2013-06-24

Experimental Illumination of Comprehensive Fitness Landscapes: A Dissertation

Ryan T. Hietpas
University of Massachusetts Medical School

EXPERIMENTAL ILLUMINATION OF COMPREHENSIVE FITNESS LANDSCAPES

A Dissertation Presented

By

RYAN THOMAS HIETPAS

Submitted to the Faculty of the

University of Massachusetts Graduate School of Biomedical Sciences, Worcester

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

June, 24 2013

BIOCHEMISTRY AND MOLECULAR PHARMACOLOGY

EXPERIMENTAL ILLUMINATION OF COMPREHENSIVE FITNESS LANDSCAPES

A Dissertation Presented
By

RYAN THOMAS HIETPAS

The signatures of the Dissertation Defense Committee signifies completion and approval as to style and content of the Dissertation

Daniel N. A. Bolon, Ph.D., Thesis Advisor

Nick Rhind, Ph.D., Member of Committee

Paul Kaufman, Ph.D., Member of Committee

Jeffrey D. Jensen, Ph.D., Member of Committee

Daniel M. Weinreich, Ph.D., Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the Dissertation Committee

C. Robert Matthews, Ph.D., Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the school.

Anthony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical Sciences

Biochemistry and Molecular Pharmacology
(June 24, 2013)

Dedication

To my wife, best friend, and confidant, Krista Bee

"I find it elevating and exhilarating to discover that we live in a universe which permits the evolution of molecular machines as intricate and subtle as we."

-Dr. Carl Sagan

Acknowledgement

My time at UMass Medical School has been a period of huge personal and intellectual growth, and there are many people I would like to thank for the opportunities I have enjoyed. First and foremost, I would like to thank my thesis advisor Dr. Daniel N. A. Bolon whose guidance and mentorship have been invaluable to both my experimental abilities and thought process. I would also like to thank the members of my Thesis Research Advisory Committee and Thesis Defense Committees, Dr. C. Robert Matthews, Dr. Jeffrey D. Jensen, Dr. Nick Rhind, Dr. Paul Kaufman, and Dr. Daniel M. Weinreich for their constructive criticism and support throughout my training.

I am also grateful to the members of the Bolon lab, Benjamin P. Roscoe, Li Jiang, Dr. Parul Mishra, and Pamela Cote as well as past members including Dr. Natalie Wayne-Pursell, Lester Pullen, and Romen Kocibelli. Their personalities, scientific expertise, and support have been essential to cope with the ups and downs of research science. A special thanks to Dr. Jeffrey D. Jensen and his laboratory members, especially Dr. Claudia Bank, for providing constructive conversation, collaboration, rigorous mathematical insight, and essential knowledge in the field of population genetics in which I am far from an expert. Thanks to the Dr. C. Robert Matthews lab including Robert (Paul) Nobrega and Yvonne Chan, as well as Dr. Konstantin Zeldovich and Dr. Kelly Thayer for their numerous contributions in various aspects of the work presented here.

A thank you to my informal mentors who have provided both tangible guidance and intangible support, Dr. William Kobertz, Dr. Kendall Knight, and Dr. Anthony Carruthers. I am also fortunate to be surrounded by a department which is deeply invested in its students and technology. I am especially grateful for the constant help supplied by Luca Leone, Irene Couture, Karen Welch, and Betty Ann Hoyt. You have all become my surrogate family during my time at UMass.

Although this work is dedicated to my wife Krista B. Hietpas, her patience, emotional support, and tolerance of my character flaws too numerous to mention here, have kept me grounded and sane throughout the entire graduate training process, and should be recognized again. Even though I have no words to describe my gratitude, I am and will always be indebted to my parents, Thomas and Sally Hietpas, who have stood behind me in all my endeavors and shaped me into the person I am today (the good parts, at least). To my brother Ross Hietpas for making sure I never forget where I came from, and his soon-to-be wife Jenna Roberts for keeping him in line while I am not around, thank you.

Abstract

Evolution is the single cohesive logical framework in which all biological processes may exist simultaneously. Incremental changes in phenotype over imperceptibly large timescales have given rise to the enormous diversity of life we witness on earth both presently and through the natural record. The basic unit of evolution is mutation, and by perturbing biological processes, mutations may alter the fitness of an individual. However, the fitness effect of a mutation is difficult to infer from historical record, and complex to obtain experimentally in an efficient and accurate manner.

We have recently developed a high throughput method to iteratively mutagenize regions of essential genes in yeast and subsequently analyze individual mutant fitness termed Exceedingly Methodical and Parallel Investigation of Randomized Individual Codons (EMPIRIC). Utilizing this technique as exemplified in Chapters II and III, it is possible to determine the fitness effects of all possible point mutations in parallel through growth competition followed by a high throughput sequencing readout. We have employed this technique to determine the distribution of fitness effects in a nine amino acid region of the Hsp90 gene of *S. cerevisiae* under elevated temperature, and found the bimodal distribution of fitness effects to be remarkably consistent with near-neutral theory. Comparing the measured fitness effects of mutants to the natural record, phylogenetic alignments appear to be a poor predictor of experimental fitness.

In Chapter IV, to further interrogate the properties of this region, library competition under conditions of elevated temperature and salinity were performed to study the potential of protein adaptation. Strikingly, whereas both optimal and elevated temperatures produced no statistically significant beneficial mutations, under conditions of elevated salinity, adaptive mutations appear with fitness advantages up to 8% greater than wild type. Of particular interest, mutations conferring fitness benefits under conditions of elevated salinity almost always experience a fitness defect in other experimental conditions, indicating these mutations are environmentally specialized. Applying the experimental fitness measurements to long standing theoretical predictions of adaptation, our results are remarkably consistent with Fisher's Geometric Model of protein evolution.

Epistasis between mutations can have profound effects on evolutionary trajectories. Although the importance of epistasis has been realized since the early 1900s, the interdependence of mutations is difficult to study *in vivo* due to the stochastic and constant nature of background mutations. In Chapter V, utilizing the EMPIRIC methodology allows us to study the distribution of fitness effects in the context of mutant genetic backgrounds with minimal influence from unintended background mutations. By analyzing intragenic epistatic interactions, we uncovered a complex interplay between solvent shielded structural residues and solvent exposed hydrophobic surface in the amino acid 582-590 region of Hsp90. Additionally, negative epistasis appears to be negatively correlated with mutational promiscuity while additive interactions are

positively correlated, indicating potential avenues for proteins to navigate fitness 'valleys'.

In summary, the work presented in this dissertation is focused on applying experimental context to the theory-rich field of evolutionary biology. The development and implementation of a novel methodology for the rapid and accurate assessment of organismal fitness has allowed us to address some of the most basic processes of evolution including adaptation and protein expression level. Through the work presented here and by investigators across the world, the application of experimental data to evolutionary theory has the potential to improve drug design and human health in general, as well as allow for predictive medicine in the coming era of personalized medicine.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

aa - Amino Acid

ADHD - Attention Deficit Hyperactivity Disorder

ADP - Adenosine Diphosphate

ATP - Adenosine Triphosphate

CD - Circular Dichroism

CFP - Cyan Fluorescent Protein

CFU - Colony Forming Units

CGV - Cryptic Genetic Variation

CI - Confidence Interval

C-terminal - Carboxy-Terminal

ddNTP - Dideoxynucleotide Triphosphate

DE - Directed Evolution

DFE - Distribution of Fitness Effects

DHFR - Dihydrofolate Reductase

DNA - Deoxyribonucleic Acid

EMPIRIC - Exceedingly Methodical and Parallel Investigation of Randomized Individual Codons

FACS - Fluorescence-Activated Cell Sorting

FGM - Fisher's Geometric Model

GFP - Green Fluorescent Protein

HIV - Human Immunodeficiency Virus

Hsp90 - Heat Shock Protein 90 kilodalton

MCMC - Markov Chain Monte Carlo

N-terminal - Amino-Terminal

NTR - Non-Translated Region

OD - Optical Density

PCR - Polymerase Chain Reaction

rmsd - Root-Mean-Square Deviation

SELEX - Systematic Evolution of Ligands by Exponential Enrichment

SERF - Self-Encoded Removable Fragment

SGA - Synthetic Genetic Analysis

TPR - Tetratricopeptide Repeat

UV - Ultraviolet

wt - Wild Type

YFP - Yellow Fluorescent Protein

Preface

Chapter I has not been previously published and is not under consideration for publication at this time. I composed this section based on the current primary literature and my understanding of the subjects presented. Li Jiang, Thomas L. Hietpas, Dr. Parul Mishra, and Robert (Paul) Nobrega graciously edited portions of this chapter.

Chapter II has previously been published as:

Hietpas R*, Roscoe B*, Jiang L, Bolon DNA. "Fitness analyses of all possible point mutations for regions of genes in yeast." Nat Protoc. 2012 Jun 21; 7(7): 1382-96.

Chapter III has previously been published as:

Hietpas RT, Jensen JD, Bolon DNA. "Experimental illumination of a fitness landscape." Proc Natl Acad Sci U S A. 2011 May 10; 108(19):7896-901.

A commentary on Chapter III has been published in *The Proceedings of the National Academy of Sciences* as:

Moses AM and Davidson AR. "In vitro evolution goes deep." Proc Natl Acad Sci U S A. 2011 May 17; 108(20):8071-2.

Chapter IV has been accepted for publication at *Evolution* as:

Hietpas RT*, Bank C*, Jensen JD[¥], Bolon DNA[¥]. "Shifting fitness landscapes in response to altered environments." In Press.

Chapter V is in preparation for submission as:

Hietpas RT, Bank C, Jensen JD, Bolon DNA. "Experimental characterization of intragenic epistatic effects"

Chapter VI has not been previously published and is not under consideration for publication at this time. Li Jiang and Thomas L. Hietpas graciously edited portions this chapter.

Forward: From Darwin to DNA

Evolution by means of natural selection is the single cohesive logical framework in which all biological processes may exist simultaneously. From the original abiogenic event giving rise to life as we understand it, to the most complex and highly regulated signaling pathway, selection for the most fit species (chemical or biological) is central. It seems peculiar then that the outline and rationale of such a basic concept underlying life itself would have relatively little concrete mechanistic detail. Although the earliest traces of evolutionary thought can be attributed to the Greeks[1], Romans[2], and Chinese[3]; it was not until the 19th century that Charles Darwin, Alfred Wallace, and Thomas Henry Huxley fundamentally shifted observations away from supernatural intervention and towards the thesis of gradual improvement of species over immense time scales. The immeasurably important work "The Origin of Species", synthesized from Charles Darwin's time in the Galapagos[4], and the letters of Alfred Wallace, generated during and after his observations in the Amazon River Basin and Malay Archipelago[5], formed the genesis of what we know today as evolutionary biology.

The earliest delineations of the process of evolution posit that organisms subject to the selective pressures of life (scarceness of resources, predation, abiotic stress, etc…) will change in incremental steps over numerous generations to adapt to these selective pressures[4, 5]. The result of natural selection, therefore, is the fittest individuals experiencing greater reproductive success and survivability, while less fit individuals are

less reproductively successful leading to eventual extinction. This method of describing the mutability of species is not only groundbreaking by its own accord, but astonishing considering the heritability of traits through the central dogma of molecular biology was not known until nearly a century later.

From the publication of "The Origin of Species" until the turn of the 20th century, little progress was made in the field as Darwin's theory was still controversial and the debate remained highly contentious. However, in 1900 Hugo de Vries, Carl Correns, and Erich von Tschermak rediscovered the work of the Augustinian friar and botanist Gregor Mendel regarding the inheritance of phenotypic traits in the pea plant (*Pisum sativum*)[6-8] . Mendel's work, originally performed between 1856 and 1863, addressed the inheritance of several phenotypic traits in the pea plant, and showed that inheritance was a reproducibly discrete phenomenon with statistically predictive powers instead of the widely held belief that inheritance was an average of parental traits (blending inheritance)[9]. American geneticist William E. Castle was the earliest biologist to recognize the importance of synthesizing Mendelian genetics with Darwinian selection, and among other achievements, Castle was able to show selection in mice could be based on small variations[10]. Combining Mendel's work within the framework of natural selection, the field of evolution began to move towards a more encompassing theory of inheritance within populations, or what is now known as the field of population genetics.

The early 1900's gave rise to some of the most influential evolutionary biologists of the 20th century including R. A. Fisher and Sewall Wright. The goal of these investigators was to combine the ideas presented by both Darwin and Mendel into a quantitative scaffold in order to make statistical inferences into the inherited characteristics of populations. The first addition to the modern synthesis movement was by R.A. Fisher in 1918 titled 'The correlation between relatives on the supposition of Mendelian inheritance'[11]. The most striking conclusion of Fisher's work was that his model could explain the continuous variation (more simply, the probability distribution) of traits in a population by invoking Mendelian inheritance of these traits as discrete units (which we now know as genes). During this time, Sewall Wright (a student of William E. Castle) was attempting to apply the same logic to a different set of assumptions based on natural observation.

