and prospective feedstock for the production of biofuels and biochemicals. In our lab, genome shuffling was previously used for the evolutionary engineering of *S. cerevisiae* strains tolerant to this complex mixture of carbon sources and microbial inhibitors. Highly tolerant mutants were successfully isolated and characterized, showing potential for the fermentation of SSL to ethanol. A strain with superior tolerance and fermentation characteristics was isolated, and it was dissected in detail by transcriptional and whole genome sequencing. The dynamics of the evolutionary engineering experiment were previously probed in a targeted way.

In this thesis, I report an exhaustive survey of this evolutionary engineering experiment by whole population whole genome sequencing. I further dissect the genetic architecture of enhanced SSL-tolerance, identifying key mutations and epistatic interactions. Together, evolutionary and phenotypic data provide a detailed understanding of the evolutionary forces that shaped the outcomes of the genome shuffling experiment, and further our grasp of the selected phenotype.

2. Materials and Methods

with excerpts from:

Biot-Pelletier D, Pinel D, Larue K, Martin VJJ (2016) The impact of historical contingency on the outcomes of evolutionary engineering. Manuscript in preparation.

Biot-Pelletier D, Martin VJJ (2016). Seamless site-directed mutagenesis of the *Saccharomyces cerevisiae* genome using CRISPR-Cas9. J Biol Eng. 10:6.

2.1 Genome shuffling by recursive population mating

In order to generate strains of *S. cerevisiae* with increased tolerance to spent sulphite liquor, haploid strains CEN.PK 113-1A (*MATa*) and CEN.PK113-7D (*MATα*) were submitted to genome shuffling. The experiment was described in detail in a previous publication[8] and is summarized by Figure 2.1. Briefly, pools of *MATα* and *MATa* haploid mutants were generated by UV irradiation, and spread onto SSL gradient agar plates (described in Figure 2.1B). Yeast growing at SSL concentrations above maximum wildtype levels was scraped off the gradient plate, and an aliquot from both pools was reserved for sequencing and kept at -80^{o}C. The two pools of selected haploid mutants were mixed for mating (as described in section 2.12), and resulting diploids were selected on SSL gradient plates as described above. This first pool of selected diploids was submitted to sporulation (as described in section 2.11) and mating, generating a first pool of shuffled mutants, propagated by overnight incubation at 30^{o}C in YPD. Four additional cycles of selection, sporulation and mating were performed. An aliquot of the propagated shuffled mutants was reserved for each of the five rounds of genome shuffling and kept at -80^{o}C.

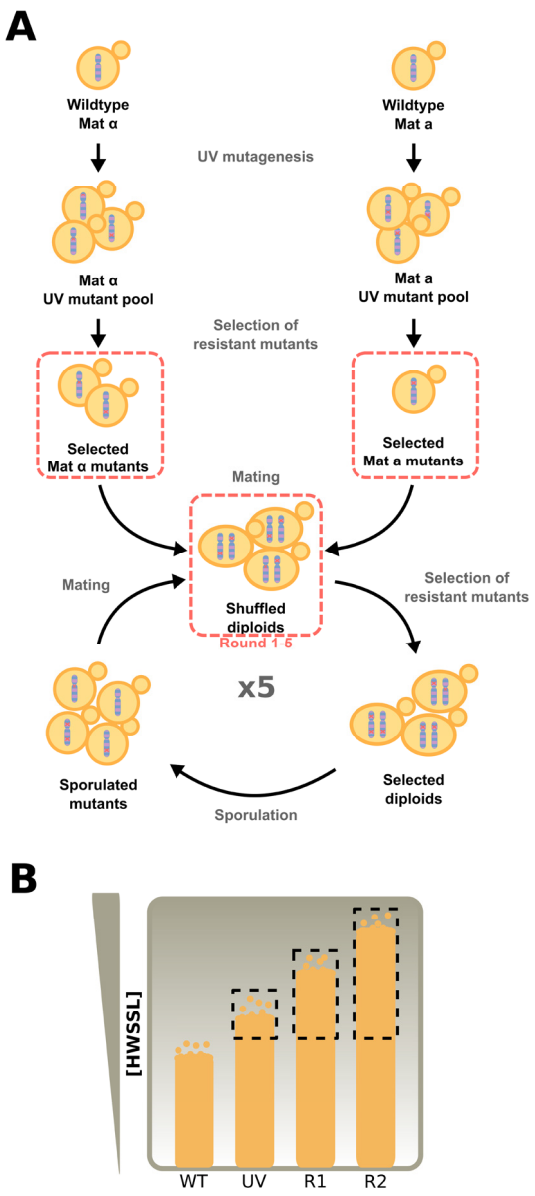


Figure 2.1 Outline of the genome shuffling experiment. (A) Wildtype haploid cells of both mating types were submitted to UV irradiation to generate a pool of haploid mutants. Mutants with tolerance to SSL superior to their wildtype ancestors were selected, and the pools of tolerant mutants were mated, generating diploids. Tolerant diploids were selected, and submitted to sporulation. Resulting spores were mated at random, effecting the shuffling of mutations. Five cycles of diploid selection, sporulation and mating were performed. Dashed red boxes indicate time points submitted to population genome sequencing. (B) After mutagenesis and each round of shuffling, selection for increased SSL tolerance was done by scraping mutants growing above wildtype level on agar plates containing a gradient of HWSSL concentration.

2.2 Pooled population genome sequencing

To obtain an exhaustive survey of mutations selected by genome shuffling and to measure their frequency throughout evolution, I performed whole genome sequencing on population samples at several time points of the evolutionary engineering experiment. Seven samples from six time points corresponding to haploid UV mutants from the initial gradient plate selection (UVa and UVα) and to diploid shuffled mutants from each of the five rounds of genome shuffling (R1 to R5) were selected for sequencing, as indicated in Figure 2.1. Aliquots of shuffled mutants reserved for sequencing were thawed at room temperature and incubated in 5 ml YPD at 30^{o}C for 1 hr. Cells were then harvested by centrifugation, and incubated for 1 hr at 37^{o}C in 50 mM Tris-HCl pH 8.0,10 mM EDTA, 5% (v/v) 2-mercaptoethanol, 200 U/ml yeast lytic enzyme (MP Biomedicals). Yeast genomic DNA was extracted from the resulting mixture using the DNeasy Blood and Tissue Kit (Qiagen) according to the instructions of the manufacturer. DNA concentration was measured using the QuantiFluor dsDNA System (Promega).

Genomic DNA purified from the mutant pools was submitted to the McGill University and Génome Québec Innovation Centre for library preparation (TrueSeq, Illumina) and sequencing (HiSeq 2500, 100 bp paired-end reads). Each mutant pool was sequenced on a separate lane of a HiSeq chip to maximize depth of coverage.

2.3 Quality control of sequencing data and read alignment

Quality control of raw sequencing data was performed using FastQC [313], and overlapping read pairs were merged with PEAR [314]. Alignment to the CEN.PK113-7D reference genome [315] was done using *bwa mem* [316], and performed separately for

overlapping and non-overlapping reads. Output SAM files for overlapping and non-overlapping reads were merged with the MergeSamFiles utility in Picard Tools [317]. Picard was next used to add read groups, sort reads, then mark and remove duplicates prior to indel realignment with the Genome Analysis Toolkit [318–320]. Alignment metrics were extracted using Picard Tools.

2.4 Base error model

To distinguish genuine SNP calls from sequencing errors, a statistical error model was built from the sequencing data itself. To this end, read counts for each genomic position were obtained from the indel-realigned output using bam-readcount [321], with minimum mapping quality and minimum base quality both set at 30. The base error model was computed from read counts using a methodology inspired from Barrick and Lenski [64]. Iterating over every nucleotide position in the reference genome, the probability of reading a base call given its quality score and the reference base was computed, generating a series of 4x4 matrices. Only positions with depth of coverage comprised between the alignment's 1st and 99th percentile were considered, to avoid regions of low coverage and because unusually high coverage was taken as an indication of incorrect alignment. The error model was then refined to correct for regions of the genome with an above average proportion of mismatches compared to the reference. To this end, the reference genome was divided into 2000 nucleotide regions, and local reference mismatch rates were compared to the genome wide rate. For regions with above average mismatch rate, the local mismatch rate was divided by the genome wide mismatch rate to yield a correction factor, which was used to multiply

mismatch probabilities predicted by the base error model. Uncorrected probabilities were used for regions with average and below average mismatch rates.

2.5 Primary SNP calling

Using the same criteria as for error model construction, I iterated through every position in the reference genome, extracting base counts and comparing them to error model predictions. Probability of the observed read counts given the error model was calculated using either a multinomial test (if depth of coverage < 170) or G-test (depth > 170). This probability was multiplied by the number of nucleotide positions in the reference genome to yield the expected number of positions with the same read counts found by chance in the dataset. If this number was less than one, a SNP was called.

2.6 SNP filtering

The raw SNP calls were filtered to account for systematic errors, misalignments, and various biases encountered in the data. The list of SNP calls was compared to our previous re-sequencing of CEN.PK113-7D [9] to filter parental mutations and potential systematic errors. The pileup was extracted using samtools mpileup [322] for each mutant position in the list of SNP calls, with a minimum quality threshold of 30 for both base reads and alignment.

From the pileup, I filtered SNP calls for quality score bias. SNP calls for which the mean quality score of mismatch reads was lower than that of reference matching reads were submitted to a Kolmogorov-Smirnov test comparing the quality score distributions of match and mismatch reads. Significantly ($p < 0.05$) different quality score distributions led to SNP call rejection.

SNP calls were further filtered for strand bias. Fisher's exact test was applied to determine if match and mismatch reads were similarly distributed between the forward and reverse strands. SNP calls for which mismatch reads showed a significant (p-value < 0.05) strand bias were filtered out.

I next reasoned that mutations were unlikely to arise independently and be encountered simultaneously in both haploid mutant pools. I tested whether SNP calls encountered in both pools (Fisher's exact test, $p < 0.05$) had the same proportion of mismatches. If they did, a systematic error was suspected, and further testing was performed. Mismatch count was compared between each haploid mutant pool and the round 1 to 5 data. Fisher's method was used to combine the probabilities of the 10 tests. Highly significantly different ($p < 0.001$) proportions of mismatches between the 7 datasets led to SNPs being conserved. Otherwise, they were discarded.

Filtering was completed by visual inspection of the alignments, using integrative genomics viewer [323,324]. High number of mismatches or secondary reads mapping in the vicinity of a SNP call led to its rejection. Mutations that were either synonymous or that could not be detected at any of the samples time points were excluded from downstream analyses.

Indel calling was attempted from pileup, applying an error model and filtering as described for SNPs. I did not detect indels using this strategy.

2.7 Mutant allele frequency, strength of selection and hierarchical clustering of evolutionary trajectories

Further analysis was performed to reveal and compare the evolutionary trajectory of each mutation and extract a measure of their selection. The proportion of mismatch

reads at given genomic coordinates was considered to reflect the frequency (p) of the associated mutant allele within the sequenced pools. The frequency of all mutations was extracted for the seven sequenced pools. The frequency of each mutation was necessarily zero in one of the haploid pools, allowing the identification of its mutant pool of origin (*MATα* or *MATa*). Furthermore, because the two haploid mutant pools represented a single evolutionary time point, their frequency was averaged to obtain the pre-shuffling allele frequencies in the full haploid mutant pool, or UV time point (i.e. $p_{UV}=(p_{\alpha}+p_{a})/2$). Allele frequencies of the round 1 to 5 pools (R1-R5) each represented time points of their own.

Strength of apparent positive or negative selection was estimated from allele frequency changes between time points ($\Delta p_{t1\text{-}t2}=p_{t2}/p_{t1}$). Allele frequency changes above one indicate frequency increase, below one indicate a decrease, and equal to one mean the absence of fluctuation. Annex I details the theoretical reasons that led to this methodological choice. The geometric mean frequency change (M) was used as a synthetic measure of selection, to smooth the effect of proportionally large frequency changes often observed between time points UV and R1. Hierarchichal clustering of evolutionary trajectories was performed by running the *clustermap* routine of the *Seaborn* Python library [325].

2.8 Multiple sequence alignment, homology modeling of *S. cerevisiae* Gdh1p and protein structure analysis

To explore the relationship between mutations identified in *GDH1* and their impact on Gdh1p, I performed multiple sequence alignment between the mutant protein sequence and close homologs, and mapped the substitutions on a homology model.

Multiple sequence alignment of Gdh1p homologs was performed with Clustal Omega [326–328]. Homology modeling was performed automatically on the ModWeb server [329]. The best scoring model was based on the structure of *Plasmodium falciparum* glutamate dehydrogenase (52% identity, PDB ID: 2BMA [330]). Exploration of the structure and mapping of Gdh1p substitutions was done with Pymol [331].

2.9 Site-directed mutagenesis of yeast by CRISPR-Cas9

To test their impact on the SSL tolerance phenotype, point mutations previously identified by sequencing of mutant R57 [9] and a subset of SNPs identified by population sequencing were reintroduced into wildtype backgrounds (CEN.PK113-1A, *MATα* or CEN.PK113-7D, *MATa*) or reverted to wildtype in mutant R57 as described in detail in Section 3.5. Briefly, gRNAs were designed to introduce double-stranded breaks (DSBs) in the vincinity of mutant positions. DSBs were then repaired by homologous recombination with donor DNAs replacing the 20 nucleotide protospacer sequence by a heterologous stuffer sequence. This stuffer was targeted by a second gRNA, and the resulting DSBs were repaired by a stuffer-free donor DNA that carried the point mutation. The result was the seamless introduction of single nucleotide changes [332].

2.10 Growth curves of mutants in presence of spent sulphite liquor and other inhibitors

To measure the tolerance of mutants to SSL, 3 ml of YNB 1% (w/v) glucose were inoculated with single colonies from freshly streaked YPD plates, and grown for 48 hrs at 30^{o}C with shaking. Cultures were then centrifuged for 5 min at 3220 x g, and the pellets were suspended in 3 ml of spent sulphite liquor. After overnight incubation in

SSL at 30°C with shaking, cells were washed three times in 10 mM sodium citrate pH 5.5, then suspended in the same buffer at a density of ($8x10^6$) cell/ml. Five µl of the citrate suspensions were used to inoculate 175 µl of YNB 1% (w/v) glucose supplemented with varying concentrations of SSL (0-85% (v/v)), acetic acid (0-203 mM) or hydrogen peroxide (0-4.9 mM), in the wells of a 96-well plate (Costar 3595, Corning). Plates were then incubated at 30°C with shaking in a Tecan Sunrise absorbance reader, measuring absorbance at 595 nm every 20 min for 8620 min. Because inhibitors may affect growth rate, lag time, maximum cell density or any combination thereof, I decided to compute the area under the growth curves and use that number as a measure of growth.

2.11 Sporulation of R57 diploid cells

Backcrossing of R57 with the wildtype required sporulation of this diploid strain. R57 cells from a 5 ml overnight YPD culture were spread on pre-sporulation medium (0.8% (w/v) yeast extract, 0.3% (w/v) peptone,10% (w/v) dextrose, 2% (w/v) agar) and incubated 24 hrs at 30°C, then 24 hrs at room temperature. Pre-sporulation agar was scraped and cells were washed twice in 10 mM sodium citrate pH 5.5 before they were spread on sporulation medium (1% (w/v) potassium acetate, 0.1% (w/v) yeast extract, 0.05% (w/v) dextrose, 2% (w/v) agar). The sporulation plate was kept at room temperature, and tetrad formation was monitored with a microscope every two days. After 6 days, I estimated that a sporulation efficiency of >50% was satisfactory. Cells were scraped and suspended in 5 ml water supplemented with 100 U of yeast lytic enzyme (MP Biomedicals) and 10 µl of 2-mercaptoethanol, then incubated overnight with gentle shaking at 30°C to digest the cell wall of asci and vegetative cells. Next, 5 ml

of 1.5% (v/v) IGEPAL was added to the digest, which was then incubated on ice for 15 min. The solution was then sonicated for three cycles of 30 sec followed by 120 sec cool down periods. Sonicated spores were harvested by centrifugation at 12 000 x g for 10 min, and suspended in 5 ml 1.5% (v/v) IGEPAL. The sonication procedure was repeated, the spores were once again recovered by centrifugation, and suspended in 250 µl of YPD broth. The spore suspension was examined under the microscope to confirm ascus disruption and lysis of vegetative cells.

2.12 Backcrossing of R57 spores with wildtype haploids

To generate derivatives of R57 with random combinations of mutations, fresh R57 spores were mixed in approximately equal amounts (as assessed by A_{660}) with CEN.PK113-1A cells (*MATα*) from an overnight YPD culture. The mix was spotted on YPD agar, and incubated overnight at 30^{o}C. The following day, the mating spot was scraped, suspended in YPD, spotted again on YPD agar, then incubated 48 hrs at 30^{o}C. Mating and schmoo formation were monitored daily by microscopy. The mating spot was scraped once again, and submitted to sporulation as described above. Resulting R57 x 1A spores were spotted on YPD agar and allowed to germinate for 48 hrs. Cell paste from the germination spot was picked with an inoculation loop and streaked for single colonies on six fresh YPD agar petri dishes. After 48 hrs, 86 colonies were picked, grown overnight in 3 ml YPD, then mixed with glycerol at a final concentration of 37.5% (v/v) and frozen at -80^{o}C. The R57 backcrossing strategy is summarized in Figure 3.5 of Chapter 3.

2.13 Preparation of yeast genomic DNA template for PCR

To generate gDNA templates for PCR amplification of mutant loci and mating type determination, the following DNA extraction procedure was adopted. Three ml cultures were inoculated from freshly streaked colonies and grown overnight with shaking at 30^{o}C. Cells from 1.5 ml of culture were transferred into a microcentrifuge tube and centrifuged at maximum speed in a tabletop microcentrifuge. Supernatant was discarded, and the pellet was suspended in 250 µl of 50 mM Tris-Cl pH 8.0 supplemented by 20 U of yeast lytic enzyme. This lysis solution was incubated at 37^{o}C for 1 hr, then vigorously mixed with 250 µl of 200 mM NaOH, 1% (w/v) SDS. After a 5-min incubation at room temperature, the mixture was neutralized by addition of 350 µl of 3M potassium acetate pH 5.5 then mixed by vortexing. Debris was removed by a 10-min centrifugation at maximum speed in a tapletop microcentrifuge. Supernatant was transferred to a new microcentrifuge tube, and mixed with 600 µl of 2-propanol to precipate DNA, which was pelleted by centrifugation at maximum speed for 10 min in a tapletop microcentrifuge. Supernatant was discarded and the pellet was air dried for 15 min at room temperature. One hundred µl of water was added to the dried pellet, and the tube was vortexed before incubation for 15 min in a water bath set at 55^{o}C. The DNA solution was vortexed again to ensure proper solubilization, then stored at -20^{o}C.

2.14 Determination of mating type and ploidy by PCR

Mating type and ploidy of R57 segregants were determined to aid downstream genotyping. For each strain, two PCR reactions were performed. A common *MAT* locus reverse primer (5’ AGTCACATCAAGATCGTTTATGG 3’) was used with a forward primer specific for either the *MATα* (5’ GCACGGAATATGGGACTACTTCG 3’) or *MATa*

(5’ ACTCCACTTCAAGTAAGAGTTTG 3’) genes. Composition of the PCR reactions was as follows: 1 µl gDNA template, 0.5 µM each primer, 200 mM dNTPs, 1X KCl Taq buffer, 1.25 U Taq DNA polymerase in a 50 µl reaction volume. The cycling conditions were 95^{o}C for 3 min, then 35 cycles of 95^{o}C for 30s, 55^{o}C for 30 s, 72^{o}C for 1 min, followed by a 10 min incubation at 72^{o}C and a final hold at 10^{o}C. Detection of an approximately 500 bp product by 0.8% (w/v) agarose gel electrophoresis in both reactions indicated a diploid strain, while haploidy was inferred if only one reaction yielded a band. Mating type of haploids was deduced from the identity of the positive reaction.

2.15 Genotyping of R57 backcrossed isolates by amplicon sequencing

To determine the set of mutations carried by each R57 backcross, massively parallel amplicon sequencing was performed. For each strain, the 18 mutant loci from strain R57 were amplified with corresponding gene specific primers. Both forward and reverse specific primers consisted of a common 5’ heel sequence (forward: 5’ CGTTCAACCTTGTCCAACAGTG 3’ reverse: 5’ GAAGCGATGACTCGAGCGTATT 3’) and a 24-28 nucleotide gene specific sequence at the 3’ end. PCR reactions contained 0.5 µM primers, 200 µM dNTPs, 1X high fidelity Phusion buffer (Thermo Fisher), 1.5% (v/v) DMSO and 1U of Phusion High Fidelity DNA polymerase (Thermo Fisher) in 50 µl. Genomic DNA template, prepared as described above, was diluted 1:10 and 1 µl of that dilution was added to the 50 µl PCR reaction. Cycling conditions were as follows: 98^{o}C for 30s, then 35 cycles of 98^{o}C for 10s, 55^{o}C for 20s, 72^{o}C for 4s, followed by a 5 min final extension at 72^{o}C, and hold at 10^{o}C. Ion torrent sequencing adapters and barcodes were added in a second PCR. This second reaction was performed using a common reverse primer consisting of the ion torrent P1 adapter (5’

CCTCTCTATGGGCAGTCGGTGAT 3') and the reverse heel sequence. The 96 forward primers consisted of the ion torrent A adapter (5' CCATCTCATCCCTGCGTGTCTCCGACTCAG 3'), a unique 8 nucleotide barcode, and the forward heel sequence. Composition of this second PCR reaction was the same as for locus specific amplification, using 1 µl of a 1:10 dilution of the first reaction as template. Cycling conditions were: 98^{o}C for 30s, then 35 cycles of 98^{o}C for 10s, 70^{o}C for 20s, 72^{o}C for 6s, followed by a 5 min final extension at 72^{o}C, and hold at 10^{o}C. All primers used for genotyping of R57 backcrossed mutants are listed in Annex II. Success of all PCR reactions was checked by electrophoresis on a 2% (w/v) agarose gel with 1X TBE running buffer. Crude PCR products carrying the same barcode were pooled. The pool was loaded on a 2% (w/v) agarose gel and submitted to electrophoresis using 1X TBE running buffer. All material between 100 bp and 400 bp was excised with a scalpel, and DNA was purified using the GeneJet gel extraction kit (Thermo Fisher). DNA concentration in each pool was measured by A_{260} in a Tecan m200 plate reader using a NanoQuant 200 plate. Equal amounts of DNA from each pool were mixed to make the sequencing library. The DNA concentration of this pool was measured on a Qubit 2.0 fluorometer (Life Technologies).

Ion torrent sequencing template was prepared from our pooled library with an Ion PGM Template OT2 400 Kit according to the manufacturer's instructions. Sequencing was performed with an Ion PGM sequencer using the Ion PGM Hi-Q sequencing kit and an Ion 316 Chip v2 BC, all following manufacturer's protocols and instructions.

2.16 Analysis of amplicon sequencing data

Analysis of raw amplicon sequencing data was required to assign genotypes to R57 backcrosses. The unaligned BAM output from the Ion PGM was converted to FASTQ with SamToFastq in Picard Tools [317]. Reads were sorted to distinct FASTQ files according to their barcodes, and then trimmed of adapter and barcode sequences using a custom Python script. A reference FASTA file was built by extracting and concatenating the sequences of the target genes, plus 1 kb upstream and downstream, from the CEN.PK113-7D reference genome [315]. The reads in FASTQ format were aligned to the reference FASTA file using bwa mem [316]. Picard was next used to add read groups and sort reads. Read counts for each of the SNP positions were extracted with bam-readcount [321] setting both minimum mapping quality and minimum base quality at 30.

Genotypes were called from read counts using a custom Python script. Heterozygotes were distinguished from homozygotes by assuming that in homozygotes, the most frequent base call would have a frequency of 0.997, and all other base calls 0.001 each. A G-test was performed to test whether the base count distribution differed significantly from this assumption. If it did, a heterozygous genotype was called. Otherwise, the identity of the most frequent base call was checked. If the most frequent base call was the reference base, a wildtype genotype was called. Otherwise a homozygous mutant genotype was called. In strains previously identified as haploid, heterozygosity calls were corrected, and genotype was called based on the identity of the most frequent base call. All genotype calls were reviewed visually.

2.17 Multiple linear regression model of SSL tolerance in R57 backcrossed mutants

Multiple linear regression was performed to assign a contribution for each R57 mutation to the SSL tolerance phenotype. Distinct linear models were computed from the haploid and diploid mutant datasets, using distinct methodologies. The full haploid dataset consisted of 52 data points, each corresponding to a single strain, including the wildtype CEN.PK113-1A strain. The diploid dataset was smaller, with 36 data points including strain R57. The general equation for both linear models is of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \cdots + \beta_k x_k$$

where *y* represents the area under the growth curve in 85% (v/v) SSL for a given strain, each x_i represents the genotype of the strain at locus *i* of *k* mutant loci of interest, and each β_i the linear coefficient associated with locus *i*, corresponding to its contribution to the SSL tolerance phenotype. Term β_0 corresponds to the value of *y* when all x_i=0 which in haploids is the area under the growth curve in 85% (v/v) SSL for wildtype cells. A straightforward biological interpretation for β_0 in diploids cannot be given.

Correlation was suspected (R^2 >> 0.5) in both datasets between pairs of mutations located at close coordinates on the same chromosomes. Those were the *aro1-C1283T* and *aro1-C1284T*, *ste5-C512T* and *ste5-T2649C*, and *gdh1-C47T* and *fit3-C(+43)T* mutations. To avoid issues with collinearity, *aro1-C1283T*, *ste5-T2649C* (silent) and *fit3-C(+43)T* (hypothesized to be inconsequential because not detected by population sequencing) were removed from my regression analyses. I also expected correlation between the *aro1* and *ste5* mutations, but decided to keep one representative of both, seeking identification of the best predictor of SSL tolerance by regression. Other correlations were suspected, especially in the diploid dataset. Unlike

the mutations mentioned above, I did not have biological reasons to exclude these variables from my analyses, and hypothesized that these correlations were coincidental.

For haploids, each mutant position has two possible genotypes, such that x_i=0 for wildtype and x_i=1 for mutant, and each β_i is a direct estimate of the effect of mutation *i* on the phenotype. My dataset thus included 18 potential explanatory variables. Using the *curve_fit* wrapper from the *Scipy* Python library, I estimated the linear coefficients for all combinations of *k*=1 to *k*=18 explanatory variables. For each of the resulting linear models, I computed R^2, variance of the residuals, Mallow's C_p statistic, and p-values that each β_i=0. For each number of *k* variables, I chose the model that minimized C_p, then plotted this minimum C_p against *k*. I found that *k*=8 explanatory variables minimized both C_p and variance of residuals, and corresponded to the point where R^2 reached a plateau. Those three observations suggested that fitting with those selected eight variables maximized predictive power of the model while minimizing overfitting. P-values for the linear coefficients were all less than to 0.01.

For diploids, a single binary variable cannot be applied to each locus, because three possible genotypes are possible (wildtype, heterozygous mutant and homozygous mutant). Potential cases of heterozygote superiority or inferiority mean that a linear relationship between the phenotype and the number of mutant alleles at each locus cannot be assumed. Instead, three binary variables with associated linear coefficients are assigned to each locus for each potential genotype. The linear model is modified such that the $\beta_i x_i$ become $\beta_{iwt}x_{iwt}+ \beta_{ihet}x_{ihet}+ \beta_{ihom}x_{ihom}$. For example, a heterozygous mutant would have x_{iwt}=0, x_{ihet}=1 and x_{ihom}=0. This procedure multiplies the number of variables by three. This means that the thorough approach used with the haploid dataset could not be applied to diploids. Indeed, the number of available data points (36) is

smaller than the number of potential explanatory variables (54). I considered that regression with all potential explanatory variables was not possible without serious risks of overfitting. Without additional and biologically motivated constraints to impose onto the data, I chose not to attempt model fitting in those conditions. The consequence is that I could not compute the C_p statistic for diploid models. However, in haploid models, I noticed a strong correlation between the C_p statistic and the mean residual sum of square (MRSS) computed when performing leave-one-out cross validation. I chose to use MRSS as a proxy for C_p in diploid models. Further, the systematic fitting of linear models for all combinations of k variables among 54 potential explanatory variables rapidly becomes impractical for relatively low values of k. I chose to proceed in a stepwise manner. I started by identifying the model with k=1 that minimized MRSS. I then sampled all models with k=2 that included the variable chosen in the first step, and chose the one that led to the greatest reduction in MRSS. I incrementally added variables to the model in the same way until k=35. At each step, I monitored p-values for each β_i. If any variable had a p-value > 0.01, I rejected it and computed all alternative models where that single variable was replaced, and selected the one that minimized MRSS. I plotted MRSS against k and selected the model that minimized MRSS.

I used an *ad hoc* methodology to consider potential interactions between pairs of mutations, and proceeded in the same way for haploids and diploids. I only considered pairs of variables already included in the final interaction-free models. Each pair of variables (i and j) was taken into account by adding a term to the linear equation of the form $\beta_{ij}x_ix_j$. Interaction models were built using the same stepwise methodology followed for the diploid models. Thus, interaction terms were added incrementally to the existing, interaction-free model, choosing the one out of all possible pairs that led to the greatest

reduction in MRSS. Addition of interaction terms was stopped when MRSS either reached a plateau or started increasing, indicating overfitting.

The selected linear models were validated using a variety of methods. Residuals were plotted against each of the model variables and against the measured values of *y* to detect fitting bias. I further tested normality of the distribution of residuals using a Kolmogorov-Smirnov test. For each data point (i.e. each strain) I computed Cook's distance to identify outliers or influential points. I did detect an outlier in each model, but in the absence of strong biological reasons or other observations to reject them, I chose to keep those points as part of the natural variability of the data.

2.18 Measurement of ROS accumulation in wildtype and *gsh1* mutant *S. cerevisiae* cells

To test the effect of the *gsh1*-A(-73)T mutation on ROS accumulation, a flow cytometry-based fluorescence assay was performed. Wildtype (CEN.PK113-1A) and *gsh1*-A(-73)T haploid cells were inoculated in triplicate in 3 ml YNB 1% (w/v) glucose from fresh colonies and grown 24 hrs at 30^{o}C with shaking. The following day, cell density was estimated by recording A_{660} of the cultures. An equal amount of cells (10^6) was inoculated intro fresh 3 ml YNB 1% (w/w) glucose cultures, and incubated at 30^{o}C with shaking. After 24 hrs in this inhibitor-free medium, 10^6 cells from each culture were assayed for ROS accumulation, as described in the next paragraph. The rest of the cells were centrifuged at maximum speed in a tabletop microcentrifuge, and suspended in undiluted SSL pH 5.5. Cells were incubated in SSL for 16 hrs at 30^{o}C with shaking. After this incubation, cells were washed twice with 10 mM citrate pH 5.5, then suspended in the same buffer.

Samples of 10^6 cells were taken for measurement of ROS accumulation. Cells from the same sample were used to inoculate fresh YNB 1% (w/v) glucose 70% (v/v) SSL cultures at a cell density of (3.33×10^5) cells/ml. Cells were incubated in this high SSL medium for 48 hrs, withdrawing samples for ROS measurement at the 24 and 48 hr time points.

ROS accumulation in yeast was measured by staining with CellROX Deep Red reagent (Molecular probes) according to manufacturer instructions. Briefly, 10^6 cells washed in 10 mM sodium citrate pH 5.5 were suspended in a 5 µM solution of CellROX Deep Red reagent and incubated for 30 min at 37^{o}C. Staining solution was then removed and cells were washed three times in 10 mM sodium citrate pH 5.5. Cells were suspended in 300 µl 10 mM sodium citrate pH 5.5, and mean cell fluorescence was measured with filter FL3 of an Accuri C6 flow cytometer.

3. Results

with excerpts from:

Biot-Pelletier D, Martin VJJ (2016). Seamless site-directed mutagenesis of the *Saccharomyces cerevisiae* genome using CRISPR-Cas9. J Biol Eng. 10:6.

Biot-Pelletier D *et al.* (2016) The impact of historical contingency on the outcomes of evolutionary engineering. Manuscript in preparation.

3.1 Whole population sequencing of mutant pools from an evolutionary engineering experiment

To gain insight into the evolutionary dynamics underlying genome shuffling and to expand our survey of mutations selected by evolutionary engineering, I investigated the heterogeneous genotype of seven mutant pools from six time points of the GS experiment [8,9] (Figure 2.1A). The experiment consisted of UV-induced mutagenesis of haploids, selection of SSL-tolerant mutants, and five rounds of iterative mating and selection. For sequencing, I selected time points covering the entire length of the experiment, including both pools of selected haploid mutants, and shuffled mutants from each round of genome shuffling (R1-R5). Each mutant pool was submitted to genome re-sequencing, generating upwards of 300 million reads, for an average of 40 billion nucleotides per sample with a mean base quality score of 35.07 (Table 4.1). The 100 nucleotide reads were aligned to the CEN.PK113-7D reference genome [315], which is one of the parent strains used in the experiment. Mean depth of coverage oscillated between 712x and 1551x, for an average of 1091x, enabling the detection of mutations represented in less than 1% of the population (Table 4.1)

Table 4.1 Population sequencing and alignment metrics

Pool	Number of reads	Read length	Mean base quality	Mean depth of coverage	Depth of coverage 1st percentile	Depth of coverage 99th percentile	Mean mapping quality
MATa	335377268	100	35.28	986	408	1827	47.23
MATα	488893080	100	35.09	1126	471	2172	44.46
R1	408135762	100	34.60	1182	460	2208	47.64
R2	311653312	100	35.15	712	295	1400	41.11
R3	388469930	100	34.73	807	328	1699	39.04
R4	382880430	100	34.46	1551	594	4008	48.58
R5	392463442	100	36.17	1270	549	2529	46.98
Average	386839032	100	35.07	1091	444	2263	45.01

A base error model was used for calling SNPs, allowing distinction of genuine mutations from sequencing errors (described in Chapter 2). Filtering and manual examination of mutation calls resulted in a list of 188 SNPs (Table 4.2). Mutations that were silent or escaped detection in at least one of the six sampled time points were excluded from downstream analyses, restricting the list to 105 SNPs. The nature, frequency, annotations and apparent selection of all SNPs is described in detail in Annex IV. Furthermore, Annex IV unambiguously distinguishes raw (188 SNPs) and filtered (105 SNPs) mutations. The list of mutations suggests a significant mutational bias, especially in favor of the equivalent G>A and C>T transitions. The majority of mutations previously identified by sequencing of the R57 shuffling mutant [9] were detected by whole population sequencing, and a significant proportion of those mutations are found at medium to high frequency throughout the genome shuffling experiment. Mutations in *TOF2*, *DOP1* and *FIT3* were previously identified in mutant R57, but could not be detected. Their frequencies are hypothesized to fall below my detection threshold.