In 1930 and 1932 Fisher and Sewall respectively published their culminating works merging the ideas of Mendelian inheritance and Darwinian natural selection. In Fisher's compendium, 'The Genetical Theory of Natural Selection', he was able to illustrate that the ideas of both Darwin and Mendel were not only compatible, but occurred at a predictable tempo[12]. Perhaps most relevant to the work described in this dissertation is Fisher's explanation of adaptive evolution whereby a population with a given (high) fitness and distance from an optimal fitness can be thought of as a point in multidimensional space (hitherto referred to as Fisher's Geometric Model or FGM). From the current fitness 'point', vectors (mutations) are generated in random directions and

with random magnitudes from the optimum. Vectors which bring the current phenotype closer to the optimum are considered adaptive, whereas vectors increasing distance from the optimum are deleterious. To come to these conclusions, Fisher assumed genetic interactions were additive in nature and species under his model were of relatively high fitness[13].

Meanwhile Sewall Wright had come to a similar but distinct conclusion about the basis of population inheritance and adaptation in what he termed an 'adaptive landscape' in "The roles of mutation, inbreeding, crossbreeding and selection in evolution"[14]. However, Wright worked under a distinctly different set of assumptions including pervasive genetic interaction, as well as low fitness populations attempting to ascend fitness peaks[13]. The adaptive landscape model places populations not in multidimensional space, but on a topographical surface of fitness hills and valleys. Through processes of fitness changes generated by mutation, populations could navigate the local fitness terrain from high fitness hills to lower fitness valleys and, in turn, allow for adaptive mutations to push populations back up to higher fitness[14].

Where Fisher and Wright had made phenomenal leaps in the understanding of genetic behavior within populations at the mathematical level, the next generation of evolutionary biologists including Theodosius Dobzhansky, Edmund Ford, and Ernst Mayr differed from their predecessors by realizing population genetic theory was not always consistent with natural observations[15-17]. Studies during this time period were

broadly focused on defining variation within populations, observing natural selection in nature, and characterizing the emergence of new species through genetic isolation and selection with the goal of placing natural events in the framework of mathematical theories developed by Fisher and Wright. Significant debates raged during this time as experimental data became available. Whereas Fisher's ideas of selection were based on the forces of Darwinian selection, Wright recognized the importance of genetic drift. This difference, in light of the mechanisms of genetic inheritance, set the stage for investigators such as Kimura and Ohta to address the relative importance of drift and selection in their theories regarding the distribution of fitness effects[18, 19].

Chapter I - Introduction

Molecular Aspects of Evolution

Evolutionary Biology in the Molecular Biology Era

The elucidation of the structure of DNA is among the greatest scientific achievements of the 20th century. DNA forms an anti-parallel double stranded polymer of a sugar-phosphate backbone covalently linked by phosphodiester bonds. Extending from each backbone sugar is one of four nitrogenous bases which interact with a base from the opposite strand through hydrogen bonding, as well as sequentially between bases within the same strand through stacking interactions[20, 21]. Interestingly, of the four possible bases of DNA, only two combinations of base pairing exist; Adenine to Thymine and Cytosine to Guanine, indicating each strand in the helix is a cognate of the other. The beauty of this structure is the simplicity, modularity, and functionality which was most certainly not lost on the discoverers who immediately suggested a mechanism for replication[20]. Not only did the solved structure of DNA result in the birth of the field of molecular biology, but gave a new fundamental insight into the heritability of traits.

For evolutionary biologists, the mechanistic explanation of the origin of polymorphism, genetic linkage, variation, and a genes-to-protein mechanism allowed for a new focus in the formation of the fields of molecular genetics and molecular evolution. With the knowledge of the chemical structure of bases, the organization of these bases into genes, and genes into genomes, it was now possible to formulate testable hypotheses

regarding the transmission of traits within and between populations. Although formal experimental evolution predated the discovery of the structure of DNA by nearly a century[22], and non-formal selection (domestication) for desirable traits for thousands of years before that, the chemical basis of inheritance allowed for more focused mechanistic studies of the loci and basis of heritable traits.

In Vivo Experimental Evolution

Experimental evolution attempts to directly address questions in the field of population genetics and bridge the gap between theoretical models and observed phenomena. The first recorded evidence of controlled experimental evolution was in 1880 when William Dallinger, apparently inspired by Darwin, applied gradual thermal stress to a population of protists over the course of approximately seven years[22, 23]. He found that non-adapted organisms would be killed by a temperature of 60°C, whereas organisms which had been slowly pre-adapted were capable of tolerating temperatures as high as 70°C. Although Dallinger's experiments were nearly identical to modern laboratory evolution experiments, the lack of knowledge in heredity and molecular biology made it nearly impossible to draw any mechanistic conclusions as to the thermal adaptation he witnessed.

Long term growth and adaptation experiments conceptualized in light of molecular biology emerged in 1981 when two early pioneers of experimental evolution, Michael Rose and Brian Charlesworth, were among many to understand the importance

of studying multiple parallel experimental lineages over evolutionarily relevant timescales. By selecting for age-specific fecundity[24], time to senescence[25, 26], and stress response[27] in *D. melanogaster,* it became clear that experimental evolution was a powerful tool for understanding the genetic basis for fundamental processes of life[28]. In the same vein as Rose, Richard Lenski began the largest evolution experiment recorded to date. Since 1988, a laboratory strain of *E. coli* has been grown and diluted in a 24 hour cycle under different environmental conditions. At the time of writing, the experimental lineage has surpassed 50,000 generations and allowed for experimental observation of adaptation[29], epistasis between deleterious and beneficial mutations[30], and estimates of the rate of the 'molecular clock'[31]. Most notably, these long term experiments resulted in the generation of a novel adaptive event, allowing *E. coli* to metabolize citrate under oxic conditions.

One physiological characteristic of the *E. coli* genus is the inability to metabolize citrate in aerobic conditions. Upon growth of Lenski's ancestral lineage for ~30,000 generations in low glucose and high citrate medium, a citrate metabolizing phenotype (cit^+) was realized[29]. Genomic analysis of cells with a cit^+ phenotype revealed a requirement for both a duplication of the anoxic *citT* operon, in conjunction with altered regulation characteristics and at least one single nucleotide polymorphism to refine the cit^+ phenotype[32]. These findings were the first experimental analyses to lend evidence to several important evolutionary observations including mechanisms of adaptation,

neofunctionalization of duplicated genes, epistasis, and background potentiation of new phenotypes.

Yeast, including *S. cerevisiae,* has also become the focus of experimental evolution due to their genetic malleability, eukaryotic cellular machinery, and utility in heterologous protein expression. Due to extensive post-translational processing and secretion, yeast have been utilized to produce active or precursor human biologic pharmaceuticals including insulin[33], interferon alpha A[34], interleukin-1 β[35], lysozyme[36], macrophage colony stimulating factor[37], and human serum albumin[38]. Although yeast do not harbor or produce infectious agents or pyrogens which could contaminate human biologics, the ability to optimize and genetically canalize heterologous protein expression is challenging. For this reason, long term experimental evolution of strains expressing a protein of interest have allowed investigators to determine the genetic characteristics influencing protein expression and secretion to aid in protein production as well as how to overcome expression obstacles[33].

Even mammals with relatively long generation times have become the focus of long term experimentation. Garland *et.al* has successfully selected a population of mice for voluntary running behavior for over 50 generations to study the physiological and genetic basis for complex traits. Physiologically, mice which have been selected for increased voluntary running have developed several new musculoskeletal phenotypes, as well as alterations in stress response such as a decrease in Sod-2 and increase in Hsp70

levels[39-42]. Selection for the 'high runner' phenotype has also become a valuable model for human ADHD treatment as the mice selected for running behavior display atypical dopaminergic responses that can be altered by treatment with psychoactive drugs such as Ritalin, Prozac, canabanoids and apormorphine[43, 44]. Currently, these traits are being mapped to specific loci to determine the underlying genetic and epigenetic changes to better understand the link between genotype and phenotype in mammals[45].

In Vitro Selection and Directed Evolution

The molecular biology 'revolution' led to the development of numerous technologies which are now invaluable to the field of experimental evolution. Polymerase chain reaction (PCR)[46, 47], cassette mutagenesis[48, 49], error prone PCR[50, 51], DNA shuffling[52], and Gateway cloning[53], increased the ability to generate mutational diversity of parental libraries to maximize the likelihood of discovering an improved protein variants. Meanwhile systems for linking genotype to phenotype for enhanced selection including phage display[54], ribosome display[55], SELEX[56], cell surface display[57], and high speed FACS analysis improved screening throughput. The advent of highly diverse libraries and means of selecting mutants for properties of interest has a variety of practical applications, as well as offering a more highly controlled system to study molecular evolution.

Mutational studies of proteins are utilized by numerous scientific disciplines to study the biophysical details of residues within a protein. Historically, mutational

analysis has been used to assess the defect introduced by a mutation and characterize residues contributing to stability[58-60], catalytic domains[61-63], binding interfaces[64, 65] and function by phosphomimicry[66, 67]. Mutations are first incorporated into the gene of interest and then introduced into a model organism or purified system to study the resulting protein properties. An organismal assessment of fitness may be desirable because the mutant protein will experience biologically relevant trafficking and modification steps which may not otherwise be apparent by biophysical measurement. However, purified protein systems are desirable when biophysical, biochemical, or structural analyses are the focus of the work. An increased throughput method of screening proteins for essential residues, known as alanine scanning, attempts to isolate positional importance by iteratively replacing residues with alanine, or glycine. Alanine scanning has been successfully applied in the characterization of numerous protein structure-function relationships including human growth hormone binding[68, 69], essential cyclotide structure[70], and sodium ion channel transport[71].

Although mutagenesis approaches have been widely successful for characterizing a variety of biological and biochemical processes, single codon substitution approaches possess some inherent disadvantages. The most obvious disadvantage is the lack of site specific biophysical requirements based on only a single residue substitution. Although alanine is chosen for many mutagenic approaches due to its inert side chain, modest volume, and secondary structural properties, the methyl group of alanine represents only a small fraction of potential amino acid chemical characteristics. For this reason, false

negatives for essential residues are common, especially when replacing chemically similar residues such as valine. Alanine also exhibits strong helical propensity which may bias results based on the local sequence in which it is placed. Ideally, each residue would be replaced with all other possible substitutions to gain insight into requirements such as charge, volume, hydrophobicity, and contribution to secondary structure.

Directed evolution (DE) is a logical extension of the fields of evolutionary biology and protein engineering used to create and select for improved variants over a wide range of mutational changes. By starting with a protein of interest, introducing mutations, and selecting for improved protein variants based on predetermined criteria, investigators have harnessed DE to improve stability[72, 73], catalytic activity[74, 75], and even neofunctionalize enzymes to recognize new substrates[76, 77]. DE is most fundamentally an *in vitro* selection experiment where a library of protein variants is generated and competed, but instead of passively observing changing populations over time, the investigator selects a protein property of interest to improve, and screens mutant libraries by biochemical assays. Improved variants (as defined within the experiment) are then subjected to several 'generations' of diversification and selection until no further improvement is possible or the protein is judged sufficiently improved. By linking a phenotype or protein property to a genetic construct, it is also possible to infer the underlying interactions between residues as well as the basis for broader biophysical properties such as stability and specific activity. However, the protein of interest is not

linked to organism fitness, so drawing broader evolutionary conclusions can prove difficult.

Nucleic Acid Sequencing Technology

Mutations are aptly referred to as the 'raw material' on which selection can act to allow for adaptive evolution, and the ability of a population to adapt is of fundamental biological interest. To understand how adaptation occurs, it is necessary to elucidate the underlying mutations which give rise to fitness changes by analyzing the linear nucleic acid sequence. Chain termination sequencing (also referred to as Sanger sequencing) was the first widely utilized technology to generate whole genome data from phage to human beings and was the state of the art from 1977 until the first accurately sequenced human genome in 2003 (see genome.gov for an in depth timeline). Briefly, Sanger sequencing relies on the stochastic incorporation of dideoxynucleotides (ddNTPs) during *in vitro* DNA dependent DNA synthesis followed by size analysis of terminated fragments. By sequentially analyzing the identity of the incorporated ddNTP at each position, the linear DNA sequence can be solved[78]. Chain termination sequencing has the advantage of being simple and accurate for single reads of less than 1,000 base pairs, and is still extensively utilized in molecular cloning and quality control prior to higher throughput methods of sequencing. The drawback of this technology is the low throughput in comparison to the size of complex genomes as well as the cost per megabase of raw DNA sequence[79]. Even with significant automation and improved processing power, chain termination sequencing is still sub-optimal for large and complex sequence identification.

The second or 'next' generation of sequencing technology has resulted in new methodologies to exploit the unique properties of the DNA molecule. By sequencing small fragments in parallel (usually 20-250 base pair fragments), sequence analysis has now been linked to numerous DNA specific processes such as base incorporation during synthesis (Illumina, Helicos, and SMRT technologies), pyrophosphate or hydrogen ion release upon base incorporation (454 and Ion Torrent technologies), and base pair hybridization/ligation (SOLiD and Nanoball). All of these technologies share a parallelized work flow which allows for enormous numbers sequence reads (from 35,000 for SMRT to approximately 3,000,000,000 for the Illumina Hi-Seq platform) and nearly a five order of magnitude cost reduction over chain termination per megabase of raw sequence[79, 80].

Experimental Fitness Measurement

One of the most fundamental values in evolutionary biology is fitness (W), yet fitness estimation continues to be a challenge for experimentalists. Originally derived by J.B.S. Haldane in 1927 to describe the probability of a phenotype to appear in the next generation, fitness can either be defined as absolute or relative[81]. Absolute fitness is a direct measurement of number of individuals possessing a genotype (N) in a population both before and after selection, or $W_{abs} = N_{after}/N_{before}$. On the other hand, relative fitness (which is the fitness term used in this dissertation) is not a direct measurement of individuals, but the fitness of an individual in a single generation compared to the

populations average fitness. Relative fitness can therefore be an ensemble measurement which mitigates many technical challenges of defining the start and end of selection as well as counting all individuals.

Even in the context of relative fitness, there are many challenges that impede the accurate and precise quantification of fitness. The first major technical hurdle is the accurate measurement of specific genotypes within an experimental population[82]. Many early experiments attempted to measure the frequency of an allele by sampling a population for phenotypic traits such as coat color[83]. By recording phenotypic characteristics over several generations, investigators were able to make an estimate of the distribution of an allele in a population and apply these data back to theoretical models. Yet, considering what has been previously discussed regarding the molecular mechanism of inheritance, many mutations within a single loci do not observably impact phenotype either due to synonymous substitution with no impact on the protein, or mutation within a region of a gene with no functional role in coat color. Therefore, by measuring only the phenotypic distribution of traits, many spontaneously occurring mutations remain unobserved due to lack of sensitivity in the assay or canalization of a trait.

To overcome the technical shortcomings of phenotypic observation, physical measurement of the genotype at a given locus is preferable. However, genotypic analysis has only been possible since the advent of nucleic acid sequencing technology (circa

1970). Since its development, DNA sequencing has progressed from sequencing hundreds of bases a day to current methods resulting in over 1 billion base calls in a single run[84]. It is now common practice to sequence entire genomes to look for polymorphisms between many populations to establish evolutionary lineage[85, 86], link phenotypes to genotypes[87, 88], and estimate the selective advantage or disadvantage of a mutation over evolutionary time[89, 90]. Now, not only are observable traits apparent to scrutiny, but also the underlying molecular mechanism.

The second major challenge to accurately calculating fitness is the potential interference of epistasis through drifting background mutations, especially during long term experimentation. It has been understood since the time of Wright's work that mutational interdependence between the genetic background and a new mutation may affect fitness. Ideally, a stable and isogenic line of organisms would be used to calculate fitness. However, the reality is that any biological system will accumulate random mutations over time due to imperfect replication and repair processes. One way to mitigate the problem of changing genetic background is to choose an organism which is known to have a low per genome mutation rate per replication. For example, *H. sapien* has an estimated per genome mutation rate of 0.49 bases per genome replication whereas *D. melanogaster* is nearly an order of magnitude lower at 0.058/replication or *S. cerevisiae* which is 2 orders of magnitude lower at 0.0027[91].

Switching experimental systems is often not a viable solution to alter the background mutation rate, so other methods are necessary to mitigate epistatic interference. The ability to propagate large populations forward during an experiment can prevent systematic background changes which may occur due to bottlenecks in the population. It is also important to consider that alterations made to the system to decrease background effects may have their own selective advantage or disadvantage, so controls must be designed appropriately to account for these fitness differences.

The final major technical challenge to experimental fitness calculation that will be addressed here is that the fitness effect of mutations can vary with environmental changes. The effects of the environment on the gradual improvement of species was one of the founding principles from Darwin's work in 'The Origin of Species', and environmental influences on adaptation were considered by both Fisher and Wright in their models of adaptation. In the FGM, environmental perturbations take the form of a shifting optimum relative to the current phenotype which changes the probability of a random mutation being beneficial. In Wright's adaptive landscape, environmental changes resulted in the landscape itself shifting, relocating a stationary population to a new local fitness landscapes[92].

Controlling the environmental milieu between experiments is perhaps the single most challenging aspect of accurately and reproducibly measuring fitness. Factors such as metabolite composition and concentration, temperature, pH, and atmosphere are all

potential fitness considerations which may or may not be easy to control. In the case of microbes such as *E. coli* and *S. cerevisiae*, placing populations into the same vessel ensures all members of that population experience the same extracellular environment. The same is true for *D. melaongaster*, where temperature and humidity controlled incubators are used for climate control. The reproducibility of fitness measurements can be high, but likely never perfect because it is not physically possible to control all variables in a biological system. The ability of an investigator to recognize sources of environmental variability and maximize reproducibility is key to successful fitness measurements.

Bacteria and yeast are extraordinary systems to study evolution in a controlled environment because of their unique physiology and genetic malleability. Most importantly, in laboratory conditions, microbes have a very short generation time relative to more complex organisms. Whereas mice have a generation time of 10 weeks[93], wild type *E. coli* double in approximately 30-40 minutes and wild type *S. cerevisiae* in 2-3 hours. This point is experimentally important because numerous generations can be propagated over relatively short timeframes. To counter variation introduced by genetic exchange, both yeast and bacteria can be genetically modified to prevent sexual reproduction and horizontal gene transfer respectively[94]. The fact that these organisms grow exponentially by binary fission allows fitness measurements to be proportional to the rate of division, and as growth rate is the most commonly affected phenotypic trait, growth rate is a broad mutational target[95, 96]. In terms of growth rate, highly fit alleles will

allow for faster growth rates whereas less fit alleles will experience slower growth rates. When populations of differential fitness are parsed over time, the ratio of the abundance of a mutant to a highly fit allele (usually wild type) allows for the calculation of fitness.

Calculation of growth rate for microbial evolution experiments is performed by several methods depending on sensitivity, efficiency, and cost effectiveness. The general theme for determining growth rate is by estimating the proportion of a specific allele over time during log phase growth of a culture (for cellular organisms). This can be accomplished by microscopic visualization (hemocytometry or flow cytometry), colony forming unit (CFU) calculation (plaque forming units in viruses), biomass measurement, but most commonly by spectroscopic analysis of turbidity (usually by absorbance near 600nm). Relative fitness of a monoculture has more recently been performed by measurement of the diameter of a yeast or bacterial colony as a proxy for growth rate[96, 97]. In yeast, colony size resulting from the dissection and growth of tetrads is commonly used for fitness analysis. According to the laws of independent assortment, half of the tetrads will be wild type with the other half harboring the mutant of interest due to segregation during reductive cell division. Colony size determination is also a very sensitive method to detect fitness differences because the mutant and wild type haploids are nearly isogenic[97].

Monoculture growth is not always an optimal method for fitness determination because although individuals may be highly isogenic and the measurements can be

automated, wild type and mutant isolates may not experience identical environmental conditions over the course of genetic manipulation and growth. To overcome this technical challenge, the use of binary competitions has become a method of choice for many applications because the genetic variant of interest and wild type strain are grown in the same vessel to mitigate environmental fitness effects. The challenge of this technique is performing an accurate quantification of the ratio of mutant to wild type cells at different times. The relative proportion of each population can be calculated by expressing variant fluorophores in the mutant and wild type strains and counting the ratio by flow cytometry or ensemble bulk fluorescence measurement (performed in chapter III)[98, 99]. Paired growth fitness measurements have been demonstrated to be exceedingly accurate, and this methodology has been utilized to accurately measured selection coefficient below 10^{-3} [82]. It is important to consider in binary competition experiments that expressing protein variants or incorporating exogenous nucleotide sequences is potentially not selectively neutral, and proper controls are necessary to account for background growth defects introduced by manipulating the system.

More recently, high throughput methods to calculate fitness have been introduced to allow for the screening of hundreds to thousands of genetic and environmental parameters in a single experiment. By automating population management and growth measurement in small scale monoculture, Jarosz *et al.* were able to measure the fitness of 102 genetically distinct wild and lab strains of yeast in 100 different conditions with and without reduction of the endogenous Hsp90 pool (>20,400 individual fitness

measurements) to determine the relationship between genetic and phenotypic variation[100]. The invention of synthetic genetic arrays (SGAs) has also contributed greatly to automated high throughput fitness measurement. SGAs seek to assess the interaction between a query mutation and the yeast gene deletion library by assessing the fitness of many genetic crosses in parallel[101]. By mating the query to the library and sporulating yeast to iteratively select for double mutants, synthetic lethal phenotypes can be detected by colony growth analysis and be used to uncover synthetic lethal genetic interactions. The SGA method has also been employed as a screening technique for a library of query mutations generated by random mutagenesis[102]. SGA methodology can also be applied to other model systems including the fission yeast *S. pombe* and *E. coli*[103, 104].

The Distribution of Fitness Effects (DFE) of New Mutations

In a microbial systems, the growth rate of a mutant compared to a highly fit reference is a quantitative measure of fitness. If a mutant growth rate is identical to the growth rate of wild type, W=1, if the mutant growth rate is higher than wild type, W>1, and if the mutant growth rate is lower than wild type, W<1. Fitness can also be defined in terms of selection coefficient (s) which represents the viability difference between a mutant and wild type and the conversion between W and s is a mathematical manipulation where W=1+s . In this mathematical construct wild type s=0, mutations more fit than wild type are positive, and mutations less fit than wild type are negative with a minimum of -1[105]. The conversion to selection coefficient has simple practical value because: (i) the slope of a line generated by plotting log2(mutant/wild type) over

generation time is by definition the selection coefficient, which makes it experimentally more useful and, (ii) the selection coefficient is a parameter of many population genetics models and (iii) the positive and negative values scheme is intuitive and valuable for communication.

In the mid 1900s, new empirical data became available regarding amino acid substitution rates[105] and the background mutation rate (molecular clock), and these results were not entirely consistent with the standing model of natural selection. In the 1950s and 1960s, many believed that new mutations had strong advantageous or deleterious fitness costs, with very few mutations being selectively neutral due to natural selection for high fitness individuals driving Darwinian evolution. To address the discontinuities between theory and observation, Dr. Motoo Kimura, armed with knowledge of genetic drift from the work of Wright's, proposed a mathematical construct where randomly occurring mutations would be either strictly fitness-neutral and be acted on only by genetic drift, or have a strongly deleterious or advantageous fitness effect and therefore be acted upon by natural selection. Under this logical scaffold, strongly deleterious mutations would be quickly eliminated from the population, and therefore Kimura assumed the primary source of genetic variation was through neutral mutations, and dictated by genetic drift[18]. This new model of molecular evolution was termed neutral theory.

One fundamental problem in the field was that neither neutral theory nor natural selection could fully account for the experimental evidence gathered regarding mutation

rate and substitutions at synonymous sites. In 1971 Dr. Tomoko Ohta, a student of Kimura, included a third category of mutations which were only very slightly deleterious or advantageous to fitness, and could therefore be acted on by both genetic drift and natural selection depending on the effective population size. This explanation of the distribution of fitness effects (DFE), called near-neutral theory, describes the relationship of selection and genetic drift on opposite ends of a continuous spectrum of allele fixation events depending on the effective population size and the fitness effect of a mutation[19]. As more experimental measurements of fitness become available (including Chapter III), mutational effects seem consistent with Ohta's near-neutral theory and remains an excellent example of the more general relationship between theory and data.

Molecular Adaptation

An organism's environment is constantly in flux, and the ability of an organism to adapt to changing environmental conditions and physiochemical insults is essential to fitness. Consequently, mutations which increase survivability and reproducibility of an organism are also propagated at an increased frequency within a population. Therefore adaptation to new environments is heritable through adaptive mutation, and broad theories regarding the distribution of beneficial mutations including frequency and magnitude have been hypothesized since Fisher and Wright, and more currently in the context of extreme value theory[106].

Adaptive mutations may be broadly defined as a mutation which causes increased fitness in a particular environment over wild type (or highly fit reference strain). Beneficial mutations vary in fitness effect and occur in populations of varying size, but the average fitness benefit and frequency of beneficial mutations are still largely unknown. Current methods for inferring selection coefficients from phylogenetic data have indicated strong disagreement of up to four orders of magnitude for the average magnitude of new beneficial mutations[107, 108], and these analyses are not necessarily highly quantitative in terms of frequency due to the assumption of neutral fitness effects of synonymous substitutions.