3.2 Clustering reveals cohorts of mutations with highly correlated evolutionary trajectories

The list of 105 single nucleotide substitutions can be divided into two unequal groups based on their pool of origin (Figure 3.1). Mutations arose during mutagenesis of either the *MATa* or *MATα* parent strains, and sequencing detected them in either one or the other. More mutations are found in the *MATa* pool (70 mutations) than in the *MATα* pool (30 mutations). Five mutations were detected in all reads of their original mutant

Table 4.2 List of all SNPs detected by population genome sequencing

Mutation name	Chromosome	Position	Mutation name	Chromosome	Position
abd1 A-79T	II	691767	*hem12 A348T*	IV	552205
aca1 A400C	V	241101	*hsh49 A-56T*	XV	912763
aim9 T1683C	V	321641	*hvg1 C-204T*	V	228985
air1 C153T	IX	210767	*idp3 T184C*	XIV	615005
aro1 C1283T	IV	705763	*ifh1 G-305T*	XII	585797
aro1 C1284T	IV	705764	*ina1 G-200C*	XII	950953
aro7 G260A	XVI	675369	*ioc4 G732A*	XIII	356114
art5 G454T	VII	626635	*ipt1 T67A*	IV	591276
atg26 A512T	XII	533884	*ira2 C7081T*	XV	178150
aus1 T32C	XV	353895	*isw1 C3280T*	II	708255
avl9 C1806G	XII	375434	*itt1 T1169A*	XIII	138718
avo1 C972T	XV	182653	*kcs1 C2709T*	IV	479556
bar1 A1465G	IX	323804	*ktr6 G1222A*	XVI	457234
bas1 T165C	XI	638089	*ldh1 T892C*	II	632485
bcs1 T-43A	IV	1226571	*lhs1 C152T*	XI	298871
bst1 A1342G	VI	85891	*los1 A790T*	XI	50841
cbk1 G424A	XIV	335293	*lys14 C209T*	IV	511898
cbs2 G143T	IV	851367	*mal11 A482T*	VII	1075338
cca1 C488T	V	522177	*mal11 G310A*	VII	1075510
cdc25 G4346A	XII	752650	*met17 C-353T*	XII	732191
cdc39 G46T	III	280159	*met6 G941A*	V	341223
cld1 C37T	VII	713749	*mlh1 A675G*	XIII	595559
coa2 G114A	XVI	188192	*mnn4 G1543A*	XI	65925
cos111 C14T	II	629176	*mpd1 G384T*	XV	852693
cos111 T429C	II	629640	*mrn1 A13G*	XVI	197775
csf1 G3487A	XII	312246	*mrp7 G990T*	XIV	621441
dbp3 G1476T	VII	360387	*mrps28 T490A*	IV	1146802
dcs2 A73G	XV	658399	*mrps9 T443C*	II	535696
def1 C-108T	XI	338505	*msh2 G580A*	XV	147961
dgr2 C481T	XI	214266	*mtm1 A943T*	VII	1006369
dic1 G-357A	XII	828229	*nan1 G186A*	XVI	313085
die2 C414G	VII	947836	*nbp35 T-66C*	VII	343111
dop1 T2829A	IV	737166	*ndc80 C1235T*	IX	79308
dur1,2 T4873A	II	637333	*ngr1 T-154C*	II	647727
efm1 C851T	VIII	22630	*nip1 T1728C*	XIII	893698
flr1 C604T	II	251960	*nop58 A25T*	XV	896797
fmc1 G357A	IX	179880	*nqm1 G781A*	VII	580660
fol1 T1668A	XIV	166291	*nrd1 A-335C*	XIV	174651
gdh1 A1232G	XV	1041809	*nrg1 C505A*	IV	542863
gdh1 A68G	XV	1042973	*nrg1 G137T*	IV	543231
gdh1 C1345T	XV	1041696	*nup120 T1749C*	XI	331865
gdh1 G1105A	XV	1041936	*ost6 A159G*	XIII	233615
gdh1 G1331A	XV	1041710	*pal1 G815A*	IV	1171005
gdh1 G299A	XV	1042742	*pat1 G-31A*	III	252656
gdh1 G47A	XV	1042994	*pbp1 T-191C*	VII	853411
glr1 T312A	XVI	375810	*pdr10 G129A*	XV	931928
gsh1 A-73T	X	236425	*pex30 C134T*	XII	779348

Table 4.2 (continued)

Mutation name	Chromosome	Position	Mutation name	Chromosome	Position
pho5 A297T	II	430649	*tl(caa)k G429A*	XI	458742
pim1 T561C	II	180718	*top1 G-114A*	XV	315502
prm4 T-85A	XVI	256851	*tp(ugg)o1 C99T*	XV	301196
psd1 A1496G	XIV	316178	*tpc1 A529T*	VII	677153
psf1 A606G	IV	473759	*tup1 G1784A*	III	260666
rad50 T1268C	XIV	176678	*ubp1 A484G*	IV	243035
rgp1 T789C	IV	729044	*ubp1 T146A*	IV	242697
rif1 A3757G	II	753345	*ubp7 T2466A*	IX	50556
rif1 T474A	II	756628	*ubx6 A149G*	X	348484
rlp7 C-140A	XIV	627284	*uip4 C-164T*	XVI	195589
rme1 C654A	VII	583243	*uls1 C732T*	XV	693207
rpc31 G-273A	XIV	348796	*utp18 G305A*	X	312397
rpl21a T-140C	II	606552	*utp20 C4359T*	II	231997
rrb1 A903G	XIII	533795	*vhr1 A141G*	IX	250129
rsc2 C1667T	XII	842996	*vhr1 A1570T*	IX	251558
rtt107 T2346A	VIII	405314	*vhr1 A192G*	IX	250180
rup1 G1555A	XV	584771	*vhr1 T1892A*	IX	251880
sec16 A4795T	XVI	391858	*vhs3 G596A*	XV	429264
sei1 T-181C	XII	928561	*vid24 A223G*	II	451741
sfa1 G913A	IV	160517	*vma13 G259A*	XVI	645528
sgo1 C575A	XV	465347	*vps13 T-48C*	XII	63692
slp1 G52A	XV	624781	*vps29 G-25A*	VIII	129450
slu7 T788C	IV	618855	*yap1801 A1650T*	VIII	420640
snr189 C-120T	III	178915	*ycr090c C309G*	III	272552
snr46 A-262T	VII	545112	*ycr095w-a A194G*	III	289988
spc97 A-238G	VIII	448097	*ydl109c G20A*	IV	265238
sps22 A564G	III	42728	*ydr249c G664A*	IV	959135
srb7 A203G	IV	1078243	*ydr278c T-382C*	IV	1017656
srb8 C3787G	III	258154	*ydr524w-c A-15G*	IV	1489499
sro77 G-160T	II	14039	*yer186w-a T-555A*	V	565112
ssa1 G91T	I	141343	*yfl054c A809T*	VI	21979
ssf1 A580T	VIII	229916	*ygr027w-a C-300T*	VII	535761
ssn2 A734T	IV	1349197	*ygr042w G415A*	VII	579894
ssn2 C1567A	IV	1348364	*ygr125w A2059T*	VII	744388
ssn2 G832A	IV	1349099	*yhr045w C-124A*	VIII	195420
ssn3 T666C	XVI	474039	*yhr045w C668T*	VIII	196211
ssn8 G589A	XIV	584704	*yhr045w T103A*	VIII	195646
ste5 C512T	IV	658858	*yil054w T-93C*	IX	254448
ste5 T2649C	IV	660995	*yjl171c A288C*	X	100601
sum1 G2358A	IV	1081955	*ykr005c T451C*	XI	449057
sxm1 C-86T	IV	1263230	*ymr244w A-247T*	XIII	757002
tef1 T-126C	XVI	700466	*ynl058c T7C*	XIV	516708
tf(gaa)n G54A	XIV	375015	*ypk1 G1954A*	XI	207304
th(gug)g2 G-7T	VII	319777	*yps3 T84C*	XII	388661
thi2 C432T	II	701406	*ysp1 T716C*	VIII	407821
thi2 T501C	II	701337	*yur1 A328G*	X	152669
ti(aau)g C61T	VII	739187	*zuo1 G348A*	VII	1062812

pool, suggesting they spontaneously arose before mutagenesis. Their frequency oscillates around 0.5 throughout the evolution (Figure 3.2). Visual examination of the allele frequency data suggested the presence of cohorts of mutations with strongly correlated evolutionary trajectories. For example, mutations *aro1-C1283-4T* and *ste5-C512T*, both on chromosome IV, display similar frequencies at all time points and originate in the same haploid mutant pool. This observation suggested the existence of subgroups of mutations originating in the same initial mutant hitchhiking on one or a few driver mutations. To test this hypothesis and identify cohorts of mutations potentially linked by origin, I performed hierarchical clustering on the evolutionary trajectories of all SNPs in my list (Figure 3.1). Nine groups of mutations were deduced from the resulting dendrograms. The majority of mutations are found at very low frequency throughout the evolutionary engineering experiment, with varying levels of apparent selection. Those mutations are assigned to cohorts **α1** and **a2** (Figure 3.1). Equality of initial frequency or a significant correlation cannot be inferred with confidence between the evolutionary trajectories of these low frequency cohorts.

Three mutations present unique trajectories that prevent grouping with other mutations, each instead forming its own group. Cohort **a1**, consisting of single mutation *mal11-G310A*, stands out of the pool as displaying the strongest apparent selection, with a mean allele frequency change of 1.685. Similarly, the *gdh1-G47A* mutation displays strong apparent selection (M=1.532) and is alone in cohort **a4**. Mutation *gdh1-A68G* is placed with **α3** mutations by the clustering algorithm, but its trajectory is markedly different from other mutations in that cohort. All mutations in cohort **α3** have an initial frequency of approximately 0.44, while *gdh1-A68G* is first detected at 0.036. Further, *gdh1-A68G* displays strong apparent selection (M =1.604) while group **α3**

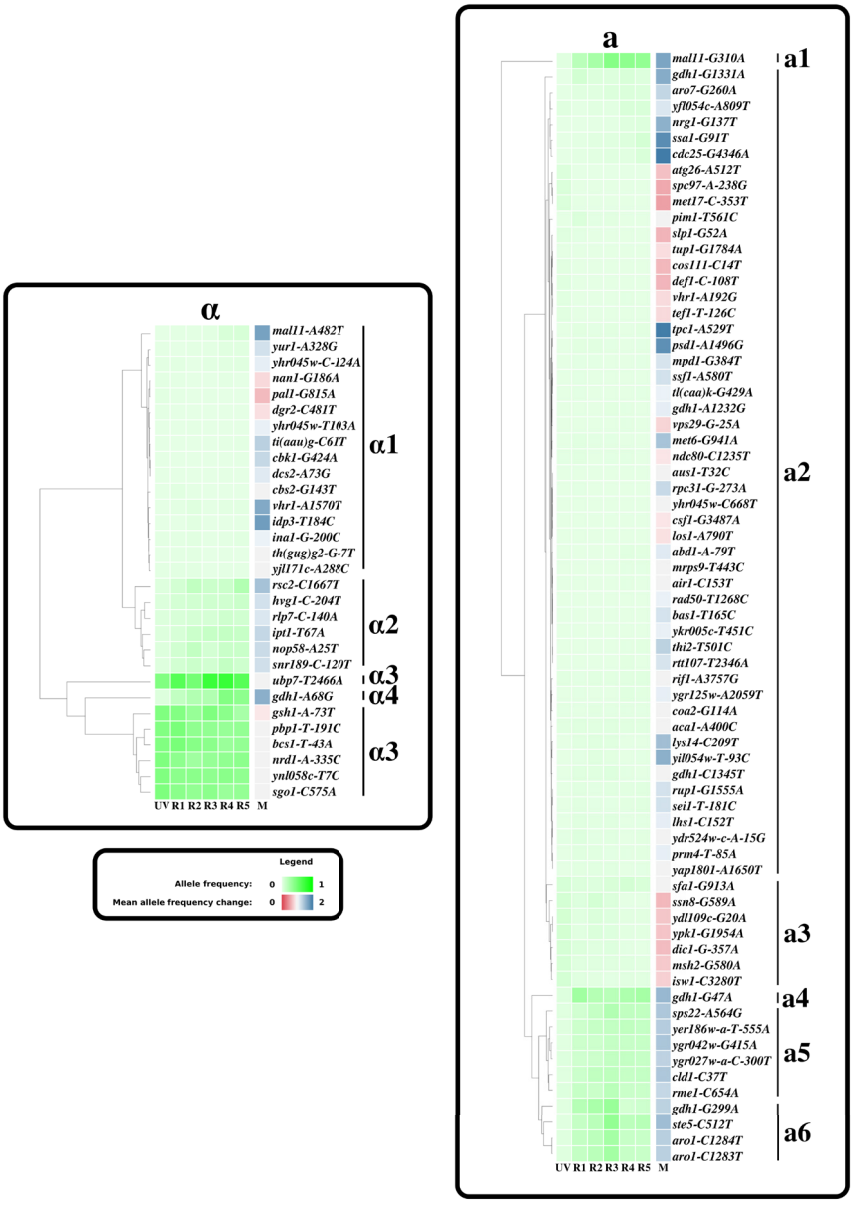


Figure 3.1 Evolutionary trajectories for all non-silent mutations identified by population genome sequencing. Mutations arose either in the *MATa* (left) or *MATα* (right) haploid pools. On the vertical axis are the names of the mutations, giving the closest gene, coordinates with respect to that gene, and the nature of the nucleotide substitution. On the horizontal axis are each of the six evolutionary timepoints (UV, R1, R2, R3, R4, R5), and the mean allele frequency change (M). Frequencies of the mutant alleles are represented by shades of green. Mean allele frequency changes are represented in shades of red (M< 1, declining frequency) or blue (M>1, increasing frequency). Hierarchical clustering of individual evolutionary trajectories is represented by dendrograms on the left. Mutations were assigned to cohorts of mutations (a1-5, α1-3) on the basis of this clustering.

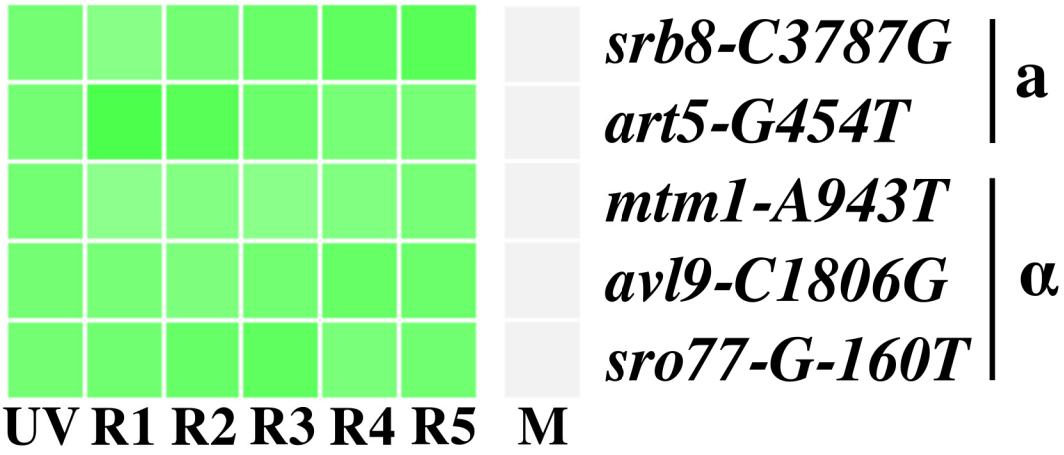


Figure 3.2 Five mutations arose in the parental strains before mutagenesis. Evolutionary trajectories and apparent selection are reported**.** Frequencies of the mutant alleles are represented by shades of green.

mutations show the near absence of selection (M =0.9-1.1). I therefore assign *gdh1-A68G* to its own cohort (**α4**).

The most frequent *MATa* derived cohorts are **a5** and **a6**. Mutations in genes *ARO1* and *STE5*, mentioned at the beginning of this section as examples of strongly correlated trajectories, belong to **a6**. Mutations in both groups start at a frequency of approximately 0.03 and increase steadily to reach maximum frequency at round 3. The general trend for both groups is decline at rounds 4 and 5. Higher frequency, in particular at round 3, is the main difference between **a5** from **a6**. Mutation *mal11-G310A* (cohort **a1**) displays the same general pattern of steady increase to R3, followed by a relative decline. However, it reaches a higher maximum frequency and its decline is less pronounced. Together, these observations pose the questions of whether **a1**, **a5** and **a6** mutations belong to a single cohort.

Cohort **α3** consists of seven mutations remarkable for their high and virtually identical initial frequency. Mutant alleles from this group display a frequency of ~0.88 in the *MATα* pool, nearly sweeping that population at early selection steps. This group benefits from the founder effect, remaining highly represented throughout the evolutionary engineering experiment. The general pattern followed by all but one mutation (*ubp7-T2466A*) suggests an absence of selection, or a slow decline (M=0.903-0.970). The mutation in gene *UBP7* diverges from the rest of the group, on average increasing in frequency throughout the evolution. Large increases in frequency (0.49 to 0.74 between R2 and R3) are compensated by phases of decline, amounting to a more modest synthetic measure of selection (M =1.070).

3.3 Certain genes are mutation hotspots

I observed several independent occurrences of mutations mapping to the same genes, suggesting an important role in the SSL tolerance phenotype. The presence of 1 out of 188 single mutations in any gene is statistically significant in a 12 million base pair genome (binomial test with p-value=0.01). The chance occurrence of more than one mutation in the same gene is even less likely. Furthermore, three lines of evidence can indicate independent mutational events: detection in a different haploid pool of origin, significantly different frequencies in the same pool of origin, and absence of colinearity between evolutionary trajectories. A systematic survey revealed eight genes to which more than one mutation was mapped (Figure 3.3). From this list I exclude four strongly correlated mutations from cohort **a6** and mapping to chromosome IV in genes *STE5* (one of which is silent) and *ARO1*, which probably resulted from the same mutation event.

Among the 25 mutations considered, 9 escape detection in at least one of the sampled time points (Annex IV). Mutations clustering in genes *SSN2*, *YHR045W*, *UBP1* and *COS111* are detected at consistently low frequencies, with negative or weakly positive apparent selection. Two of the *VHR1* mutations are positively selected, while the other two are negatively selected, and all four remain at relatively low frequencies throughout the experiment.

Both of the mutations mapping to *NRG1* are rare, but display strong and sustained positive selection, and one of them (G137T) was found in the R57 mutant. While correlations are difficult to establish with confidence at such low frequencies, these mutations display highly similar frequencies at several time points, perhaps indicating that they arose during the same mutational event.

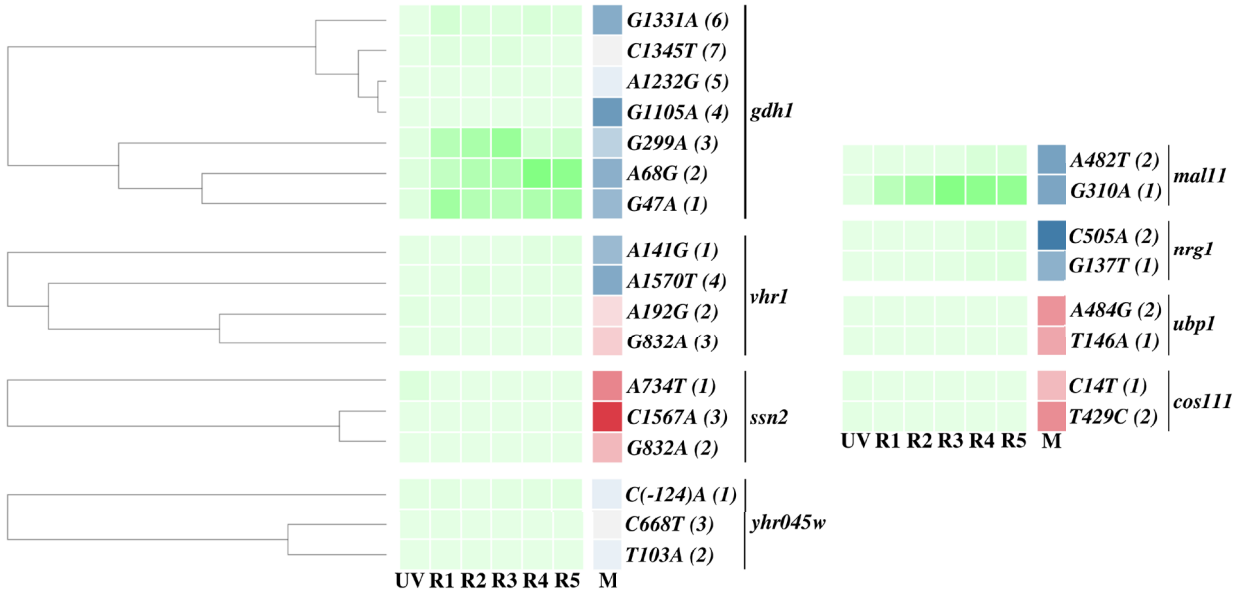


Figure 3.3. Evolutionary trajectories, apparent selection and clustering of all mutation hotspots identified by population sequencing. Mean allele frequency changes are represented in shades of red (M< 1, declining frequency) or blue (M>1, increasing frequency). Hierarchical clustering of individual evolutionary trajectories is represented by dendrograms on the left.

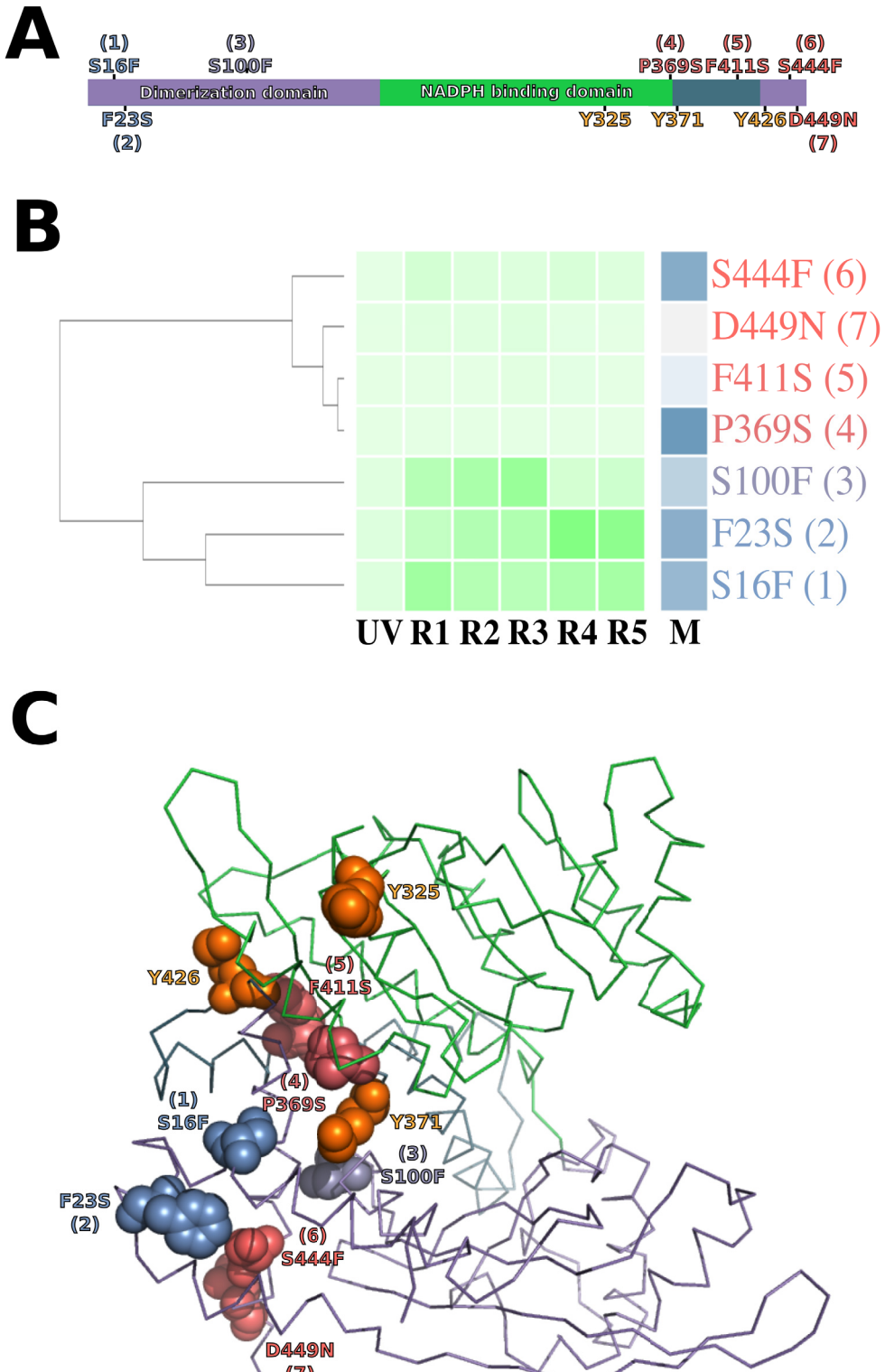

A
(1)
S16F
(3)
S100F
(4)
P369S
(5)
F411S
(6)
S444F
Dimerization domain
NADPH binding domain
F23S
(2)
Y325
Y371
Y426
D449N
(7)
B
S444F (6)
D449N (7)
F411S (5)
P369S (4)
S100F (3)
F23S (2)
S16F (1)
UV R1 R2 R3 R4 R5 M
C
Y325
Y426
(5)
F411S
(4)
P369S
(1)
S16F
Y371
(3)
S100F
F23S
(2)
(6)
S444F
D449N
(7)

Figure 3.4. Mutations mapping to gene *GDH1* cluster in specific areas of the protein. Mutations are numbered from 1 to 7 and color-coded according to their position from N- (blue) to C-terminal (red). The dimerization domain is in purple and the NADPH binding domain is in green. (A) Mutations affect either the amino or carboxy-terminal portions of Gdh1p. (B) Cluster map, as in Figure 3.1, of the Gdh1p amino acid substitutions, showing allele frequencies at each evolutionary time point (green heatmap), mean allele frequency change (blue heatmap), and clustering of evolutionary trajectories (left dendrogram). (C) Gdh1p substitutions were mapped onto a homology model of the protein based on the structure of *Plasmodium falciparum* glutamate dehydrogenase (PDB ID: 2BMA).

Two independent mutations are detected in gene *MAL11*. Both display positive selection and are found in R57. Mutation *mal11-G310A* was mentioned earlier as displaying a high frequency and the strongest observed positive selection.

The most remarkable cluster of mutations maps seven non-silent hits to gene *GDH1*. Strong positive selection is observed for most of these mutations, and three of them rapidly rise to prominence in the course of the evolution. Together, these observations convincingly argue for the critical role of *GDH1* in the phenotype of interest or at least in the evolutionary dynamics of the genome shuffling experiment. The seven substitutions are found to cluster to the amino- and carboxy-terminal portions of Gdh1p (Figure 3.4A), with the most frequent ones found at the N-terminus (Figure 3.4B).

Mapping of Gdh1p substitutions on a homology model derived from the structure of the highly homologous *P. falciparum* glutamate dehydrogenase 1 [330] shows both N- and C-terminal substitutions grouped near the hinge region separating the two structural domains of the protein (Figure 3.4C). More precisely, the substitutions are close to a groove and adjoining tunnel found at this junction, adjacent to the active site cleft. In particular, the strongly selected S16F substitution maps precisely at the entrance of the tunnel, forming part of the surface of the groove. The prominent N-terminal substitutions

cluster the closest to the hinge, while the less frequent C-terminal ones are found at the periphery of the hinge area.

3.4 Genotyping of backcrossed isolates of R57 suggests a linear model for the contribution of individual mutations to the SSL tolerance phenotype

Sporulation was induced in strain R57 and the resulting haploids were recursively backcrossed with their wildtype haploid parent CEN.PK113-1A (as described in Figure 3.5) to generate strains with random combinations of mutations. A total of 86 segregants were isolated after two rounds of sporulation and mating. Growth curves in the absence and presence of SSL were recorded for all isolates, revealing a seemingly continuous distribution in their level of tolerance (Figure 3.6). A Kolmogorov-Smirnov test (α=0.01 and 0.05) suggests a normal distribution for the growth of the isolates in SSL, in agreement with the hypothesis of a polygenic quantitative trait (Annex V: Kolmogorov-Smirnoff test for normality on the distribution of SSL tolerance scores among R57 backcrossed isolates).

To estimate the contribution of each R57 mutation to the SSL tolerance phenotype, I used amplicon sequencing to genotype the 86 segregants (Figure 3.7). For the vast majority of strains and loci, I achieved depth of coverage well above 30 (average of 794), enabling their confident genotyping (Table 4.3). Multiple linear regression was applied separately on the haploid and diploid datasets to generate models for the predicted effect of each SNP on the phenotype (summarized in Figure 3.7, and Tables 4.4 and 4.5). Fewer variables and more data points mean that I have higher confidence in the haploid model than in the diploid model (see Materials and

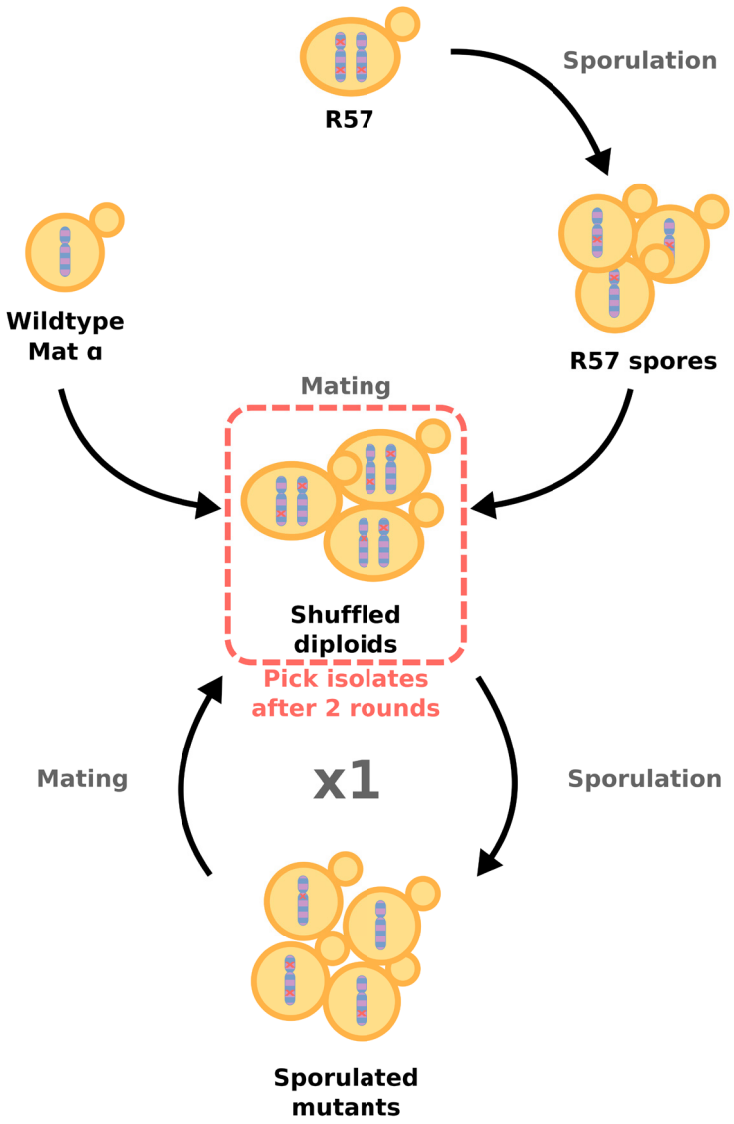


Figure 3.5 Outline of the R57 backcrossing strategy. The diploid R57 strain was sporulated to generate a pool of haploid derivatives, which were mated at random with *MATα* haploid strain CEN.PK113-1A, thus excluding all mutant *MATα* spores. Resulting diploids were sporulated and mated against each other to generate diploid recombinants.

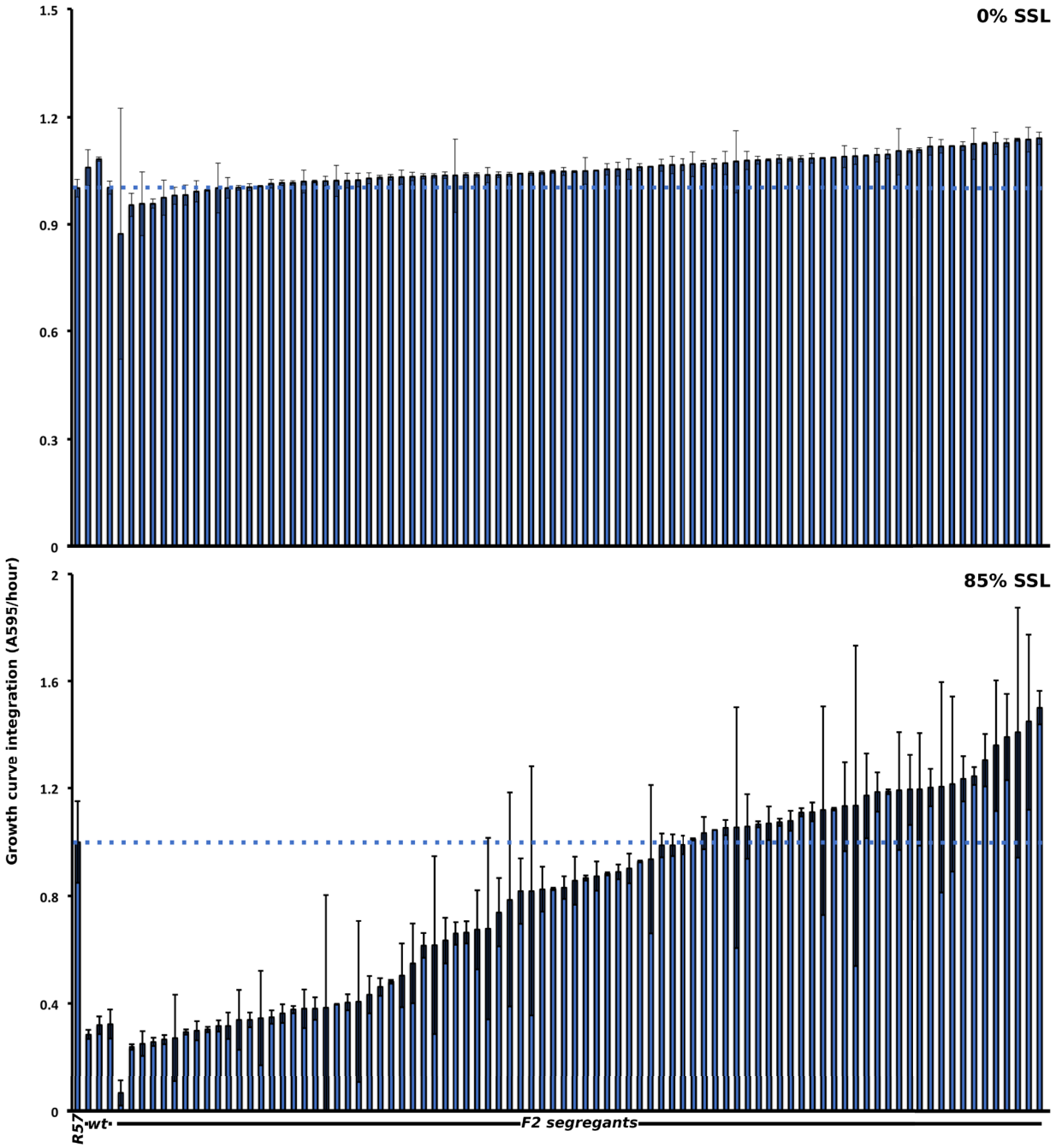


Figure 3.6 Backcrossing of R57 with wildtype cells generates strains presenting a wide spectrum of tolerance to SSL. Growth in the presence and absence of SSL is reported for R57, various wildtype cell types, and 86 F2 isolates from backcrossing of R57 and CEN.PK113-1A**.**

Table 4.3 Amplicon sequencing and alignment metrics for genotyping of R57 segregants

Locus	Average depth	Max depth	Min depth	Nb below 30
ARO1-1283	1082	1801	1	1
ARO1-1284	259	736	0	1
ART5-454	1193	1826	76	0
BCS1-(-43)	126	214	5	5
DOP1-40	807	1588	0	1
FIT3-(-42)	229	657	79	0
GDH1-68	1149	1635	227	0
GDH1-47	1033	1488	181	0
GSH1-(-73)	497	875	5	3
MAL11-482	1546	2253	3	1
MAL11-310	711	1816	0	1
NOP58-25	506	772	1	2
NRG1-137	672	1436	39	0
PBP1-(-191)	258	606	0	2
SGO1-575	1087	1530	6	2
SSA1-91	890	1711	0	3
STE5_1512	268	541	2	1
STE5_3649	1090	1712	1	6
TOF2-2141	1238	1861	11	2
UBP7-2466	395	741	0	2
YNL058C-7	1590	3157	3	2
ALL	794	3157	0	35

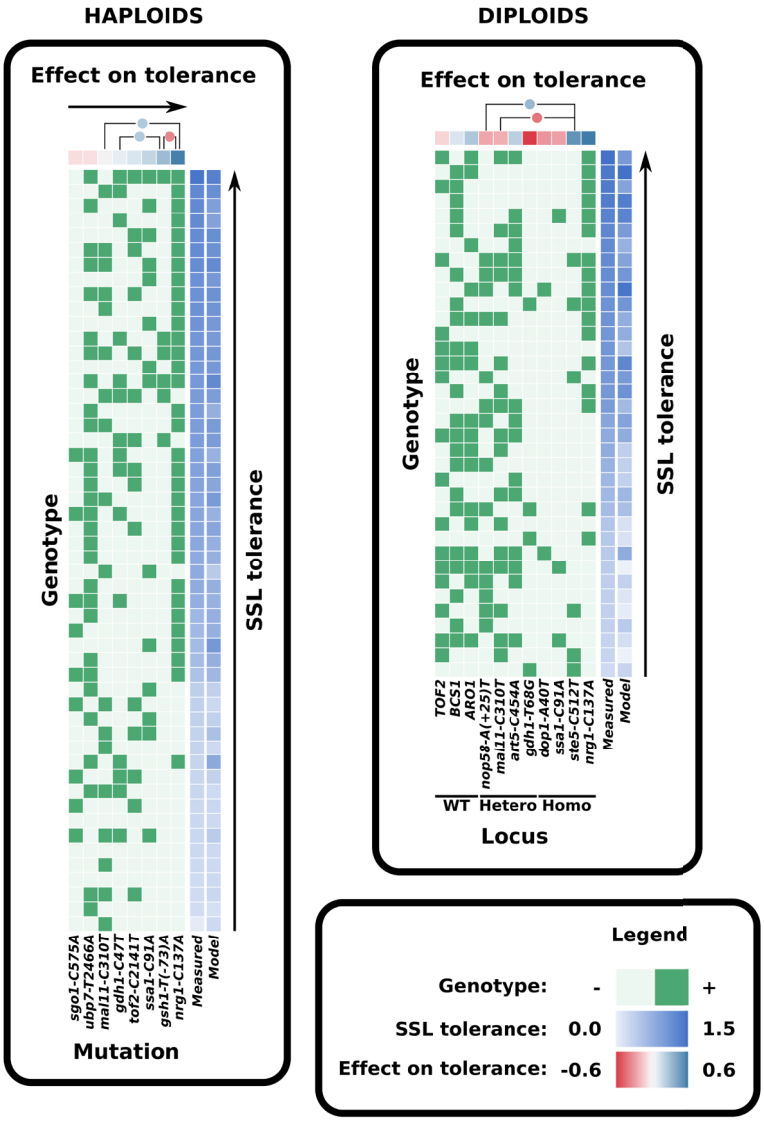


Figure 3.7 Genotyping of second-generation mutants from backcrossing of R57 and wildtype yeast suggest a model of SNP contributions to the SSL tolerance phenotype. Haploid (left heatmap) and diploid (right heatmap) isolates are scored in green for the genotypes indicated at the bottom. Growth in 85% (v/v) SSL is scored in shades of blue on the right. Each row represents a single strain. Contribution to the phenotype of the indicated genotypes was inferred by multiple linear regression, yielding coefficients represented at the top in shades of red (diminishes tolerance) to blue (increases tolerance). A model of genetic interaction was created, and the resulting coefficients are represented as circles at the top of the heatmaps. Growth in SSL predicted by the linear model is reported in shades of blue in the rightmost column, showing the level of agreement between the model and the data. Genotypes not proposed to influence the phenotype are omitted. See Annex V for full dataset.