Conversely, experimental evolution techniques which measure the underlying fitness effects of mutations have been informative as to the proportion of beneficial mutations occurring in a population, but not necessarily the magnitude of their fitness effects. Currently, the most sensitive experimental fitness measurement allows for calculations of $s > 10^{-4}$ whereas current estimations predict selection coefficients of $s > 10^{-7}$ to be acted upon by selection (as opposed to drift)[82]. As previously mentioned, mutant accumulation experiments are the main source of data for the proportion of observed fitness effects, but do not shed light on the entire DFE of new mutations. Only relatively fit mutations that remain in the population for many generations are measured, and this broad focus may be advantageous for studying beneficial mutations even though only a small proportion of all mutagenic events are detected. The main findings from mutant accumulation experiments indicate previous theoretical predictions may be correct in

assuming beneficial mutations are rare due to gradual improvement over evolutionary timescales[109, 110], but quantitative analysis of beneficial mutant frequency remains unknown.

Although beneficial mutations are commonly isolated in studies of antibiotic resistance[111], antiparasitic resistance[112], antiretroviral resistance[113], and chemotherapeutic resistance[114], the results are convoluted by inherent measurement problems. Under strong selection such as that of drug resistance, the wild type genotype is very unfit, therefore the calculation of fitness compared to wild type may yield misleading values as the fitness of wild type approaches zero. The fitness of the wild type allele must also be scrutinized when choosing an analytical framework, because as mentioned previously, Fisher's Geometric model and Wright's adaptive landscape make distinctly different predictions regarding the initial fitness of a population.

Fitness in the Context of Expression Level

A central paradox in mammalian evolutionary biology became evident in the 1970s when several investigators concluded the genetic difference between chimps and humans was likely too small to account for the gross phenotypic differences between the species[115]. Work by King *et al.* demonstrated that not only were changes in protein coding region contributing to species differences, but relative expression levels played a key role in evolutionary processes. More recent analyses of expression differences indicate expression level variation in coat color proteins in mice act to drive adaptive

evolution by changing predation patterns[116]. The study of expression level tuning to generate optimal fitness as well as how expression level acts in conjunction with new mutations has become of central importance to evolutionary biology.

The fitness contribution of a protein is dependant both on biochemical properties which are dictated by the amino acid composition and protein structure, as well as the number of protein molecules expressed in the cell. Previous analyses suggest that protein expression level is optimized for maximal fitness[117-120] and experiments investigating tuning of the Lac operon of *E. coli* have demonstrated that expression level can be quickly optimized to meet cellular metabolic needs[121]. However, these results seem to contradict findings that many essential proteins can be significantly reduced in expression without significant fitness defects[122, 123]. To address the question as to why proteins are expressed at a particular level, the relationship between fitness and expression has been previously characterized into the framework of a cost-benefit relationship[121]. The energy and materials necessary to transcribe and synthesize a nascent polypeptide represents a cost to the cell whereas the fitness improvement garnered through function per molecule represents a benefit to the cell. Additionally, the fitness cost of gene duplication events has been estimated at $s > 10^{-5}$ [124] , and a fitness defect of this magnitude is predicted to be subject to selection as opposed to drift in large microbial populations, making expression level changes a key to understanding molecular evolution[125].

Just as mutations within coding regions can fix in a population either by drift or by selection depending on the population size and fitness effect, changes in protein expression can be fixed by either process. Mechanistically, this is because protein expression is modulated by mutations in promoter regions[126], NTRs[127], transcriptional regulators[128], gene duplication or deletion events[124], and epigenetic factors[129, 130]. In broader terms, any mutation in the genome is subject to the same evolutionary forces as mutations within a coding region depending on fitness effect, including mutations which affect expression level. Therefore the fitness cost of gene expression changes, may be vastly different between genes and the relationship between expression, fitness, and mutation is of obvious interest.

Epistatic Interaction

With the advent of systems biology, it has become apparent that phenotypic traits are often due to the interaction of multiple genes, and therefore complex mutational landscapes. The interdependence of mutations, or epistasis, is most generally classified as the observation of interdependent fitness effects from multiple mutations[131]. In a mathematical framework, a non-zero difference between the observed fitness of a double mutant and the product of independent fitness of each individual mutation defines both the magnitude and directionality of epistatic interactions. Epistasis between two or more mutations within or between genes can fundamentally alter fitness distributions by suppressing or exacerbating the fitness effects of secondary mutations, and this phenomenon is of particular relevance to human disease[132].

Epistasis is a multidimensional process, and the magnitude of epistasis is not necessarily informative without directionality. Unidimensional epistasis (also known as mean epistasis or directional epistasis) is a description of double mutant fitness within the null hypothesis of purely independent fitness effects. In this framework, mutations which interact to produce a smaller fitness effect than the independent prediction are termed negatively epistatic, whereas larger-than-predicted fitness effects are termed positively epistatic. However, the positive or negative denotation of epistasis makes no assumption as to overall fitness in the system. For example, two mutations can be defined as negatively epistatic yet result in a net beneficial fitness effect, or be defined as positively epistatic and result in a deleterious fitness effect. Multidimensional descriptions of epistasis are an essential tool to assess magnitude, sign, *and* fitness effect of genetic interactions to draw more relevant evolutionary conclusions[131, 133]. Combined with ancestral reconstruction techniques, it is possible to 'rewind the tape of life'[134] to assess accessible evolutionary intermediates and discover the root biochemical causes of epistasis[135-137].

Epistatic interactions appear frequently in biological systems ranging from viruses to humans[138-144]. Epistasis may occur between two genes, known as intergenic epistasis, or within a single gene, known as intragenic epistasis. The causes of intergenic epistasis include differential regulation by transcription factors[145], direct protein interaction changes[146], and redundancy introduced by gene duplication events[147] whereas intragenic

epistasis is best explained in the context of biophysical properties produced by protein engineering, rational design, and directed evolution.

Intergenic and intragenic epistatic interactions may occur by distinct mechanisms, but both processes can be realized in the context of fitness. Previous studies of intergenic epistasis indicate the majority of interdependent fitness effects are negative, whereas intragenic epistasis has been characterized as having a greater preponderance of synergistic positive epistasis[148]. The difference in directionality of intragenic epistasis is interesting not only from the viewpoint of directed evolution and the creation of more active or more stable enzymes, but because compensatory secondary mutations have been shown to rescue the fitness defect introduced by a conditionally adaptive primary mutation under chemotherapeutic regamines[149-152].

The mechanisms of intragenic epistasis bridge the fields of molecular evolution and biochemistry due to the mechanisms by which protein properties change with mutation, and how multiple mutations in redundant genes neofunctionalize to produce new protein functions. In biochemically oriented literature, the study of non-additive effects on protein function has previously been studied as 'double mutant cycles'[153]. In this experimental framework, two mutations are made independently as well as together in the same molecule. The biophysical properties of each of these three species are analyzed to search for non-additive biochemical changes, and therefore indicate interdependence between two residues. Double mutant cycles have been used to detect

numerous structural interactions including critical functional residues[154], long-range interactions[155], exposed and buried salt bridges[156, 157], and hydrogen bond networking[158] as well as protein-protein interactions[159].

Mechanistically, intragenic epistasis may alter the stability of a protein molecule by either exacerbating stability defects past a molecule specific stability threshold, or through suppression of stability defects[160]. In the absence of stability perturbations, conformational epistasis of binding or protein docking sites may alter the specificity, rate, or maximal activity of an enzyme. Conformational epistasis has been observed in the divergence of specificity between the mineralocorticoid receptor and glucocorticoid receptor through several mutations of the ancestral corticoid receptor[161]. Of potential explanatory power discussed within chapter V, multiple mutations may alter both stability and conformation of a protein (coined intramolecular pleiotropy[162]). The relationship between stability and conformation has best been described in directed evolution experiments in systems such as β-lactamase[163], cytochrome P450[164], Tre recombinase[165], and lipase A[166]. The common thread linking these studies is the tradeoff between protein stability and specific activity through mutational pathways that may not always be accessible *in vivo*.

In vitro systems such as directed evolution are capable of sampling nearly all possible protein variants regardless of their impact on organism fitness. However, natural systems are constrained to mutational pathways which support viability and reproduction.

Ancestral reconstruction studies of β-lactamase are a prime example for the hypothesized vs. realized potential of adaptive evolution. Five mutations in the bacterial β-lactamase gene are necessary to increase cefotaxime resistance by a factor of ~100,000 and the same five mutations are capable of combining through 120 mutational pathways[167]. However, Weinreich *et al* found many of these pathways to be inaccessible to adaptive evolution due to the inability of most mutant combinations to increase cefotaxime resistance. Not only is this an interesting example of the importance of epistasis in evolutionary trajectories, but additionally highlights the considerations which must be made when deciding between methodologies to study evolution experimentally.

The Hsp90 Molecular Chaperone

Rationale

The work presented in this dissertation is focused on a nine amino acid region in the yeast Hsp90 protein. This region was chosen based on a variety of interesting evolutionary and biochemical characteristics. For instance, two aromatic residues projecting into solvent separated by a glycine residue immediately led us to the hypothesis that the amino acid 582-590 region could be a putative docking interface when combined with the knowledge of extremely high phylogenetic conservation of these energetically unfavorable residue positions. Additionally, Hsp90 has long been known to participate in an ensemble of protein-protein interactions, but little biochemical characterization of these interactions exists in the literature, with no studies of the amino

acid 582-590 region to date. In short, an unstudied and biochemically interesting region of a highly conserved protein is an extremely interesting target for investigation.

Structure, Function and Biochemistry

Heat shock protein 90kDa (Hsp90, Hsp82 in yeast) is a conserved, highly expressed, and thoroughly networked protein chaperone. In yeast, Hsp90 is expressed in the cytosol as two nearly identical isoforms from two loci which are ~97% identical: Hsp90 and Hsc90. Whereas Hsc90 is constitutively expressed, Hsp90 is inducible under conditions of proteotoxic stress[123]. Hsp90 is one of the most highly expressed proteins in the cell, constituting 1-2% of all cytosolic protein under normal growth conditions, and 5-6% under conditions of cellular stress[168]. Expression regulation of Hsp90 occurs through interaction with the transcription factor Hsf1which is normally associated with chaperone machinery including Hsp40/70 as well as Hsp90 as an inactive monomer. Under conditions of cellular stress, Hsf1 dissociates from Hsp90, homotrimerizes, and translocates to the nucleus where it drives expression from three heat shock element (HSE) motifs in the Hsp90 promoter[169, 170]. The Hsf1 response is also negatively regulated by the re-association of Hsf1 to Hsp90 when proteotoxic conditions subside, indicating an interesting biochemical link between Hsf1 and the proteins it induces.

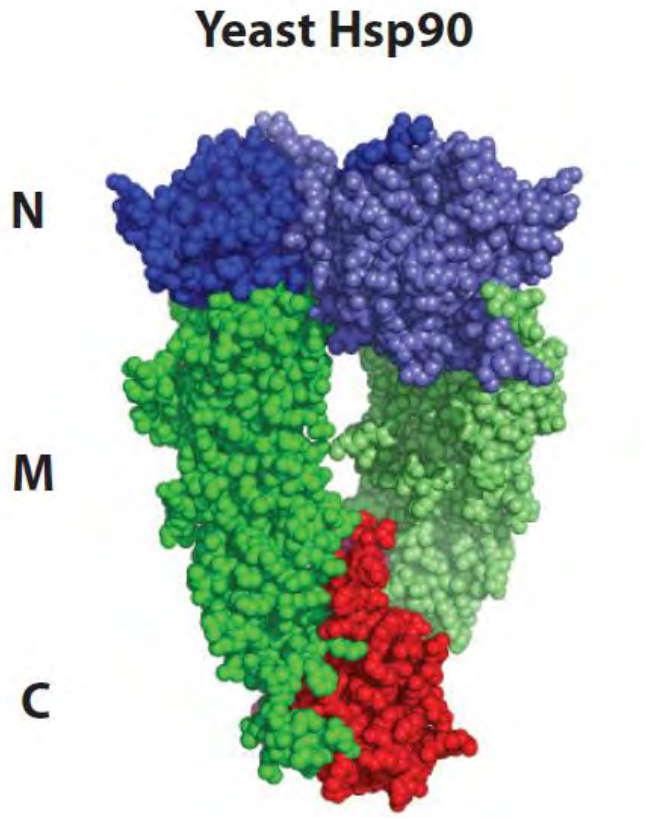
Structurally, yeast Hsp90 (also referred to as Hsp82) is a C-terminal homodimer composed of two 709 amino acid (81.4kDa) monomers. Each monomer can be subdivided into three distinct domains referred to as the N-terminal domain (N), middle

domain (M), and C-terminal domain (C) (Figure 1.1)[171]. The N-domain of Hsp90 is the most thoroughly studied due to its potential clinical relevance and known interacting partners. The N-domain contains an ATP binding pocket, and in conjunction with conformational changes of the dimer, slowly hydrolyzes ATP (relative to other characterized ATPases) to perform indispensible intracellular functions[172, 173]. The conformational changes along with ATP hydrolysis is referred to as the "ATPase cycle". Although mechanistic detail of the ATPase cycle is still a topic of study, the general features of this process are ATP binding (residue D79 is essential), closing of the "ATP lid", N-terminal dimerization and intramonomer N-M contact, compaction of the Hsp90 dimer and ATP hydrolysis (residues Glu33 and Arg380 are essential), and finally ADP release[174].