Table 4.4 Metrics and linear coefficients of the haploid linear regression model

R^2	MRSS*	σ(residuals)
0.836	0.032	0.027
Variable	**Coefficient**	**p-value†**
β_0	0.278	0.000
nrg1-C137A	0.579	0.000
gsh1-T(-73)A	0.300	0.000
ssa1-C91A	0.188	0.000
tof2-C2141T	0.110	0.000
gdh1-C47T	0.071	0.000
mal11-C310T	-0.007	0.254
ubp7-T2466A	-0.081	0.000
sgo1-C575A	-0.093	0.000
mal11-C310T x nrg1-C137A	0.245	0.000
gdh1-C47T x gsh1-T(-73)A	0.242	0.000
gsh1-T(-73)A x nrg1-C137A	-0.368	0.000

*Mean residual sum of squares
†p-value that linear coefficients are equal to 0

Table 4.5 Metrics and linear coefficients of the diploid linear regression model

	R^2	MRSS*	σ(residuals)
	0.921	0.018	0.025
	Variable	**Coefficient**	**p-value†**
	β_0	0.598	0.000
wildtype	*ARO1*	0.251	0.000
	BCS1	0.108	0.000
	TOF2	-0.119	0.000
heterozygous	*art5-C454A*	0.212	0.000
	mal11-C310T	-0.235	0.000
	nop58-A(+25)T	-0.258	0.000
homozygous	*nrg1-C137A*	0.643	0.000
	ste5-C512T	0.488	0.000
	ssa1-C91A	-0.284	0.000
	dop1-A40T	-0.315	0.000
	gdh1-T68G	-0.757	0.000
	nop58-A(+25)T (het) x *ste5-C512T* (hom)	0.319	0.000
	mal11-C310T (het) x *ste5-C512T* (hom)	-0.418	0.000

*Mean residual sum of squares
†p-value that linear coefficients are equal to 0

Methods). Accordingly, there is better agreement between predictions of the haploid model and measured phenotypes (Figure 3.7).

In both haploids and diploids, the strongest predictor for enhanced SSL tolerance is the *nrg1-C137A* mutation. Examination of the genotype heatmaps in Figure 3.7 shows a clear clustering of *nrg1* mutants at higher SSL tolerance levels. In haploids, the *gsh1-T(-73)A* mutation is strongly associated with high SSL tolerance, while modeling proposes negative epistasis between the *nrg1* and *gsh1* alleles. Mutations *ssa1-C91A*, *tof2-C2141T* and *gdh1-C47T* are also associated with modest increases in haploid tolerance to SSL, albeit the effect of the latter is proposed to be enhanced by interaction with *gsh1-T(-73)A*. Mutations *mal11-C310T*, *ubp7-T2466A* and especially *sgo1-C575A* are deleterious for haploids. Mutation *mal11-C310T* has virtually no effect (p-value=0.254 that linear coefficient is equal to 0), but is proposed to interact with *nrg1-C137A* to further increase tolerance.

In diploids, each locus is associated with three possible genotypes each interpreted as a separate variable. Aside from *nrg1-C137A*, mutants homozygous for *ste5-C512T* are predicted to display increased SSL-tolerance, while homozygous *ssa1-C91A*, *gdh1-T68G* and *dop1-A40T* are associated with a loss of fitness in the presence of SSL. My model predicts that heterozygosity leads to a loss of tolerance in *mal11-C310T* and *nop58-A(+25)T*, while it causes an enhanced phenotype for *art5-C454A*. Wildtype *ARO1* and *BCS1* alleles are associated with better growth in SSL, suggesting mutations at these loci are deleterious. Conversely, absence of a mutation in gene *TOF2* predicts reduced tolerance to SSL, arguing for a beneficial effect of *tof2-C2141T*. The model proposes that *ste5-C512T* interacts with two other mutations: positively with *nop58-A(+25)T* and negatively with *mal11-C310T*.

Constants between the haploid and diploid models include the strong positive effect of *nrg1-C137A*, the weakly negative effect of *mal11-C310T* in the absence of interactions, and the benefits of *tof2-C2141T* for SSL tolerance.

3.5 A CRISPR-Cas9 method for the seamless site-directed mutagenesis of the *S. cerevisiae* genome

Following the first reported application of CRISPR-Cas9 in *Saccharomyces cerevisiae* [333], several methods exploiting the potential of this technology for yeast genome editing were published enabling gene disruption [333–336], gene deletion [337,338], heterologous sequence integration [334,336,337,339,340] and insertion of point mutations [333,337–339]. I therefore reasoned that recent developments in CRISPR-Cas9 technology should permit the seamless introduction of point mutations discovered in R57 back into the wildtype parental background for testing phenotype to genotype associations. Similarly, I set to revert each of these point mutations to wildtype in mutant strain R57. However, strategies reported for the introduction of a point mutation using CRISPR-Cas9 suffered several caveats that restrict the range of mutations that can be introduced at any given locus. One difficulty is that sequencing is required to detect the successful integration of a point mutation. However, the main challenge to using CRISPR-Cas9 for the introduction of point mutations is the risk of repeated cutting by Cas9 after homology directed repair (HDR) of the initial double stranded break (DSB). Indeed, point mutations may not be located within the protospacer sequence, leaving it intact after HDR. Even if the mutation is located close enough to a protospacer adjacent motif (PAM) to modify the gRNA target sequence, a single substitution is generally insufficient to prevent

recognition by the gRNA/Cas9 complex [341]. Several strategies have therefore been devised to prevent Cas9 from cutting repeatedly at the site of interest. Mutation of the PAM along with the target point mutation position abolishes target recognition by the gRNA/Cas9 complex, and has allowed the successful introduction of premature stop codons [333,338]. This strategy remains confined to cases where the PAM site mutation is either silent or deemed inconsequential. An alternative is the insertion of so-called heterology blocks in addition to the mutation of interest [339]. A heterology block consists in a number of additional silent mutations meant to abolish gRNA recognition. While heterology blocks change codon usage in an open reading frame (ORF) and may potentially affect mRNA translation, they represent a quick and convenient means of introducing point mutations. Moreover, their successful integration is easily detected by PCR. However, the concept of a silent mutation is meaningless in untranslated regions of the genome, such as non-coding RNAs and intergenic sequences.

Mans and coworkers [337] demonstrated successful insertion of a point mutation without altering the PAM or resorting to a heterology block. The inserted mutation eliminated a restriction site and replaced it with another, providing for easy detection of successful mutants. Sequencing revealed that several restriction positive clones displayed additional unwanted mutations, likely due to repeated cutting by Cas9. This direct strategy therefore requires the screening of several clones by sequencing – a comparatively time consuming and costly process. These authors suggested an alternative two-step strategy for the seamless site-directed mutagenesis of the yeast genome using CRISPR-Cas9, but did not demonstrate it experimentally. A similar proposition was made shortly after by Lee and coworkers [342]. Here, I propose three

variations on this general method, and report its successful application at 17 positions across the genome of *S. cerevisiae* haploid strains CENPK113-1A, CEN.PK113-7D and the R57 mutant diploid strain.

Using two successive CRISPR events, the method enables the introduction of point mutations without altering the PAM or inserting additional silent mutations (Figure 3.8). In the first CRISPR event, the Cas9-induced DSB is repaired by a homologous repair fragment which replaces the 20 nucleotide protospacer by a heterologous sequence of the same length (termed the “stuffer”), preventing repeated cutting by Cas9 (Figure 3.8). After curing of the initial guide, a second gRNA targeting the stuffer is introduced. The DSB is repaired by a DNA fragment carrying the desired point mutation, thereby removing the stuffer and abolishing recognition by the second gRNA. Stuffer insertion and removal is conveniently detected by colony PCR. This is in contrast to single-step methods that make use of sequencing to identify clones both devoid of unwanted secondary mutations, and harboring the desired point mutation, unless the point mutation coincidentally creates or removes a restriction site [337]. In the two-step method described here, the only modification introduced in the parent strain is a single point mutation (or any desired modification).

In a recent study, a similar approach was used in human induced pluripotent stem cells for the correction of heterozygous β-thalassemia mutations [343]. The piggyBac transposon system, carrying antibiotic resistance markers and acting as a stuffer, was inserted into the hemoglobin B gene by CRISPR assisted HDR. The transposon was then excised with the help of a specialized transposase, and the mutation corrected by homologous recombination with the non-mutant copy of the gene. Use of a two-step

procedure for the seamless alteration of the yeast genome is reminiscent of the Delitto perfetto method, whereby successive rounds of positive and negative selection are used to transiently introduce a marker cassette [344].

In the present method, a single stuffer with a random unique sequence is employed in most instances, which allows for the repeated use of the same targeting gRNA sequence and PCR primers for confirming the presence of the stuffer. For inserting or removing single point mutations, protospacer replacement was attempted for 15 out of 17 positions using the sequence 5'-AGATGCGGGAGAGGTTCTCG-3' as a stuffer. Screening by PCR of three clones per position revealed that the stuffer sequence was successfully inserted in at least one of the three clones tested in all but five positions (in genes *MAL11*, *UBP7* and *GDH1* for all strains, and *STE5 566* and *PBP1* in R57) (Figure 3.8). However convenient, I suspected that the transient disruption of important genes by a standard stuffer could reduce or abolish cell viability. For example, the mutant strain R57 carries mutations in or near essential genes *DOP1* and *NOP58* [345,346]. In addition, I observed that insertion of the stuffer in the *ARO1* gene of *S. cerevisiae* considerably reduced its growth rate on YPD medium. I therefore hypothesized that failure to insert the stuffer sequence in genes *MAL11*, *UBP7*, *GDH1*, *STE5* (at position 566) and *PBP1* could be due to similar viability issues.

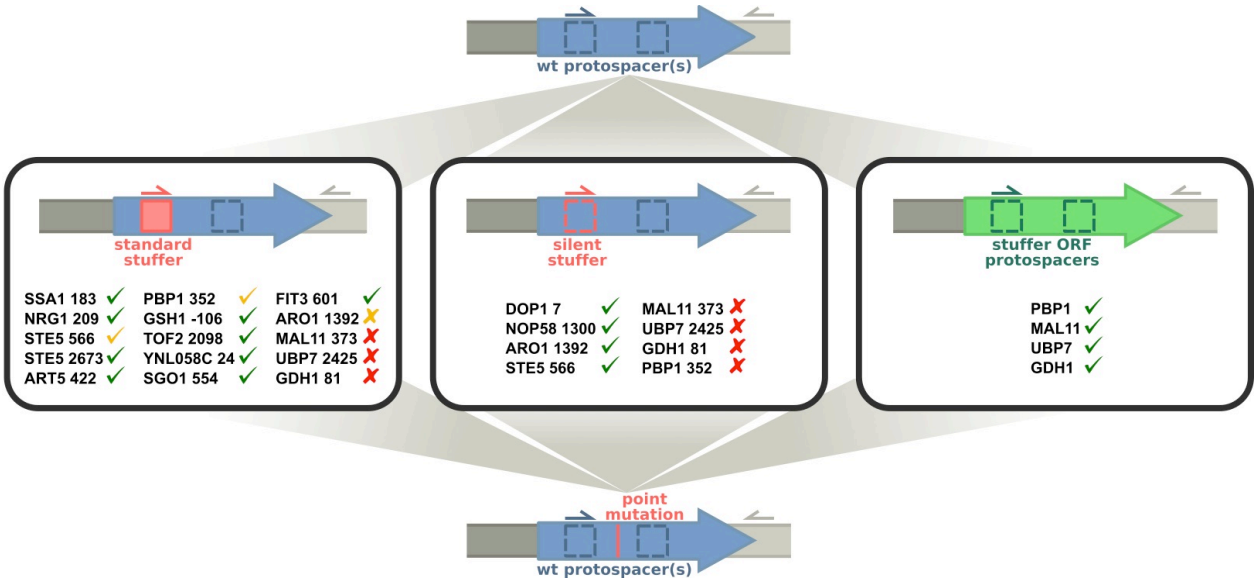


Figure 3.8 Outline of the two-step, stuffer-assisted genome site-directed mutagenesis strategy. Two variations of the strategy were applied. In the stuffer strategy a protospacer target sequence located near the site to mutagenize is replaced by a heterologous 20-nucleotide sequence (the stuffer) by CRISPR-Cas9 assisted homologous recombination, leaving the PAM site intact. The stuffer may be a standard, randomly generated sequence (the standard stuffer, left box) or a degenerate sequence bearing at least seven mismatches with the original protospacer (a silent stuffer, middle box). The second CRISPR step uses the stuffer as a protospacer, restoring the original protospacer sequence and introducing the desired mutation in a single homologous recombination event. A second variation on the strategy replaces the entire target ORF – or nearby ORF if an intergenic region is the target of mutagenesis – by a heterologous stuffer ORF (e.g. GFP), which is targeted by one or more gRNAs in the second step (right box). Homologous recombination restores the original ORF with mutations. Successful integration is easily assessed by PCR. The strategies were tested at the positions indicated in the boxes. Positions are identified by the coordinate of the first nucleotide of the PAM site (NGG) with respect to the nearest ORF. For each position, stuffer insertion was successful in either all strains tested (green check), two out of three strains (yellow check), in all strains but led to a slow growth phenotype (yellow-x), or in none of the strains (red-x). Once a stuffer was inserted, its removal and replacement by the point mutant sequence was successful in all cases.

For the two essential genes (*DOP1*, *NOP58*) and the six previously unsuccessful positions (*ARO1*, *MAL11*, *UBP7*, *GDH1*, *STE5*-566, *PBP1*), I designed custom stuffers (and stuffer-targeting gRNAs) that did not disrupt the coding region using degenerate sequences. Similar to heterology blocks, these silent stuffers introduced at least seven nucleotide substitutions to protect against repeated cutting by Cas9.

I was able to insert the silent stuffers at *DOP1*, *NOP58*, *ARO1* and *STE5* (Figure 3.8), and growth defects were not observed in the resulting strains. However, the silent stuffer insertion method failed for *MAL11*, *UBP7*, *GDH1* and *PBP1*. Suspecting my choice of gRNA target sequences to be the cause, I designed, for each of the four genes, two or three additional gRNAs with targets evenly spaced along the ORF to increase chances of DSBs. To avoid having to design stuffer fragments for each target, I designed donor DNAs containing the yeGFP sequence with, at their 5' and 3' ends, 50-bp homology to the promoter and terminator of the target genes. The expected result was the precise replacement of the native ORFs by yeGFP (Figure 3.8). Not presuming of the success of any one individual guide, this strategy prevents further recognition by the gRNA/Cas9 complex anywhere in the gene by replacing the entire target ORF. The new guides were simultaneously transformed into yeast with the yeGFP stuffer. Integrants were identified in all four loci (Figure 3.8), suggesting at least one guide per locus was functional. I suggest that stuffer ORFs can prove useful when the selection of a functional gRNA target is problematic. However, I note that it is not suitable in genes that are essential or strongly affect viability when deleted, in both cases preventing downstream transformation and CRISPR events.

In strains containing the short stuffers, the second CRISPR event used DNA fragments averaging 500 bp for DSB repair and introduction of point mutations. Longer

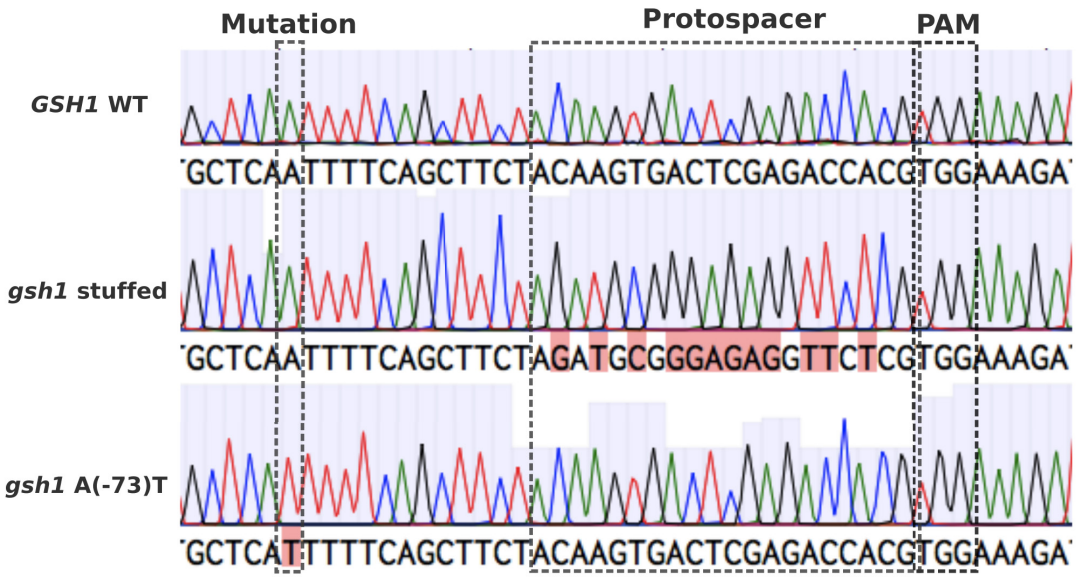


Figure 3.9 Sequencing shows successful insertion of the stuffer and subsequent introduction of a point mutation in the *GSH1* promoter sequence.

fragments spanning the promoter, ORF and terminator were required at loci stuffed with yeGFP. For all stuffer-containing strains, replacement of the stuffed sequence by the point mutant sequence was successful (Figure 3.8). Introduction of point mutations was confirmed by Sanger sequencing revealing the absence additional unwanted mutations in the targeted loci (see Figure 3.9 for an example). While the efficiency of stuffer insertion was highly variable and rarely at 100%, I observe that for most positions considered, all clones screened for stuffer removal and point mutation insertion were positive. CRISPR efficiency was high at positions bearing the standard short and yeGFP stuffers, but lower on average for positions carrying the custom silent stuffers. These observations suggest that a standard stuffer is useful in reducing the variability of recognition and cutting by the gRNA/Cas9 complex during the second CRISPR event. Whenever feasible, I propose that it should be the preferred method for CRISPR assisted genomic insertion of point mutations into the genome.

This section reports a strategy to introduce precise changes at the single nucleotide level in the genome of *S. cerevisiae*. I believe that this two-step procedure can be applied to any organism with suitable HDR machinery at virtually any genomic coordinates to modify coding and non-coding sequences, in essential and non-essential genes. Furthermore, it is not constrained by the precise location and sequence of the PAM and protospacer. Because it does not require the introduction of large transposons or selection cassettes, this method is less disruptive than similar two-step methods reported previously [342,344],. Rather, it transiently introduces a few potentially silent mutations. However, the implementation of this method requires the generation of an intermediate stuffed mutant that is submitted to a second cycle of transformation, PCR

verification, sequencing and gRNA curing. Welcome improvements would allow stuffer integration and removal from a single transformation using, for example, transient or inducible gRNAs. Because of its wide applicability, I believe this seamless, genome-level site-directed mutagenesis procedure will prove useful to a wide range of researchers interested in the precise genome editing of *S. cerevisiae* and other organisms.

3.6 Introduction of point mutations in wildtype backgrounds identifies alleles contributing to the SSL tolerance phenotype

To further dissect the effect of individual mutations on the SSL-tolerance phenotype, I introduced the R57 mutations in isolation into the *MATα* and *MATa* wildtype backgrounds. In our hands, the baseline growth rate of the parent *MATa* strain (CEN.PK113-7D) is consistently much slower than for the *MATα* (CEN.PK113-1A) parent. Addition of SSL exacerbated this phenotype, leading to barely detectable growth. Therefore, I only report growth data for the *MATα* mutants (Figure 3.10A). In this background, increased ability to grow in the presence of SSL was observed for *gsh1-T(-73)A* and *nrg1-G137T* mutants, in agreement with my observations on backcrossed R57 derivatives (Figure 3.7). Among mutations identified by population sequencing, a subset of the most positively selected were introduced into the *MATα* haploid, but an increase in SSL tolerance was not observed for those strains (Figure 3.10B). Similarly, none of the single diploid mutants showed an increase in SSL tolerance (Figure 3.10, C and D). A growth defect was instead observed in most heterozygous and all homozygous mutants. Together, these results suggested that the *gsh1-T(-73)A* and

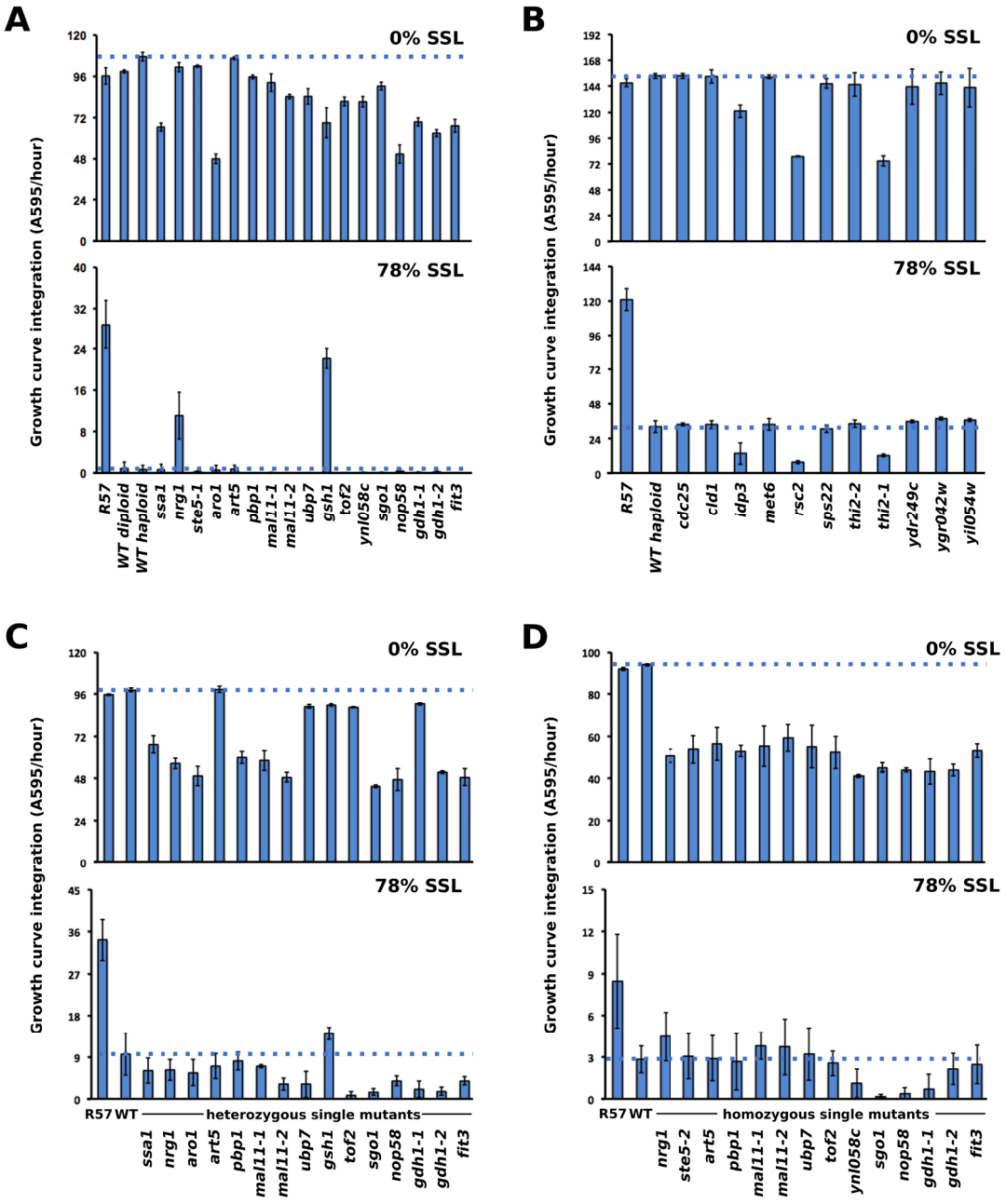


Figure 3.10 Growth in presence and absence of SSL of single mutants identifies the contribution of mutations *nrg1-G137T* and *gsh1-T(-73)A* to the tolerance phenotype. Area under the growth curve in the presence and absence of SSL is reported for single haploid mutants carrying R57 mutations (A), highly selected mutations identified by population sequencing (B), heterozygous (C), and homozygous (D) diploids and revertant derivatives of R57 *SGO1 gdh1-2/2*, wildtype for the indicated genes.

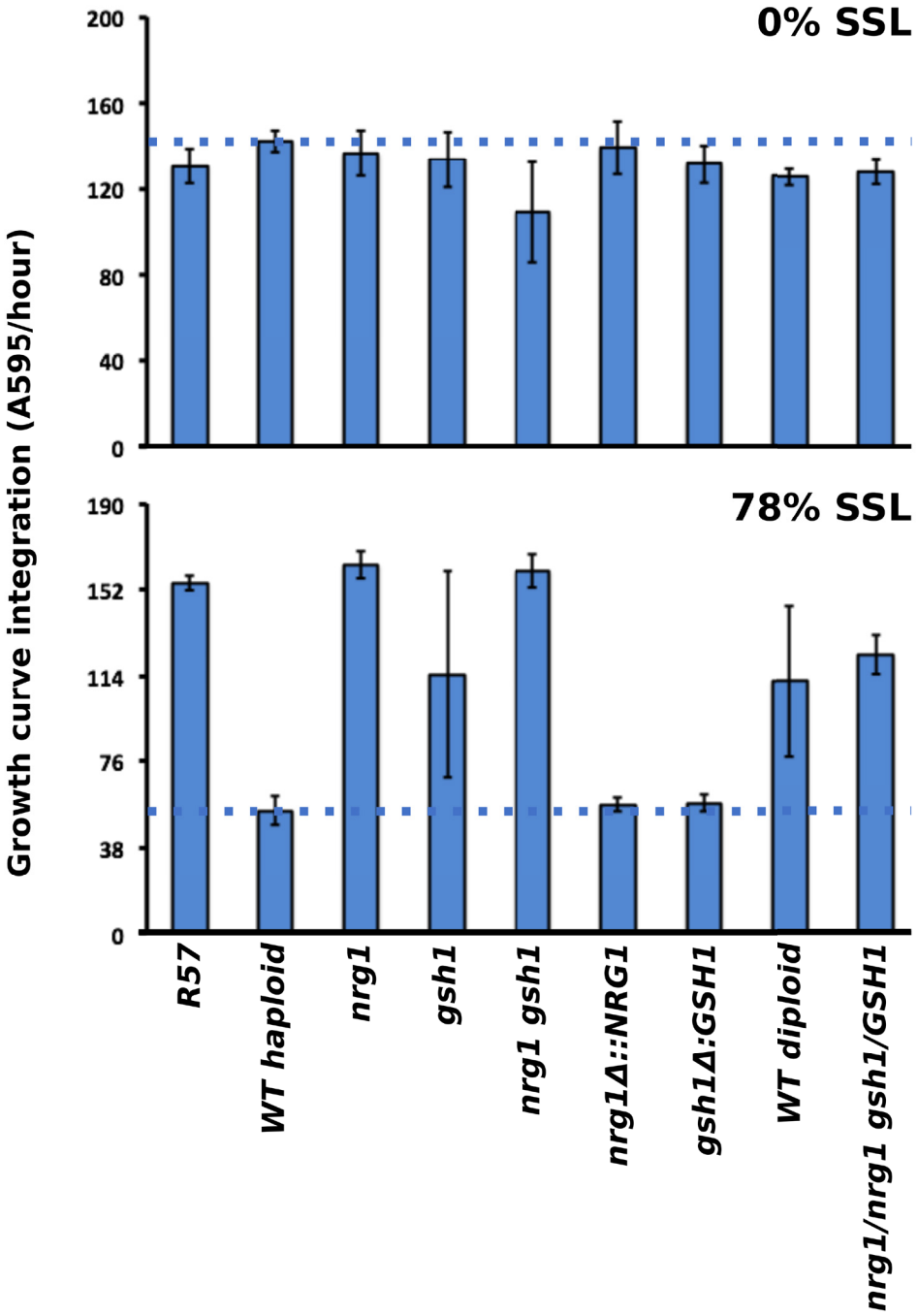


Figure 3.11 Reversion of *nrg1* and *gsh1* mutations leads to loss of the SSL tolerance phenotype in haploid single mutants. Area under the growth curve for *nrg1* and *gsh1* double mutants, haploid (*nrg1 gsh1*) and diploid (*nrg1*/*nrg1 gsh1*/*GSH1*) is reported.

nrg1-G137T alleles could increase the tolerance of haploid cells to SSL, but were not sufficient in a diploid background.

To confirm that the *gsh1-T(-73)A* and *nrg1-G137T* mutations caused the observed increases in SSL tolerance, I reverted these positions wildtype in the point mutants. This reversion to wildtype abolished the enhanced phenotype of haploid single mutants (Figure 3.11). I also tested whether the effects of these mutations was additive or epistatic by generating double mutants, in both haploid and diploid backgrounds. I did not detect additional gains or losses of tolerance in haploid and diploid *nrg1 gsh1* double mutants.. This observation is consistant with the linear model of haploid backcrosses, which predicts negative epistasis between these two mutants (Figure 4.7).

3.7 Mutations in *NRG1* confer tolerance to acetic acid and oxidative stress

In an attempt to identify the physicochemical stresses to which *nrg1* and *gsh1* mutations conferred tolerance, I compared the growth of haploid single mutants with wildtype and R57 cells in the presence of increasing concentrations of acetic acid (Figure 3.12A) and hydrogen peroxide (Figure 3.12B). Haploid *nrg1-G137T* mutants display increased tolerance to both compounds, growing faster in the presence of higher concentrations than both wildtype and R57 cells. The *gsh1-T(-73)A* mutation does not seem to confer the same advantage. This latter result is surprising considering the well-characterized role of *GSH1* in the oxidative stress response. However, the concentration range used in the experiment was largely uninformative: most conditions were inhibitory to all genetic backgrounds including R57. A narrower concentration range may have revealed an enhanced phenotype in *gsh1-T(-73)A* mutants.

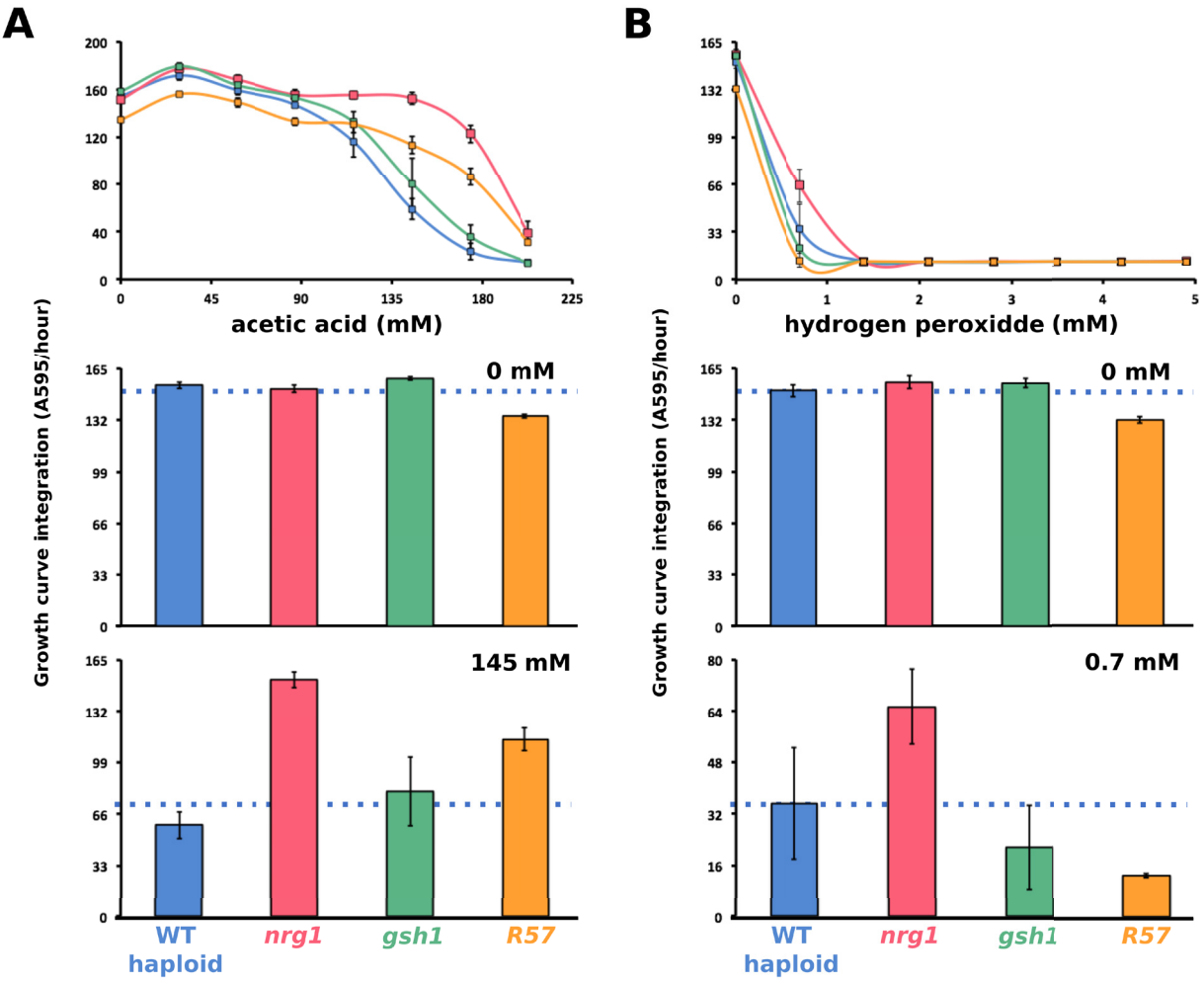


Figure 3.12 The *nrg1-C137A* allele confers tolerance to increased acetic acid and hydrogen peroxide concentrations. Area under the growth curve of wildtype haploid (blue), *nrg1* (red), *gsh1* (green) and R57 (orange) cells in presence of increasing concentrations of acetic acid (A) and hydrogen peroxide (B) is reported. Data for selected points of the dose response curves (top panel) are reported in the middle and bottom panels.

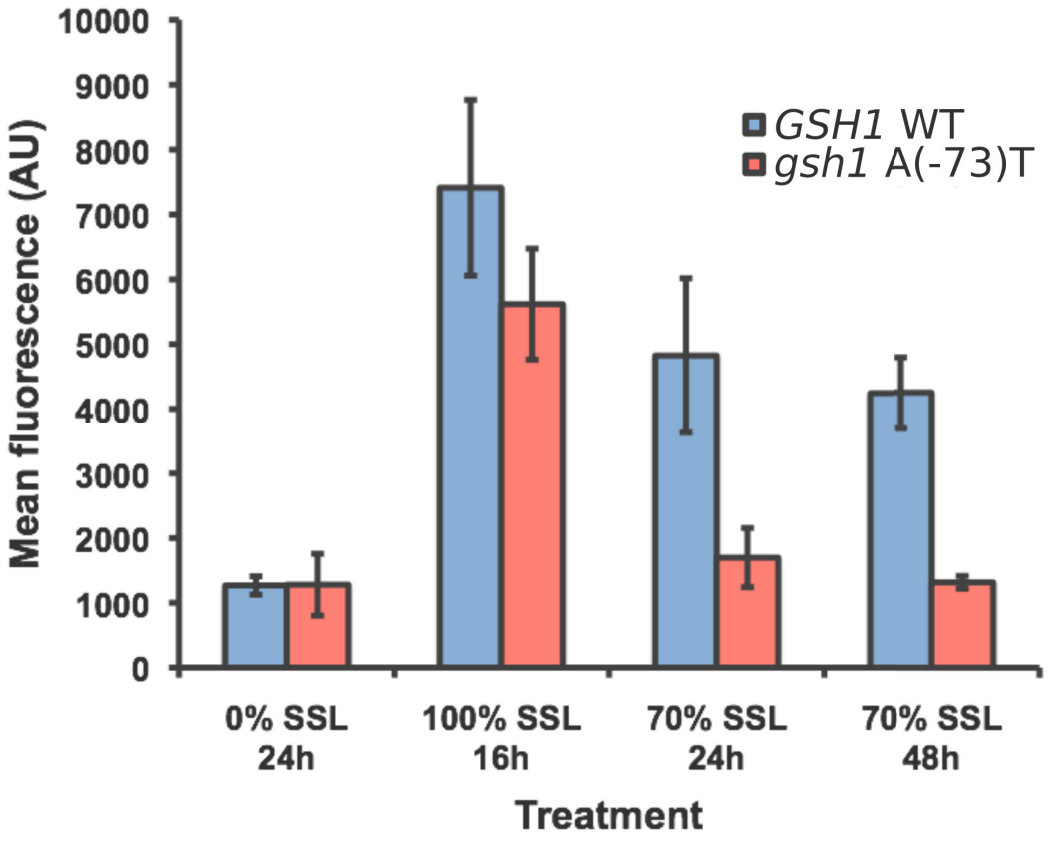


Figure 3.13 Mutants carrying the *gsh1-A(-73)T* allele accumulate less reactive oxygen species than the wildtype in the presence of SSL. ROS accumulation induced by exposure to SSL was compared between a *gsh1-A(-73)T* point mutant, and its parent wildtype strain (WT). ROS accumulation was assessed using flow cytometry, measuring the mean fluorescence of cells treated with CellROX Deep Red reagent. ROS were measured 16h after inoculation in minimal medium (Mid-log), after overnight incubation in undiluted SSL (acute stress), or after 24h and 48h in minimal medium containing 70% (v/v) SSL**.**

3.8 Mutation *gsh1-T(-73)A* reduces ROS accumulation upon exposure to SSL.

Because the synthesis of reduced glutathione and the recycling of its oxidized form play a major role in tolerance to oxidative stress, I hypothesized that a mutation of the *GSH1* promoter would modulate glutathione synthesis in the cell and thereby affect levels of reactive oxygen species (ROS). Since the mutant strain R57 was selected for its tolerance to SSL, I used the fluorescent CellROX Deep Red Reagent and flow cytometry to assess cytosolic ROS accumulation in wildtype and *gsh1 A(-73)T* cells upon exposure to SSL. When grown in non-toxic medium (YNB 1% glucose), the wildtype and mutant stains accumulate comparably low levels of ROS (Figure 3.13). Subsequent exposure to undiluted SSL similarly increases ROS levels in both strains. However, following acute stress induction in 100% SSL and transfer to YNB 1% glucose supplemented with 70% SSL, the mutant accumulates markedly lower amounts of ROS after 24 and 48 hrs incubation, suggesting the *gsh1 A(-73)T* mutation affects cell response to oxidative stress.

3.9 Reversion of single mutations in R57 suggests a compensatory role for mutant *gdh1* alleles

Each mutation in the R57 mutant strain was reverted to wildtype (Figure 3.14A). The majority of R57 revertants behave like their parent in the presence of SSL. Reversion of the *gdh1-G47A* mutation in R57 and related derivatives wildtype at the *GDH1-68* and *SGO1-575* positions leads to a pronounced decrease in fitness in the presence of SSL, also noticed in inhibitor-free medium. From this observation, I hypothesized that *gdh1* mutations complement secondary deleterious mutations found in R57. I therefore generated series of R57 derivatives from the *GDH1* or *SGO1 gdh1-*

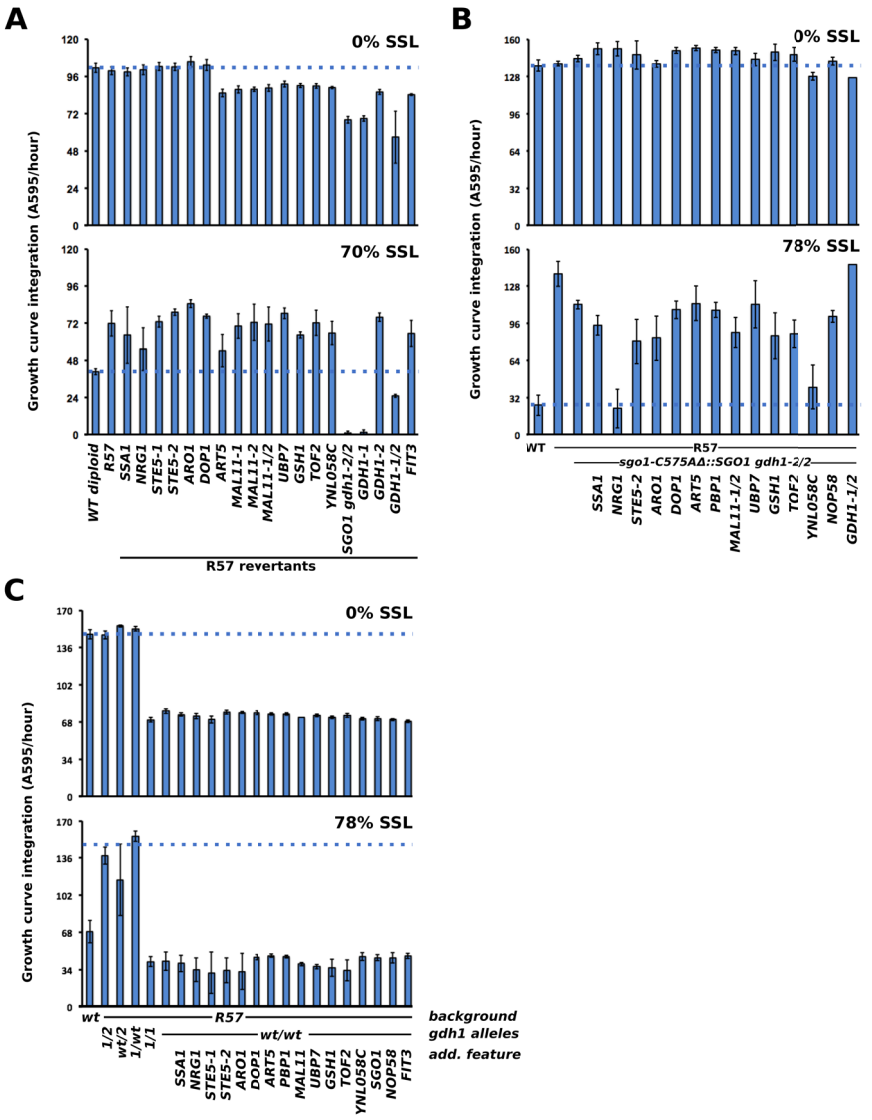


Figure 3.14 Reversion of single mutations in R57 suggests a compensatory role for mutant *gdh1* alleles and identifies single mutations involved in SSL tolerance. Area under the growth curve in the presence and absence of SSL is reported for revertant derivatives of R57 (A), R57 *SGO1 gdh1-2/2* (B) and R57 *GDH1* (C) wildtype for the indicated genes.

T68G backgrounds, wildtype at each secondary mutation. The R57 *GDH1* growth defect could not be rescued by reversion of single secondary mutations (Figure 3.14B). While I could not fully reproduce the R57 *SGO1 gdh1-T68G* growth defect, possibly because of batch-to-batch variations in SSL composition, I identified triple revertants with reduced tolerance to SSL (Figure 3.14C). Removal of mutation *ynl058c-A7G* in this background confers wildtype tolerance to SSL. The same phenotype is observed in the R57 *SGO1 gdh1-T68G NRG1* triple revertant, consistant with other observations concerning mutation *nrg1-C137A*. Intriguingly, the sole reversion of *nrg1-C137A* to wildtype in R57 is not sufficient to reduce tolerance to SSL (Figure 3.14A), while its presence in heterozygous and homozygous diploid mutants does not enhance the ability to grow in the presence of this lignocellulosic hydrolysate (Figure 3.10 C and D). Together, these results suggest that expression of the *nrg1-C137A* phenotype in diploids depends on epistasis with other mutations, possibly in gene *GDH1*. Alternatively, the presence of several tolerance-conferring mutations in R57 may ensure the robustness of its phenotype.

3.10 SSL tolerance of *gdh1* mutants

To test the effect of *gdh1* mutations on the SSL tolerance phenotype, various point mutants were generated by CRISPR-Cas9 from the wildtype and R57 backgrounds (Figure 3.15). Growth curves of these mutants in the presence of various concentrations of SSL were recorded. In the absence of SSL, most haploid *gdh1* point mutants grow as well as their wildtype parent, with the exception of A1232G, which displays reduced growth. Addition of SSL reveals decreased fitness compared to wildtype. Growth defects are observed both in the presence and absence of SSL in

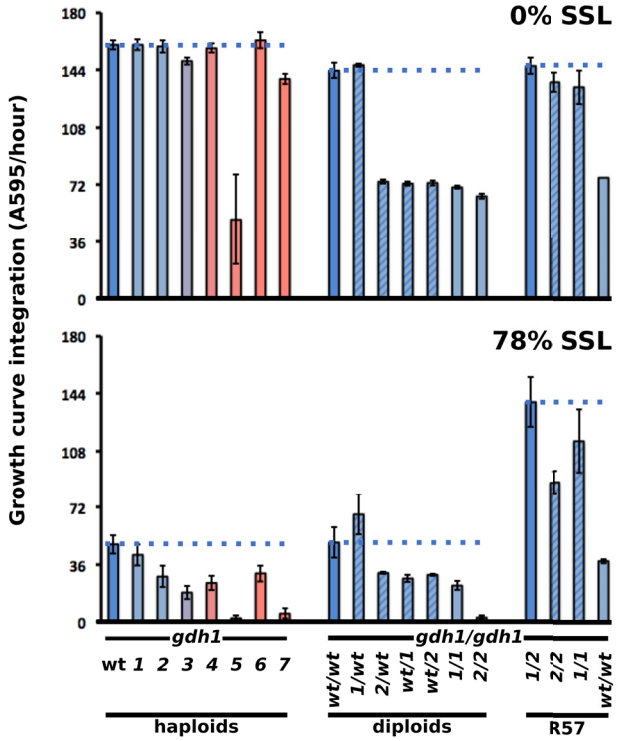


Figure 3.15 SSL-tolerance of *gdh1* mutants. Area under the growth curve of haploid and diploid *gdh1* mutants was compared to wildtype in presence (0% (v/v) SSL) and absence (78% (v/v) SSL) of SSL. Derivatives of R57 with various *gdh1* genotypes were similarly compared to their parent strain. Color coding of the mutations is the same as in Figure 3.4.

diploid *gdh1* point mutants, both heterozygous and homozygous. Reversion of either *gdh1-G47A* or *gdh1-A68G* in R57 does not appear to alter growth in the absence of SSL, but the *GDH1/GDH1* genotype is associated with a growth defect in this strain. Mutations in gene *GDH1* appear to contribute to the growth of R57 in SSL, with reduced growth most obvious when *gdh1-G47A* is reverted either alone or in combination with *gdh1-A68G*.

4. Discussion

with excerpts from:

Biot-Pelletier D et al (2016) The impact of historical contingency on the outcomes of evolutionary engineering. Manuscript in preparation.

4.1 Physiology of selected mutations

4.1a Mutations in genes *NRG1* and *GSH1* are the main determinants of SSL tolerance in our genome shuffling experiment

Much evidence indicates that mutations *nrg1-C137A* and *gsh1-A(-73)T* confer the strongest direct gains in SSL tolerance among all SNPs considered in this study. In both the haploid and diploid models of backcrossed mutants, mutation *nrg1-C137C* contributes the largest increases in tolerance (Figure 3.7). The diploid model further predicts that cells homozygous for *nrg1-C137A* are superior to heterozygotes, suggesting that this mutation is either incompletely dominant or recessive. My models predict that mutation *gsh1-T(-73)A* makes the second largest contribution to SSL tolerance, but in haploids only.

In agreement with predictions from the model, growth curves of haploid single mutants detect gains in SSL tolerance for *nrg1-C137A* and *gsh1-T(-73)A* mutants only (Figure 3.10). These observations support the hypothesis that, at least in the absence of additional interacting mutations, those two mutations allow the strongest increases in SSL tolerance. Further supporting this claim, reversion of *nrg1-C137A* to wildtype in R57 *SGO1*/*SGO1 gdh1-A68G*/*gdh1-A68G* cells reduces tolerance to SSL (Figure 3.14B).

Additional evidence supports a strong role for *NRG1* in the phenotype of interest. Previous RNAseq results in R57 showed considerable upregulation of five targets of transcriptional regulator Nrg1p, including *NRG1* itself [9]. Population sequencing identified this gene as a mutation hotspot, with one additional mutation mapping to the *NRG1* open reading frame. Low frequency characterizes both *nrg1* alleles at all time points, but highly positive apparent selection is consistent with a beneficial effect on the tolerance phenotype (Figure 3.3). From my data, I cannot rule out the possibility that both *nrg1* mutations arose during the same mutational event. Considering that the *nrg1-C505A* introduces a stop codon, one may hypothesize that it precluded full expression of the beneficial phenotype associated with the *nrg1-C137T* mutation, perhaps explaining its persistently low frequency. Only the low-probability uncoupling of both mutations by recombination would have allowed expression of the SSL tolerance trait enabled by *nrg1-C137T*.

I have shown that haploids carrying the mutant *gsh1-T(-73)A* allele accumulate lower levels of reactive oxygen species (ROS) than their wildtype parents when exposed to high concentrations of SSL (Figure 3.13). This is consistent with the role of Gsh1p in the synthesis of glutathione, a tripeptide of glutamate, cysteine and glycine involved in the scavenging of ROS. Condensation of glutamate and cysteine, catalyzed by Gsh1p, is the committed step in the synthesis of glutathione. Reaction with ROS oxidizes glutathione, which is reduced form by the NADPH-dependent Glr1p enzyme [347,348]. The mutation identified in our mutants is located 73 bp upstream of the start codon, likely altering the *GSH1* promoter. Upregulation of *GSH1* by this modified promoter is proposed to increase glutathione synthesis and reduce accumulation of ROS in *gsh1-T(-73)A* mutants. Intriguingly, I did not detect increased tolerance to hydrogen peroxide in

gsh1-T(-73)A haploid mutants (Figure 3.12). I hypothesize that I tested a range of peroxide concentrations that was both too wide and not fine enough to properly characterize the dose-response relationship. Further investigation would be required to fully describe the tolerance of *gsh1-T(-73)A* mutants to oxidative stress. Nevertheless, the near sweep of the *MATα* mutant pool by the *gsh1-T(-73)A* allele suggests a significant selective advantage in the presence of high concentrations of SSL.

Together, my results suggest that *nrg1-C137A* and *gsh1-T(-73)A* are the core determinants of SSL tolerance in our experiment. Prevalence of the *gsh1* allele and scarcity of *nrg1* argues that the former had a greater impact on the evolutionary dynamics of our experiment. I favour a model in which *nrg1-C137A* and *gsh1-T(-73)A* confer basic tolerance, with other mutations responsible for epistatic phenomena.

Repression of stress response genes by transcriptional regulator Nrg1p

NRG1 and its close paralog *NRG2*, encode C2H2 zinc finger DNA-binding proteins, and were first identified as mediators of glucose repression [349,350]. As such, *NRG1* transcript levels are downregulated 6-fold in the presence of ethanol and glycerol as carbon sources, yet are induced 2.7-fold during diauxic shift [351]. Both Nrg1p and Nrg2p appear to be regulated by the Snf1p protein kinase, with which they interact and is involved in the response to changes in carbon sources. This regulation is not localization dependent, as Nrg1p constitutively localizes to the nucleus [352]. Snf1p/Nrg1/2p pairs were further demonstrated to regulate invasive growth via protein Flo11p in conditions of glucose limitation [353]. The Nrg1/2p repressors have been implicated in the response to stresses other than nutrient limitation. The earliest evidence has shown that *NRG2* is upregulated by zinc limitation [354], as well as alkaline

pH in a Rim101p-dependent manner [355]. *NRG1* is similarly regulated by Rim101p [249] and is known to repress *ZPS1*, involved in response to low zinc [354], as well as *ENA1*, a sodium/lithium pump involved in salt tolerance [249,356].

Identification of transcripts with altered expression in null mutants of *NRG1* and *NRG2* established their role in the regulation of the general stress response [357]. Deletion of *NRG1* or *NRG2* changes the transcription of 150 genes, almost two-thirds of which are glucose-repressed and show increased transcription upon glucose limitation. However, many have STREs or STRE-like sequences in their promoter regions. Meta-analysis of expression studies on these genes suggests general regulation by stress or by general stress response regulators Msn2p/Msn4p. The binding sites of Nrg1/2p and Msn2/4p seem to overlap, further supporting the hypothesis that Nrg proteins are responsible for the repression of general stress response genes [349,358]. Vyas and coworkers further showed that substitutions in either Nrg1p or Nrg2p increase tolerance to various stresses, notably salt and oxidative stresses [357]. More recently, *NRG1* was identified among acetic acid responsive genes of the Haa1p regulon [257]. Together, this evidence argues for a role of Nrg1p in repression of general stress response genes. I therefore propose that *nrg1-C137A* is a loss of function mutation that leads to the upregulation of various protective pathways, as evidenced by its protective effects against both acetic acid and hydrogen peroxide in the haploid CEN.PK113-1A background (Figure 3.12).

Selection of mutations affecting global gene regulation, like *nrg1-C137A*, is a common theme of experimental evolution in stressful conditions. For example, Dettman and coworkers reviewed loci of evolution experiments conducted in yeast, and found a prevalence of genes involved in the SAGA-mediated stress response pathway [21].

Another well-documented example is the selection of mutations affecting the *rpoS* stress response sigma factor in *E. coli*, which was identified as selected in response to five different environments [72,359–363]. Loss of this stress response factor is well documented in nature where it is hypothesized to arise from a trade-off between stress-resistance and growth rate [364]. Mutations that diminish the function of stress response systems are a common outcome of experimental evolution. It is hypothesized that such genetic changes relieve the halted growth phenotype typically associated with stress to enable proliferation in the presence of constant stress [363,365,366]. The repressor function of *NRG1* rather suggests that the mutant alleles of this gene selected by our genome shuffling experiment lead to a general upegulation of stress response functions.

Glutathione metabolism and regulation of GSH1 *in* Saccharomyces cerevisiae

The role of glutathione in response to oxidative stress was introduced in Chapter 1 of this thesis. We have seen how the oxidation of this tripeptide of glutamate, cysteine and glycine into its disulfide form participates in the scavenging of ROS, either via direct reaction or through the action of glutathione peroxidases [273]. This molecule is essential for growth, as shown by the fact that mutants lacking *GSH1* are auxotrophic for glutathione [367]. Glutathione is synthesized from its component amino acids in a two-step pathway. The first and rate-limiting step is the condensation of glutamate and cysteine by Gsh1p [347,348,368]. The later addition of glycine to form glutathione is catalyzed by Gsh2p. The biosynthesis intermediate γ-glutamylcysteine displays antioxidant activity of its own, suggesting that it is the evolutionary precursor of glutathione [368]. The recycling of glutathione from the oxidized to the reduced state is catalyzed by glutathione reductase, encoded by *GLR1* [273].

GSH1 is directly repressed by the presence of glutathione. However, in the absence of glutathione, transcription activator Met4p, involved in the biosynthesis of sulfur-containing amino acids, induces the expression of *GSH1*. Under oxidative stress, expression of *GSH1* is driven by Yap1p, a major oxidative stress response regulator in yeast [369,370]. Hydrogen peroxide responsive elements independent of Yap1p have also been identified in the *GSH1* promoter, further illustrating its role in oxidative stress response [371]. The *gsh1-T(-73)A* mutation, located in the promoter, reduces the level of reactive oxygen species accumulated upon exposure to SSL. This argues for elevated glutathione and Gsh1p levels in the mutant caused by alteration of the promoter. Interestingly, the mutation falls outside of characterized Yap1p and peroxide responsive elements [371], arguing for a regulatory mechanism that is independent of oxidative stress signals. The position of this SNP in the region proximal to the start codon identifies alteration of the basal promoter as the most likely mechanism.

4.1b Epistasis between *gdh1* and *gsh1* alleles

Several lines of evidence point to *GDH1* as a key element of either the SSL tolerance phenotype or the evolutionary dynamics of the genome shuffling experiment. Indeed, this gene is the most populated mutation hotspot identified by population genome sequencing, with seven non-silent mutations. Among those, three show both high frequency and strong apparent selection (Figure 3.4). Moreover, reversion of *gdh1* alleles to wildtype in the R57 background leads to a strong loss of fitness, both in the presence and absence of SSL, suggesting a compensatory role (Figure 3.14C). Reversion of *gdh1* alleles in R57 also enables the detection of the tolerance-conferring effects of *nrg1* and *ynl058c* alleles.

GDH1 encodes one of three glutamate dehydrogenase enzymes found in *Saccharomyces cerevisae*. This group of enzymes catalyzes the interconversion of glutamate into α-ketoglutarate and ammonia. The NAD^+-dependent Gdh2p enzyme catalyzes the deamination reaction [372], while $NADP^+$-dependent Gdh1p and Gdh3p facilitate the reverse [373]. Gdh1p is thus involved in the synthesis of glutamate and the fixation of inorganic nitrogen. Glutamate biosynthesis is also effected by Gdh3p, or by the combined action of glutamine synthetase (Gln1p) and glutamate synthase (Glt1p) [374,375]. Gdh1p and Gdh3p are highly homologous and catalyze identical reactions, but display different expression patterns. Under fermentative conditions, Gdh1p is the dominant glutamate dehydrogenase isoform, while carbon limitation, non-fermentable carbon sources and entry into stationary phase induce the expression of Gdh3p [376,377].

While transcription of *GDH1* is sustained at all phases of growth, cells that enter the stationary phase specifically degrade Gdh1p via ubiquitination at the specific Y426 residue. On the contrary, Gdh3p is specifically expressed during the stationary phase. Gdh1p is a faster enzyme (3-fold higher rate of α-ketoglutarate utilization), suited for growth-sustaining glutamate synthesis. On the other hand, Gdh3p is slower, better suited to sustain glutatione synthesis during the stationary phase and under stressful conditions [376]. Accordingly, Gdh3p has been implicated in stress tolerance in yeast. Deletion of *GDH3* leads to increased sensitivity to heat and hydrogen peroxide, as well as aging. The *gdh3Δ* mutant displays apoptotic markers under stress, and accumulates reduced levels of glutathione and glutamate. It also accumulates higher levels of ROS. These defects are suppressed by glutamate or glutathione supplementation, and by the secondary deletion of Gdh2p, which catalyzes the reverse reaction. In contrast, the

sharp losses of stress tolerance observed in *gdh3Δ* mutants are not reproduced upon deletion of close homolog *GDH1*. Transient loss of tolerance to hydrogen peroxide is nevertheless observed in the early phases of growth in *gdh1Δ* mutants [377].

The seven mutations of the *gdh1* hotspot selected by genome shuffling cluster near known targets of ubiquitination, suggesting they may increase the stability of the protein by restricting access of ubiquitinases to target lysine residues (Figure 3.4). This should increase the supply of glutamate to Gsh1p in actively growing *gsh1-T(-73)A* mutants. An alternative hypothesis on the molecular mechanism underlying the effect of *gdh1* mutations involves interdomain flexibility of the Gdh1p enzyme. The seven amino acid substitutions identified in this protein are found to cluster in and around the hinge region separating the two structural domains of Gdh1p. Of those, the three N-terminal substitutions map the closest to the groove found at the junction of both domains. Structural studies of glutamate dehydrogenase from *Clostridium symbiosum* demonstrated movement of these domains upon substrate binding, transitioning from an open to a closed conformation [378]. This transition is critical for efficient catalysis in glutamate dehydrogenase enzymes. It could be that steric changes in the Gdh1p hinge region impact baseline glutamate dehydrogenase activity. Regardless of the precise molecular mechanism underlying these mutations, I propose that the *gdh1* hotspot was selected to compensate the strong pull on glutamate exerted by upregulation of glutathione biosynthesis in *gsh1-T(-73)A* mutants. This hypothesis would explain the strong growth defect incurred by R57 upon reversion of *gdh1* mutations. In those cells, enhanced glutathione biosynthesis diverts glutamate from other essential physiological functions, leading to delayed growth. In support of this hypothesis, genetic interaction is suggested by the haploid model of backcrossed mutants between the *gdh1-C47T* and

gsh1-T(-73)A mutations (Figure 3.7). Increase in frequency of all *gdh1* mutations is observed at R1, or after the first recombination event, perhaps indicating that the *GDH1* locus is under strong selection as a consequence of the early **α3** sweep.

The *gdh1* phenotype is strongly epistatic. While we have seen that *gdh1* alleles are critical to the fitness of R57, their introduction into wildtype backgrounds is associated with growth defects. Increased glutamate dehydrogenase activity must result in accelerated consumption of $NADP^+$ and accumulation of NADPH and glutamate. This redox imbalance and glutamate surplus must have deleterious effects in the absence of increased glutathione synthesis. The context dependence of *gdh1* phenotypes is further illustrated by the superiority of heterozygous R57, carrying two different *gdh1* alelles, over its homozygous mutant derivatives. These results are hints that heterogeneous multimers of glutamate dehydrogenase confer a distinct competitive advantage in the presence of SSL (Figure 3.14 and 4.15). The dependence of *gdh1* phenotypes on the genetic background of their carrier strain is further demonstrated by the response of homozygous *gdh1-A68G* derivatives of R57 to the reversion of *nrg1-C137A* and *ynl058c-T7C* (Figure 3.14B). Indeed, reversion of these mutations alone in the original R57 background is not sufficient to lead to a loss of tolerance. Neither is this effect observed in a R57 *GDH1* background, perhaps because loss of both *gdh1* alleles leads to a loss of fitness that is too important to observe the effect of reverting to *NRG1* or *YNL058C*. I therefore propose that the partial loss of fitness in the homozygous *gdh1-A68G* derivatives of R57 leaves room to observe the additional effect of reverting other mutations.

Results presented in this thesis show a strong evolutionary response of the *GDH1* locus in response to the prominence of the *gsh1-T(-73)A* allele. Deleterious in the

wildtype background, mutations in this gene appear required in the R57 mutant, indicating a role in complementation of defects caused by other mutant alleles. I propose that their role is to increase glutamate supply in glutathione-overproducing cells.

4.1c Hypotheses on the role of *mal11* mutations

In *S. cerevisiae*, maltose utilization is encoded by the *MAL* complexes. Anyone of these complexes consists of three genes. Gene 1 of these complexes encodes a permease responsible for maltose import, while gene 2 encodes for a maltase, which hydrolyzes the disaccharide to glucose. Gene 3 encodes a regulatory protein responsible for activation of maltose utilization functions [379]. The genome of strain CEN.PK113-7D used for the preparation of this thesis contains four *MAL* complexes numbered 1 to 4 [380]. Gene *MAL11*, a mutational hotspot in our genome shuffling experiment, encodes the maltose permease of the *MAL1* complex. Mal11p is a proton-driven symporter of the major facilitator superfamily [381,382]. It was initially described as a maltose transporter, but was later shown to possess broad specificity for diverse α-glucosides, notably the stress protectant molecule trehalose [383]. The link between genes involved in maltose metabolism and trehalose accumulation is well-established, suggesting a potential role for *MAL11* in stress tolerance [384]. The two *mal11* mutations identified by our sequencing efforts code for amino acid substitutions to the N-terminal portion of the protein. These substitutions are located outside the core MFS fold predicted for Mal11p. These substitutions may therefore affect the regulation of Mal11p activity rather than directly modulate its transport function.

MAL genes are glucose-repressed, notably by transcriptional regulator Mig1p in a Snf1p-dependent manner [385,386]. However, Mig1p is not sufficient for complete glucose repression of *MAL* complexes, and other components downstream of the Snf1p kinase appear to be involved [387]. Glucose repression of Mig1p target *STA1* by Nrg1p was demonstrated, suggesting other glucose-repressed genes, such as *MAL11*, may be similarly regulated [349]. My model of haploid R57 segregants suggests a modest effect of *mal11-C310T* in the absence of *nrg1-C137A*, but predicts positive epistasis between the two mutations. I therefore hypothesize the regulation of *MAL11* by Nrg1p. Interestingly, while most of the sequence upstream of *MAL11* is highly similar to that of other *MAL* permease genes, the portion proximal to the start codon is unique, echoing the unique substrate specificity of Mal11p [383]. This further supports the possibility of a distinct regulatory mechanism, perhaps involving Nrg1p.

From the predicted epistasis between *mal11-C310T* and *nrg1-C137A*, it is tempting to draw a parallel with the *gdh1*/*gsh1* pair described earlier. However, from an evolutionary point of view, the selection of *mal11* alleles in response to *nrg1* mutations appears unlikely. Indeed, *mal11-C310T* rises to prominence during the early rounds of genome shuffling, and remains considerably more abundant than either mutations of the *nrg1* hotspot. In other words, unlike *gsh1-T(-73)A*, mutant alleles of *NRG1* do not benefit from a strong founder effect: they are therefore unlikely to drive large scale compensatory evolution, as would otherwise be inferred from the high frequency of *mal11-C310T*. While the *nrg1-C137A* allele confers a large increase in tolerance on its own in haploid cells, interaction with *mal11* alleles may have promoted its selection. I thus favour a view in which *mal11* alleles are selected either for a direct structural effect on the tolerance phenotype, or in complementation of secondary mutations, like those of

the *nrg1* hotspot. In both cases, I propose that the effect of *mal11* mutations is caused by an alteration of trehalose transport, a molecule that promotes survival under environmental stress.

4.1d Genes involved in protein homeostasis

Sequencing identified high-frequency mutations in genes *ART5* and *UBP7*, as well as a mutational hotspot mapping to *UBP1*. All three genes are involved in protein homeostasis, and my results suggest their potential implication in the response to membrane protein misfolding. Misfolding at the yeast membrane triggers ubiquitin-dependent endocytosis and vacuolar degradation of misfolded membrane proteins [388,389]. This process requires the action of the ubiquitin ligase encoded by *RPS5* [390–392]. Rps5p ubiquitinates its targets either directly or with the assistance of adapter proteins of the *ART* family. Cells defective in the function of *ART* genes display increased sensitivity under stress [393]. Among coincidental founding mutations identified by population genome sequencing is a single nucleotide substitution in gene *ART5*, detected in all mutants of the initial *MATa* pool (Figure 3.2). This mutation seems to have had a profound impact on the evolutionary dynamics of the genome shuffling experiment, as it can be related to the highly frequent *ubp7-T2466A* mutation and to the *ubp1* mutational hotspot. Both *UBP7* and *UBP1* encode for ubiquitin-specific proteases, involved in protein deubiquitination. In particular, Ubp7p has been directly involved in clathrin-mediated endocytosis via interaction with endosomal sorting complex protein Hse1p [394,395]. Recent findings indicate an additional role for Ubp7p in the DNA damage response [396]. Association of Ubp2p, another yeast deubiquitinase, with Art5p partner Rsp5p has been documented [397–399]. More generally, it has been demonstrated that the

association of membrane ubiquitin ligases with deubiquitinating enzymes increases their discriminating power and ensures specific degradation of terminally damaged proteins [400].

Together, these considerations suggest that repeated selection of *ubp1* mutations, and the persistence of *ubp7-T2466A* are a direct response to the coincidental presence of the *art5-G454T* mutation in the *MATa* founding mutants. I propose that alteration of Art5p in these mutants deregulated membrane protein homeostasis by either stimulating or reducing Rsp5p recruitment to misfolded targets. In this perspective, ubiquitin protease substitutions would have a complementing role by antagonizing the deleterious effect of the *art5-G454T* allele. The sensitivity of *art* mutants to stress [393] makes this rescue mechanism all the more relevant in the context of selection for increased SSL-tolerance, and could explain the persistence and high frequency of the *ubp7-T2466A* allele, as well as the repeated selection of *ubp1* mutations.

Another mutation identified by sequencing and suggested to promote SSL-tolerance can be related to protein homeostasis. Mutation *ssa1-C91A* is detected at low frequency throughout the evolutionary engineering experiment, but displays stongly positive apparent selection, in a manner reminiscent of *nrg1* mutations (Figure 3.1). It maps to one of four highly homologous *SSA* genes in yeast, encoding for cytosolic chaperones of the Hsp70 family [401,402]. The main roles of these chaperones are to bind newly-translated proteins to prevent misfolding and aggregation, and to assist in proper folding [403,404]. *SSA1* is constitutively expressed at significant levels, but induced by heat-shock [405,406]. In addition to this well-documented role in heat shock, *SSA1* has been implicated in the response to oxidative stress [407] and DNA damage [408,409]. In this

respect, the recent implication of *UBP7* in the response to DNA damage may be a hint that *ssa1-C91A* and *ubp7-T2466A* address a common evolutionary problem in the context of our genome shuffling experiment [396]. Ssa1p also has been implicated in the disassembly of clathrin-coated vesicles, which directly relates to the function of Art5p [410]. However, *UBP7*- and *ART5*-independent selection at the *SSA1* locus would not be surprising given its well-documented role in protein folding and the stress response. One may note that the *ssa1-C91A* phenotype appears epistatic. The linear models of R57 backcrosses predict a modest benefit to SSL tolerance in haploids, but the contrary is true in diploids (Figure 3.7). This context dependence mirrors findings on the effect of Ssa1p overexpression on prion toxicity, where it was found to promote the formation of some prions [411,412] and cure cells from others [413].

Above, I have argued that the selection of mutations in genes *ART5*, *UBP7* and *UBP1*, involved in protein homeostasis, is largely an evolutionary artefact. I believe it is a case of compensatory evolution in response to the contingent presence of *art5-G454T* in half of the founding mutants of the genome shuffling experiment. Whether the *ssa1-C91A* mutation serves a similar purpose is open to debate. Protein homestasis is a likely target of evolution during adaptation to high stress, and the beneficial effect of some or all of these mutations on tolerance to SSL cannot be excluded. For example, loss-of-function mutations at the *ART5*, *UBP1* and *UBP7* loci could lead to a reduction in the endocytosis of membrane proteins in conditions of stress. The resulting maintenance of certain membrane functions upon SSL exposure could have confered a selective advantage to our genome shuffled mutants. Both hypotheses are not mutually exclusive, and it may be that compensatory evolution led to a more robust SSL tolerance phenotype.

4.1e Rationalizing the effect of other potential contributing genes

My results contain evidence for the contribution of other genes and mutations to SSL tolerance. While this evidence is not as strong as that discussed above, it may identify genes that significantly impact tolerance to inhibitors found in SSL. This section attempts to succinctly present the affected genes and provide tentative mechanisms for their contribution to the phenotype.

YNL058C

Reversion of the *ynl058c-A7G* mutation in R57 does not cause a detectable loss of tolerance to SSL. However, in a derivative of R57 wildtype at the *SGO1* locus and homozygous for the *gdh1-A68G* allele, reversion to wildtype at the *YNL058C* locus leads to a loss of SSL tolerance. This suggests a role for this gene in tolerance to SSL. The function of *YNL058C* is essentially unknown. The protein it encodes appears to localize to the vacuole [414]. It has a paralog, *PRM5*, predicted to have one transmembrane segment. Both *YNL058C* and *PRM5* are induced via the cell wall integrity pathway, indicating a role in the response to cell wall damage [415,416]. Downregulation of *YNL058C* was also observed upon DNA damage [417]. Together, these reports indicate a role for *YNL058C* in the response to stress and cell damage, in agreement with an involvement in SSL tolerance.

TOF2 and SGO1

Two mutations mapping to genes involved in chromosome segregation, *sgo1-C575A* and *tof2-C2141T*, were identified by our sequencing efforts. Mutation *sgo1-C575A* is part of the prominent **α3** cohort, and is thus proposed to have hitchhiked with

the *gsh1-T(-73)A* and *ynl058c-T7C* alleles. The model of haploid backcrosses leaves little doubt about the deleterious effects of *sgo1-C575A*. Results further suggest that reversion to the wildtype *SGO1* genotype facilitates detection of the beneficial effects of *nrg1-C137A* and *ynl058c-T7C* alleles. The *tof2-C2141T* mutation was identified by sequencing of the R57 genome, but escaped detection by whole population sequencing, suggesting a very low frequency throughout the evolution process. While this latter observation would suggest a modest influence on the selected phenotype, both models of R57 backcrosses suggest a beneficial effect for the *tof2* mutant allele (Figure 3.7). Phenotypic evidence thus indicates antagonistic effects of *sgo1-C575A* and *tof2-C2141T* on the SSL tolerance phenotype.

SGO1 encodes shugoshin, a pericentromeric adaptor protein involved in the integration of multiple important functions required for chromosome segregation [418–420]. It was initially identified for its role in the regulation of chromosome segregation during meiosis [421–423]. Yeast shugoshin was notably implicated in sensing of mitotic spindle tension [424–426]. Tof2p was first characterized as a protein that interacts with topoisomerase I [427]. It has been mainly implicated in the silencing, condensation and segregation of rDNA during cell replication [428–431]. Both shugoshin and Tof2p have been shown to recruit condensin, but in different contexts [420,428]. Together, implication of both proteins in chromosomal segregation during cell replication, the strong hitchhiking-driven founder effect favouring *sgo1-C575A*, and the positive effect of *tof2-C2141T* hint that both mutant alleles may be involved in a compensatory mechanism, whereas mutation in *TOF2* complements a lesion in *SGO1*.

STE5

Mutation *ste5-C512T* dislays the highest apparent selection and average frequency in cohort **a6**. The model of R57 backcrosses indicates that it makes the second highest contribution to SSL tolerance in diploids. *STE5* encodes a scaffold protein involved in facilitation and intregration of pheromone-induced MAPK signalling [432]. In haploid cells, binding of mating pheromones to their cognate G protein coupled receptors (GPCRs) triggers nucleotide exchange and the release of the G protein α subunit from the αβγ complex formed by Gpa1p, Ste4p and Ste18p. The free Gβγ complex becomes available to interact with Ste5p, which it recruits at the plasma membrane. Ste5p forms a signalling complex with MAP kinase pathway components Ste11p (MAPKKK), Ste7p (MAPKK) and Fus3p/Kss1p (MAPK). This complex continuously cycles between the nucleus and cytoplasm, but interaction of Ste5p with the Gβγ complex causes its enrichment at the membrane and activates the phosphorylation cascade via conformational changes [432,433].

Sterile (*STE*) genes were identified for their involvement in mating, but components of the pheromone-induced MAPK pathway also were implicated in stress signalling. For example, Ste20p (MAPKKKK) and Ste11p are involved in the HOG pathway to activate osmotic stress response [434,435]. Those same proteins, along with Ste7p and Kss1p, mediate the nutrient starvation signals that lead to invasive and pseudohyphal growth in haploids and diploids [436,437]. This pathway was similarly implicated in promoting cell wall integrity [438]. Specificity for the mating pathway appears to be conferred by the identity of the effector MAPK. Fus3p is proposed to specifically mediate mating response, and Ste5p acts as its specific scaffold [439–442]. Thus, Ste5p

imparts specificity to the MAPK output in addition to integrating and faciliting transmission of the signal. However, because Ste5p can bind the Kss1p MAPK, it mediates signals involved in stress. It is thus likely that the *ste5-C512T* mutation modulates the Ste20p-Ste11p-Ste7p-Kss1p pathway to stimulate the execution of a stress response program. However, considering the heavy reliance of our genome shuffling experiment on yeast's sexual reproduction cycle, it cannot be excluded that selection favoured mutants with increased mating efficiency, and *STE5* appears as a likely target for such improvements.