Besides being conformationally dynamic, Hsp90 is also highly interactive. By 2-hybrid and SGA analysis, Hsp90 has been shown to interact with approximately 3% of the yeast proteome, leading the classification of an interaction hub[175-177]. However, the mechanistic detail and biological relevance of the majority of these interactions remains an open question in the field. One challenge to studying the interactome of Hsp90 by standard biochemical methods is not only the large number of identified interacting partners, but the transient nature with which many substrates bind[178]. For this reason, *in vivo* assays have been developed to investigate well characterized functions of Hsp90 including kinase activation and hormone receptor maturation[179]. A limited number of *in*

Figure 1.1 Yeast Hsp90 solved structure[171]. The yeast Hsp90 homodimer divided by domain where the N-domain is represented in blue, the M-domain in green, and the C-domain in red. Subdued colors indicate the second monomer domains.

Figure 1.1



vitro assays have also been developed to assess the chaperone activity of Hsp90 including aggregation assays[178, 180, 181] and ATPase rate measurement[182].

Hsp90 interacts with numerous other proteins which can be broadly characterized as either co-chaperones or clients (substrates). Briefly, co-chaperones interact with Hsp90 to modulate ATPase activity, client specificity, or chaperone complex formation[183], whereas clients are acted upon by Hsp90 to aid in refolding, become activated in the case of kinases, or mature into substrate binding conformations in the case of steroid hormone receptors. The association of Hsp90 with oncogenic kinases as well as pathogenic proteins is particularly interesting to biomedical science because the ATP binding site of Hsp90 has become a target for chemotherapeutics to augment cancer therapy and antimicrobial treatment. It also stands to reason that during treatment with Hsp90 specific inhibitors, adaptive mutations conferring resistance to these inhibitors will develop, so an advanced understanding of the mutations which may confer such resistance is the logical next step to the work presented in this dissertation.

The M-domain of Hsp90 spans residues 253-524 and is connected to the N-domain through a flexible and transferable charged linker[171, 180]. The M-domain has been shown to interact with several clients and co-chaperones including Akt, eNOS, and Aha1[184, 185] as well as self-associations with the N-domain to putatively stabilize closed complexes for ATP hydrolysis. The ATPase activity of Hsp90 has been described as 'split' because residues in the M domain, including Arg380 and Gln384, appear to make

contacts with the N-domain ATP binding pocket in the closed conformation, and mutation of these residues does not support growth in yeast and reduces ATPase activity *in vitro*[174, 186]. An amphipathic loop containing residue Trp300 does not exhibit ATPase activity defects when mutated, but causes severe growth defects in yeast indicating this loop is potentially involved in client binding[186]. Additionally, the ATPase modulator AhaI binds a large portion of the M-domain (and portions of the N-domain) to stabilize an ATP hydrolysis competent state, and stimulate ATP hydrolysis[184].

As previously stated, Hsp90 predominantly exists as a dimer in which C-domains from two monomers strongly associate (K_d=60nM) to assume a dynamic molecular 'pincer'[187]. Furthermore, dimerization of full length monomers is essential for yeast viability as well as *in vivo* and *in vitro* functionality[188]. The most C-terminal residues of Hsp90 are MetGluGluValAsp (MEEVD), and this polypeptide sequence has been well characterized as a recognition peptide for tetratricopeptide repeat (TPR) protein interaction domains. Hsp90 has been shown to bind to a variety of proteins including FKBP51/54, FKBP52, Cyp40, p60, and others through interactions between the TPR domain of these proteins and the MEEVD of Hsp90[189].

The C-domain of Hsp90 also contains the region with which this dissertation is primarily concerned. The region of amino acids 582-590 forms a solvent exposed loop with two aromatic residues projecting into solvent (Phe583 and Trp585) indicative of a putative protein binding interface[171, 180]. Furthermore, residue Ser586 projects into the

core of the protein where the γ-hydroxyl of serine comes within close proximity to main chain atoms and may participate in a stabilizing hydrogen bond (see Figure S3.4). This region was first characterized as significant for Hsp90 functionality during mutational studies which showed mutations to Ala587 caused slight temperature sensitivity as well as defects in glucocorticoid receptor maturation[179]. The following chapters seek to extend our knowledge of this region of Hsp90 as well as apply mutational studies of this region to evolution biology and population genetics.

Hsp90 and Evolution

Originally discovered as a protein upregulated after heat shock of *D. melanogaster* cells[190], Hsp90 has recently been implicated as a capacitor and potentiator of molecular evolution. Stocks of *D. melanogaster* heterozygous with lethal Hsp83 (*D. melanogaster* Hsp90) mutations have a much higher prevalence of phenotypic abnormalities compared to wild type stocks. Interestingly, the frequency of morphological defects can be reproduced not only in mutants of Hsp90, but by elevating thermal stress or treatment with an Hsp90 specific inhibitor which have the same effect of lowering the available pool of functional Hsp90[191]. This initial work concludes that large pools of functional Hsp90 are capable of suppressing underlying genetic variation, whereas conditions of stress may lead to decreased canalization (expression of a phenotype despite mutation) due to expression of cryptic genetic variation (CGV).

The role of Hsp90 as an evolutionary capacitor has since been described in yeast[100], plants[192], worms[193] , and fish[194]. Current mechanistic hypotheses posit genetic variation in Hsp90 clients will generally cause protein destabilization, but the chaperone activity of Hsp90 will suppress misfolding and allow CGV to accumulate. When environmental conditions shift significantly enough to warrant response from Hsf1, the pool of Hsp90 is depleted, releasing misfolded protein variants to generate a range of new phenotypic effects. This release of cryptic genetic variation and subsequent expression of new traits may be a novel mechanism for organisms to 'sense' environmental perturbations, and respond with novel and potentially adaptive phenotypes which can be acted on by selection.

Conversely, Hsp90 has also been found to act as a potentiator of adaptive evolution by promoting genomic instability[195]. Aneuploidy is a known mechanism by which eukaryotes can generate adaptive potential under acute stress[196, 197], and is also frequently observed in human cancers. Hsp90 contributes to aneuploidy by interacting with kinetochore proteins in a conserved mechanism to ensure high fidelity chromosome segregation, and depletion of the Hsp90 pool under stress perturbs the fidelity of segregation, leading to aneuploidy[195, 198, 199]. Aneuploid cells may then be able to adapt to a particular stress, or become sensitized to perturbations, making Hsp90 a viable drug target in the treatment of cancer.

Due to its role as both a capacitor and potentiator of adaptive evolution, Hsp90 has become a logical target for chemotherapeutic intervention. Treatment with Hsp90 specific inhibitors alone have produced only modest clinical results, however, due to numerous interactions with known oncogenic proteins, combinations of small molecule inhibitors with standard chemotherapeutic regimes has the mechanistic potential to fundamentally change cancer treatment[200, 201]. Cancer is a disease characterized by genomic instability and increased mutation rate, so depleting the cellular pool of Hsp90 has the potential to both to expose deleterious mutations in cancer cells while acting in a capacitor role and potentiate hyper-instability in cancer genomes to effectively sensitize cells to broad spectrum chemotherapy.

Standing Questions and the Scope of this Dissertation

The field of evolutionary biology has historically been rife with theoretical literature, but scarce in experimental evidence to support it. The overarching goal of my work has been to experimentally examine theoretical questions in evolutionary biology asked as early as 1930. The field of experimental evolution has the potential to answer questions ranging from the origins of life to the treatment of disease. As whole genome sequencing becomes faster and cheaper, polymorphism based personalized medicine is likely to give individuals in depth knowledge of genetic factors affecting their own health and wellness. However, without a foundational knowledge of the effects of mutations, the

interaction of these mutations, and how mutational fitness effects are perturbed by environment, personalized medicine by sequence analysis can be little more than a predictive tool.

Chapter II and III

To address standing questions regarding the distribution of fitness effects of new mutations, we have developed a technique coined Exceedingly Methodical and Parallel Investigation of Randomized Individual Codons (EMPIRIC) which allows us to accurately measure the fitness effects of all possible codon substitutions in regions of genes in yeast (method described in detail in Chapter II). Utilizing the EMPIRIC technique in the context of amino acids 582-590 of Hsp90, Chapter III addresses predictions made by Ohta and Kimura[18, 19] regarding the distribution of fitness effects in the context of the near neutral model of evolution. Although a bimodal distribution of fitness effects has been presented for a limited number of single mutations in vesicular stomatitis virus[90] and other organisms, the EMPIRIC technique has allowed us to iteratively saturate this region of Hsp90 with new mutations. By introducing systematic mutant libraries instead of sporadic mutagenesis or mutant accumulation, we are also able to extend these findings to the use of phylogenetic conservation a predictive tool for mutational fitness effects as well as the optimization of the genetic code.

Chapter IV

An essential but experimentally elusive property of biological systems is the relationship between environmental perturbations and the strength and nature of

selection. Since Fisher and Wright, there have been extensive disagreements about the frequency and magnitude of new beneficial mutations. Two central models to describe adaptation have been central to this debate: Fisher's Geometric Model and Wright's Adaptive Landscape. In Chapter IV, I present findings to address the distribution of fitness effects under non-optimal environmental conditions to calculate the frequency and magnitude of beneficial mutations under non-optimal conditions. This study is unique because the environmental perturbations (and therefore selection) are relatively modest as demonstrated by high fitness of the wild type sequence. We then address the relationship of Fisher's Geometric Model to the studied environmental perturbations, the quantitative cost of adaptation, and the magnitude and frequency of new beneficial mutations.

Chapter V

One substantial difference between the theories of Fisher and Wright was the assumptions each made regarding epistasis. Whereas Fisher essentially ignored epistasis, Wright considered epistatic interactions to be frequent and of substantial importance to adaptive evolution. In chapter V, I discuss the fact that the EMPIRIC technique allows us to logically extend our understanding of epistasis by probing perturbations in the fitness landscape due to differing genetic backgrounds. Due to the comprehensive nature of our mutagenesis technique, we are able to examine the full distribution of intragenic epistastic effects of seven non-wild type backgrounds. This approach to epistasis is not only relevant to evolutionary biology because, to my knowledge, a comprehensive

distribution of epistatic effects has not been reported, but it allows us to deconstruct the biochemical details of intragenic epistasis in this region.

Chapter II – Fitness Analyses of All Possible Point Mutations for Regions of Genes in Yeast

Alternatively: Broadly Applicable Materials and Methods for the EMPIRIC Technique

This work has been published previously as *Hietpas R.*, Roscoe B.*, Jiang L., Bolon DNA. "Fitness analyses of all possible point mutations for regions of genes in yeast." Nat Protoc. 2012 Jun 21; 7(7): 1382-96.*

The work presenting in the following chapter was a collaborative effort. I, Benjamin P. Roscoe, Li Jiang and Dr. Daniel N. A. Bolon all contributed to the development and optimization of the protocol as well as preparing the manuscript. I prepared the initial draft for the section regarding generating mutant libraries. Benjamin P. Roscoe prepared the initial draft for the section on growth competition. Benjamin P. Roscoe and Li Jiang prepared the initial draft for the section on preparing samples for deep sequencing. Dr. Daniel N. A. Bolon prepared the initial draft for the section on processing sequencing data. Dr. Daniel N. A. Bolon supervised the work and prepared the final version of the manuscript.

Abstract

Deep sequencing can accurately measure the relative abundance of hundreds of mutations in a single bulk competition experiment, which can give a direct readout of the fitness of each mutant. Here we describe a protocol that we previously developed and optimized to measure the fitness effects of all possible individual codon substitutions for 10-aa regions of essential genes in yeast. Starting with a conditional strain (i.e., a temperature-sensitive strain), we describe how to efficiently generate plasmid libraries of point mutants that can then be transformed to generate libraries of yeast. The yeast libraries are competed under conditions that select for mutant function. Deep-sequencing analyses are used to determine the relative fitness of all mutants. This approach is faster and cheaper per mutant compared with analyzing individually isolated mutants. The protocol can be performed in ~4 weeks and many 10-aa regions can be analyzed in parallel.

Introduction

Evolution is a critical principle for interpreting and understanding biology. Evolutionary processes have shaped life in its present state and continue to mediate future population trajectories. The basic rule of evolution is competition, and fitness is the measure of individual competitive advantage/disadvantage. Genetic mutations are a dominant mechanism impacting fitness. The relationship between genetic mutations and fitness describes the raw evolutionary potential available to organisms. Here we describe a method that we refer to as EMPIRIC (Exceedingly Methodical and Parallel Investigation of Randomized Individual Codons) to systematically generate all possible point mutations in regions of important genes and quantify the fitness effect of each mutant[202].

Many previous methods have been utilized to analyze the fitness effects of mutations. These methods fall broadly into two classes: population-genetic based inferences from sequence analyses of naturally-evolving populations[203, 204], and direct fitness measurements of mutants[99, 100]. Population genetic models combined with polymorphism data provide routes to understand recent selection in all current organisms. However, mutations that cause a selectable fitness effect can be challenging to distinguish from hitchhiking mutants at linked genetic loci[205]. In contrast, experimental fitness competitions have the benefit of directly measuring fitness effects of specific mutations, though they cannot be applied to all organisms.

Experimental fitness measurements often involve isolating specific mutants and following their growth properties for multiple generations. These analyses are ideally suited for organisms that can be easily manipulated genetically and that have short generation times such as microbes. Indeed, the ability to genetically manipulate *S. cerevisiae* enabled systematic analyses of the fitness of single gene knockouts and the identification of essential genes[206]. The yeast deletion strains were generated with a unique DNA sequence or barcode bracketed by common primer sites for each gene knockout. These bar codes enable the relative abundance of each mutant to be monitored using PCR and sequencing. This approach enables quantitative analyses of relative fitness from bulk cultures of knockout strains. The fitness effects of knockouts provides useful insights, but the knockout approach does not provide direct information on the fitness effects of many types of mutations that occur during natural evolution including point mutations.

Analyzing the fitness effects of point mutations is relevant to biology because they are a common form of mutation in evolution. Point mutations that lead to drug resistance have been extensively analyzed[167]. Drug-resistance mutations can be readily identified from both natural/clinical isolates and from laboratory selection experiments. The fitness effects of mutations in drug-resistant genes are frequently analyzed based on a dose-response curve. The large magnitude of growth changes associated with drug-resistant mutations facilitates their analysis and represents a stringent selection pressure.

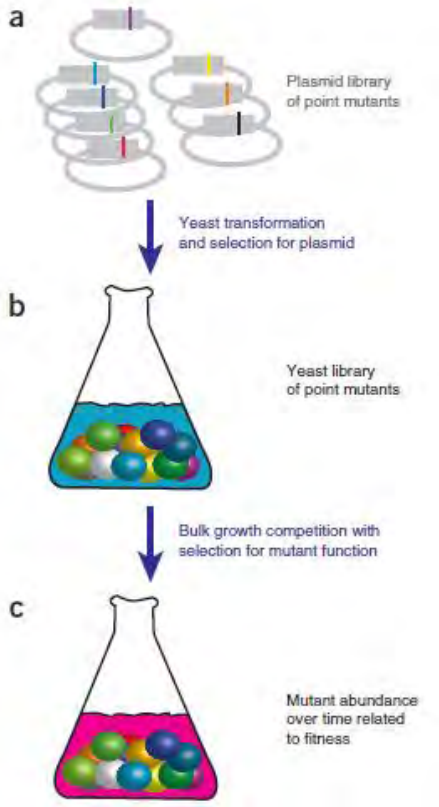
Many genes are involved in adaptation to less stringent selection pressures[207] than drug-resistance. Because the fitness changes are small relative to drug-resistance mutations, analyzing mutant fitness effects in the majority of genes requires accurate measurements of relative growth, and careful control of genetic background. Both growth curves of individual strains[100] and binary competition experiments between fluorescently-labeled strains[99] enable accurate measurement of the fitness effects of one mutant per culture. Using isolated individual mutations, alanine scanning has been used to identify hot-spots for protein function[208].

We developed the EMPIRIC approach to monitor the relative abundance of saturation point-mutants in a single bulk culture (Figure 2.1) with very high signal to noise[202]. This sequencing approach is similar in concept to the bar coded knockout collection[206], as well as methods developed to analyze binding function of larger and more complex libraries using affinity isolation approaches[209-211]. In all three of these approaches, sequencing is utilized to monitor the relative abundance of mutants after the application of a selective pressure. In the knockout collection, mutants are identified by a unique barcode between universal primer binding sites. The affinity isolation approaches have been able to interrogate larger mutant libraries including many double mutants[209], and are well-suited where broad sampling of double mutants are desired.

In the EMPIRIC approach, the libraries are constructed to contain only point-mutants that are quantified directly using focused deep sequencing of mutated regions.

Figure 2.1 Bulk competition of libraries of point mutants in yeast. (a) Plasmid libraries are transformed into yeast. (b) Yeast that have taken up plasmid are selected for and amplified. (c) Selection pressure is applied to the library copy of the mutated gene and samples are collected over time in bulk competition.

Figure 2.1


a
Plasmid library
of point mutants
Yeast transformation
and selection for plasmid
b
Yeast library
of point mutants
Bulk growth competition with
selection for mutant function
c
Mutant abundance
over time related
to fitness

This approach results in a strong sequencing signal for all possible point mutants, and it is ideally suited for applications where accurate and systematic measurements of point-mutant function or fitness are desired. While the initial application of EMPIRIC was analyzing the effect of mutants in Hsp90 on yeast growth[202], the protocol could be modified to analyze growth in other genetically tractable systems (i.e. cancer cells and viruses) as well as for *in vitro* function utilizing display approaches[209]. In addition, we have found that throughput can be dramatically increased by analyzing multiple regions in parallel. We have performed parallel analyses of 8 separate 10-aa regions in the same four week time period required to analyze one region[212]. The maximum size of region that we have analyzed by EMPIRIC is currently 10 amino acids. The size of a region that can be accurately analyzed is constrained by sequencing read-length and accuracy

Overview of the EMPIRIC method

The EMPIRIC method is designed to measure the competitive advantage or disadvantage of point mutants in high-throughput. Efficient analyses of mutants are facilitated by three main components: a rapid strategy to generate saturation mutants at consecutive amino acid positions in a gene; synchronized application of selection pressure to all mutants in a mixed competition experiment; and accurate measurement of the relative abundance of each mutant using deep-sequencing. We use a cassette ligation strategy to efficiently generate mutant libraries. This stage involves DNA manipulations including PCR, ligations, and bacterial transformations. In order to synchronize selection

pressure, we use a conditional yeast strain such as a temperature-sensitive strain. This stage involves yeast microbiological techniques, including transformation and growth in liquid culture. We use a deep-sequencing approach to measure the abundance of each mutant. This stage involves isolation of DNA from yeast, DNA manipulations including PCR to generate focused libraries for sequencing, and bioinformatic analyses of the resulting sequencing data.

Experimental design

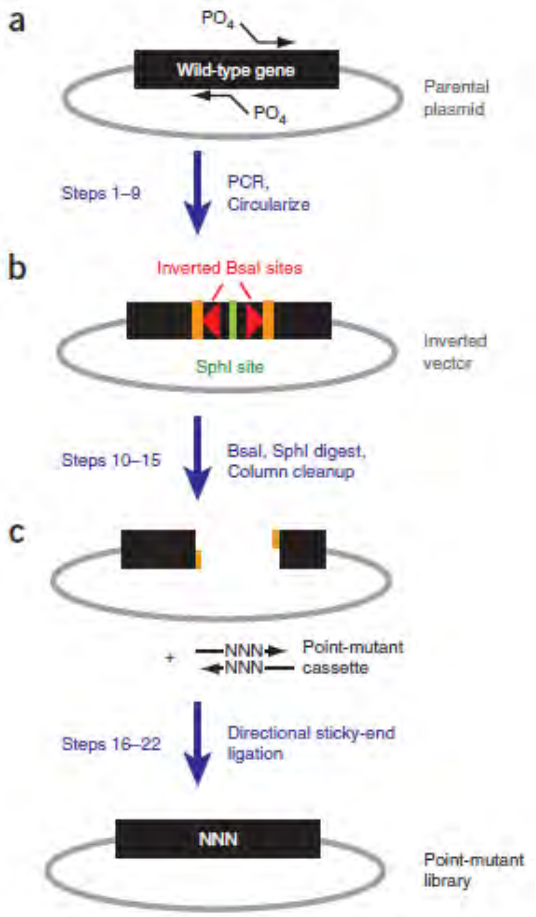
In order to accurately measure the relative abundance of all possible point mutants for regions of genes using deep-sequencing, the EMPIRIC approach was developed with careful consideration of signal-to-noise. Signal is the relative abundance of a mutant in a library. Noise comes from misreads that distort the measured abundance of a mutant from its actual abundance in the library. In library generation, the goal is to have all mutants present at similar abundance. In the growth competition, the goal is to rapidly analyze mutants under selection while minimizing the potential for secondary adaptive mutations. In library analysis, the goal is to minimize noise from misreads.

Mutant abundance

The primary factor that can be manipulated to maximize signal is the relative abundance of each mutant in the starting plasmid library. Ideally, all mutants will be present at a relative abundance well above the noise that comes from misreads. We optimized a cassette ligation strategy (Figure 2.2) that can be applied iteratively and in

Figure 2.2 Steps to generate plasmid libraries of point mutants. (a) Whole-plasmid PCR to generate inverted BsaI vector. (b) Digestion of this vector to generate directional sticky ends. (c) Cassette ligation to introduce point mutants.

Figure 2.2


a
PO4
Wild-type gene
PO4
Parental plasmid
Steps 1–9
PCR, Circularize
b
Inverted BsaI sites
SphI site
Inverted vector
Steps 10–15
BsaI, SphI digest, Column cleanup
c
+
NNN
NNN
Point-mutant cassette
Steps 16–22
Directional sticky-end ligation
NNN
Point-mutant library

parallel to generate libraries of point mutants in which all variants are present at similar relative abundance. Alternative methods exist to generate point mutant libraries, including Quickchange mutagenesis and gene synthesis, but in our experience the cassette ligation strategy has resulted in the most efficient and reproducible results.

Design of oligonucleotides for generating vectors with inverted type IIS restriction sites

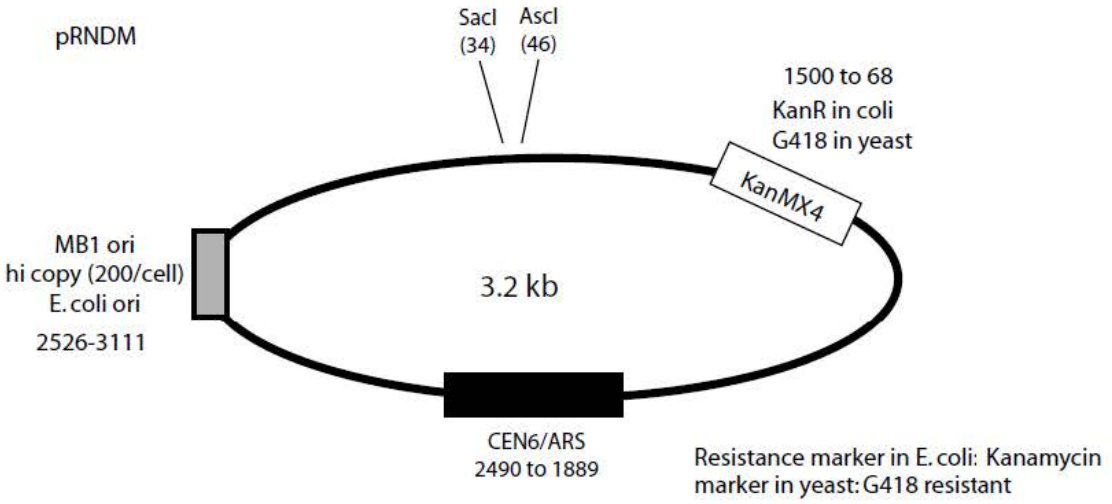
Oligonucleotides should be designed as primers for whole-plasmid PCR in order to generate vectors with inverted type IIS (i.e., BsaI) restriction sites for cloning (Figure 2.2a and Table 2.1 – vector forward (for) and vector reverse (rev)). We have used this approach to amplify vectors up to 10 kb. Whole-plasmid primers should have ~20 bases that are complementary to the target plasmid and 5' extensions to encode restriction site. The inclusion of additional unique restriction sites in the 5' extensions (immediately upstream of the BsaI site) can be used to reduce background during subsequent cassette ligations (i.e., SphI in Figure 2.2). For this strategy to succeed, it is important to have a parental plasmid construct that lacks BsaI sites. We generated a minimal yeast and bacterial shuttle plasmid that we refer to as pRNDM with a KanMX4 marker[220], which confers kanamycin resistance in bacteria, confers G418 resistance in yeast, and lacks BsaI sites (Supporting Figure S2.1).