4.2 Evolutionary dynamics of genome shuffling experiments

4.2a. Clustering of highly correlated mutations suggests widespread genetic hitchhiking

From the clustering of mutations in cohorts of correlated evolutionary trajectories, I propose that a large proportion of mutant alleles arose together in a few founding individuals. While these mutants carried one or a few beneficial mutations that increased tolerance to SSL, the rest of their SNPs hitchhiked and rose to prominence despite their neutral or deleterious effects. From the list of mutations detected by population sequencing, a restricted subset is thus expected to contribute to the phenotype of interest. In support of this claim, I have observed that few single mutants display an enhanced phenotype (Figure 3.10). Interaction with secondary mutations may be required for some SNPs to express a phenotype, but their selection before recombination is not expected unless epistatic pairs are found by luck in initial mutants. Cohorts **a5**, **a6** and **α3** are the most prominent examples of clusters of correlated evolutionary trajectories, owing to the high frequency of their mutations (Figure 3.1). The

trajectories and apparent selection of mutations in these groups point to candidate driver mutations.

Cohort **α3** clusters the most frequent mutations detected by population sequencing. Their correlated trajectories, added to their high and identical frequencies in the initial mutant pool are strong evidence for a common origin. The distinct trajectory and positive apparent selection of mutation *ubp7-T2466A* would suggest that it is the driver for cohort **α3**. While this hypothesis would agree with our previous observations on the growth of *ubp7-T2466A* single mutants on SSL gradient plates [9], it does not align with results of the current thesis. The haploid model of backcrossed mutants suggests a deleterious effect for *ubp7-T2466A* (Figure 3.7). The same model proposes that *gsh1-T(-73)A*, an **α3** mutation, enhances tolerance to SSL, as confirmed by the phenotype of single haploid mutants (Figure 3.10). Reversion of *ynl058c-T7C* also leads to a loss of tolerance in R57-derived mutants. The fact that I was not able to reproduce the SSL tolerance phenotype previously detected in *ubp7-T2466A* mutants underlines potential differences between growth in liquid and solid media. It may suggest that our growth curve assays trade sensitivity for throughput when compared to gradient plates. In this perspective, the serendipitous union of three beneficial mutations in **α3** could explain its sweep and near fixation at early selection steps.

In cohort **a6**, decline in frequency is sharp after round 3, but is milder for *ste5-C512T*, leading to a larger mean frequency change than other SNPs in the same cluster. The diploid model of R57 backcrossed mutants suggests a positive contribution of this allele to the SSL-tolerance phenotype. Another candidate driver for **a6** is *gdh1-G299A*, which maps to the most populated hotspot identified by population sequencing. The common evolutionary pattern between cohorts **a1**, **a5** and **a6** may suggest that they in

fact form a single large cohort of mutations sharing a common origin. Following this hypothesis, and assuming a single driver mutation for the entire group, the strongly positively selected *mal11-G310A* mutation, mapping to a hotspot, is the most likely candidate driver. Evidence for the effect of all three mutations is provided by the linear models (Figure 3.7) and from growth curves of mutants in SSL (Figures 4.10 and 4.14), arguing against this hypothesis. These contradicting explanations may indicate that the pattern observed in those three groups is not due to a common origin, but rather to underlying evolutionary mechanisms similarly affecting all three cohorts. For example, a recombination event at R4 may have generated a superior strain, leading to the displacement of most **a5** and **a6** mutations. Otherwise, the negative interaction proposed by the diploid model between *ste5-C512T* and *mal11-C310T* suggests that recombination is necessary to separate both mutations and allow the full expression of their beneficial effects. Alternatively, differences in apparent selection may indicate varying degrees of contribution to the SSL tolerance phenotype within the cohort. Again, encounter of several driver mutations in a single initial mutant is not expected, but cannot be excluded.

Genetic hitchhiking is expected to have important consequences on the outcomes of genome shuffling. It is expected to negatively affect the fitness of mutants, because the majority of mutations are expected to be neutral or deleterious [443]. Prevalent hitchhiking represents a major confounding factor when attempting to link genotype to phenotype by sequencing of single mutants. Identifying productive mutations and genes critical to a phenotype of interest can become a challenge in mutants containing few productive mutations hidden among a large number of

hitchhikers. Population genome sequencing to track evolutionary trajectories and direct phenotypic measurements help identify driver mutations.

Our results argue for a tight control of initial mutagenesis conditions. In our experiment, we aimed for an average of one point mutation per mutant by tuning the UV dose to a specified kill rate. Nevertheless, mutants with multiple mutations appear to dominate the pool. On the other hand, genome shuffling aims at selecting productive epistatic interactions, and efficient selection or screening methods may be able to take advantage of fortuitous combinations that may appear early from more aggressive mutagenesis. Whatever the preferred strategy, our results demonstrate the impact that genetic hitchhiking can exert on genome shuffling, and this information should guide the design of future evolutionary engineering endeavors.

4.2b Mutation hotspots indicate convergent evolution and identify key selective pressures

We have identified seven hotspots to which two or more mutations were mapped (Figure 3.3). Distinct evolutionary trajectories and different pools of origin indicate independent mutation events. The repeated selection of independent mutations mapping to the same gene is compelling evidence of a role in the phenotype of interest. Additional evidence suggests an influence of *nrg1*, *gdh1* and *mal11* alleles on the SSL tolerance phenotype, either alone or via genetic interactions (Figures 4.7, 4.10 and 4.14). Mutations among the most frequent and positively selected are mapped to the latter two hotspots. Together, these observations argue for a critical role of genes *NRG1*, *GDH1* and *MAL11* in our evolutionary engineering experiment and suggest that mutations from other hotspots may be involved in the phenotype of interest.

Three hotspots show striking correspondence with the so-called “founder” mutations identified by whole population sequencing (Figure 3.2). I relate mutations found in hotspot genes *YHR045W*, *SSN2* and *UBP1* with founder mutations in *MTM1*, *SRB8* and *ART5,* respectively. All cells in the initial *MATα* mutant pool carry the A943T nucleotide substitution in gene *MTM1*, encoding a mitochondrial manganese transporter involved in pyridoxal 5’-phosphate trafficking and iron metabolism [444,445]. Gene *YHR045W* has similarly been implicated in iron metabolism, with transcriptional profiling of *yhr045wΔ* mutants showing upregulation of genes involved in this metabolic process [446]. Similarly, the *ssn2* hotspot can be linked to the *srb8-C3787G* substitution. Both genes encode subunits of the RNA polymerase II mediator complex [447–451], suggesting that complementation of *srb8*-associated defects by *ssn2* alleles was selected by genome shuffling. The *ubp1* hotspot is notable, owing to the high frequency of the *ubp7-T2466A* mutation in our pool. Both genes encode ubiquitin-specific proteases [452,453]. The prevalence of the founding *art5-G454T* mutation, mapping to a gene involved in the regulation of membrane protein homeostasis [390], argues that complementing mutations involved in this cell process were selected during our experiment. The *yhr045w*, *ssn2* and *ubp1* mutations all remain at low frequency, with either weakly positive or negative apparent selection. This would suggest that epistasis with founder mutations confers marginal competitive advantages, and that mutations *mtm1-A943T*, *srb8-C3787G* and *art5-G454T* do not have strongly crippling effects on fitness. However, the prevalence and persistence of mutation *ubp7-T2466A* is hypothesized to result from the same epistatic dynamics. The coincidental presence of this mutation in the same mutant as the **α3** driver *gsh1-T(-73)A* likely caused the selection of *ubp7-T2466A* by hitchhiking,

while its arguably minor role in complementing *art5-G454T* was seemingly sufficient to ensure its persistence.

Other hotspots have been involved in biotin metabolism (*VHR1*) [446,454], and drug resistance (*COS111*) [455]. Considering their low abundance, frequently negative apparent selection and the absence of additional evidence, I cannot confidently ascribe a role for these mutations in the SSL tolerance phenotype.

4.2c The founder effect and compensatory evolution: impact of historical contingency

Above, I have discussed the five mutations present in the *MATa* and *MATα* parent strains before mutagenesis (Figure 3.2). Their presence in all reads of their respective pools of origin and their remarkable stability throughout the experiment argues for a modest effect with respect to the SSL tolerance phenotype, despite the slight heterozygote superiority predicted by the diploid model for the *art5-G454T* allele (Figure 3.7), and the compensatory mechanism inferred from mutational hotspots (refer to Section 4.2b above). At any rate, these mutations benefit from the founder effect and remain prominent throughout evolution despite a seemingly modest influence on the phenotype. The second founding feature of the experiment is the near fixation of cohort **α3** in the initial *MATα* pool resulting from an aggressive initial selection. I have evidence of a contribution to the SSL tolerance phenotype for mutations *ynl058c-A7G* and *gsh1-T(-73)A*. Mutation *ubp7-T2466A* is also proposed to play a role in relation to the founding *art5-G454T* allele. Other mutations in cohort **α3** are proposed to be neutral or deleterious (e.g. *sgo1-C575A*, Figure 3.7) yet persist at high frequency up to round 5

(Figure 3.1). By round five, **α3** mutations are still prominently represented in the population despite a decline from an initial average of 0.44 to a low of 0.34, illustrating the advantage they enjoyed by hitchhiking with strongly beneficial alleles.

The detection of seven independent mutations in gene *GDH1* and the reduced fitness of *gdh1-G47A* revertants of R57 suggest a major role for the complementation of deleterious mutations in the evolutionary dynamics of our genome shuffling experiment. The repeated and independent selection of mutations in *GDH1* indicates that the deleterious alleles they complement are frequent in the population. Once again, because of its early sweep and role in glutamate consumption, the **α3** mutation *gsh1-T(-73)A* is an obvious candidate counterpart for *gdh1* mutations. Accordingly, my model of backcrossed haploid mutants suggests a positive interaction between *gsh1-T(-73)A* and *gdh1-C47T*. The data presented above suggest several other instances of complementation between pairs of mutations. Previously, I have discussed the correspondence between founding mutations and several mutational hotspots. Those are the *ubp1-7*/*art5*, *yhr045w*/*mtm1* and *ssn2*/*srb8* pairs. Complementation of deleterious *sgo1-C575A* by *tof2-C2141T* was also suggested. Together, these links drawn between salient founding features of the experiment and loci under selective pressure suggest that compensatory evolution was a major driving force of our evolutionary engineering efforts.

Several lines of evidence indicate that the founding features of our genome shuffling experiment and the compensatory mechanisms that they elicited strongly biased the path of evolution towards specific evolutionary solutions. The salient features of this contingent starting point are the presence of spontaneous mutations in both

parental strains, and the near fixation of cohort **α3** in the *MATα* pool. These initial conditions led to a strong founder effect, which proved difficult to overcome. Moreover, it influenced the selection of other mutations, favoring compensatory evolution and effectively marginalizing genuine beneficial mutations.

Comparison of the evolutionary trajectories of *nrg1-G137T* and *gsh1-T(-73)A* further illustrates the impact of historical contingency. In spite of its apparent superiority over *gsh1-T(-73)A* (Figures 4.7, 4.10 and 4.12) and strong positive selection, mutation *nrg1-G137T* is still found at low frequency at the end of the experiment. The negative epistasis predicted by the linear model between these two mutant alleles may have further mitigated the selection of *nrg1* alleles. This comparison highlights the sensitivity of genome shuffling to the founder effect, and confirms the intuition that aggressive selection regimens early in evolution restrict genetic diversity at the expense of rare beneficial mutations. It is in agreement with the classical model of adaptation in asexual populations, which predicts that clonal interference precludes the persistence of genetic diversity, because it leads to periodic selective sweeps that purge it from the population. Possible linkage with the *nrg1-C505A* mutation could explain the low frequency of the demonstrably beneficial *nrg1-C137A* mutation. Both hypotheses raise concerns over historical contingency. Aggressive selection may have forced the selection of a "quick fix", effectively confining evolution to a local fitness maximum. If *nrg1-C137A* was indeed affected by linkage disequilibrium, then its low frequency would further illustrate the consequences of historical contingency. These observations resonate with Ersnt Mayr's concept of peripatric speciation as it relates to the founder effect, which predicts that in large and diverse populations, most selected alleles are those that play well with many others, because of strong and prevalent epistasis [456]. On the contrary in small isolated

populations, often subject to the founder effect, restricted diversity favours alleles that have strong effect on their own, or in a narrower range of genetic backgrounds. This result is reminiscent of a previous study which showed how sign epistasis between mutually exclusive beneficial alleles leads to a rugged fitness landscape [67,72,73]. Similarly, I suggest that negative epistasis coupled to aggressive selection forced selection towards a local fitness maximum, deviating the path of evolution from the global maximum.

I finally propose that the repeated positive selection and high frequency of *gdh1* and *mal11* mutations are explained by compensatory evolution in response to the **α3** sweep. I have shown how epistasis potentiates the effect of *gdh1* alleles. These alleles are detrimental in wildtype backgrounds, but are necessary for the expression of the SSL-tolerant phenotype in strain R57. Moreover, increases in tolerance imparted by *nrg1-G137T* and *ynl058c-T7C* require the presence of *gdh1-G47A* in the R57 background. In agreement with a background-dependent phenotype, I observe a stark increase in frequency of all *gdh1* mutations at the onset of recombination. A similar pattern is observed for *mal11* alleles: strong selection of *mal11-G310A* appears to depend on recombination, and our linear models propose dependence on genetic interaction with other alleles. Mutation hotspots at the *gdh1* and *mal11* loci suggest compensatory evolution to highly prevalent defects. Mutations from cohort **α3** most closely fit this profile. I therefore propose that the evolutionary signal detected at the *GDH1* and *MAL11* loci is a response to the early **α3** sweep. Together, our results argue that the historically contingent presence of spontaneous mutations in parental strains and the near fixation of a cohort of hitchhikers with few driver mutations had a profound effect on the outcomes of evolution. It prominently selected for epistasis to complement

the mutational burden brought by deleterious hitchhikers and founder mutations, while it reserved a relatively marginal role for additional beneficial mutations.

5. Conclusions and future directions

with excerpts from:

Biot-Pelletier D et al (2016) The impact of historical contingency on the outcomes of evolutionary engineering. Manuscript in preparation.

5.1 Proposed models for SSL tolerance and its evolution by genome shuffling

I propose a model of SSL tolerance based on the results of this study (Figure 4.1). In this model, mutations *nrg1-C137A*, *gsh1-T(-73)A*, *ynl058c-T7C*, *ste5-C512T* and possibly *mal11-C310T* directly confer increased tolerance to SSL. Mutations in genes *GDH1* and possibly in *MAL11* each complement the deleterious effects of a subset of mutations. Mutant alleles from the *yhr045w*, *ssn2* and *ubp1* hotspots, along with the *ubp7-T2466A* and *tof2-C2141T* mutations complement the coincidental defects in iron metabolism, transcriptional regulation, membrane protein homeostasis and chromosome segregation caused by founder and sweeping alleles. Other mutations are hypothesized to hitchhike on the restricted number of beneficial mutations, probably causing a burden. Minor contributions to SSL tolerance are likely from other less frequent mutations like *ssa1-C91A* as predicted by our linear models of backcrossed mutants.

I also propose a model of the dynamics of our genome shuffling experiment, based on the proposed model of SSL tolerance and on the allele frequency time series obtained from population sequencing (Figure 3.1). Mutagenesis generated equally diverse pools of mutants from the *MATα* and *MATa* parental strains. An aggressive selection regimen restricted genetic diversity, leading to a near sweep of the *MATα* pool

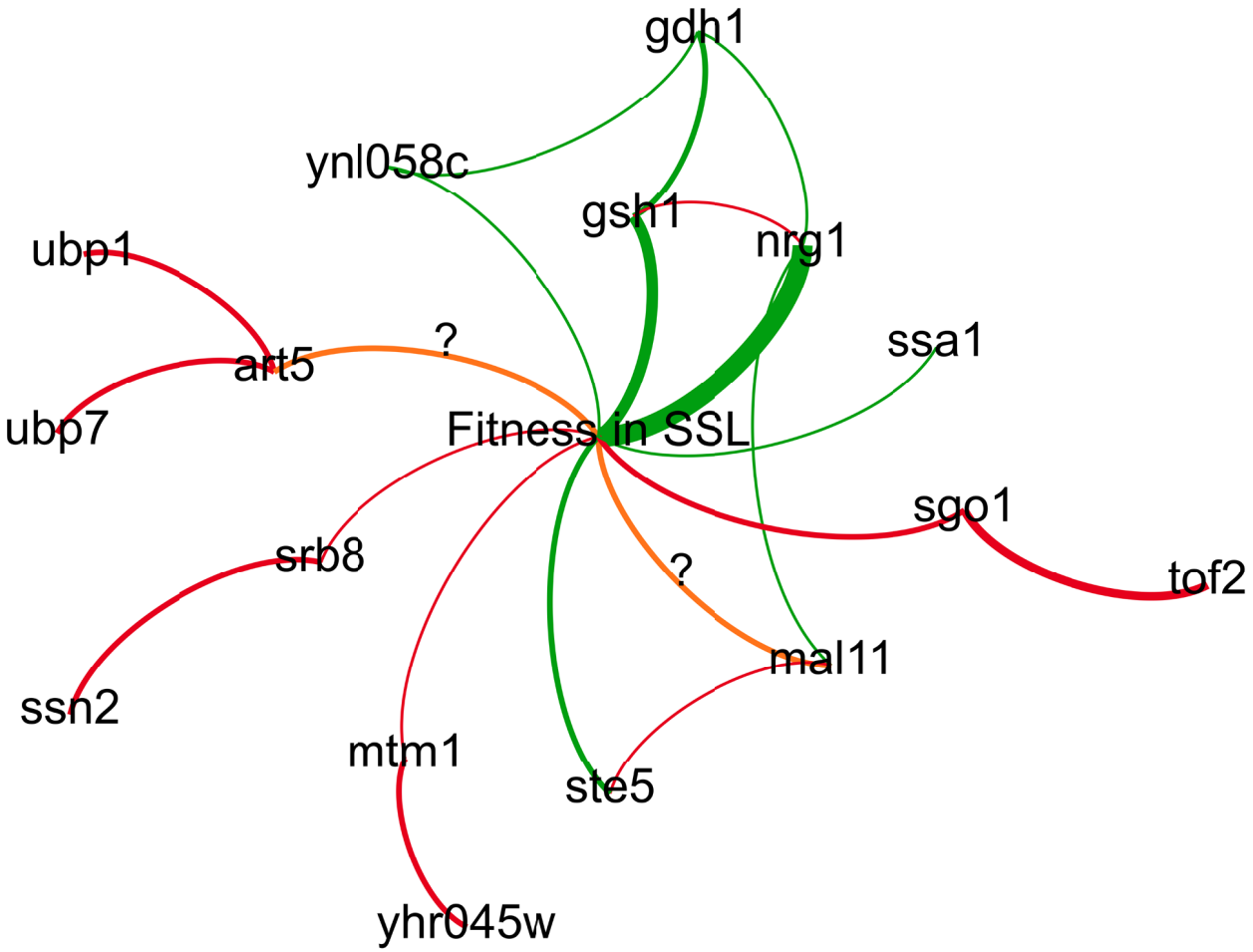


Figure 4.1. Network map model of SSL tolerance in genome shuffled mutants. Nodes represent mutations or phenotypes, and edges represent effects or interactions. Curvature of the edges is clockwise with respect to effector nodes and counter-clockwise with respect to target nodes. Green edges indicate a stimulating or enhancing effect, while red lines indicate an inhibition of the target. Orange lines indicate persisting doubts on the relationship between the nodes that they connect. Thickness of the edges is related to the amount of evidence available for the indicated interactions.

by mutants carrying the **α3** mutations, which includes the serendipitous pair consisting of tolerance enhancing *gsh1-T(-73)A* and *ynl058c-T7C*. Mutations *mal11-A482T* and *gdh1-A68G* were found at a low frequency in this pool. A more relaxed selection generated a more diverse pool of *MATa* mutants, among which were the tolerance enhancing *nrg1-C137A* and *ste5-C512T* mutations. Cohorts **a6** and perhaps **a5** hitchhiked on this latter mutation. The *gdh1* mutations (with the exception of A68G) and *mal11-G310A* were also selected into this initial pool.

Initial recombination created the first epistatic pairs between tolerance-conferring and complementing mutations. Combinations of founders and complementing mutations occurred on a large scale at this stage. Selection on these shuffled mutants brought a large increase in the frequency of complementing mutations thanks to the competitive advantage they imparted onto SSL-tolerant but metabolically imbalanced mutants. Further shuffling and selection combined several tolerance-conferring and helper mutations into single cells, increasing their fitness in the presence of SSL. The strong selective advantage of *nrg1-C137A* led to its steady increase in frequency. I expect that additional rounds of shuffling would have witnessed the rise of *nrg1* alleles to prominence.

5.2 Lessons for the design of genome shuffling experiments

From my data I draw several conclusions on the conditions that favor optimal genome shuffling outcomes. I have shown that the evolutionary solutions found by this method are critically linked to initial conditions. I therefore suggest that special attention be paid to these initial conditions during experimental design. While the genetic makeup of parental strains is generally known with accuracy, it is unrealistic to control for

spontaneous mutations in parental clones. At any rate, our results suggest that the effect of founder mutations will generally be modest. Indeed, crippling mutations affect viability, and I expect that they will be underrepresented in fresh laboratory cultures. However, the prevalence of the *ubp7-T2466A* mutation illustrates how the combination of genetic hitchhiking and aggressive initial selection amplified the otherwise limited effect of founder mutations.

Initial selection was performed before any recombination had taken place. By selecting asexually reproducing yeast, our experiment proved vulnerable to clonal interference. In this context, beneficial mutations competed against each other, and aggressive selection led to a considerable loss of diversity. Beneficial mutations were probably eliminated, and were no longer available for downstream recombination. Moreover, clonal interference favored the near fixation of a single cohort of mutations, which influenced the evolutionary dynamics of the entire experiment. I therefore advocate for more relaxed selection at steps that precede recombination. Alternatively, generating several pools of mutants may maximize beneficial diversity. A range of selection regimens may be applied to further favor diversity. Sexual recombination reduces clonal interference and effectively purifies hitchhikers [57], therefore I expect that an aggressive selection regimen will have less dramatic effects once shuffling has started. Therefore, selection may even be postponed until cycles of recombination have begun.

In our experiment, we aimed for an average of one point mutation per cell by tuning the mutagen dose to a specified kill rate. Nevertheless, several mutants with multiple mutations were selected. Hitchhiking during the early steps may prove difficult

to avoid. Our results show that recursive recombination and selection slowly clean away hitchhikers. Genome shuffling studies typically report three to five cycles of recombination, but performing additional rounds may prove rewarding. Combined with a diverse pool of beneficial mutations, this approach should enable the evolutionary engineering of optimized microbial strains that combine several beneficial mutations, profit from positive epistasis, and contain minimal numbers of deleterious hitchhikers.

My results illustrate some of the advantages of genome shuffling. The pleiotropic effects of beneficial mutations can lead to a trade-off, either in the presence or the absence of stress. Recursive recombination and selection allows the construction of strains where delicate complementation networks operate to offset the fitness cost of core beneficial mutations. The strong fitness defects of revertant *GDH1* derivatives of R57, and the parallel crippling effect of *gdh1* alleles in wildtype diploids are illustrations of this phenomenon. Also, while most mutations identified in R57 cause a mutational burden when isolated, the growth of the full strain in the absence of SSL does not display a strong defect. The prominence and strong positive selection of compensatory alleles (notably *gdh1*, *mal11* and *ubp7* mutations) leads me to conclude that epistasis is a major driving force of evolutionary engineering by genome shuffling. Genome shuffling thus uses the advantages of sexual recombination: it can be used to reduce clonal interference and combine beneficial mutations, but also imparts a kind of hybrid vigor to its offspring. This advantage, combined with induced diversity, repeated artificial selection and the fast doubling rate and large populations of microbes, allows the rapid exploration of large mutational and combinatorial spaces.

5.3 Future directions

The discussion above raised many questions and formulated a large number of hypotheses, both on the evolutionary dynamics of our genome shuffling experiment, and on the basic cell and molecular biology of underlying mutations identified by sequencing. Many have the potential to inform our understanding of stress tolerance in yeast and could help the rational engineering of microbes tolerant to lignocellulosic hydrolysates. Among those questions and hypotheses, I list some of the most prominent and suggest experiments to explore and address them.

Mutations of the *nrg1* hotspot are loss-of-function alleles, which lead to the derepression of a general stress response regulon

RNAseq data from previous studies in our lab identified the most prominently upregulated targets of Nrg1p in R57. Those should be confirmed in *nrg1-C137A* single mutants by RT-PCR. If *nrg1* alleles identified by our sequencing efforts indeed lead to loss of function, modulation of Nrg1p expression by CRISPRi or other methods should have similar effects on the expression of these upregulated genes. Conversely, the effect of overexpressing these Nrg1p targets should be tested for their effect on the SSL-tolerance phenotype. I have shown a beneficial effect of *nrg1-C137A* on tolerance to SSL, acetic acid and hydrogen peroxide. This is in agreement with the hypothesis that Nrg1p acts as a general repressor of stress response genes. I suggest that the range of inhibitory conditions to which *nrg1-C137A* confers increased tolerance be explored. Osmotic stress, ethanol and furfural are examples of inhibitors that easily could be tested.

The *gsh1-T(-73)C* allele leads to increased transcription of *GSH1*, and thus to elevated γ-glytamylcysteine synthetase activity and intracellular glutathione concentration.

Levels of *GSH1* transcript and Gsh1p could be compared between wildtype and *gsh1* mutant by RT-PCR and Western blotting to confirm that increased expression results from the promoter mutation. Intracellular glutathione and γ-glutamylcysteine levels can be measured to establish that reduced ROS accumulation in *gsh1-T(-73)C* cells is linked to elevated levels of antioxidant.

Mutations mapping to *GDH1* increase the supply of glutamate to Gsh1p by diminishing Gdh1p ubiquitination.

Intracellular glutamate, γ-glutamylcysteine, glutathione and $NADP^+$/NADPH should be determined in wildtype, *gsh1*, *gdh1* and *gsh1*/*gdh1* mutants to test whether these mutations have a significant impact on the level of these compounds. Western blotting should be performed to compare levels of Gdh1p in wildtype and mutant strains. Enzyme assays on cell lysates may provide complementary information about glutamate dehydrogenase activity in *gdh1* mutants. Direct mutagenesis of ubiquitination lysines, and the effects of these mutations on glutamate, glutathione and Gdh1p levels should be probed. Mutations mapping to *GDH1* may affect glutamate dehydrogenase enzyme kinetics, and characterization of purified Gdh1p variants is an option to consider.

Mutations in *UBP1* and *UBP7* were selected to complement defects in membrane protein endocytosis of *art5-G454A* mutant.

To test the role of *ubp* alleles in the complementation of *art5-G454A*, association of the wildtype and mutant proteins should be measured. Using classical approaches, such as co-immunoprecipitation, yeast two-hybrid assays or pull-down methods, a physical interaction between the Ubp1/7 and Art5 proteins could be tested. However, purely genetic interactions cannot be excluded. The involvement of Art5p and Ubp7p in protein endocytosis warrants experiments on the effect of associated mutations on this process. For example, there is evidence indicating that Ubp7p deubiquitinates endocytic protein Ede1p [395]. Using antibodies directed against Ede1p and ubiquitin, its ubiquitination state in the presence of various combinations of the *ubp* and *art5* mutations could be assessed by Western blotting. Mating pheromone receptors have been used as model systems for the study of endocytosis. They may be an appropriate system for the study of the *ubp-art5* interplay.

The *mal11* alleles exert their effect by altering intracellular trehalose concentration.

A first step to investigate this hypothesis would be to measure trehalose levels in *mal11* mutants. The proposed positive epistasis between *mal11-C310T* and *nrg1-C137A* warrants the same measurement in the presence and absence of the latter.

The *ste5-C512T* mutation increases SSL tolerance by enhancing the activity of MAPK-mediated stress response pathways

Phosphorylation of components of the pheromone-induced MAPK pathway (Ste20p, Ste11p, Ste7p, Kss1p and Fus3p) can be determined by Western blotting using

antibodies specific for the phosphorylated forms. In the absence of these antibodies, mass spectrometry methods allow investigation of the phosphorylation state of proteins [457]. Measuring the effect of *ste5-C512T* on the phosphorylation state of these proteins, and some of the main targets of MAP kinases Kss1p and Fus3p would help test its influence on the activation of its associated pathways.

Are competition assays more sensitive to differences in fitness than growth curve assays?

For preparation of this thesis, tolerance of mutants to SSL was tested by measuring growth curves. This approach was successful in detecting the effect of a few mutations, notably *nrg1-C137A* and *gsh1-T(-73)C*, but suggests the absence of an effect for most alleles. Whether this is a question of sensitivity or an accurate representation of the effect of the tested alleles is open to debate. Competition assays using fluorescent markers have been applied to directly compare the fitness of mutations identified by experimental evolution (see [458] for a study on the merits of this method). I propose that it may be applied to detect differences in fitness not detected by performing simple growth curves.

Can the evolutionary dynamics inferred from the population sequencing data be reproduced?

Repetition of the genome shuffling experiment by purposefully modulating the initial conditions would allow testing of my hypotheses on the dynamics of evolutionary engineering, notably those about compensatory evolution and historical contingency.

The paths of evolution in the presence and the absence of the *gsh1-T(-73)C* allele in the parent strains could be compared to test for the selection of *gdh1* alleles. The same could be applied to all mutations hypothesized to have benefitted from the founder effect. The influence of the founder effect on the fate of hitchhikers can similarly be probed. A driver mutation (e.g. *gsh1-T(-73)C*) and a hypothesized hitchhiker (e.g. *sgo1-C575A*) can be introduced in a parental population, either in the same starting cells, or in separate ones. The fate of these alleles may subsequently be monitored through the genome shuffling experiment to confirm that the hitchhiker owed its high frequency solely to the driver mutation.

References

1. Darwin, C. *On the Origins of Species by Means of Natural Selection*. (John Murray, 1860).
2. Halliburton, R. *Introduction to population genetics*. (Pearson Prentice Hall, 2004).
3. Bennett, A. F. & Hughes, B. S. Microbial experimental evolution. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **297,** R17-25 (2009).
4. Jerison, E. R. & Desai, M. M. Genomic investigations of evolutionary dynamics and epistasis in microbial evolution experiments. *Current Opinion in Genetics and Development* **35,** 33–39 (2015).
5. Blount, Z. D., Borland, C. Z. & Lenski, R. E. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 7899–7906 (2008).
6. Pál, C., Papp, B. & Pósfai, G. The dawn of evolutionary genome engineering. *Nat. Rev. Genet.* **15,** 504–12 (2014).
7. Biot-Pelletier, D. & Martin, V. J. J. Evolutionary engineering by genome shuffling. *Appl. Microbiol. Biotechnol.* **98,** 3877–3887 (2014).
8. Pinel, D. *et al. Saccharomyces cerevisiae* genome shuffling through recursive population mating leads to improved tolerance to spent sulfite liquor. *Appl. Environ. Microbiol.* **77,** 4736–4743 (2011).
9. Pinel, D., Colatriano, D., Jiang, H., Lee, H. & Martin, V. J. Deconstructing the genetic basis of spent sulphite liquor tolerance using deep sequencing of genome-shuffled yeast. *Biotechnol. Biofuels* **8,** 53 (2015).
10. Garland, T. & Rose, M.R. (ed). *Experimental evolution: concepts, methods and*

applications of selection experiments. (University of California Press, 2009).

11. Rice, W. R. & Chippindale, A. K. Sexual recombination and the power of natural selection. *Science* **294,** 555–559 (2001).
12. Lenski, R. Phenotypic and genomic evolution during a 20,000-generation experiment with the bacterium *Escherichia coli*. *Plant breeding reviews* **24,** 225–265 (2004).
13. Riehle, M. M., Bennett, A. F. & Long, A. D. Differential patterns of gene expression and gene complement in laboratory-evolved lines of *E. coli*. in *Integrative and Comparative Biology* **45,** 532–538 (2005).
14. Maughan, H., Masel, J., Birky, C. W. & Nicholson, W. L. The roles of mutation accumulation and selection in loss of sporulation in experimental populations of *Bacillus subtilis*. *Genetics* **177,** 937–948 (2007).
15. Velicer, G. J., Lenski, R. E. & Kroos, L. Rescue of social motility lost during evolution of *Myxococcus xanthus* in an asocial environment. *J. Bacteriol.* **184,** 2719–2727 (2002).
16. Ferenci, T. The spread of a beneficial mutation in experimental bacterial populations: the influence of the environment and genotype on the fixation of *rpoS* mutations. *Heredity (Edinb).* **100,** 446–452 (2008).
17. Dallinger, W. H. The President's Address. *J. R. Microsc. Soc.* **7,** 185–199 (1887).
18. Novick, a & Szilard, L. Experiments with the Chemostat on spontaneous mutations of bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **36,** 708–719 (1950).
19. Hartl, D. L., Dykhuizen, D. E. & Dean, A. M. Limits of adaptation: The evolution of selective neutrality. *Genetics* **111,** 655–674 (1985).
20. Dykhuizen, D. E., Dean, A. M. & Hartl, D. L. Metabolic flux and fitness. *Genetics*

115, 25–31 (1987).

21. Dettman, J. R. *et al.* Evolutionary insight from whole-genome sequencing of experimentally evolved microbes. *Mol. Ecol.* **21,** 2058–77 (2012).
22. Lenski, R. E. *Escherichia coli* long-term experimental evolution project. (2016). at <http://myxo.css.msu.edu/index.html>
23. Lenski, R. E., Rose, M. R., Simpson, S. C. & Tadler, S. C. Long-Term Experimental Evolution in *Escherichia coli* . I . Adaptation and Divergence During. *Am. Nat.* **138,** 1315–1341 (1991).
24. Vasi, F., Travisano, M. & Lenski, R. E. Long-Term Experimental Evolution in *Escherichia coli*. II. Changes in Life-History Traits During Adaptation to a Seasonal Environment. *Am. Nat.* **144,** 432–456 (1994).
25. Travisano, M., Vasi, F. K. & Lenski, R. E. Long-Term Experimental Evolution in *Escherichia coli*. III. Variation Among Replicate Populations in Correlated Responses to Novel Environments. *Evolution (N. Y).* **49,** 189–200 (1995).
26. Travisano, M. & Lenski, R. E. Long-term experimental evolution in *Escherichia coli*. IV. Targets of selection and the specificity of adaptation. *Genetics* **143,** 15–26 (1996).
27. Souza, V., Turner, P. E. & Lenski, R. E. Long-term experimental evolution in *Escherichia coli*. V. Effects of recombination with immigrant genotypes on the rate of bacterial evolution. *J. Evol. Biol.* **10,** 743–769 (1997).
28. Travisano, M. Long-term experimental evolution in *Escherichia coli*. VI. Environmental constraints on adaptation and divergence. *Genetics* **146,** 471–479 (1997).
29. Elena, S. F. & Lenski, R. E. Long-Term Experimental Evolution in *Escherichia coli*

. VII . Mechanisms Maintaining Genetic Variability Within Populations. *Evolution (N. Y).* **51,** 1058–1067 (1997).

30. Rozen, D. E. & Lenski, R. E. Long-Term Experimental Evolution in *Escherichia coli*. VIII. Dynamics of a Balanced Polymorphism. *Am. Nat.* **155,** 24–35 (2000).

31. Schneider, D., Duperchy, E., Coursange, E., Lenski, R. E. & Blot, M. Long-term experimental evolution in *Escherichia coli*. IX. Characterization of insertion sequence-mediated mutations and rearrangements. *Genetics* **156,** 477–488 (2000).

32. Cooper, V. S. Long-term experimental evolution in *Escherichia coli*. X. Quantifying the fundamental and realized niche. *BMC Evol. Biol.* **2,** 12 (2002).

33. Visser, J. A. G. & Lenski, R. E. Long-term experimental evolution in *Escherichia coli* . XI. Rejection of non-transitive interactions as cause of declining rate of adaptation. *BMC Evolutionary Biology* **2,** 1 (2002).

34. Crozat, E., Philippe, N., Lenski, R. E., Geiselmann, J. & Schneider, D. Long-term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection. *Genetics* **169,** 523–532 (2005).

35. Rozen, D. E., Schneider, D. & Lenski, R. E. Long-term experimental evolution in *Escherichia coli*. XIII. Phylogenetic history of a balanced polymorphism. *J. Mol. Evol.* **61,** 171–180 (2005).

36. Lenski, R. E. & Travisano, M. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 6808–6814 (1994).

37. Cooper, V. S., Schneider, D., Blot, M. & Lenski, R. E. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli*

B. *J. Bacteriol.* **183,** 2834–2841 (2001).

38. Cooper, T. F., Rozen, D. E. & Lenski, R. E. Parallel changes in gene expression after 20, 000 generations of evolution in *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 1072–1077 (2003).

39. Maughan, H. *et al.* The population genetics of phenotypic deterioration in experimental populations of *Bacillus subtilis*. *Evolution* **60,** 686–695 (2006).

40. Kvitek, D. J. & Sherlock, G. Whole Genome, Whole Population Sequencing Reveals That Loss of Signaling Networks Is the Major Adaptive Strategy in a Constant Environment. *PLoS Genet.* **9,** (2013).