Table 2.1

Oligo name	Oligo sequence (5′–3′)	Key features	Modifications required	Purpose
Vector for	ggtggtggtgcatgc**ggtctc** a**ATTA**CTCAGTTGATGAGTTT	Capitals represent nucleotides complementary to plasmid sequence, bold letters represent BsaI overhangs, and italics indicate BsaI restriction site	5′ phosphorylation (Step 1)	Used with 'Vector rev' for whole-plasmid PCR (Steps 1–9)
Vector rev	gcagcagcagcatgc**ggtctc**a-**CATA**GTATTCTATTTTTCTC	Capitals represent nucleotide complementary to plasmid sequence, bold letters represent BsaI overhangs and italics indicate BsaI restriction site	5′ phosphorylation (Step 1)	Used with 'Vector for' for whole-plasmid PCR (Steps 1–9)
Cas for	**tatg**NNNagtgaaacttttgaatttcaagctgaa	Underlined text represents the ten-codon region of interest. N indicates a mixture of A,C,T,G at the randomized codon. Bold text indicates overhangs complementary to BsaI sites		Annealed with 'Cas rev' and ligated to vector to create a saturation library (Steps 10–22). Need a different oligo for each codon to be randomized
Cas rev	**taat**ttcagcttgaaattcaaaagtttcactNNN	Underlined text represents the ten-codon region of interest. N indicates a mixture of A,C,T,G at the randomized codon. Bold text indicates overhangs complementary to BsaI sites		Annealed with 'Cas for' and ligated to vector to create a saturation library by ligation (Steps 10–22). You need a different oligo for each codon to be randomized
PCR1a for	AAGACGGTAGGTATTGATTGT	Complementary to the promoter region of the library version of the gene of interest. This primer is specific to the vector		Used with 'PCR1a rev' to amplify the library version of the gene of interest (Step 43)
PCR1a rev	GGGACCTAGACTTCAGGTTGTC	Complementary to the 3′ UTR region of the library version of the gene of interest. This primer is specific to the vector		Used with 'PCR1a for' to amplify the library version of the gene of interest (Step 43)
PCR1b for	gggaccaccacc**tccgac**ACAC-CCCAATCATGTTGCAG	Capitals indicate nucleotides complementary to the template. The MmeI site is indicated in bold		Used with 'PCR1b rev' to amplify randomized region. Designed to add an upstream MmeI site to the amplicon (Step 44)
PCR1b rev	N_{25}-GATAAAGACATTAATGGTTG	Capitals indicate nucleotides complementary to the template. N_{25} indicates 25-nt binding site for a 3′ deep-sequencing primer—check with sequencing provider for current recommendations		Used with 'PCR1b for' to amplify randomized region. Designed to add a downstream primer binding site to the amplicon for deep sequencing (Step 44)
Adapt for	N_{25}-ACGTag	Capitals indicate a bar code and N_{25} indicates binding site for a 5′ deep-sequencing primer—check with sequencing provider for current recommendations. Lowercase nucleotides are complementary to the MmeI site		Annealed with 'Adapt rev' and ligated to MmeI-digested PCR1b product to add a bar code and an upstream (5′) primer binding site for deep sequencing (Step 46). A different oligo with a unique bar code is needed for each sequencing sample
Adapt rev	ACGT-N_{25}	Capitals indicate a bar code and N_{25} indicates binding site for a 5′ deep-sequencing primer—check with sequencing provider for current recommendations		Annealed with 'Adapt for' and ligated to MmeI-digested PCR1b product to add a bar code and an upstream primer binding site for deep sequencing (Step 46). A different oligo with a unique bar code is needed for each sequencing sample

Figure S2.1 Features and sequence of bacterial-yeast shuttle plasmid pRNDM. This plasmid was derived from pRS414 with the tryptophan marker replaced by KanMX4 and the beta-lactamase gene removed.

Figure S2.1



Sequence:

TTTTTACGGTTCCTGGGAACAAAAGCTGGAGCTCgtttaaaccgGCGCGCCTTAGCTC-
GTTTTCGACACTGGATGGCGGCGTTAGTATCGAATCGACAGCAGTATAGCGACCAGCATTCACATACGATTGACGCATGATATTACTTTCTGCGCACTTAACTTCGC
ATCTGGGCAGATGATGTCGAGGCGAAAAAAAATATAAATCACGCTAACATTTGATTAAAATAGAACAACTACAATATAAAAAAACTATACAAATGACAAGTTCTTGA
AAACAAGAATCTTTTTATTGTCAGTACTGATTAGAAAAACTCATCGAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAG
CCGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTA
ATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAGAAATCACCATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGCTTATGCATTTCTTTCCAGACTTGTTCA
ACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGATCGCTGTTAAAAG
GACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGT
TTTGCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGGCATAAATTCCGTCAGCCAGTTTAGTCTGACC
ATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCG
ACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTAATCGCGGCCTCGAAACGTGAGTCTTTTCCTTACCCATGGTTGTTTATGTTCGGA
TGTGATGTGAGAACTGTATCCTAGCAAGATTTTAAAAGGAAGTATATGAAAGAAGAACCTCAGTGGCAAATCCTAACCTTTTATATTTCTCTACAGGGGCGCGGCGT
GGGGACAATTCAACGCGTCTGTGAGGGGAGCGTTTCCCTGCTCGCAGGTCTGCAGCGAGGAGCCGTAATTTTTGCTTCGCGCCGTGCGGCCATCAAAATGTATGG
ATGCAAATGATTATACATGGGGATGTATGGGCTAAATGTACGGGCGACAGTCACATCATGCCCCTGAGCTGCGCACGTCAAGACTGTCAAGGAGGGTATTCTGGGC
CTCCATGTCGCTGGCCGGGTGACCCGGCGGGGACGAGGCAAGCTAAACAGATCGGCCGCGTTCTATAGTGTCACCTAAATCGTATGTGTATGATACATAAGGTTATG
TATTAATTGTAGCCGCGTTCTAACGACAATATGTCCATATATGCGTATATATACCAATCTAAGTCTGTGCTCCTTCCTTCGTTCTTCCTTCTGTTCGGAGATTACCGAATC
AAAAAAATTTCAAGGAAACCGAAATCAAAAAAAAGAATAAAAAAAAAAATGATGAATTGAAAAGGTGGTATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCA
TAGTTAAGCCAGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGC
TGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGACGAAAGGGCCTCGTGATACGCCTATTTTTATAGGTTAATGTCATGATAATAATGGTTTCTTAG
ACGGATCGCTTGCCTGTAACTTACACGCGCCTCGTATCTTTTAATGATGGAATAATTTGGGAATTTACTCTGTGTTTATTTATTTTTATGTTTTGTATTTGGATTTTAGA
AAGTAAATAAAGAAGGTAGAAGAGTTACGGAATGAAGAAAAAAAAAATAAACAAAGGTTTAAAAAATTTCAACAAAAAGCGTACTTTACATATATATTTATTAGACA
AGAAAAGCAGATTAAATAGATATACATTCGATTAACGATAAGTAAAATGTAAAATCACAGGATTTTCGTGTGTGGTCTTCTACACAGACAAGATGAAACAATTCGGC
ATTAATANNGAGAGCAGGAAGAGCAAGATAAAAGGTAGTATTTGTTGGCGATCCCCCTAGAGTCTTTTACATCTTCGGAAAACAAAAACTATTTTTTCTTTAATTTC
TTTTTTTACTTTCTATTTTTAATTTATATATTTATATTAAAAAATTTAAATTATAATTATTTTTATAGCACGTGATGAAAAGGACCCAGGTGGCACTTTTCGGGGAAATGT
GCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTAC
CAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCTTCTAGTGTAGCCGTA
GTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGT
TGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAG
ATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAG
GGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTA
TGGAAAAACGCCAGCAACGCGGCC

Design of oligonucleotides cassettes with individual codons randomized

Oligonucleotides for the cassette mutagenesis step (Cas for and Cas rev in Table 2.1) should have cohesive ends that are complementary to the BsaI 5' overhangs in the vector. These oligos will be annealed to each other to form a double stranded cassette – they are not used for priming amplification/mutagenesis. For each amino acid position that you would like to randomize, design a cassette with a degenerate codon (i.e., NNN) on both strands. We have obtained consistent results using cassettes where each oligonucleotide is 40 bases in length (30 bases for the 10 amino acid region, 3 bases on either side of this region that improve ligation efficiency for the randomization of edge positions, and the 4-base 5' overhangs).

Design of oligonucleotides to amplify the library gene

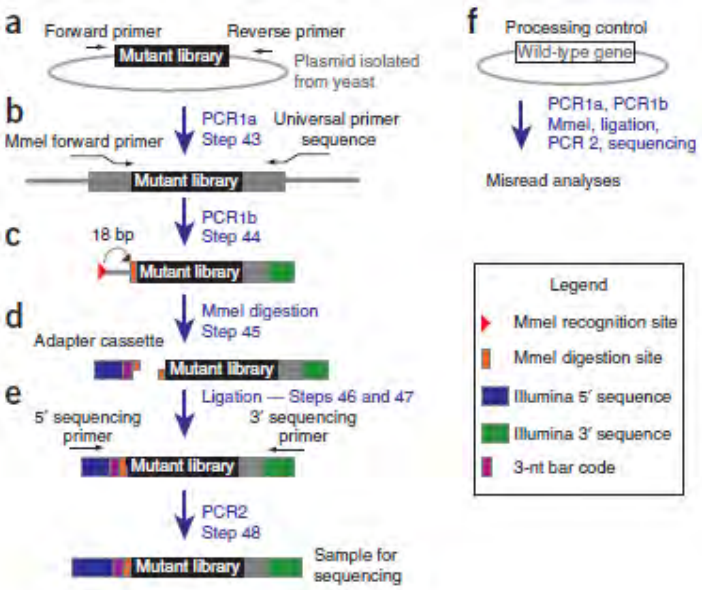
PCR1a primers (Figure 2.3a, PCR1a for and PCR1a rev in Table 2.1) should be designed to specifically amplify the library version of the gene of interest (and not the conditional genomic copy also present in cells). The optimal size range of the amplicon is 250 bases. Standard primer design approaches should be used. PCR1a primers should be 18-22 bases and anneal uniquely to the library plasmid (i.e. to unique regions upstream and downstream of the gene of interest). These are primers are gene-specific.

Design of oligonucleotides to focus sequencing on the randomized region

PCR1b primers (Figure 2.3b) should be designed to amplify the randomized region, add an upstream MmeI site (the purpose of the MmeI site is to provide a site for

Figure 2.3 Steps to prepare DNA for deep sequencing. (a) PCR amplify mutant library using primers specific to the plasmid library. (b) Perform a second PCR step to add an MmeI site to the 5' end and an Illumina universal primer sequence to the 3' end. (c) Perform MmeI digestion to create a sticky end adjacent to the randomized region of the mutant library. (d) Ligate an adapter to the 5' end containing a bar code. (e) PCR with universal deep-sequencing primers. (f) Parallel analyses of a wild type plasmid provides information on misreads.

Figure 2.3


a
Forward primer
Reverse primer
Mutant library
Plasmid isolated from yeast
b
MmeI forward primer
PCR1a Step 43
Universal primer sequence
Mutant library
c
18 bp
PCR1b Step 44
Mutant library
d
Adapter cassette
MmeI digestion Step 45
Mutant library
e
5′ sequencing primer
Ligation — Steps 46 and 47
3′ sequencing primer
Mutant library
PCR2 Step 48
Mutant library
Sample for sequencing
f
Processing control
Wild-type gene
PCR1a, PCR1b MmeI, ligation, PCR 2, sequencing
Misread analyses
Legend
MmeI recognition site
MmeI digestion site
Illumina 5′ sequence
Illumina 3′ sequence
3-nt bar code

adapter ligation), and add a downstream Illumina sequencing site. The MmeI PCR1b primer (PCR1b for in Table 2.1) should have 20 bases of complementarity to the region immediately upstream from the randomized region (gene specific) and a 5' extension encoding a restriction site for MmeI. The downstream PCR1b primer (PCR1b rev in Table 2.1) should have 18-22 bases of complementarity that target binding 200 bases downstream of the randomized region (in order to generate a 200 base amplicon) and a 5' extension of 25 bases complementary to Illumina sequencing primers.