41. Cooper, V. S. & Lenski, R. E. The population genetics of ecological specialization in evolving *Escherichia coli* populations. *Nature* **407,** 736–739 (2000).

42. Bennett, A. F. & Lenski, R. E. Evolutionary adaptation to temperature. V. Adaptive mechanisms and correlated responses in experimental lines of *Escherichia coli*. *Evolution (N. Y).* **50,** 493–503 (1996).

43. Hughes, B. S., Cullum, A. J. & Bennett, A. F. Evolutionary adaptation to environmental pH in experimental lineages of *Escherichia coli*. *Evolution (N. Y).* **61,** 1725–1734 (2007).

44. Mongold, J. A., Bennett, A. F. & Lenski., R. E. Evolutionary adaptation to temperature. IV. Adaptation of *Escherichia coli* at a niche boundary. *Evolution (N. Y).* 35–43 (1996).

45. Bennett, A. F. & Lenski, R. E. An experimental test of evolutionary trade-offs during temperature adaptation. *Proc. Natl. Acad. Sci. U. S. A.* **104 Suppl,** 8649–54 (2007).

46. Treves, D. S., Manning, S. & Adams, J. Repeated evolution of an acetate-

crossfeeding polymorphism in long-term populations of *Escherichia coli*. *Mol. Biol. Evol.* **15,** 789–97 (1998).

47. Rosenzweig, R. F., Sharp, R. R., Treves, D. S. & Adams, J. Microbial evolution in a simple unstructured environment: Genetic differentiation in *Escherichia coli*. *Genetics* **137,** 903–917 (1994).

48. Rainey, P. B. & Travisano, M. Adaptive radiation in a heterogeneous environment. *Nature* **394,** 69–72 (1998).

49. Xu, J. Estimating the spontaneous mutation rate of loss of sex in the human pathogenic fungus Cryptococcus neoformans. *Genetics* **162,** 1157–1167 (2002).

50. Zeyl, C., Curtin, C., Karnap, K. & Beauchamp, E. Antagonism between sexual and natural selection in experimental populations of *Saccharomyces cerevisiae*. *Evolution (N. Y).* **59,** 2109–2115 (2005).

51. Dasilva, J. & Bell, G. The Ecology and Genetics of Fitness in *Chlamydomonas* .6. Antagonism Between Natural-Selection and Sexual Selection. *Proc. R. Soc. London Ser. B-Biological Sci.* **249,** 227–233 (1992).

52. Greig, D., Borts, R. H. & Louis, E. J. The effect of sex on adaptation to high temperature in heterozygous and homozygous yeast. *Proc. Biol. Sci.* **265,** 1017–1023 (1998).

53. Klapholz, S., Waddell, C. S. & Easton Esposito, R. The role of the *SPO11* gene in meiotic recombination in yeast. *Genetics* **110,** 187–216 (1985).

54. Steele, D. F., Morris, M. E. & Jinks-Robertson, S. Allelic and ectopic interactions in recombination-defective yeast strains. *Genetics* **127,** 53–60 (1991).

55. Goddard, M. R., Godfray, H. C. J. & Burt, A. Sex increases the efficacy of natural selection in experimental yeast populations. *Nature* **434,** 636–640 (2005).

56. Colegrave, N., Kaltz, O. & Bell, G. The ecology and genetics of fitness in *Chlamydomonas*. VIII. The dynamics of adaptation to novel environments after a single episode of sex. *Evolution* **56,** 14–21 (2002).

57. Mcdonald, M. J., Rice, D. P., Desai, M. M. & Daniel, P. Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature* **531,** 233–6 (2016).

58. Brown, C. J., Todd, K. M. & Rosenzweig, R. F. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol. Biol. Evol.* **15,** 931–42 (1998).

59. Dunham, M. J. *et al.* Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 16144–16149 (2002).

60. Ferea, T. L., Botstein, D., Brown, P. O. & Rosenzweig, R. F. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc Natl Acad Sci U S A* **96,** 9721–9726 (1999).

61. Gresham, D. *et al.* Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. *Science* **311,** 1932–1936 (2006).

62. Segrè, A. V., Murray, A. W. & Leu, J. Y. High-resolution mutation mapping reveals parallel experimental evolution in yeast. *PLoS Biol.* **4,** 1372–1385 (2006).

63. Barrick, J. E. *et al.* Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461,** 1243–1247 (2009).

64. Barrick, J. E. & Lenski, R. E. Genome-wide Mutational Diversity in an Evolving Population of *Escherichia coli* Genome-wide Mutational Diversity in an Evolving Population of *Escherichia coli*. 119–129 (2009). doi:10.1101/sqb.2009.74.018

65. Tenaillon, O. *et al.* Tempo and mode of genome evolution in a 50,000 - generation

experiment. *Nature* **536,** 165–170 (2016).

66. Lang, G. I. *et al.* Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500,** 571–4 (2013).

67. Tenaillon, O. *et al.* The Molecular Diversity of Adaptive Convergence. *Science (80-.).* **335,** 457–461 (2012).

68. Herron, M. D. & Doebeli, M. Parallel Evolutionary Dynamics of Adaptive Diversification in *Escherichia coli*. *PLoS Biol.* **11,** (2013).

69. Cooper, V. S., Staples, R. K., Traverse, C. C. & Ellis, C. N. Parallel evolution of small colony variants in *Burkholderia cenocepacia* biofilms. *Genomics* **104,** 1–6 (2014).

70. Burke, M. K., Liti, G. & Long, A. D. Standing genetic variation drives repeatable experimental evolution in outcrossing populations of *Saccharomyces cerevisiae*. *Mol. Biol. Evol.* **31,** 3228–3239 (2014).

71. Lee, D. H. & Palsson, B. O. Adaptive evolution of *Escherichia coli* K-12 MG1655 during growth on a nonnative carbon source, L-l,2-propanediol. *Appl. Environ. Microbiol.* **76,** 4158–4168 (2010).

72. Conrad, T. M. *et al.* Whole-genome resequencing of *Escherichia coli* K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. *Genome Biol.* **10,** R118 (2009).

73. Kvitek, D. J. & Sherlock, G. Reciprocal Sign Epistasis between Frequently Experimentally Evolved Adaptive Mutations Causes a Rugged Fitness Landscape. **7,** (2011).

74. Kimura, M. The role of compensatory neutral mutations in molecular evolution. *J. Genet.* **64,** (1985).

75. Weinreich, D. & Chao, L. Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution (N. Y).* **59,** 1175–1182 (2005).

76. Burch, C. & Chao, L. Evolution by small steps and rugged landscapes in the RNA virus phi6. *Genetics* **151,** 921–927 (1999).

77. Poon, A. & Chao, L. The rate of compensatory mutation in the DNA bacteriophage φX174. *Genetics* **170,** 989–999 (2005).

78. Poon, A. & Chao, L. Functional origins of fitness effect-sizes of compensatory mutations in the DNA bacteriophage phiX174. *Evolution (N. Y).* **60,** 2032–2043 (2006).

79. Harcombe, W., Springman, R. & Bull, J. Compensatory evolution for a gene deletion is not limited to its immediate functional network. *BMC Evol. Biol.* **9,** (2009).

80. Moore, F., Rozen, D. & Lenski, R. Pervasive compensatory adaptation in *Escherichia coli*. *Proc. Biol. Sci.* **267,** 515–522 (2000).

81. Björkman, J., Nagaev, I., Berg, O., Hughes, D. & Andersson, D. Effects of environment on compensatory mutations to ameliorate costs of antibiotic resistance. *Science (80-.).* **287,** 1479–1482 (2000).

82. Blank, D., Wolf, L., Ackermann, M. & Silander, O. K. The predictability of molecular evolution during functional innovation. *Proc Natl Acad Sci U S A* **111,** 3044–3049 (2014).

83. Szamecz, B. *et al.* The Genomic Landscape of Compensatory Evolution. *PLoS Biol.* **12,** (2014).

84. Filteau, M. *et al.* Evolutionary rescue by compensatory mutations is constrained by genomic and environmental backgrounds. *Mol. Syst. Biol.* **11,** (2015).

85. Estes, S. & Lynch, M. Rapid fitness recovery in mutationally degraded lines of Caenorhabditis elegans. *Evolution* **57,** 1022–1030 (2003).

86. Estes, S., Phillips, P. C. & Denver, D. R. Fitness recovery and compensatory evolution in natural mutant lines of *C. elegans*. *Evolution (N. Y).* **65,** 2335–2344 (2011).

87. Orr, H. The genetic theory of adaptation: a brief history. *Nat. Rev. Genet.* **6,** 119–127 (2005).

88. Doniger, S. *et al.* A catalog of neutral and deleterious polymorphisms in yeast. *PLoS Genet.* **4,** (2008).

89. Iwasa, Y., Michor, F. & Nowak, M. Stochastic tunnels in evolutionary dynamics. *Genetics* **166,** 1571–1579 (2004).

90. Ashworth, A., Lord, C. & Reis-Filho, J. Genetic interactions in cancer progression and treatment. *Cell* **145,** 30–38 (2011).

91. Mills, D. R., Peterson, R. L. & Spiegelman, S. An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Natl. Acad. Sci. U. S. A.* **58,** 217–24 (1967).

92. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10,** 866–76 (2009).

93. Tsien, R. Y. Constructing and exploiting the fluorescent protein paintbox (Nobel Lecture). *Angew. Chem. Int. Ed. Engl.* **48,** 5612–5626 (2009).

94. Campbell, R. E. *et al.* A monomeric red fluorescent protein. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 7877–82 (2002).

95. Shaner, N. C. *et al.* Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nat. Biotechnol.* **22,**

1567–72 (2004).

96. Sarkar, I., Hauber, I., Hauber, J. & Buchholz, F. HIV-1 Proviral DNA Excision Using an Evolved Recombinase. *Science (80-.).* **316,** 1912–1915 (2007).

97. Fasan, R., Chen, M. M., Crook, N. C. & Arnold, F. H. Engineered alkane-hydroxylating cytochrome P450BM3 exhibiting nativelike catalytic properties. *Angew. Chemie - Int. Ed.* **46,** 8414–8418 (2007).

98. Reetz, M. T., Carballeira, J. D. & Vogel, A. Iterative saturation mutagenesis on the basis of b factors as a strategy for increasing protein thermostability. *Angew. Chemie - Int. Ed.* **45,** 7745–7751 (2006).

99. Crameri, A., Raillard, S.-A., Bermudez, E. & Stemmer, W. P. C. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391,** 288–91 (1998).

100. Ostermeier, M., Shim, J. H. & Benkovic, S. J. A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat. Biotechnol.* **17,** 1205–1209 (1999).

101. Lutz, S., Ostermeier, M., Moore, G. L., Maranas, C. D. & Benkovic, S. J. Creating multiple-crossover DNA libraries independent of sequence identity. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 11248–11253 (2001).

102. Hiraga, K. & Arnold, F. H. General method for sequence-independent site-directed chimeragenesis. *J. Mol. Biol.* **330,** 287–296 (2003).

103. Fischbach, M. A., Lai, J. R., Roche, E. D., Walsh, C. T. & Liu, D. R. Directed evolution can rapidly improve the activity of chimeric assembly-line enzymes. *Proc. Natl. Acad. Sci.* **104,** 11951–11956 (2007).

104. Reetz, M. T., Bocola, M., Carballeira, J. D., Zha, D. & Vogel, A. Expanding the range of substrate acceptance of enzymes: Combinatorial active-site saturation

test. *Angew. Chemie - Int. Ed.* **44,** 4192–4196 (2005).

105. Park, S. *et al.* Focusing mutations into the *P. fluorescens* esterase binding site increases enantioselectivity more effectively than distant mutations. *Chem. Biol.* **12,** 45–54 (2005).

106. Paramesvaran, J., Hibbert, E. G., Russell, A. J. & Dalby, P. A. Distributions of enzyme residues yielding mutants with improved substrate specificities from two different directed evolution strategies. *Protein Eng. Des. Sel.* **22,** 401–411 (2009).

107. Hansson, L. O., Bolton-Grob, R., Massoud, T. & Mannervik, B. Evolution of differential substrate specificities in Mu class glutathione transferases probed by DNA shuffling. *J. Mol. Biol.* **287,** 265–276 (1999).

108. Voigt, C. A., Martinez, C., Wang, Z.-G., Mayo, S. L. & Arnold, F. H. Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9,** 553–558 (2002).

109. Esvelt, K. M., Carlson, J. C. & Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **472,** 499–503 (2011).

110. Romanini, D. W., Peralta-Yahya, P., Mondol, V. & Cornish, V. W. A heritable recombination system for synthetic darwinian evolution in yeast. *ACS Synth. Biol.* **1,** 602–609 (2012).

111. Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460,** 894–8 (2009).

112. Wang, H. H. *et al.* Genome-scale promoter engineering by coselection MAGE. *Nat. Methods* **9,** 591–3 (2012).

113. Ma, N. J., Moonan, D. W. & Isaacs, F. J. Precise manipulation of bacterial chromosomes by conjugative assembly genome engineering. *Nat. Protoc.* **9,** 2285–300 (2014).

114. Dicarlo, J. E. *et al.* Yeast oligo-mediated genome engineering (YOGE). *ACS Synth. Biol.* **2,** 741–749 (2013).

115. Isaacs, F. J. *et al.* Precise Manipulation of Chromosomes in Vivo Enables Genome-Wide Codon Replacement. *Science.* **333,** 348–353 (2011).

116. Alper, H. & Stephanopoulos, G. Global transcription machinery engineering: A new approach for improving cellular phenotype. *Metab. Eng.* **9,** 258–267 (2007).

117. Alper, H., Moxley, J., Nevoigt, E., Fink, G. R. & Stephanopoulos, G. Engineering Yeast Transcription Machinery for Improved Ethanol Tolerance and Production. *Science.* **314,** 1565–1568 (2006).

118. Warner, J. R., Reeder, P. J., Karimpour-Fard, A., Woodruff, L. B. A. & Gill, R. T. Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides. *Nat. Biotechnol.* **28,** 856–62 (2010).

119. Sandoval, N. R. *et al.* Strategy for directing combinatorial genome engineering in *Escherichia coli*. *Proc. Natl. Acad. Sci.* **109,** 10540–10545 (2012).

120. Crook, N. & Alper, H. S. in *Engineering complex phenotypes in industrial strains* (ed. Patnaik, R.) 1–33 (John Wiley & sons, 2013). doi:10.1002/9781118433034.ch1

121. DelCardayre, S. *et al.* Evolution of whole cells and organisms by recursive sequence recombination. 112 (2013).

122. Zhang, Y.-X. X. *et al.* Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* **415,** 644–646 (2002).

123. Tobin, M., Stemmer, W. P. C., Ness, J. E. & Minshull, J. S. Evolution of whole cells and organisms by recursive sequence recombination. 92 (2001).

124. Demeke, M. M. *et al.* Development of a D-xylose fermenting and inhibitor tolerant

industrial *Saccharomyces cerevisiae* strain with high performance in lignocellulose hydrolysates using metabolic and evolutionary engineering. *Biotechnol. Biofuels* **6,** 89 (2013).

125. Jingping, G., Hongbing, S., Gang, S., Hongzhi, L. & Wenxiang, P. A genome shuffling-generated *Saccharomyces cerevisiae* isolate that ferments xylose and glucose to produce high levels of ethanol. *J. Ind. Microbiol. Biotechnol.* **39,** 777–87 (2012).
126. Bajwa, P. K., Pinel, D., Martin, V. J. J., Trevors, J. T. & Lee, H. Strain improvement of the pentose-fermenting yeast *Pichia stipitis* by genome shuffling. *J Microbiol Methods* **81,** 179–186 (2010).
127. Zhang, W. & Geng, A. Improved ethanol production by a xylose-fermenting recombinant yeast strain constructed through a modified genome shuffling method. *Biotechnol. Biofuels* **5,** 46 (2012).
128. El-Bondkly, A. M. A. Molecular identification using ITS sequences and genome shuffling to improve 2-deoxyglucose tolerance and xylanase activity of marine-derived fungus, *Aspergillus* sp. NRCF5. *Appl. Biochem. Biotechnol.* **167,** 2160–73 (2012).
129. Xu, F. *et al.* Strain improvement for enhanced production of cellulase in Trichoderma viride. *Prikl Biokhim Mikrobiol* **47,** 61–65 (2011).
130. Cheng, Y., Song, X., Qin, Y. & Qu, Y. Genome shuffling improves production of cellulase by *Penicillium decumbens* JU-A10. *J Appl Microbiol* **107,** 1837–1846 (2009).
131. Li, W., Chen, G., Gu, L., Zeng, W. & Liang, Z. Genome Shuffling of *Aspergillus niger* for Improving Transglycosylation Activity. *Appl. Biochem. Biotechnol.* **172,**

50–61 (2013).

132. Cao, X., Song, Q., Wang, C. & Hou, L. Genome shuffling of *Hansenula anomala* to improve flavour formation of soy sauce. *World J. Microbiol. Biotechnol.* **28,** 1857–62 (2012).

133. Patnaik, R. *et al.* Patnaik et al (2002) GS of *Lactobacillus* for improved acid tolerance. *Nat. Biotechnol.* **20,** 707–712 (2002).

134. Zheng, P. *et al.* Genome shuffling improves thermotolerance and glutamic acid production of *Corynebacteria glutamicum*. *World J. Microbiol. Biotechnol.* **28,** 1035–43 (2012).

135. Li, W. F. F. *et al.* Oxidative stress-resistance assay for screening yeast strains overproducing heterologous proteins. *Genetika* **47,** 1175–1183 (2011).

136. Zheng, D.-Q. *et al.* Screening and construction of *Saccharomyces cerevisiae* strains with improved multi-tolerance and bioethanol fermentation performance. *Bioresour. Technol.* **102,** 3020–7 (2011).

137. Wei, P., Li, Z., He, P., Lin, Y. & Jiang, N. Genome shuffling in the ethanologenic yeast *Candida krusei* to improve acetic acid tolerance. *Biotechnol. Appl. Biochem.* **49,** 113–120 (2008).

138. Li, S. *et al.* Genome shuffling enhanced ε-poly-L-lysine production by improving glucose tolerance of *Streptomyces graminearus*. *Appl. Biochem. Biotechnol.* **166,** 414–23 (2012).

139. Zheng, D. Q. *et al.* Drug resistance marker-aided genome shuffling to improve acetic acid tolerance in *Saccharomyces cerevisiae*. *J Ind Microbiol Biotechnol* **38,** 415–422 (2011).

140. Dai, M. & Copley, S. D. C.-P. Genome shuffling improves degradation of the

anthropogenic pesticide pentachlorophenol by *Sphingobium chlorophenolicum* ATCC 39723. *Appl Env. Microbiol* **70,** 2391–2397 (2004).

141. Lee, B. U., Cho, Y. S., Park, S. C. & Oh, K. H. Enhanced degradation of TNT by genome-shuffled *Stenotrophomonas maltophilia* OK-5. *Curr Microbiol* **59,** 346–351 (2009).
142. Clermont, N., Lerat, S. & Beaulieu, C. Genome shuffling enhances biocontrol abilities of *Streptomyces* strains against two potato pathogens. *J Appl Microbiol* **111,** 671–682 (2011).
143. Wang, Z. G., Wu, X. H. & Friedberg, E. C. Nucleotide excision repair of DNA by human cell extracts is suppressed in reconstituted nucleosomes. *J. Biol. Chem.* **266,** 22472–8 (1991).
144. Setlow, R. B. Cyclobutane-type pyrimidine dimers in polynucleotides. *Science* **153,** 379–86 (1966).
145. Singer, B. in *Molecular and cellular mechanisms of mutagenesis* (eds. Lemontt, J. & Generoso, W.) 1–42 (Plenum Press, 1981).
146. Cupples, C. G. & Miller, J. H. A set of lacZ mutations in *Escherichia coli* that allow rapid detection of each of the six base substitutions. *Proc. Natl. Acad. Sci. U. S. A.* **86,** 5345–9 (1989).
147. Hampsey, M. A Tester System for Detecting Each of the Six Base-Pair Substitutions in *Saccharomyces cerevisiae* by Selecting for an Essential Cysteine in Iso- 1 -Cytochrome c. *Genetics* **128,** 59–67 (1991).
148. John, R. P., Gangadharan, D. & Madhavan Nampoothiri, K. Genome shuffling of *Lactobacillus delbrueckii* mutant and *Bacillus amyloliquefaciens* through protoplasmic fusion for L-lactic acid production from starchy wastes. *Bioresour.*

Technol. **99,** 8008–15 (2008).

149. Li, S. *et al.* Combining genome shuffling and interspecific hybridization among *Streptomyces* improved ε-poly-L-lysine production. *Appl. Biochem. Biotechnol.* **169,** 338–50 (2013).

150. Fodor, K., Demiri, E. & Alföldi, L. Polyethylene glycol-induced fusion of heat-inactivated and living protoplasts of *Bacillus megaterium*. *J. Bacteriol.* **135,** 68–70 (1978).

151. Zhao, K. *et al.* Screening and breeding of high taxol producing fungi by genome shuffling. *Sci China C Life Sci* **51,** 222–231 (2008).

152. Hopwood, DA & Wright HM. Bacterial protoplast fusion: recombination in fused protoplasts of *Streptomyces coelicolor*. *Mol. Gen. Genet.* **162,** 307–317 (1978).

153. Dai, M., Ziesman, S. & Copley, S. D. Visualization of protoplast fusion and quantitation of recombination in fused protoplasts of auxotrophic strains of *Escherichia coli*. *Metab. Eng.* **7,** 45–52 (2005).

154. Gong, G. L. *et al.* Mutation and a high-throughput screening method for improving the production of Epothilones of *Sorangium*. *J Ind Microbiol Biotechnol* **34,** 615–623 (2007).

155. Wang, H., Zhang, J., Wang, X., Qi, W. & Dai, Y. Genome shuffling improves production of the low-temperature alkalophilic lipase by Acinetobacter johnsonii. *Biotechnol. Lett.* **34,** 145–51 (2012).

156. Zheng, P., Zhang, K., Yan, Q., Xu, Y. & Sun, Z. Enhanced succinic acid production by *Actinobacillus succinogenes* after genome shuffling. *J. Ind. Microbiol. Biotechnol.* **40,** 831–40 (2013).

157. Tao, X. *et al.* A novel strategy to construct yeast *Saccharomyces cerevisiae*

strains for very high gravity fermentation. *PLoS One* **7,** e31235 (2012).

158. Zhao, J. *et al.* Genome shuffling of *Bacillus amyloliquefaciens* for improving antimicrobial lipopeptide production and an analysis of relative gene expression using FQ RT-PCR. *J. Ind. Microbiol. Biotechnol.* **39,** 889–96 (2012).

159. Jin, Q., Jin, Z., Zhang, L., Yao, S. & Li, F. Probing the molecular mechanisms for pristinamycin yield enhancement in *Streptomyces pristinaespiralis*. *Curr. Microbiol.* **65,** 792–8 (2012).

160. Xu, B. *et al.* Evolution of *Streptomyces pristinaespiralis* for resistance and production of pristinamycin by genome shuffling. *Appl Microbiol Biotechnol* **80,** 261–267 (2008).

161. Cao, X., Hou, L., Lu, M., Wang, C. & Zeng, B. Genome shuffling of *Zygosaccharomyces rouxii* to accelerate and enhance the flavour formation of soy sauce. *J Sci Food Agric* **90,** 281–285 (2010).

162. Wei, Y., Wang, C., Wang, M., Cao, X. & Houa, L. Comparative analysis of salt-tolerant gene *HOG1* in a *Zygosaccharomyces rouxii* mutant strain and its parent strain. *J. Sci. Food Agric.* (2013). doi:10.1002/jsfa.6096

163. Yin, H. *et al.* Genome shuffling of *Saccharomyces cerevisiae* for enhanced glutathione yield and relative gene expression analysis using fluorescent quantitation reverse transcription polymerase chain reaction. *J. Microbiol. Methods* **127,** 188–192 (2016).

164. Zhao, J., Zhang, C., Lu, J. & Lu, Z. Enhancement of fengycin production in *Bacillus amyloliquefaciens* by genome shuffling and relative gene expression analysis using RT-PCR. *Can. J. Microbiol.* **62,** 431–436 (2016).

165. Zhang, Y., Liu, J.-Z., Huang, J.-S. & Mao, Z.-W. Genome shuffling of

Propionibacterium shermanii for improving vitamin B12 production and comparative proteome analysis. *J. Biotechnol.* **148,** 139–43 (2010).

166. Zhao, J. *et al.* Differential proteomics analysis of *Bacillus amyloliquefaciens* and its genome-shuffled mutant for improving surfactin production. *Int. J. Mol. Sci.* **15,** 19847–19869 (2014).
167. Zheng, D.-Q. *et al.* Genomic structural variations contribute to trait improvement during whole-genome shuffling of yeast. *Appl. Microbiol. Biotechnol.* (2013). doi:10.1007/s00253-013-5423-7
168. Harner, N. K. *et al.* Determinants of tolerance to inhibitors in hardwood spent sulfite liquor in genome shuffled *Pachysolen tannophilus* strains. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* **108,** 811–834 (2015).
169. Bloom, J. D. *et al.* Thermodynamic prediction of protein neutrality. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 606–611 (2005).
170. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444,** 929–932 (2006).
171. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* **19,** 596–604 (2009).
172. Fasan, R., Meharenna, Y. T., Snow, C. D., Poulos, T. L. & Arnold, F. H. Evolutionary History of a Specialized P450 Propane Monooxygenase. *J. Mol. Biol.* **383,** 1069–1080 (2008).
173. Wang, X., Minasov, G. & Shoichet, B. K. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J. Mol. Biol.* **320,** 85–95 (2002).

174. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 5869–5874 (2006).

175. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science.* **312,** 111–114 (2006).

176. Bloom, J. D., Romero, P. a, Lu, Z. & Arnold, F. H. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol. Direct* **2,** 17 (2007).

177. Amitai, G., Gupta, R. D. & Tawfik, D. S. Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP J.* **1,** 67–78 (2007).

178. Liu, C. C. *et al.* Protein evolution with an expanded genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 17688–17693 (2008).

179. Li, X. & Liu, C. C. Biological applications of expanded genetic codes. *ChemBioChem* **15,** 2335–2341 (2014).

180. Walter, K. U., Vamvaca, K. & Hilvert, D. An active enzyme constructed from a 9-amino acid alphabet. *J. Biol. Chem.* **280,** 37742–37746 (2005).

181. Müller, M. M. *et al.* Directed Evolution of a Model Primordial Enzyme Provides Insights into the Development of the Genetic Code. *PLoS Genet.* **9,** (2013).

182. Limayem, A. & Ricke, S. C. Lignocellulosic biomass for bioethanol production: Current perspectives, potential issues and future prospects. *Progress in Energy and Combustion Science* **38,** 449–467 (2012).

183. Mitchell, D. *A Note on Rising Food Prices*. *World Bank Development Prospects Group* **6,** (2008).

184. Kendall, A. & Yuan, J. Comparing life cycle assessments of different biofuel

options. *Curr. Opin. Chem. Biol.* **17,** 439–443 (2013).

185. Palmqvist, E. & Hahn-Hägerdal, B. Fermentation of lignocellulosic hydrolysates. I: Inhibition and detoxification. *Bioresource Technology* **74,** 17–24 (2000).
186. Richardson, T. L., Harner, N. K., Bajwa, P. K., Trevors, J. T. & Lee, H. Approaches to deal with toxic inhibitors during fermentation of lignocellulosic substrates. in *ACS Symposium Series* **1067,** 171–202 (2011).
187. Attfield, P. V. Stress tolerance: the key to effective strains of industrial baker's yeast. *Nat. Biotechnol.* **15,** 1351–1357 (1997).
188. Gibson, B. R., Lawrence, S. J., Leclaire, J. P. R., Powell, C. D. & Smart, K. A. Yeast responses to stresses associated with industrial brewery handling. *FEMS Microbiology Reviews* **31,** 535–569 (2007).
189. Klinke, H. B., Thomsen, A. B. & Ahring, B. K. Inhibition of ethanol-producing yeast and bacteria by degradation products produced during pre-treatment of biomass. *Applied Microbiology and Biotechnology* **66,** 10–26 (2004).
190. Liu, Z. L. Genomic adaptation of ethanologenic yeast to biomass conversion inhibitors. *Applied Microbiology and Biotechnology* **73,** 27–36 (2006).
191. Fischer, C. R., Klein-Marcuschamer, D. & Stephanopoulos, G. Selection and optimization of microbial hosts for biofuels production. *Metab. Eng.* **10,** 295–304 (2008).
192. Lynd, L. R. Overview and evaluation of fuel ethanol from cellulosic biomass: technology, economics, the environment, and policy. *Annu. Rev. Energy. Environ.* **21,** 403–465 (1996).
193. Pereira, S. R., Portugal-Nunes, D. J., Evtuguin, D. V., Serafim, L. S. & Xavier, A. M. R. B. Advances in ethanol production from hardwood spent sulphite liquors.

Process Biochemistry **48,** 272–282 (2013).

194. Chirat, C., Lachenal, D. & Sanglard, M. Extraction of xylans from hardwood chips prior to kraft cooking. *Process Biochem.* **47,** 381–385 (2012).

195. Balat, M., Balat, H. & Öz, C. Progress in bioethanol processing. *Progress in Energy and Combustion Science* **34,** 551–573 (2008).

196. Shima, J. & Nakamura, T. in *Stress Biology of Yeasts and Fungi: Applications for Industrial Brewing and Fermentation* (eds. Kitagaki, H. & Takagi, H.) 93–106 (Springer, 2015). doi:10.1007/978-4-431-55248-2

197. Delgenes, J. P., Moletta, R. & Navarro, J. M. Effects of lignocellulose degradation products on ethanol fermentations of glucose and xylose by *Saccharomyces cerevisiae*, *Zymomonas mobilis*, *Pichia stipitis*, and *Candida shehatae*. *Enzyme Microb. Technol.* **19,** 220–225 (1996).

198. Mussatto, S. I. & Roberto, I. C. Alternatives for detoxification of diluted-acid lignocellulosic hydrolyzates for use in fermentative processes: A review. *Bioresource Technology* **93,** 1–10 (2004).

199. Palmqvist, E. & Hahn-Hägerdal, B. Fermentation of lignocellulosic hydrolysates. II: Inhibitors and mechanisms of inhibition. *Bioresource Technology* **74,** 25–33 (2000).

200. Liu, Z. L. *et al.* Adaptive response of yeasts to furfural and 5-hydroxymethylfurfural and new chemical evidence for HMF conversion to 2,5-bis-hydroxymethylfuran. *J. Ind. Microbiol. Biotechnol.* **31,** 345–352 (2004).

201. Marques, A. P. ., Evtuguin, D. V., Magina, F. M. L. & Prates, A. A. Chemical Composition of Spent Liquors from Acidic Magnesium–Based Sulphite Pulping of Eucalyptus globulus. *J. Wood Chem. Technol.* **29,** 322–336 (2009).

202. Parajó, J. C., Domínguez, H. & Domínguez, J. M. Biotechnological production of xylitol. Part 3: Operation in culture media made from lignocellulose hydrolysates. *Bioresour. Technol.* **66,** 25–40 (1998).

203. Cortez, D. V. & Roberto, I. C. Individual and interaction effects of vanillin and syringaldehyde on the xylitol formation by *Candida guilliermondii*. *Bioresour. Technol.* **101,** 1858–1865 (2010).

204. Panizzi, L., Caponi, C., Catalano, S., Cioni, P. L. & Morelli, I. In vitro antimicrobial activity of extracts and isolated constituents of *Rubus ulmifolius*. *J. Ethnopharmacol.* **79,** 165–168 (2002).

205. Upadhyay, G., Gupta, S. P., Prakash, O. & Singh, M. P. Pyrogallol-mediated toxicity and natural antioxidants: Triumphs and pitfalls of preclinical findings and their translational limitations. *Chemico-Biological Interactions* **183,** 333–340 (2010).

206. Chandel, A. K., da Silva, S. S. & Singh, O. V. Detoxification of Lignocellulose Hydrolysates: Biochemical and Metabolic Engineering Toward White Biotechnology. *Bioenergy Research* **6,** 388–401 (2013).

207. Russell, J. B. Another explanation for the toxicity of fermentation acids at low pH - Anion accumulation versus uncoupling. *J. Appl. Bacteriol.* **73,** 363–370 (1992).

208. Olsson, L. & Hahn-Hägerdal, B. Fermentation of lignocellulosic hydrolysates for ethanol production. *Enzyme Microb. Technol.* **18,** 312–331 (1996).

209. Xavier, A. M. R. B., Correia, M. F., Pereira, S. R. & Evtuguin, D. V. Second-generation bioethanol from eucalypt sulphite spent liquor. *Bioresour. Technol.* **101,** 2755–2761 (2010).

210. Hahn-Hägerdal, B., Karhumaa, K., Fonseca, C., Spencer-Martins, I. & Gorwa-

Grauslund, M. F. Towards industrial pentose-fermenting yeast strains. *Applied Microbiology and Biotechnology* **74,** 937–953 (2007).

211. Moysés, D. N., Reis, V. C. B., de Almeida, J. R. M., de Moraes, L. M. P. & Torres, F. A. G. Xylose fermentation by *Saccharomyces cerevisiae*: Challenges and prospects. *Int. J. Mol. Sci.* **17,** 1–18 (2016).
212. Jeffries, T. W. *et al.* Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nat. Biotechnol.* **25,** 319–326 (2007).
213. Pereira, S. R., Ivanuša, Š., Evtuguin, D. V., Serafim, L. S. & Xavier, A. M. R. B. Biological treatment of eucalypt spent sulphite liquors: A way to boost the production of second generation bioethanol. *Bioresour. Technol.* **103,** 131–135 (2012).
214. Nigam, J. N. Ethanol production from hardwood spent sulfite liquor using an adapted strain of *Pichia stipitis*. *J. Ind. Microbiol. Biotechnol.* **26,** 145–150 (2001).
215. Pereira, S. R. *et al.* Adaptation of *Scheffersomyces stipitis* to hardwood spent sulfite liquor by evolutionary engineering. *Biotechnol. Biofuels* **8,** 50 (2015).
216. Bajwa, P. K. *et al.* Ethanol production from selected lignocellulosic hydrolysates by genome shuffled strains of *Scheffersomyces stipitis*. *Bioresour. Technol.* **102,** 9965–9 (2011).
217. Ruis, H. & Schüller, C. Stress signaling in yeast. *BioEssays* **17,** 959–965 (1995).
218. Martinez-Pastor1, M. T. *et al.* The *Saccharomyces cerevisiae* zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress-response element (STRE). *EMBO J.* **15,** 2227–2235 (1996).
219. Schmitt, A. P. & McEntee, K. Msn2p, a zinc finger DNA-binding protein, is the transcriptional activator of the multistress response in *Saccharomyces cerevisiae*.

Proc. Natl. Acad. Sci. U. S. A. **93,** 5777–82 (1996).

220. Gasch, a P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11,** 4241–4257 (2000).
221. Causton, H. C. *et al.* Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell* **12,** 323–37 (2001).
222. Treger, J. M., Schmitt, A. P., Simon, J. R. & McEntee, K. Transcriptional factor mutations reveal regulatory complexities of heat shock and newly identified stress genes in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **273,** 26875–26879 (1998).
223. Hohmann, S. Osmotic stress signaling and osmoadaptation in yeasts. *Microbiol. Mol. Biol. Rev.* **66,** 300–372 (2002).
224. Lindquist, S. The heat-shock response. *Ann. Rev. Biochem.* **55,** 1151–91 (1986).
225. Bose, S., Dutko, J. A. & Zitomer, R. S. Genetic factors that regulate the attenuation of the general stress response of yeast. *Genetics* **169,** 1215–1226 (2005).
226. Winderickx, J. *et al.* Regulation of genes encoding subunits of the trehalose synthase complex in *Saccharomyces cerevisiae*: Novel variations of STRE-mediated transcription control? *Mol. Gen. Genet.* **252,** 470–482 (1996).
227. Guillou, V., Plourde-Owobi, L., Parrou, J. L., Goma, G. & François, J. Role of reserve carbohydrates in the growth dynamics of *Saccharomyces cerevisiae*. *FEMS Yeast Res.* **4,** 773–787 (2004).
228. Mansure, J. J. C., Panek, A. D., Crowe, L. M. & Crowe, J. H. Trehalose inhibits ethanol effects on intact yeast cells and liposomes. *BBA - Biomembr.* **1191,** 309–316 (1994).
229. Sola-Penna, M. & Meyer-Fernandes, J. R. Stabilization against Thermal

Inactivation Promoted by Sugars on Enzyme Structure and Function : Why Is Trehalose More Effective Than Other Sugars? *Arch. Biochem. Biophys.* **360,** 10–14 (1998).

230. Singer, M. a & Lindquist, S. Multiple effects of trehalose on protein folding in vitro and in vivo. *Mol. Cell* **1,** 639–648 (1998).
231. Panek, A. Function of trehalose in Baker's yeast (*Saccharomyces cerevisiae*). *Arch. Biochem. Biophys.* **100,** 422–425 (1963).
232. Gibney, P. a., Schieler, A., Chen, J. C., Rabinowitz, J. D. & Botstein, D. Characterizing the in vivo role of trehalose in *Saccharomyces cerevisiae* using the AGT1 transporter. *Pnas* 1506289112- (2015). doi:10.1073/pnas.1506289112
233. Tapia, H., Young, L., Fox, D., Bertozzi, C. R. & Koshland, D. Increasing intracellular trehalose is sufficient to confer desiccation tolerance to *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 6122–7 (2015).
234. Mira, N. P., Teixeira, M. C. & Sá-Correia, I. Adaptive response and tolerance to weak acids in *Saccharomyces cerevisiae*: a genome-wide view. *OMICS* **14,** 525–540 (2010).
235. Mollapour, M. & Piper, P. W. Hog1 mitogen-activated protein kinase phosphorylation targets the yeast Fps1 aquaglyceroporin for endocytosis, thereby rendering cells resistant to acetic acid. *Mol. Cell. Biol.* **27,** 6446–56 (2007).
236. Fernandes, A. R., Mira, N. P., Vargas, R. C., Canelhas, I. & Sá-Correia, I. *Saccharomyces cerevisiae* adaptation to weak acids involves the transcription factor Haa1p and Haa1p-regulated genes. *Biochem. Biophys. Res. Commun.* **337,** 95–103 (2005).
237. Teixeira, M. C., Santos, P. M., Fernandes, A. R. & Sá-Correia, I. A proteome

analysis of the yeast response to the herbicide 2,4-dichlorophenoxyacetic acid. *Proteomics* **5,** 1889–1901 (2005).

238. Ullah, A., Orij, R., Brul, S. & Smits, G. J. Quantitative analysis of the modes of growth inhibition by weak organic acids in *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.* **78,** 8377–8387 (2012).

239. Martínez-Muñoz, G. A. & Kane, P. Vacuolar and plasma membrane proton pumps collaborate to achieve cytosolic pH homeostasis in yeast. *J. Biol. Chem.* **283,** 20309–20319 (2008).

240. Tenreiro, S. *et al. AQR1* gene (ORF *YNL065*W) encodes a plasma membrane transporter of the major facilitator superfamily that confers resistance to short-chain monocarboxylic acids and quinidine in *Saccharomyces cerevisiae*. *Biochem. Biophys. Res. Commun.* **292,** 741–748 (2002).

241. Abbott, D. A., Suir, E., van Maris, A. J. & Pronk, J. T. Physiological and transcriptional responses to high concentrations of lactic acid in anaerobic chemostat cultures of *Saccharomyces cerevisiae*. *Appl Env. Microbiol* **74,** 5759–5768 (2008).

242. Hazelwood, L. A., Walsh, M. C., Pronk, J. T. & Daran, J. M. Involvement of vacuolar sequestration and active transport in tolerance of *Saccharomyces cerevisiae* to hop iso-alpha-acids. *Appl. Environ. Microbiol.* **76,** 318–328 (2010).

243. Mira, N. P., Lourenco, A. B., Fernandes, A. R., Becker, J. D. & Sa-Correia, I. The *RIM101* pathway has a role in *Saccharomyces cerevisiae* adaptive response and resistance to propionic acid and other weak acids. *FEMS Yeast Res.* **9,** 202–216 (2009).

244. Viegas, C. A. *et al.* Yeast adaptation to 2,4-dichlorophenoxyacetic acid involves

increased membrane fatty acid saturation degree and decreased *OLE1* transcription. *Biochem. Biophys. Res. Commun.* **330,** 271–278 (2005).

245. Almeida, B. *et al.* Yeast protein expression profile during acetic acid-induced apoptosis indicates causal involvement of the *TOR* pathway. *Proteomics* **9,** 720–732 (2009).

246. Schuller, C. *et al.* Global phenotypic analysis and transcriptional profiling defines the weak acid stress response regulon in *Saccharomyces cerevisiae*. *Mol Biol Cell* **15,** 706–720 (2004).

247. Gulshan, K. & Moye-Rowley, W. S. Multidrug resistance in fungi. *Eukaryotic Cell* **6,** 1933–1942 (2007).

248. Teixeira, M. C. & Sá-Correia, I. *Saccharomyces cerevisiae* resistance to chlorinated phenoxyacetic acid herbicides involves Pdr1p-mediated transcriptional activation of *TPO1* and *PDR5* genes. *Biochem. Biophys. Res. Commun.* **292,** 530–7 (2002).

249. Lamb, T. M. & Mitchell, A. P. The Transcription Factor Rim101p Governs Ion Tolerance and Cell Differentiation by Direct Repression of the Regulatory Genes *NRG1* and *SMP1* in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **23,** 677–686 (2003).

250. Castrejon, F., Gomez, A., Sanz, M., Duran, A. & Roncero, C. The *RIM101* pathway contributes to yeast cell wall assembly and its function becomes essential in the absence of mitogen-activated protein kinase Slt2p. *Eukaryot. Cell* **5,** 507–517 (2006).

251. Peñalva, M. A., Tilburn, J., Bignell, E. & Arst, H. N. Ambient pH gene regulation in fungi: making connections. *Trends in Microbiology* **16,** 291–300 (2008).

252. Kren, A. *et al.* War1p, a novel transcription factor controlling weak acid stress response in yeast. *Mol. Cell. Biol.* **23,** 1775–1785 (2003).

253. Stratford, M. *et al.* Weak-acid preservatives: pH and proton movements in the yeast *Saccharomyces cerevisiae*. *Int. J. Food Microbiol.* **161,** 164–171 (2013).

254. Ludovico, P. *et al.* Cytochrome c release and mitochondria involvement in programmed cell death induced by acetic acid in *Saccharomyces cerevisiae*. *Mol. Biol. Cell* **13,** 2598–2606 (2002).

255. Mollapour, M., Shepherd, A. & Piper, P. W. Presence of the Fps1p aquaglyceroporin channel is essential for Hog1p activation, but suppresses Slt2(Mpk1)p activation, with acetic acid stress of yeast. *Microbiology* **155,** 3304–3311 (2009).

256. Mroczek, S. & Kufel, J. Apoptotic signals induce specific degradation of ribosomal RNA in yeast. *Nucleic Acids Res.* **36,** 2874–2888 (2008).

257. Mira, N. P., Becker, J. D. & Sá-Correia, I. Genomic expression program involving the Haa1p-regulon in *Saccharomyces cerevisiae* response to acetic acid. *OMICS* **14,** 587–601 (2010).

258. Sugiyama, M. *et al.* Nuclear localization of Haa1, which is linked to its phosphorylation status, mediates lactic acid tolerance in *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.* **80,** 3488–3495 (2014).

259. Tanaka, K., Ishii, Y., Ogawa, J. & Shima, J. Enhancement of acetic acid tolerance in *Saccharomyces cerevisiae* by overexpression of the *HAA1* gene, encoding a transcriptional activator. *Appl. Environ. Microbiol.* **78,** 8161–8163 (2012).

260. Halliwell, B. & Aruoma, O. I. DNA damage by oxygen-derived species: Its mechanism and measurement in mammalian systems. *FEBS Letters* **281,** 9–19

(1991).

261. Dawes, I. W. in *The metabolism and molecular physiology of Saccharomyces cerevisiae* (eds. Dickinson, J. R. & Schweizer, Mi.) 376–438 (CRC Press, 2004).

262. Salmon, T. B., Evert, B. A., Song, B. & Doetsch, P. W. Biological consequences of oxidative stress-induced DNA damage in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **32,** 3712–23 (2004).

263. Ribeiro, G. F., Corte-Real, M. & Johansson, B. Characterization of DNA Damage in Yeast Apoptosis Induced by Hydrogen Peroxide, Acetic Acid, and Hyperosmotic Shock. *Mol. Biol. Cell* **17,** 4584–4591 (2006).

264. Cabiscol, E., Piulats, E., Echave, P., Herrero, E. & Ros, J. Oxidative stress promotes specific protein damage in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **275,** 27393–8 (2000).

265. Girotti, A. W. Lipid hydroperoxide generation, turnover, and effector action in biological systems. *J. Lipid Res.* **39,** 1529–1542 (1998).

266. Ames, B. N. & Gold, L. S. Endogenous mutagens and the causes of aging and cancer. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **250,** 3–16 (1991).

267. Holliday, R. The induction of recombination by mitomycin C in *Ustilago* and *Saccharomyces*. *Genetics* **50,** 323–335 (1964).

268. Parry, J. M. The induction of gene conversion in yeast by herbicide preparations. *Mutat. Res.* **21,** 83–91 (1973).

269. Olinski, R. *et al.* Oxidative DNA damage: Assessment of the role in carcinogenesis, atherosclerosis, and acquired immunodeficiency syndrome. *Free Radic. Biol. Med.* **33,** 192–200 (2002).

270. Wolff, S. P. & Dean, R. T. Fragmentation of proteins by free radicals and its effect

on their susceptibility to enzymic hydrolysis. *Biochem. J.* **234,** 399–403 (1986).

271. Alic, N., Higgins, V. J. & Dawes, I. W. Identification of a *Saccharomyces cerevisiae* gene that is required for G1 arrest in response to the lipid oxidation product linoleic acid hydroperoxide. *Mol. Biol. Cell* **12,** 1801–10 (2001).
272. Levine, R. L. Carbonyl modified proteins in cellular regulation, aging, and disease. *Free Radical Biology and Medicine* **32,** 790–796 (2002).
273. Grant, C. M. Role of the glutathione/glutaredoxin and thioredoxin systems in yeast growth and response to stress conditions. *Molecular Microbiology* **39,** 533–541 (2001).
274. Ernster, L. & Dallner, G. Biochemical, physiological and medical aspects of ubiquinone function. *BBA - Mol. Basis Dis.* **1271,** 195–204 (1995).
275. Kalén, A., Norling, B., Appelkvist, E. L. & Dallner, G. Ubiquinone biosynthesis by the microsomal fraction from rat liver. *BBA - Gen. Subj.* **926,** 70–78 (1987).
276. Huh, W. K. *et al.* D-erythroascorbic acid is an important antioxidant molecule in *Saccharomyces cerevisiae*. *Mol. Microbiol.* **30,** 895–903 (1998).
277. Benaroudj, N., Lee, D. H. & Goldberg, A. L. Trehalose Accumulation during Cellular Stress Protects Cells and Cellular Proteins from Damage by Oxygen Radicals. *J. Biol. Chem.* **276,** 24261–24267 (2001).
278. Herdeiro, R. S., Pereira, M. D., Panek, A. D. & Eleutherio, E. C. A. Trehalose protects *Saccharomyces cerevisiae* from lipid peroxidation during oxidative stress. *Biochim. Biophys. Acta - Gen. Subj.* **1760,** 340–346 (2006).
279. Sebollela, A. *et al.* Inhibition of yeast glutathione reductase by trehalose: Possible implications in yeast survival and recovery from stress. *Int. J. Biochem. Cell Biol.* **36,** 900–908 (2004).

280. Pedreño, Y., Gimeno-Alcañiz, J. V., Matallana, E. & Argüelles, J. C. Response to oxidative stress caused by H_2O_2 in *Saccharomyces cerevisiae* mutants deficient in trehalase genes. *Arch. Microbiol.* **177,** 494–499 (2002).

281. Gralla, E. B. & Kosman, D. J. Molecular genetics of superoxide dismutases in yeasts and related fungi. *Adv. Genet.* **30,** 251–319 (1992).

282. Longo, V. D., Gralla, E. B. & Valentine, J. S. Superoxide dismutase activity is essential for stationary phase survival in *Saccharomyces cerevisiae*: Mitochondrial production of toxic oxygen species in vivo. *J. Biol. Chem.* **271,** 12275–12280 (1996).

283. Seah, T. & Kaplan, J. Purification and properties of the catalase of baker's yeast. *J. Biol. Chem.* **248,** 2889–2893 (1973).

284. Cohen, G., Fessl, F., Traczyk, A., Rytka, J. & Ruis, H. Isolation of the catalase A gene of *Saccharomyces cerevisiae* by complementation of the cta1 mutation. *MGG Mol. Gen. Genet.* **200,** 74–79 (1985).

285. Wieser, R. *et al.* Heat shock factor-independent heat control of transcription of the CTT1 gene encoding the cytosolic catalase T of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **266,** 12406–12411 (1991).

286. Avery, A. M. & Avery, S. V. *Saccharomyces cerevisiae* Expresses Three Phospholipid Hydroperoxide Glutathione Peroxidases. *J. Biol. Chem.* **276,** 33730–33735 (2001).

287. Kagi, J. & Schaffer, A. Biochemistry of metallothionein. *Biochemistry* **27,** 8509–8515 (1988).

288. Jamieson, D. J. Oxidative stress responses of the yeast *Saccharomyces cerevisiae*. *Yeast* **14,** 1511–1527 (1998).

289. Morano, K. a., Grant, C. M. & Moye-Rowley, W. S. The Response to Heat Shock and Oxidative Stress in *Saccharomyces cerevisiae*. *Genetics* **190,** 1157–1195 (2012).

290. Kuge, S. *et al.* Regulation of the yeast Yap1p nuclear export signal is mediated by redox signal-induced reversible disulfide bond formation. *Mol. Cell. Biol.* **21,** 6139–50 (2001).

291. Gulshan, K., Rovinsky, S. A., Coleman, S. T. & Moye-Rowley, W. S. Oxidant-specific folding of Yap1p regulates both transcriptional activation and nuclear localization. *J. Biol. Chem.* **280,** 40524–40533 (2005).

292. Kuge, S. & Jones, N. *YAP1* dependent activation of *TRX2* is essential for the response of *Saccharomyces cerevisiae* to oxidative stress by hydroperoxides. *EMBO J.* **13,** 655–64 (1994).

293. Morgan, B. A. *et al.* The Skn7 response regulator controls gene expression in the oxidative stress response of the budding yeast *Saccharomyces cerevisiae*. *EMBO J.* **16,** 1035–44 (1997).

294. Lee, J. *et al.* Yap1 and Skn7 control two specialized oxidative stress response regulons in yeast. *J. Biol. Chem.* **274,** 16040–16046 (1999).

295. Krems, B., Charizanis, C. & Entian, K. D. The response regulator-like protein Pos9/Skn7 of *Saccharomyces cerevisiae* is involved in oxidative stress resistance. *Curr. Genet.* **29,** 327–34 (1996).

296. Schüller, C., Brewster, J. L., Alexander, M. R., Gustin, M. C. & Ruis, H. The *HOG* pathway controls osmotic regulation of transcription via the stress response element (STRE) of the *Saccharomyces cerevisiae CTT1* gene. *EMBO J.* **13,** 4382–9 (1994).

297. Rep, M. *et al.* The *Saccharomyces cerevisiae* Sko1p transcription factor mediates *HOG* pathway-dependent osmotic regulation of a set of genes encoding enzymes implicated in protection from oxidative damage. *Mol. Microbiol.* **40,** 1067–1083 (2001).

298. Singh, K. K. The *Saccharomyces cerevisiae* Sln1p-Ssk1p two-component system mediates response to oxidative stress and in an oxidant-specific fashion. *Free Radic. Biol. Med.* **29,** 1043–1050 (2000).

299. Csonka, L. N. & Hanson, A. D. Prokaryotic Osmoregulation. *Annu Rev Microbiol* **45,** 569–606 (1991).

300. Sharma, S. C., Raj, D., Forouzandeh, M. & Bansal, M. P. Salt-induced changes in lipid composition and ethanol tolerance in *Saccharomyces cerevisiae*. *Appl. Biochem. Biotechnol.* **56,** 189–95 (1996).

301. Nass, R. & Rao, R. The yeast endosomal Na+/H+ exchanger, Nhx1, confers osmotolerance following acute hypertonic shock. *Microbiology* **145,** 3221–3228 (1999).

302. Poolman, B. & Glaasker, E. Regulation of compatible solute accumulation in bacteria. *Molecular Microbiology* **29,** 397–407 (1998).

303. Albertyn, J., Hohmann, S., Thevelein, J. M. & Prior, B. A. *GPD1*, which encodes glycerol-3-phosphate dehydrogenase, is essential for growth under osmotic stress in *Saccharomyces cerevisiae*, and its expression is regulated by the high-osmolarity glycerol response pathway. *Mol. Cell. Biol.* **14,** 4135–44 (1994).

304. Norbeck, J. & Blomberg, A. Protein expression during exponential growth in 0.7 M NaCl medium of *Saccharomyces cerevisiae*. *FEMS Microbiol. Lett.* **137,** 1–8 (1996).

305. Eriksson, P., André, L., Ansell, R., Blomberg, a & Adler, L. Cloning and characterization of *GPD2*, a second gene encoding sn-glycerol 3-phosphate dehydrogenase (NAD+) in *Saccharomyces cerevisiae*, and its comparison with GPD1. *Mol. Microbiol.* **17,** 95–107 (1995).

306. Lidén, G. *et al.* A glycerol-3-phosphate dehydrogenase-deficient mutant of *Saccharomyces cerevisiae* expressing the heterologous *XYL1* gene. *Appl. Environ. Microbiol.* **62,** 3894–3896 (1996).

307. Ansell, R., Granath, K., Hohmann, S., Thevelein, J. M. & Adler, L. The two isoenzymes for yeast NAD+-dependent glycerol 3-phosphate dehydrogenase encoded by *GPD1* and *GPD2* have distinct roles in osmoadaptation and redox regulation. *EMBO J.* **16,** 2179–2187 (1997).

308. Hounsa, C. G., Brandt, E. V., Thevelein, J., Hohmann, S. & Prior, B. A. Role of trehalose in survival of *Saccharomyces cerevisiae* under osmotic stress. *Microbiology* **144,** 671–680 (1998).

309. Saito, H. & Tatebayashi, K. Regulation of the osmoregulatory *HOG* MAPK cascade in yeast. *Journal of Biochemistry* **136,** 267–272 (2004).

310. Karlgren, S. *et al.* Conditional osmotic stress in yeast: A system to study transport through aquaglyceroporins and osmostress signaling. *J. Biol. Chem.* **280,** 7186–7193 (2005).

311. Gualtieri, T., Ragni, E., Mizzi, L., Fascio, U. & Popolo, L. The cell wall censor Wsc1p is involved in reorganization of actin cytoskeleton in response to hypo-osmotic shock in *Saccharomyces cerevisiae*. *Yeast* **21,** 1107–1120 (2004).

312. García-Rodríguez, L. J., Valle, R., Durán, Á. & Roncero, C. Cell integrity signaling activation in response to hyperosmotic shock in yeast. *FEBS Lett.* **579,** 6186–

6190 (2005).

313. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010). at <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

314. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30,** 614–20 (2014).

315. Otero, J. M. *et al.* Whole genome sequencing of *Saccharomyces cerevisiae*: from genotype to phenotype for improved metabolic engineering applications. *BMC Genomics* **11,** 723 (2010).

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

317. Broad Institute. Picard: A set of command line tools for manipulating high-throughput sequencing (HTS) data and formats. (2016). at <http://broadinstitute.github.io/picard/>

318. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

319. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

320. Van der Auwera, G. A. *et al.* From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* (2013). doi:10.1002/0471250953.bi1110s43

321. Larson, D. Bam-readcount: program to generate metrics at single nucleotide positions from BAM files. (2016). at <https://github.com/genome/bam-readcount>

322. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics*

25, 2078–2079 (2009).

323. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).
324. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–192 (2013).
325. Waskom, M. Seaborn: statistical data visualization. (2016).
326. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).
327. Goujon, M. *et al.* A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.* **38,** (2010).
328. McWilliam, H. *et al.* Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.* **41,** (2013).
329. Eswar, N. Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.* **31,** 3375–3380 (2003).
330. Werner, C., Stubbs, M. T., Krauth-Siegel, R. L. & Klebe, G. The crystal structure of *Plasmodium falciparum* glutamate dehydrogenase, a putative target for novel antimalarial drugs. *J. Mol. Biol.* **349,** 597–607 (2005).
331. Schrödinger LLC. The PyMOL Molecular Graphics System, Version 1.8. (2016).
332. Biot-Pelletier, D. & Martin, V. J. J. Seamless site-directed mutagenesis of the *Saccharomyces cerevisiae* genome using CRISPR-Cas9. *J. Biol. Eng.* **10,** 6 (2016).
333. DiCarlo, J. E. *et al.* Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic Acids Res.* **41,** 4336–43 (2013).

334. Ryan, O. W. *et al.* Selection of chromosomal DNA libraries using a multiplex CRISPR system. *Elife* e03703 (2014). doi:10.7554/eLife.03703

335. Bao, Z. *et al.* Homology-Integrated CRISPR-Cas (HI-CRISPR) System for One-Step Multigene Disruption in *Saccharomyces cerevisiae*. *ACS Synth. Biol.* **4,** 585–594 (2014).

336. Stovicek, V., Borodina, I. & Forster, J. CRISPR–Cas system enables fast and simple genome editing of industrial *Saccharomyces cerevisiae* strains. *Metab. Eng. Commun.* **2,** 13–22 (2015).

337. Mans, R. *et al.* CRISPR/Cas9: a molecular Swiss army knife for simultaneous introduction of multiple genetic modifications in *Saccharomyces cerevisiae*. *FEMS Yeast Res.* **15,** fov004 (2015).

338. Jakočiūnas, T. *et al.* Multiplex metabolic pathway engineering using CRISPR/Cas9 in *Saccharomyces cerevisiae*. *Metab. Eng.* **28,** 213–222 (2015).

339. Horwitz, A. A. *et al.* Efficient Multiplexed Integration of Synergistic Alleles and Metabolic Pathways in Yeasts via CRISPR-Cas. *Cell Syst.* **1,** 88–96 (2015).

340. Ronda, C. *et al.* CrEdit: CRISPR mediated multi-loci gene integration in *Saccharomyces cerevisiae*. *Microb. Cell Fact.* **14,** 97 (2015).

341. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31,** 822–6 (2013).

342. Lee, M. E., Deloache, W. C., Cervantes, B. & Dueber, J. E. A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly. *ACS Synth Biol* **4,** 975–986 (2015).

343. Xie, F. *et al.* Seamless gene correction of beta-thalassemia mutations in patient-specific iPSCs using CRISPR/Cas9 and piggyBac. *Genome Res.* gr.173427.114-

(2014). doi:10.1101/gr.173427.114

344. Storici, F. & Resnick, M. A. Delitto perfetto targeted mutagenesis in yeast with oligonucleotides. *Genet. Eng. (N. Y).* **25,** 189–207 (2003).

345. Pascon, R. C. & Miller, B. L. Morphogenesis in *Aspergillus nidulans* requires Dopey (DopA), a member of a novel family of leucine ripper-like proteins conserved from yeast to humans. *Mol. Microbiol.* **36,** 1250–1264 (2000).

346. Gautier, T., Bergès, T., Tollervey, D. & Hurt, E. Nucleolar KKE/D repeat proteins Nop56p and Nop58p interact with Nop1p and are required for ribosome biogenesis. *Mol. Cell. Biol.* **17,** 7088–98 (1997).

347. Kistler, M., Maier, K. & Eckardt-Schupp, F. Genetic and biochemical analysis of glutathione-deficient mutants of *Saccharomyces cerevisiae*. *Mutagenesis* **5,** 39–44 (1990).

348. Ohtake, Y. & Yabuuchi, S. Molecular cloning of the gamma-glutamylcysteine synthetase gene of *Saccharomyces cerevisiae*. *Yeast* **7,** 953–61 (1991).

349. Park, S. H., Koh, S. S., Chun, J. H., Hwang, H. J. & Kang, H. S. *NRG1* is a transcriptional repressor for glucose repression of *STA1* gene expression in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **19,** 2044–2050 (1999).

350. Zhou, H. & Winston, F. *NRG1* is required for glucose repression of the *SUC2* and *GAL* genes of *Saccharomyces cerevisiae*. *BMC Genet.* **2,** 5 (2001).

351. DeRisi, J. L. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science.* **278,** 680–686 (1997).

352. Vyas, V. K., Kuchin, S. & Carlson, M. Interaction of the repressors *NRG1* and Nrg2 with the Snf1 protein kinase in *Saccharomyces cerevisiae*. *Genetics* **158,** 563–572 (2001).

353. Kuchin, S., Vyas, V. K. & Carlson, M. Snf1 protein kinase and the repressors *NRG1* and Nrg2 regulate *FLO11*, haploid invasive growth, and diploid pseudohyphal differentiation. *Mol. Cell. Biol.* **22,** 3994–4000 (2002).

354. Lyons, T. J. *et al.* Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 7957–7962 (2000).

355. Lamb, T. M., Xu, W., Diamond, A. & Mitchell, A. P. Alkaline response genes of *Saccharomyces cerevisiae* and their relationship to the *RIM101* pathway. *J. Biol. Chem.* **276,** 1850–1856 (2001).

356. Haro, R., Garciadeblas, B. & Rodriguez-Navarro, A. A novel P-type ATPase from yeast involved in sodium transport. *FEBS Lett.* **291,** 189–191 (1991).

357. Vyas, V. K., Berkey, C. D., Miyao, T. & Carlson, M. Repressors *NRG1* and Nrg2 regulate a set of stress-responsive genes in *Saccharomyces cerevisiae*. *Eukaryot. Cell* **4,** 1882–1891 (2005).

358. Murad, A. M. A. *et al. NRG1* represses yeast-hypha morphogenesis and hypha-specific gene expression in Candida albicans. *EMBO J.* **20,** 4742–4752 (2001).

359. Zambrano, M. M., Siegele, D. a, Almirón, M., Tormo, A. & Kolter, R. Microbial competition: *Escherichia coli* mutants that take over stationary phase cultures. *Science* **259,** 1757–1760 (1993).

360. Herring, C. D. *et al.* Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nat Genet* **38,** 1406–1412 (2006).

361. Maharjan, R., Seeto, S., Notley-McRobb, L. & Ferenci, T. Clonal adaptive radiation in a constant environment. TL - 313. *Science.* **313,** 514–517 (2006).

362. Kinnersley, M. A., Holben, W. E. & Rosenzweig, F. *E unibus plurum*: Genomic

analysis of an experimentally evolved polymorphism in *Escherichia coli*. *PLoS Genet.* **5,** (2009).

363. Wang, L. *et al.* Divergence involving global regulatory gene mutations in an *Escherichia coli* population evolving under phosphate limitation. *Genome Biol. Evol.* **2,** 478–487 (2010).
364. Ferenci, T. What is driving the acquisition of *mutS* and *rpoS* polymorphisms in *Escherichia coli*? *Trends in Microbiology* **11,** 457–461 (2003).
365. Philippe, N., Crozat, E., Lenski, R. E. & Schneider, D. Evolution of global regulatory networks during a long-term experiment with *Escherichia coli*. *BioEssays* **29,** 846–860 (2007).
366. Conrad, T. M. *et al.* RNA polymerase mutants found through adaptive evolution reprogram *Escherichia coli* for optimal growth in minimal media. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 20500–5 (2010).
367. Lee, J. C. *et al.* The essential and ancillary role of glutathione in *Saccharomyces cerevisiae* analysed using a grande gsh1 disruptant strain. *FEMS Yeast Res.* **1,** 57–65 (2001).
368. Grant, C. M., MacIver, F. H. & Dawes, I. W. Glutathione synthetase is dispensable for growth under both normal and oxidative stress conditions in the yeast *Saccharomyces cerevisiae* due to an accumulation of the dipeptide gamma-glutamylcysteine. *Mol. Biol. Cell* **8,** 1699–1707 (1997).
369. Stephen, D. W. & Jamieson, D. J. Amino acid-dependent regulation of the *Saccharomyces cerevisiae GSH1* gene by hydrogen peroxide. *Mol. Microbiol.* **23,** 203–210 (1997).
370. Sugiyama, K. I., Izawa, S. & Inoue, Y. The Yap1p-dependent induction of

glutathione synthesis in heat shock response of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **275,** 15535–15540 (2000).

371. Dormer, U. H., Westwater, J., Stephen, D. W. S. & Jamieson, D. J. Oxidant regulation of the *Saccharomyces cerevisiae* GSH1 gene. *Biochim. Biophys. Acta - Gene Struct. Expr.* **1576,** 23–29 (2002).
372. Miller, S. M. & Magasanik, B. Role of NAD-linked glutamate dehydrogenase in nitrogen metabolism in *Saccharomyces cerevisiae*. *J. Bacteriol.* **172,** 4927–4935 (1990).
373. Avendano, A., Deluna, A., Olivera, H., Valenzuela, L. & Gonzalez, A. *GDH3* encodes a glutamate dehydrogenase isozyme, a previously unrecognized route for glutamate biosynthesis in *Saccharomyces cerevisiae*. *J. Bacteriol.* **179,** 5594–5597 (1997).
374. Benjamin, P. M., Wu, J. I., Mitchell, A. P. & Magasanik, B. Three regulatory systems control expression of glutamine synthetase in *Saccharomyces cerevisiae* at the level of transcription. *Mol. Gen. Genet.* **217,** 370–7 (1989).
375. Filetici, P., Martegani, M. P., Valenzuela, L., González, A. & Ballario, P. Sequence of the *GLT1* gene from *Saccharomyces cerevisiae* reveals the domain structure of yeast glutamate synthase. *Yeast* **12,** 1359–1366 (1996).
376. DeLuna, A., Avendaño, A., Riego, L. & González, A. NADP-glutamate dehydrogenase isoenzymes of *Saccharomyces cerevisiae*: Purification, kinetic properties, and physiological roles. *J. Biol. Chem.* **276,** 43775–43783 (2001).
377. Lee, Y. J., Kim, K. J., Kang, H. Y., Kim, H. R. & Maeng, P. J. Involvement of *GDH3*-encoded NADP+-dependent glutamate dehydrogenase in yeast cell resistance to stress-induced apoptosis in stationary phase cells. *J. Biol. Chem.*

287, 44221–44233 (2012).

378. Stillman, T. J., Baker, P. J., Britton, K. L. & Rice, D. W. Conformational flexibility in glutamate dehydrogenase. Role of water in substrate recognition and catalysis. *J. Mol. Biol.* **234,** 1131–9 (1993).

379. Wanke, V., Vavassori, M., Thevelein, J. M., Tortora, P. & Vanoni, M. Regulation of maltose utilization in *Saccharomyces cerevisiae* by genes of the *RAS*/protein kinase A pathway. *FEBS Lett.* **402,** 251–255 (1997).

380. Nijkamp, J. F. *et al.* De novo sequencing, assembly and analysis of the genome of the laboratory strain *Saccharomyces cerevisiae* CEN.PK113-7D, a model for modern industrial biotechnology. *Microb. Cell Fact.* **11,** 36 (2012).

381. Plourde-Owobi, L. *et al. AGT1*, encoding an α-glucoside transporter involved in uptake and intracellular accumulation of trehalose in *Saccharomyces cerevisiae*. *J. Bacteriol.* **181,** 3830–3832 (1999).

382. Jules, M., Guillou, V., François, J. & Parrou, J. L. Two distinct pathways for trehalose assimilation in the yeast *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.* **70,** 2771–2778 (2004).

383. Han, E. K., Cotty, F., Sottas, C., Jiang, H. & Michels, C. A. Characterization of Agt1 Encoding a General Alpha-Glucoside Transporter From *Saccharomyces*. *Mol. Microbiol.* **17,** 1093–1107 (1995).

384. de Oliveira, D. E., Rodrigues, E. G. C., Mattoon, J. R. & Panek, A. D. Relationships between trehalose metabolism and maltose utilization in *Saccharomyces cerevisiae* - II. Effect of Constitutive *MAL* Genes. *Curr. Genet.* **3,** 235–242 (1981).

385. Hu, Z., Nehlin, J. O., Ronne, H. & Michels, C. A. *MIG1*-dependent and *MIG1*-

independent glucose regulation of *MAL* gene expression in *Saccharomyces cerevisiae*. *Curr. Genet.* **28,** 258–266 (1995).

386. Zhang, C. Y., Bai, X. W., Lin, X., Liu, X. E. & Xiao, D. G. Effects of *SNF1* on Maltose Metabolism and Leavening Ability of Baker's Yeast in Lean Dough. *J. Food Sci.* **80,** M2879–M2885 (2015).
387. Hu, Z. *et al.* Analysis of the mechanism by which glucose inhibits maltose induction of *MAL* gene expression in *Saccharomyces*. *Genetics* **154,** 121–132 (2000).
388. Okiyoneda, T. *et al.* Peripheral protein quality control removes unfolded CFTR from the plasma membrane. *Science* **329,** 805–10 (2010).
389. Apaja, P. M., Xu, H. & Lukacs, G. L. Quality control for unfolded proteins at the plasma membrane. *J. Cell Biol.* **191,** 553–70 (2010).
390. Lin, C. H., MacGurn, J. A., Chu, T., Stefan, C. J. & Emr, S. D. Arrestin-Related Ubiquitin-Ligase Adaptors Regulate Endocytosis and Protein Turnover at the Cell Surface. *Cell* **135,** 714–725 (2008).
391. Nikko, E. & Pelham, H. R. B. Arrestin-mediated endocytosis of yeast plasma membrane transporters. *Traffic* **10,** 1856–1867 (2009).
392. Nikko, E., Sullivan, J. A. & Pelham, H. R. B. Arrestin-like proteins mediate ubiquitination and endocytosis of the yeast metal transporter Smf1. *EMBO Rep.* **9,** 1216–21 (2008).
393. Zhao, Y., MacGurn, J. A., Liu, M. & Emr, S. The *ART*-Rsp5 ubiquitin ligase network comprises a plasma membrane quality control system that protects yeast cells from proteotoxic stress. *Elife* **2013,** (2013).
394. Ren, J., Kee, Y., Huibregtse, J. M. & Piper, R. C. Hse1, a Component of the Yeast

Hrs-STAM Ubiquitin-sorting Complex, Associates with Ubiquitin Peptidases and a Ligase to Control Sorting Efficiency into Multivesicular Bodies. *Mol. Biol. Cell* **18,** 324–335 (2007).

395. Weinberg, J. S. & Drubin, D. G. Regulation of Clathrin-mediated endocytosis by dynamic ubiquitination and deubiquitination. *Curr. Biol.* **24,** 951–959 (2014).
396. Böhm, S. *et al.* The budding yeast ubiquitin protease Ubp7 is a novel component involved in s phase progression. *J. Biol. Chem.* **291,** 4442–4452 (2016).
397. Lam, M. H. Y. *et al.* Interaction of the deubiquitinating enzyme Ubp2 and the E3 ligase Rsp5 is required for transporter/receptor sorting in the multivesicular body pathway. *PLoS One* **4,** (2009).
398. Kee, Y., Muñoz, W., Lyon, N. & Huibregtse, J. M. The deubiquitinating enzyme Ubp2 modulates Rsp5-dependent Lys 63-linked polyubiquitin conjugates in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **281,** 36724–36731 (2006).
399. Kee, Y., Lyon, N. & Huibregtse, J. M. The Rsp5 ubiquitin ligase is coupled to and antagonized by the Ubp2 deubiquitinating enzyme. *EMBO J.* **24,** 2414–2424 (2005).
400. Zhang, Z. R., Bonifacino, J. S. & Hegde, R. S. Deubiquitinases sharpen substrate discrimination during membrane protein degradation from the ER. *Cell* **154,** 609–622 (2013).
401. Boorstein, W. R., Ziegelhoffer, T. & Craig, E. A. Molecular evolution of the *HSP70* multigene family. *J. Mol. Evol.* **38,** 1–17 (1994).
402. Werner-Washburne, M., Stone, D. E. & Craig, E. A. Complex interactions among members of an essential subfamily of *HSP70* genes in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **7,** 2568–77 (1987).

403. Bukau, B. & Horwich, A. L. The Hsp70 and Hsp60 chaperone machines. *Cell* **92,** 351–366 (1998).

404. Becker, J. & Craig, E. A. Heat-shock proteins as molecular chaperones. *Eur. J. Biochem.* **219,** 11–23 (1994).

405. Werner-Washburne, M., Becker, J., Kosic-Smithers, J. & Craig, E. A. Yeast Hsp70 RNA levels vary in response to the physiological status of the cell. *J. Bacteriol.* **171,** 2680–2688 (1989).

406. Slater, M. R. & Craig, E. A. Transcriptional regulation of an *HSP70* heat shock gene in the yeast *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **7,** 1906–16 (1987).

407. Stephen, D. W. S., Rivers, S. L. & Jamieson, D. J. The role of the *YAP1* and *YAP2* genes in the regulation of the adaptive oxidative stress responses of *Saccharomyces cerevisiae*. *Mol. Microbiol.* **16,** 415–423 (1995).

408. Young, M. R. & Craig, E. a. *Saccharomyces cerevisiae HSP70* heat shock elements are functionally distinct. *Mol. Cell. Biol.* **13,** 5637–46 (1993).

409. Gilbert, C. S. *et al.* The budding yeast Rad9 checkpoint complex: chaperone proteins are required for its function. *EMBO Rep.* **4,** 953–8 (2003).

410. Krantz, K. C. *et al.* Clathrin coat disassembly by the yeast hsc70/Ssa1p and auxilin/Swa2p proteins observed by single-particle burst analysis spectroscopy. *J. Biol. Chem.* **288,** 26721–26730 (2013).

411. Allen, K. D. *et al.* Hsp70 chaperones as modulators of prion life cycle: Novel effects of Ssa and Ssb on the *Saccharomyces cerevisiae* prion [PSI+]. *Genetics* **169,** 1227–1242 (2005).

412. Gokhale, K. C., Newnam, G. P., Sherman, M. Y. & Chernoff, Y. O. Modulation of prion-dependent polyglutamine aggregation and toxicity by chaperone proteins in

the yeast model. *J. Biol. Chem.* **280,** 22809–22818 (2005).

413. Schwimmer, C. & Masison, D. C. Antagonistic interactions between yeast [PSI(+)] and [URE3] prions and curing of [URE3] by Hsp70 protein chaperone Ssa1p but not by Ssa2p. *Mol. Cell. Biol.* **22,** 3590–3598 (2002).
414. Huh, W.-K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425,** 686–691 (2003).
415. Arias, P. *et al.* Genome-wide survey of yeast mutations leading to activation of the yeast cell integrity MAPK pathway: novel insights into diverse MAPK outcomes. *BMC Genomics* **12,** 390 (2011).
416. Jung, U. S. & Levin, D. E. Genome-wide analysis of gene expression regulated by the yeast cell wall integrity signalling pathway. *Mol. Microbiol.* **34,** 1049–1057 (1999).
417. Maas, N., Miller, K., DeFazio, L. & Toczyski, D. Cell cycle and checkpoint regulation of histone H3 K56 acetylation by Hst3 and Hst4. *Mol. Cell* **23,** 109–119 (2006).
418. Gutiérrez-Caballero, C., Cebollero, L. R. & Pendás, A. M. Shugoshins: From protectors of cohesion to versatile adaptors at the centromere. *Trends in Genetics* **28,** 351–360 (2012).
419. Rattani, A. *et al.* Sgol2 provides a regulatory platform that coordinates essential cell cycle processes during meiosis I in oocytes. *Elife* **2,** e01133 (2013).
420. Verzijlbergen, K. F. *et al.* Shugoshin biases chromosomes for biorientation through condensin recruitment to the pericentromere. *Elife* **2014,** (2014).
421. Marston, A. L., Tham, W. H., Shah, H. & Amon, A. A genome-wide screen identifies genes required for centromeric cohesion. *Science.* **303,** 1367–1370

(2004).

422. Kitajima, T. S., Kawashima, S. a & Watanabe, Y. The conserved kinetochore protein shugoshin protects centromeric cohesion during meiosis. *Nature* **427,** 510–517 (2004).

423. Katis, V. L., Galova, M., Rabitsch, K. P., Gregan, J. & Nasmyth, K. Maintenance of cohesin at centromeres after meiosis I in budding yeast requires a kinetochore-associated protein related to MEI-S332. *Curr. Biol.* **14,** 560–572 (2004).

424. Indjeian, V., Stern, B. & Murray, A. The centromeric protein Sgo1 is required to sense lack of tension on mitotic chromosomes. *Science.* **303,** 1367–1370 (2005).

425. Kiburz, B. M., Amon, A. & Marston, A. L. Shugoshin promotes sister kinetochore biorientation in *Saccharomyces cerevisiae*. *Mol. Biol. Cell* **19,** 1199–1209 (2008).

426. Kawashima, S. A. *et al.* Shugoshin enables tension-generating attachment of kinetochores by loading Aurora to centromeres. *Genes Dev.* **21,** 420–435 (2007).

427. Park, H. & Sternglanz, R. Identification and characterization of the genes for two topoisomerase I-interacting proteins from *Saccharomyces cerevisiae*. *Yeast* **15,** 35–41 (1999).

428. Johzuka, K. & Horiuchi, T. The cis Element and Factors Required for Condensin Recruitment to Chromosomes. *Mol. Cell* **34,** 26–35 (2009).

429. Waples, W., Chahwan, C., Cichonska, M. & Lavoie, B. Putting the Brake on FEAR: Tof2 Promotes the Biphasic Rlease of Cdc14 Phosphotase during Mitotic Exit. *Mol. Biol.* **20,** 3083–3092 (2009).

430. Geil, C., Schwab, M. & Seufert, W. A Nucleolus-Localized Activator of Cdc14 Phosphatase Supports rDNA Segregation in Yeast Mitosis. *Curr. Biol.* **18,** 1001–1005 (2008).

431. Huang, J. *et al.* Inhibition of homologous recombination by a cohesin-associated clamp complex recruited to the rDNA recombination enhancer. *Genes Dev.* **20,** 2887–2901 (2006).

432. Elion, E. a. The Ste5p scaffold. *J. Cell Sci.* **114,** 3967–3978 (2001).

433. Alvaro, C. G. & Thorner, J. Heterotrimeric-G-protein-coupled Receptor Signaling in Yeast Mating Pheromone Response. *J. Biol. Chem.* **291,** jbc.R116.714980 (2016).

434. Gustin, M. C., Albertyn, J., Alexander, M. & Davenport, K. MAP kinase pathways in the yeast *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* **62,** 1264–1300 (1998).

435. O'Rourke, S. M. & Herskowitz, I. The Hog1 MAPK prevents cross talk between the *HOG* and pheromone response MAPK pathways in *Saccharomyces cerevisiae*. *Genes Dev.* **12,** 2874–2886 (1998).

436. Liu, H., Styles, C. A. & Fink, G. R. Elements of the yeast pheromone response pathway required for filamentous growth of diploids. *Science.* **262,** 1741–1744 (1993).

437. Roberts, R. L. & Fink, G. R. Elements of a single map kinase cascade in *Saccharomyces cerevisiae* mediate two developmental programs in the same cell type: Mating and invasive growth. *Genes Dev.* **8,** 2974–2985 (1994).

438. Lee, B. N. & Elion, E. A. The MAPKKK Ste11 regulates vegetative growth through a kinase cascade of shared signaling components. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 12679–12684 (1999).

439. Elion, E. A., Grisafi, P. L. & Fink, G. R. *FUS3* encodes a cdc2+/CDC28-related kinase required for the transition from mitosis into conjugation. *Cell* **60,** 649–664 (1990).

440. Elion, E. a, Brill, J. a & Fink, G. R. *FUS3* represses *CLN1* and *CLN2* and in concert with *KSS1* promotes signal transduction. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 9392–6 (1991).

441. Madhani, H., Styles, C. & Fink, G. MAP kinases with distinct inhibitory functions impart signaling specificity during yeast differentiation. *Cell* **91,** 673–684 (1997).

442. Farley, F. W., Satterberg, B., Goldsmith, E. J. & Elion, E. A. Relative dependence of different outputs of the *Saccharomyces cerevisiae* pheromone response pathway on the MAP kinase Fus3p. *Genetics* **151,** 1425–1444 (1999).

443. Ohta, T. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 16134–16137 (2002).

444. Luk, E., Carroll, M., Baker, M. & Culotta, V. C. Manganese activation of superoxide dismutase 2 in *Saccharomyces cerevisiae* requires *MTM1*, a member of the mitochondrial carrier family. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 10353–10357 (2003).

445. Whittaker, M. M., Penmatsa, A. & Whittaker, J. W. The Mtm1p carrier and pyridoxal 5'-phosphate cofactor trafficking in yeast mitochondria. *Arch. Biochem. Biophys.* **568,** 64–70 (2015).

446. Jo, W. J. *et al.* Novel insights into iron metabolism by integrating deletome and transcriptome analysis in an iron deficiency model of the yeast *Saccharomyces cerevisiae*. *BMC Genomics* **10,** 130 (2009).

447. Carlson, M., Osmond, B. C., Neigeborn, L. & Botstein, D. A suppressor of *SNF1* mutations causes constitutive high-level invertase synthesis in yeast. *Genetics* **107,** 19–32 (1984).

448. Balciunas, D. & Ronne, H. Three subunits of the RNA polymerase II mediator

complex are involved in glucose repression. *Nucleic Acids Res.* **23,** 4426–4433 (1995).

449. Kornberg, R. D. Mediator and the mechanism of transcriptional activation. *Trends in Biochemical Sciences* **30,** 235–239 (2005).

450. Larschan, E. & Winston, F. The *Saccharomyces cerevisiae* Srb8-Srb11 complex functions with the SAGA complex during Gal4-activated transcription. *Mol. Cell. Biol.* **25,** 114–23 (2005).

451. Song, W., Treich, I., Qian, N., Kuchin, S. & Carlson, M. SSN genes that affect transcriptional repression in *Saccharomyces cerevisiae* encode *SIN4*, *ROX3*, and *SRB* proteins associated with RNA polymerase II. *Mol. Cell. Biol.* **16,** 115–20 (1996).

452. Tobias, J. W. & Varshavsky, A. Cloning and functional analysis of the ubiquitin-specific protease gene *UBP1* of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **266,** 12021–12028 (1991).

453. Amerik, A. Y., Li, S. J. & Hochstrasser, M. Analysis of the deubiquitinating enzymes of the yeast *Saccharomyces cerevisiae*. *Biol. Chem.* **381,** 981–92 (2000).

454. Weider, M., Machnik, A., Klebl, F. & Sauer, N. Vhr1p, a new transcription factor from budding yeast, regulates biotin-dependent expression of *VHT1* and *BIO5*. *J. Biol. Chem.* **281,** 13513–13524 (2006).

455. Leem, S.-H. *et al.* The Possible Mechanism of Action of Ciclopirox Olamine in the Yeast *Saccharomyces cerevisiae*. *Mol. Cells* **15,** 55–61 (2003).

456. Mayr, E. in *Evolution as a process* (eds. Huxley, J., Hardy, A. & Ford, E.) 157–180 (Allen and Unwin, 1954).

457. Dephoure, N., Gould, K. L., Gygi, S. P. & Kellogg, D. R. Mapping and analysis of phosphorylation sites: a quick guide for cell biologists. *Mol. Biol. Cell* **24,** 535–42 (2013).

458. Wiser, M. J. & Lenski, R. E. A comparison of methods to measure fitness in *Escherichia coli*. *PLoS One* **10,** (2015).

459. Neyman, J. & Pearson, E. S. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **231,** 289–337 (1933).

Annexes

I. Modeling mutant allele selection using a Markov chain Monte Carlo method

In this thesis, I use the mean allele frequence change between sampled evolutionary time points as a synthetic measure of the strength of selection. I found that this measure, noted M, constitutes the best estimate of the true strength of selection given the available data. This choice is valid under the assumption that genome shuffling was performed on a population with a large effective size (N_e). I came to this conclusion after a lengthy thought process, which I expose below. The motivation for this work came from the realization that genuine selection could theoretically be distinguished from genetic drift using a probabilistic approach.

The starting point is that of a mutant allele ***a***, the frequency ***p*** of which is known at two time points, t_1 and t_2 between which a full genome shuffling cycle occurs. The genome shuffling cycle is modeled as a two-step process. The first step is selection, during which alleles under selection will witness a change in frequency proportional to selection pressure. This leads to a change from the initial frequency p_1 to an intermediate, *a priori* unknown allele frequency noted p_{1sel}. Actual shuffling follows selection, and corresponds to post-selection propagation of mutants, sporulation and mating. For simplicity, this is modeled as a number of generations (estimated to 35) of pure drift, resulting in the allele frequency at t_2, noted p_2. The resulting allele frequency change is $\Delta p = p_2/p_1$ which proceeds through intermediate changes $\Delta p_{sel} = p_{sel}/p_1$ and $\Delta p_2 = p_2/p_{sel}$. For simplicity of notation, $\Delta p_{sel} = M$. This simple model can be formalized as follows:

$$p_1 \xrightarrow{selection\ (M)} p_{1sel} \xrightarrow{shuffling\ (\Delta p_2)} p_2$$

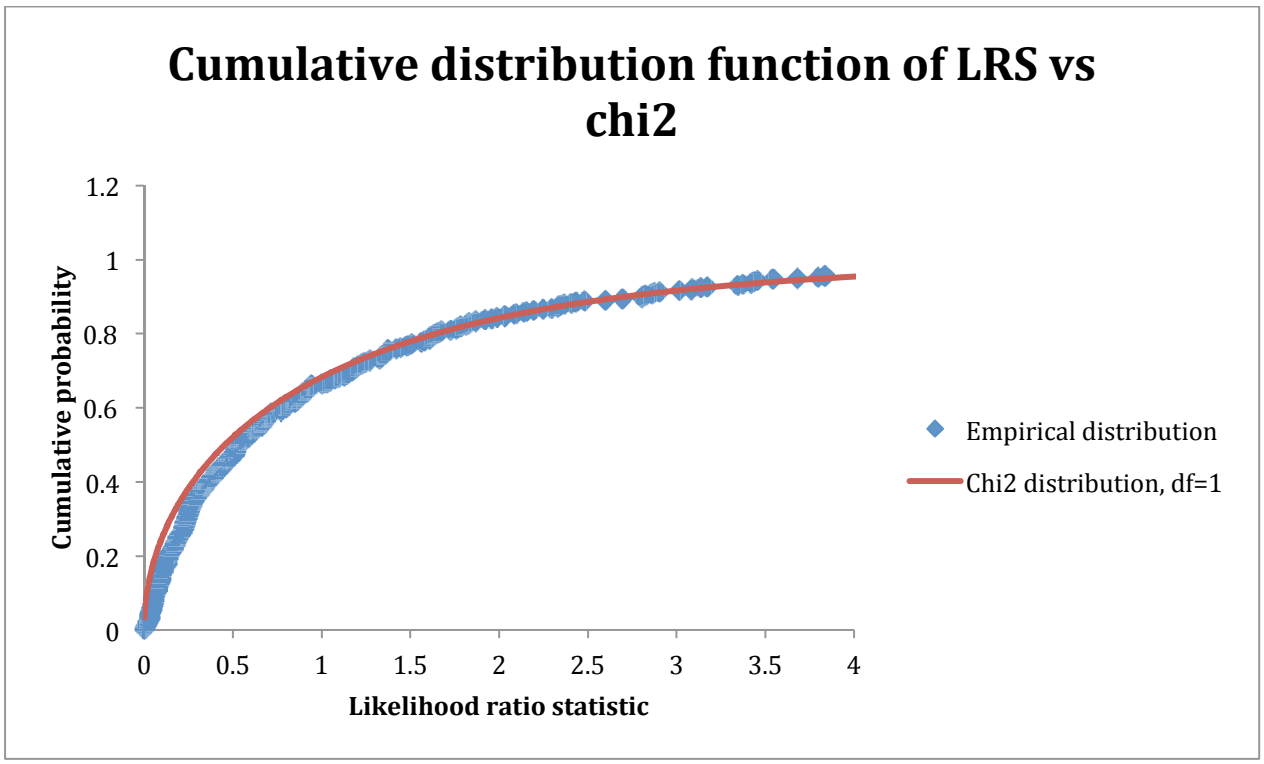
The question is to tell whether p_2 can be reached from p_1 purely by chance from the sole action of genetic drift, or if the change is better explained by actual selection. By definition, $\Delta p_2 = 1$ on average in all circumstances. Under the null hypothesis (H_0) of pure drift, $M = \Delta p = 1$ as well on average. If selection is acting, then the alternate hypothesis (H_1) is that $M = \Delta p \neq 1$ on average. Here, one may realize that H_0 is a special case of H_1, in which selection is fixed at a neutral value. In other words, H_0 is nested within H_1, with one parameter (M) assigned a fixed value. In this context, the Neyman-Pearson lemma states that the likelihood ratio test is the most powerful test at all significance levels for this problem [459]. The test statistic is the likelihood ratio (LRS), defined as follows:

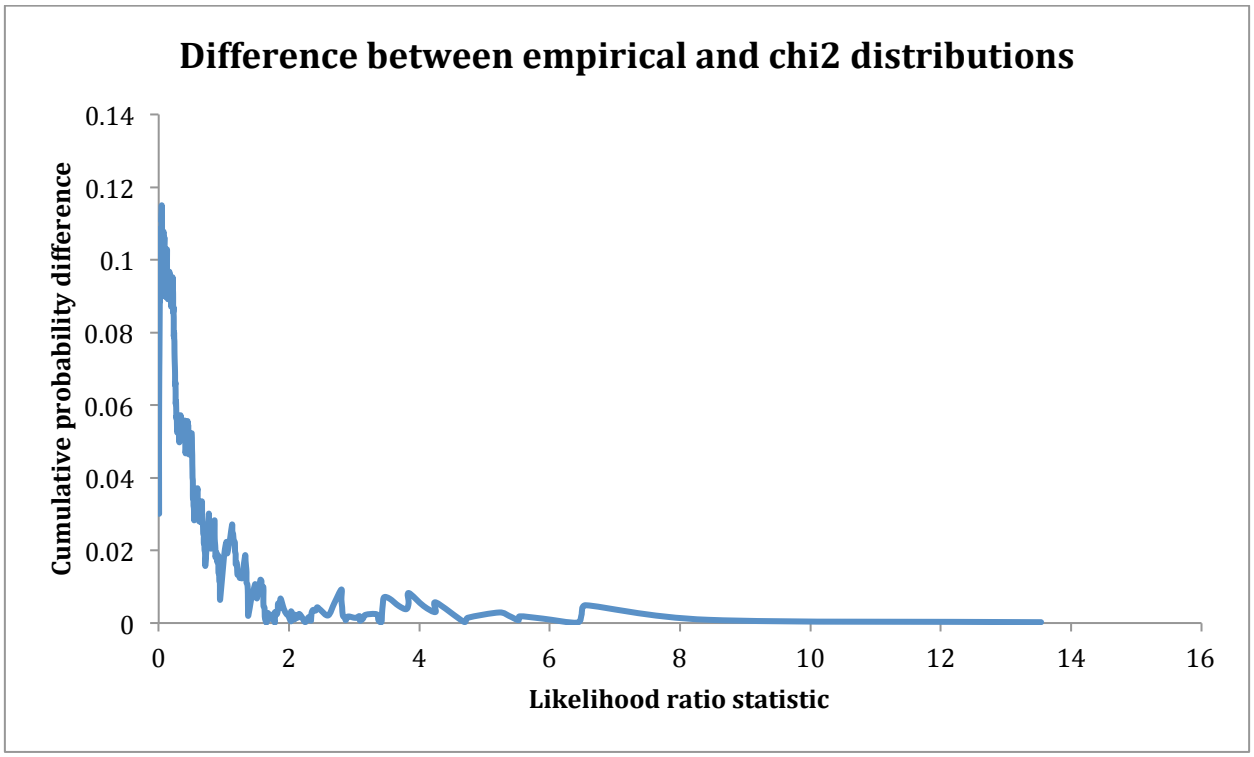
$$LRS = -2 \ln \frac{L(p_2 | p_1, N_e, M = 1)}{L(p_2 | p_1, N_e, M \neq 1)}$$

The problem therefore becomes to determine the likelihood (L) of the data under both hypotheses. The likelihood of the data under both hypotheses depends on an additional and critical parameter, the effective population size (N_e). The determination of N_e is non-trivial in our context, but I suggest means of inferring it further below. For the time being, let us assume that this information is known.

Determination of the likelihood of p_2 given the initial frequency p_1, effective population size N_e and selection M requires the knowledge of the distribution of p_2. No known parametric distribution function *a priori* describes the distribution of p_2. It must be determined empirically. A Markov chain Monte Carlo (MCMC) method can be applied to simulate the genome shuffling cycle a large number of times (1000, 10 000 times or

more) to get a distribution of p_2 values. Simulated values of p_{1sel} are obtained by drawing randomly from a binomial distribution of shape $B(N_e, Mp_1)$. From this simulated p_{1sel} value, a frequency after a generation of drift is simulated by similarly drawing from binomial distribution, this time of shape $B(N_e, p_{sel})$. This operation is repeated on this new simulated frequency value for the number of generations of drift estimated for one cycle of shuffling (i.e. 35) to yield a simulated value of p_2. With a large number of simulated p_2 values, an empirical distribution is obtained by kernel density estimation (KDE). Probability density of the real p_2 value is extracted from the empirical distribution, yielding its likelihood given the specified parameters. This operation is performed for $M=1$ (likelihood of data under null hypothesis) and $M=\Delta p$ (likelihood of alternate hypothesis). Wilk's theorem states that the likelihood ratio statistic follows a chi-squared distribution with 1 degree of freedom, enabling extraction of a p-value. However, using simulated data, I found that the true LRS distribution deviates significantly from the chi-squared distribution, indicating violation of basic assumptions. This is especially true at values of LRS close to 0. The graph below shows a comparison of the cumulative distribution for the chi-squared and empirical LRS distributions, with an inset showing the deviation.



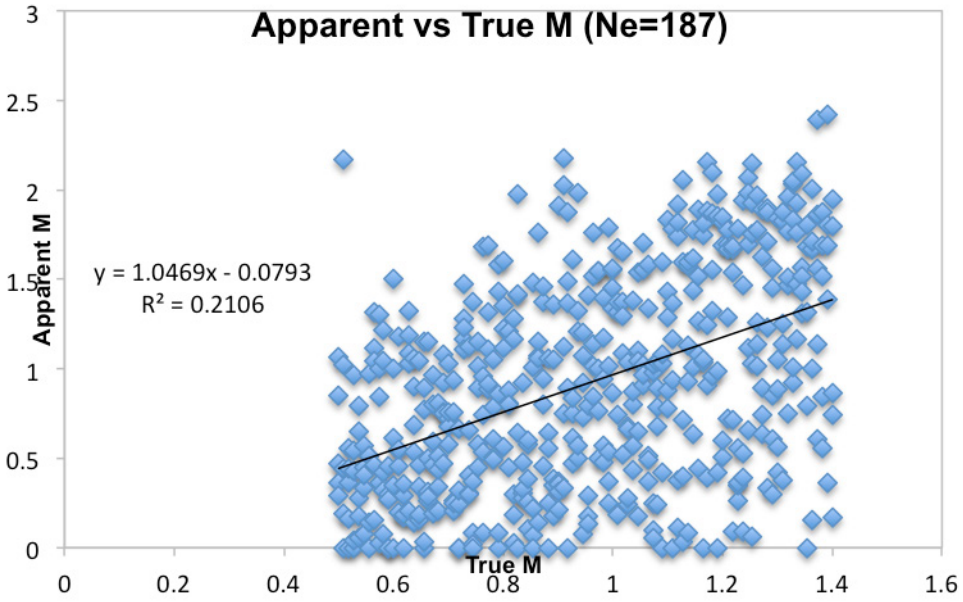


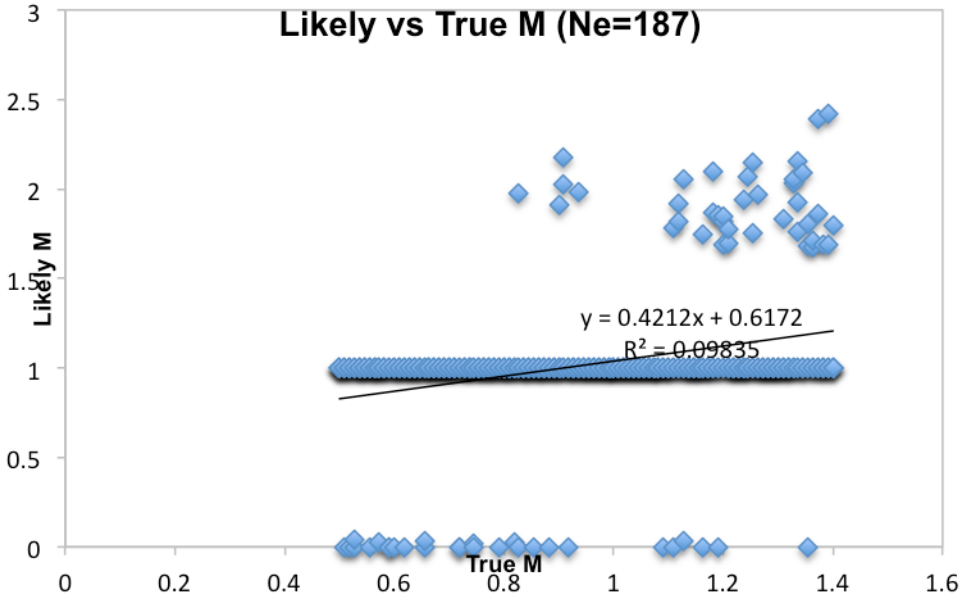
This deviation means that the chi2 distribution overestimates the probability of low LRS values, and may therefore not reject the hypothesis of drift in instances where it actually

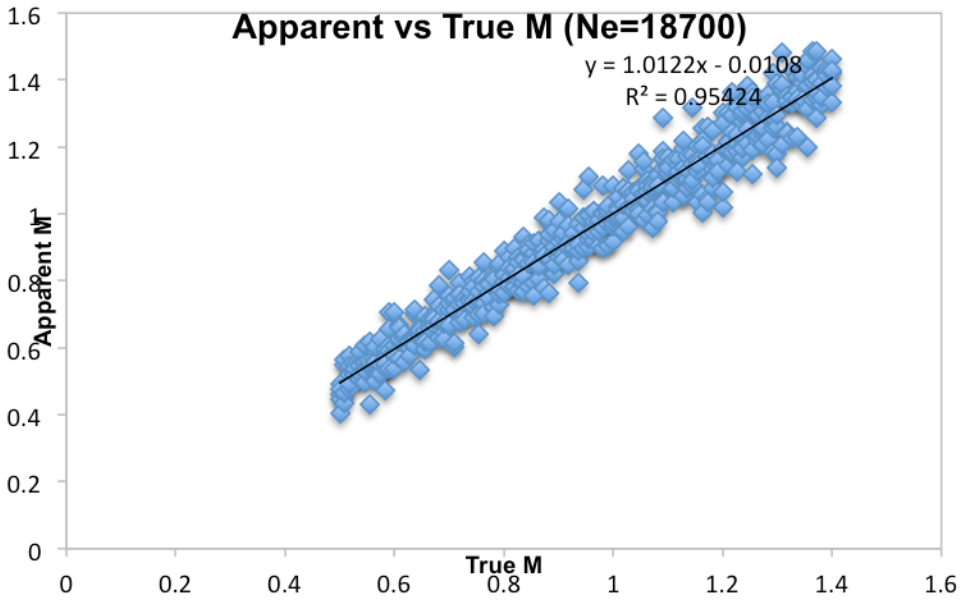
should. This is problematic because it means that an empirical distribution ideally has to be computed each time an LRS has to be estimated. Considering that the distribution of p_2 under both the hypotheses of drift and selection also has to be computed each time, this renders the whole strategy impractical.

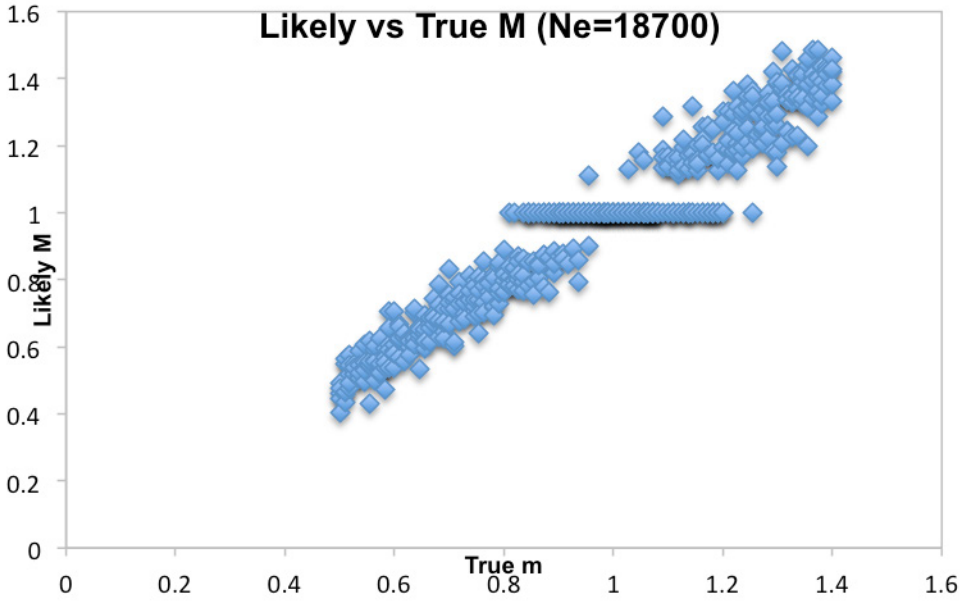
Personal communications with Pr. Lea Popovic at the Concordia Department of Mathematics about the use of approximations to speed computations has reached mitigated conclusions. The use of the chi2 distribution to compute the probability of the LRS is one such approximation. While a Kolmogorov-Smirnoff test estimates that empirical LRS distributions diverge significantly from the chi2 distribution, both show very similar trends, Pr. Popovic does not find the deviation dramatic, mirroring the opinion of Pinheiro and Bates (2000) in similar simulations with small numbers of degrees of freedom. Another approximation that could speed computations is the diffusion approximation, which can be applied to estimate the distribution of allele frequencies over long periods of drift. However, Pr. Popovic warned that this approximation is only valid over long time periods in large populations, a condition that is not met in our experiment. To ensure timely completion of computations, I resorted to using the chi2 distribution to approximate the LRS distribution, while remaining aware of the deviation it induces. Yet, I decided to continue to estimate the p_2 distribution empirically.

I generated simulated data over a range of M values for small (187) and large (18700) values of Ne, and asked the strategy outlined above to predict the strength of selection from the p_1 and p_2 values, and compared the result with a simple measurement of apparent selection using allele frequency change. The result is summarized in the graphs below.


Apparent vs True M (Ne=187)
y = 1.0469x - 0.0793
R² = 0.2106
Apparent M
True M
0
0.5
1
1.5
2
2.5
3
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6


Likely vs True M (Ne=187)
y = 0.4212x + 0.6172
R² = 0.09835
Likely M
True M
0
0.5
1
1.5
2
2.5
3
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6





These simulation results show that capacity to accurately predict the strength of selection improves as effective population size increases. Values of ***M*** close to 0 are considered drift by the algorithm, but over a narrower range at large N_e. Also note that, as N_e increases, the correlation between apparent and true ***M*** tightens, as expected. In

general the algorithm does not seem to be a better predictor than allele frequency change, and both methods are expected to converge on the same result at infinite N_e. For this reasons, the computationally intensive strategy outlined above was not retained. Allele frequency change, or apparent selection, was used to assess strength of selection. The validity of this methodological choice increases with population size. In microbial populations this condition is easily met, and it is not unreasonable to assume large N_e in this context. Means of estimating effective population size are nevertheless desirable. The next paragraph outlines a method to do so from allele frequency data.

Neutral polymorphisms are not affected by selection: their selection is a direct function of drift. Provided such markers are present in the data, effective population size can be estimated. Founder mutations (Figure 3.2) were initially believed to be neutral markers. Indeed, they were present in the evolving population before selection was exerted, and fluctuated on average close to their starting frequency of 0.5. Realization that they could be related to prominently selected features of the evolution changed this judgement, but they were nevertheless used to test the method. Ideally, one would purposefully introduce several such neutral markers in the starting population to allow N_e estimation. To this end, heterology blocks of silent mutations can easily be introduced by CRISPR.

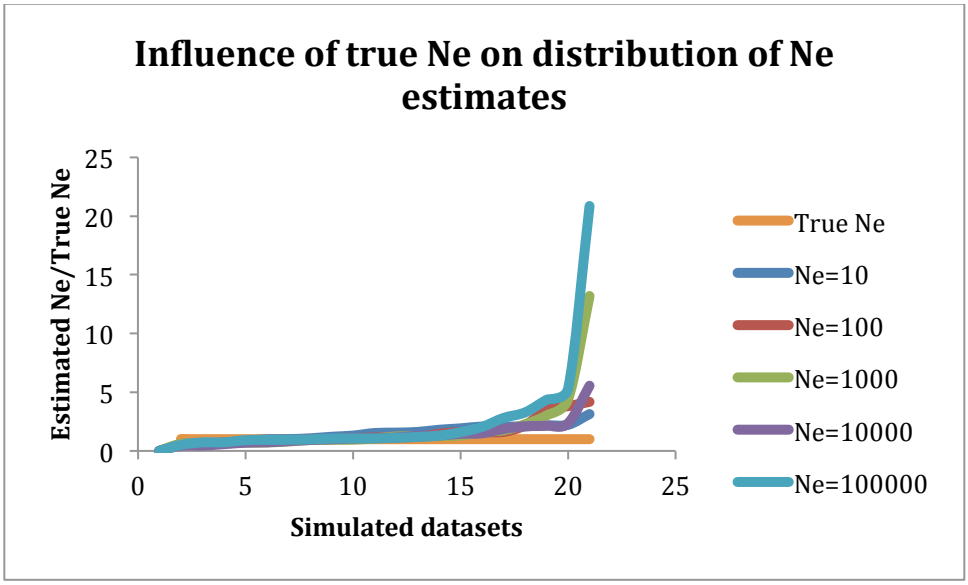
For the sake of discussion, let us assume five neutral markers (a somewhat low number) noted in vector $X= [a, b, c, d, e]$. Their initial frequencies are noted in vector $P_1=[p_{1a}, p_{1b}, p_{1c}, p_{1d}, p_{1e}]$ and their final frequencies after a round of shuffling are noted in vector $P_2=[p_{2a}, p_{2b}, p_{2c}, p_{2d}, p_{2e}]$. The likelihood of P_2 given P_1, $M=1$ (i.e. drift, since I am assuming neutral markers) and any value of N_e is computed as the product of the

likelihoods of each of the markers. A likelihood function of effective population size ***f(N$_e$)*** thus has the form:

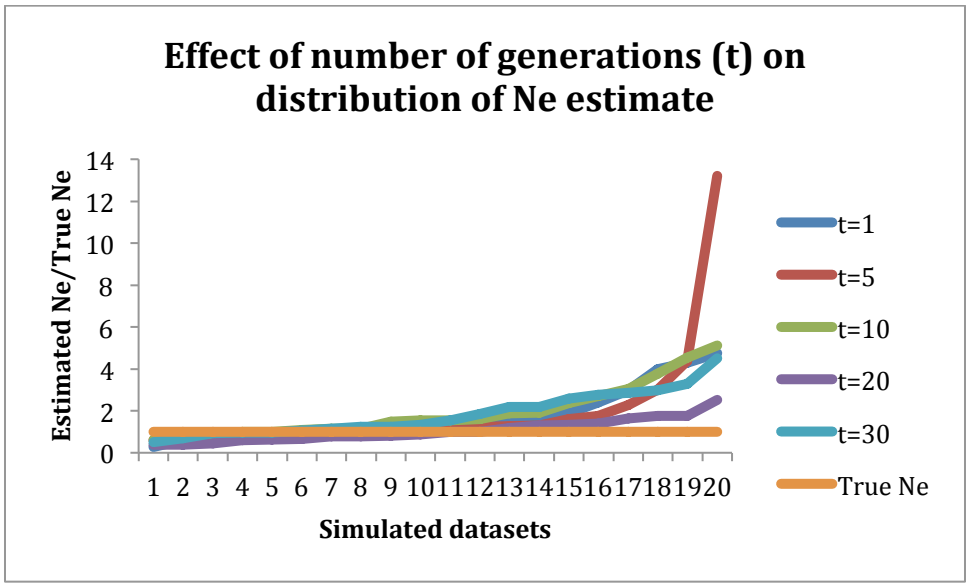
$$f(N_e) = L(P_2|P_1, M = 1, N_e) = \Pi\, L(p_{2x}|p_{1x}, M = 1, N_e)$$

The individual likelihoods are computed from empirical distributions of $\boldsymbol{p_2}$, generated by the MCMC method described above. Finding the effective population size that maximizes the likelihood function is a trivial optimization problem. The only obstacle is that the optimization search requires a large number of empirical $\boldsymbol{p_2}$ distribution evaluations, which is computationally intensive. Feeding a realistic starting point for minimization algorithms can help speed the process. I did so by sampling the function at Ne=10^1-10^9 and feeding the algorithm the order of magnitude that yielded the highest result. Examining of the shape of the likelihood function for some example data, it appears monomodal and smooth, facilitating maximization considerably.

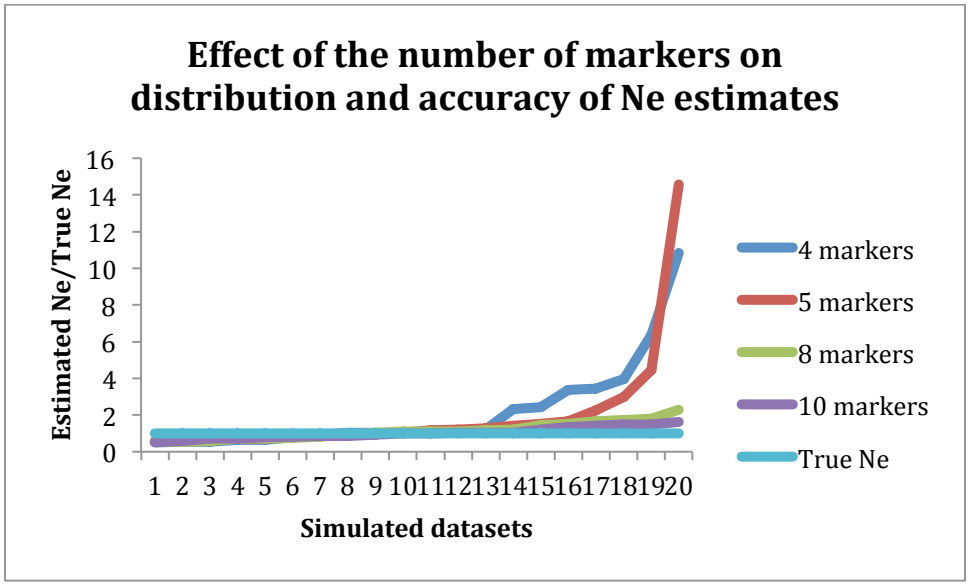
Below, I report estimation of effective population size from simulated data with known $\boldsymbol{N_e}$. I show the effect of population size, the number of neutral markers and the number of generations of drift.



Extreme overestimation of N_e are increasingly observed as true Ne increases, but in all cases the majority of estimates are close to the real value.



The number of generations of drift does not seem to have an effect on population size estimation.



The number of markers seems to have the strongest effect on Ne estimation. Experimenters should introduce as many neutral markers in their starting populations as practically feasible.

II. Primers used for generation of ion torrent sequencing libraries from R57 backcrossed isolates

III. List of generated strains

IV. Mutations uncovered by population genome sequencing

V. Kolmogorov-Smirnoff test for normality on the distribution of SSL tolerance scores among R57 backcrossed isolates

VI. Full genotyping results for R57 backcrossed isolates

Annexes II to VI are provided as separate files to lighten this text. All files available at:

http://tinyurl.com/zpcoq86