Design of bar coded adapter oligonucleotide cassette

Oligonucleotides should be designed that when annealed form a double stranded adapter. This adapter cassette should have a double stranded region including 25 bases complementary to Illumina sequencing primers and a barcode of 3-4 bases followed by a two-base single-stranded 3' overhang complementary to the overhang created by MmeI digestion of the PCR1b PCR product (Figure 2.3d). Care should be taken in designing barcodes such that all samples in a sequencing reaction can be uniquely identified. A single sequencing sample represents uniquely barcoded timepoint samples for a library of 10 randomized codons, as single-codon libraries are pooled prior to analysis. Ideally, each barcode will differ from all other barcodes at multiple positions to minimize the potential for mis-reads to cause barcode switching. With Illumina sequencing, the base composition at each position in the sequencing library is an important parameter because it impacts the ability to distinguish the position of individual clusters. For this reason, it is valuable to blend samples such that each position in the sequencing mix, including the

barcode region, has a broad distribution of bases. This challenge can also be mitigated by further blending with other sequencing samples or generating a lower density of clusters during sequencing and should be discussed with your sequencing provider.

Conditional strain

It is important to have a conditional strain that grows robustly on its own under permissive conditions, and whose growth rapidly slows or stalls in non-permissive conditions unless provided with a rescue copy of the gene of interest. In our proof-of-principle studies[202], we utilized a temperature sensitive Hsp90 strain. This strain grows robustly at 25°C, which allowed all possible point mutants in our library to be transformed into cells and propagated under this condition. This strain rapidly stalls growth at moderately elevated temperature (36°C), which was used to synchronize growth competition dependent on the function of the library version. Before starting competition experiments with libraries, it is important to identify appropriate permissive and non-permissive conditions. A wild type rescue plasmid and a null rescue plasmid can be used to determine these conditions. Ideally, you want the permissive condition to support equivalent growth rates for strains harboring either the wild type or the null rescue plasmid. For the non-permissive condition, cells harboring the wild type rescue plasmid should grow robustly (i.e. similar to the parental strain), while cells harboring the null plasmid should stall in growth.

Sources of noise

The primary cause of noise is mis-reads that can be caused by either PCR steps in processing samples, or in the sequencing reaction itself. We have found it extremely useful to include internal controls to assess the mis-read noise in each EMPIRIC experiment and sequencing run (these controls are described in the Procedure). In our experience, ~90% of Illumina sequencing runs have resulted in data quality sufficient to accurately assess the relative abundance of all point mutants in EMPIRIC experiments. The careful generation of point mutant libraries causes the majority of mis-reads to appear as double-mutants that are readily filtered out of datasets, dramatically reducing noise in subsequent analyses.

Genetic background is another potential source of noise that is important to consider in EMPIRIC experiments. In order to control for genetic background, we design entire experiments such that all required libraries are transformed into the same batch of yeast, thus minimizing potential secondary genetic differences. If secondary mutations of strong benefit sweep through a mutant population, it will cause a bi-phasic trajectory in the fitness data that can be readily identified. In this case only time-points prior to this sweep should be analyzed. If appropriate, secondary mutations of strong benefit can be minimized by pre-adapting the parental strain to the desired environmental conditions. We have also found that eliminating the 2μ plasmid that is endogenous in most yeast strains[213] can reduce the frequency of secondary adaptive genetic changes.

Materials

Reagents

▲**CRITICAL:** All media and reagents are prepared by standard methods[214], and are stored as recommended by the manufacturers. All enzymes are stored at -20°C. SAM and chemically competent bacteria are stored at -80 °C. Unless otherwise noted, all reagents are stored at room temperature.

- A conditional yeast strain (i.e. a temperature sensitive or shutoff strain) whose growth can be rescued by a plasmid-borne copy of the gene of interest. Conditional yeast strains can be generated de novo, or located in previously published work and requested.
- A starting plasmid to generate libraries that does not contain sites for the type IIS endonuclease that you plan to use for the cassette ligation strategy, such as pRNDM (Figure S2.1). The pRNDM plasmid will be provided on request.
- T4 DNA ligase (New England Biolabs, cat. no.M0202)
- T4 DNA Ligase Buffer 10x (New England Biolabs, cat. no. B0202S)
- T4 Polynucleotide Kinase (New England Biolabs, cat. no. M0201)
- Deoxyribonucleotide triphosphates (dNTPs; 10 mM each nucleotide; New England Biolabs, cat. no. N0447)
- DpnI restriction endonuclease (New England Biolabs, cat. no. R0176)
- Taq polymerase (New England Biolabs, cat. no. M0273)
- Phusion® High-Fidelity DNA polymerase (New England Biolabs, cat. no. M0530S)

▲**CRITICAL** – a high-fidelity polymerase should be used for amplification products intended for use in downstream deep-sequencing to limit PCR errors.

- BsaI restriction endonuclease (New England Biolabs, cat. no.R0535)
- SphI restriction endonuclease (New Englan Biolabs, cat. no.R0182)
- MmeI restriction endonuclease (New England Biolabs, cat. no. R0637L)
- S-adenosyl methionine (SAM; New England Biolabs, cat. no. B9003S)
- NEB3 buffer (10x with 100x BSA; New England Biolabs, cat. no.B7003)
- NEB4 buffer (10×; New England Biolabs, cat. no. B7004S)
- Agarose, PCR grade (Fisher Bioreagents, cat. no. 9012-36-6)
- Ethidium bromide (Sigma, cat. no. E1510)

  ! CAUTION Ethidium bromide is toxic and a DNA mutagen; handle properly and avoid contact using appropriate Personal Protective Equipment.
- SYBR Green I (10000×; Invitrogen, cat. no. S-7563)
- Tris Base (Fisher Bioreagents, cat. no. BP152-500)
- Acetic acid, glacial (Fisher Scientific, cat. no. A38-500)
- Bromophenol Blue (Sigma-Aldrich, cat. no. B0126)
- Ethylenediaminetetraacetic acid (EDTA; Sigma-Aldrich, cat. no. E6758)
- DNA ladder – 1 KB (New England Biolabs, cat. no. N3232)
- DNA ladder – 100 BP(New England Biolabs, cat. no. N3231)
- Zymoclean Gel DNA Recovery Kit (Zymoresearch, cat. no. D4001)
- ZR Plasmid Miniprep Kit (Zymoresearch, cat. no. D4015)
- OmniMax competent *E. coli* strain (Invitrogen, cat. no. C854003)

- Kanamycin-A monosulfate (or bacterial antibiotic matching vector marker)(Sigma-Aldrich, cat. no. K4000)
- Ampicillin sodium salt (Sigma-Aldrich, cat. no. A9518-100G)
- G418 disulfate salt (Sigma-Aldrich, cat. no. A1720)
- Polyethylene Glycol 3350 (PEG 3350; Hampton Research cat. no. HR2-591)
- Lithium acetate dihydrate (Sigma-Aldrich, cat. no. L4158)
- Salmon Sperm DNA (Sigma-Aldrich, cat. no. D1626)
- Yeast nitrogenous base without Amino Acids (VWR, cat. no. 61000-200)
- Ammonium Sulfate (Sigma-Aldrich, cat. no. A5132)
- Sodium Chloride (Fisher Bioreagents, cat. no. 5271-3)
- Zymolyase (Zymoresearch, cat. No E1004)
- Bacto- Tryptone (Becton Dickison, cat. no. 211705)
- Bacto- Peptone (Becton Dickison, cat. no. 211677)
- Bacto- Yeast Extract (Becton Dickison, cat. no. 212750)
- Bacto- Agar (Becton Dickison, cat. no. 214010)
- Adenine Hemisulfate (Sigma-Aldrich, cat. no. A9126-100g)
- Glucose (Sigma-Aldrich, cat. no. G7528-5kg)
- L-Aspartic acid (Sigma-Aldrich, cat. no. A8949)
- L-Arginine (Sigma-Aldrich, cat. no. A5006)
- L-Valine (Sigma-Aldrich, cat. no. V0513)
- L-Glutamic Acid (Sigma-Aldrich, cat. no. G1251)
- L-Serine (Sigma-Aldrich, cat. no. S4311)

- L-Threonine (Sigma-Aldrich, cat. no. T8625)
- L-Isoleucine (Sigma-Aldrich, cat. no. I2752)
- L-Phenylalanine (Sigma-Aldrich, cat. no. P2126)
- L-Tyrosine (Sigma-Aldrich, cat. no. T8566)
- L-Histidine (Sigma-Aldrich, cat. no. H8000)
- L-Methionine (Sigma-Aldrich, cat. no. M5308)
- L-Leucine (Sigma-Aldrich, cat. no. L8000)
- L-Lysine (Sigma-Aldrich, cat. no. L5501)
- Oligonucleotides (IDT DNA Technologies) see Table 2.1 for oligonucleotides used to study a 10 amino acid sequence of Hsp90 (DNA sequence: 5' GCTAGTGAAACTTTTGAATTTCAAGCTGAA 3') in pRNDM Hsp82
- Custom bio-informatics software (available from www.labs.umassmed.edu\Bolonlab).

Equipment

- Incubator set to 37°C (Fisher Scientific, Model 655D)
- 1.7 ml microcentrifuge tubes (Sorenson Biosciences, cat. no. 16070)
- Microcentrifuge (Beckman Coulter, Microfuge 18)
- UV trans-illuminator (UVP, Model M-15)
- Razor blades (VWR, cat. no. 55411-050)
- Heatblock set to 42°C (VWR, cat. no. 13259-030)
- Shaking incubator (Infors HT, Multitron Standard)

- Spectrophotometer capable of measuring absorbance at 600 nm. (Cary, 50 UV)
- Thermocycler for PCR (Applied Biosystems, cat. no. 2720)
- -80°C freezer for storage of yeast pellets (Sanyo, cat. no. MDF-U76VC)
- Heat block set at 50°C (VWR, cat. no. 13259-030)
- Autoclave (Brinkmann, cat. no. 023210100)
- 100x15mm Petri dishes (VWR, cat. no. 25384-088)
- 125ml flasks (Corning, cat. no. 29136-048)
- BD Falcon 14ml culture tubes (BD Falcon cat. no.352057)
- Tabletop centrifuge capable of spinning 14ml culture tubes at 3000*g* (Sorvall, Legend RT)
- Electrophoresis power supply (Fisher Scientific, cat. no. FB300Q)
- Agaraose gel system (Hoefer, cat. no. HE33)
- Nanodrop spectrometer (Thermo Scientific, Nanodrop2000)

Reagent Setup

PEG 3350 50% (w/v) solution – dissolve 50 grams of solid powder in water to a final volume of 100 ml. Sterilize by vacuum filtration and store at room temperature for up to one year.

Lithium acetate 1.0 M solution – dissolve 102 grams into water to a final volume of 1 L. Sterilize by vacuum filtration and store at room temperature for up to one year.

Salmon Sperm DNA, 2 mg/ml solution in TAE – dissolve 100 mg of lyophilized powder in 50 ml of TAE. Make 1.0 ml aliquots, place in boiling water bath for 10 min, place on ice for 10 min, and store at -20°C for up to one year.

G418 antibiotic, 250X solution – dissolve 500 mg of G418 in water to a final volume of 10 ml. Filter sterilize and store at -20°C for up to one year.

Kanamycin stock solution – dissolve 250 mg of Kanamycin in 10 ml of water and filter sterilize. Store at -20°C for up to one year.

LB medium -dissolve 10 g of Trytpone, 5 g of Yeast Extract, and 5 g of Sodium Chloride in 1 L of water and autoclave. Store at room temperature for up to one year.

LB+Kanamycin medium – Add 1.2 ml of Kanamycin stock solution to 1 L of LB and store at 4°C for up to one week.

LB+Kanamycin plates - prepare 1 L of LB, add 15 g of Bacto agar and autoclave. Cool to 60°C and add 1.2 ml of Kanamycin stock solution. Pour into petri dishes and cool to solidify. Store at 4 °C for up to two months.

40% Glucose (w/v) – dissolve 400 g of Glucose in water to a final volume of 1 L. Filter sterilize and store at room temperature for up to one year.

YPDA medium – dissolve 10 g of Yeast Extract, 20 g Bacto Peptone, and 0.1 g of Adenine Hemisulphate in 1 L of water. Autoclave and allow to cool to room temperature. Add 50 ml of 40% glucose. Store at 4°C for up to one month.

G418 stock solution – dissolve 500 mg of G418 in water to a final volume of 10 ml. Filter sterilize and store at -20°C for up to one year.

YPDA+G418 medium – add 4 ml of G418 stock solution to 1 L of YPDA. Store at 4°C for up to one month.

2X YPDA medium – dissolve 20 g of Yeast Extract, 40 g Peptone, and 0.1 g of Adenine Hemisulphate in 1 L of water. Autoclave and allow to cool to room temperature. Add 50 ml of 40% glucose. Store at 4°C for up to one month.

Procedure

Generating Plasmid Libraries of Point Mutants ●TIMING ~1 week

1 | Add 5' phosphates to each whole plasmid PCR primer (e.g. 'Vector for' and 'Vector rev' primers in Table 2.1): Dissolve primers in water to 100 µM and setup an individual phosphorylation reaction for each primer as tabulated below. Incubate for 30 min at 37°C. No further purification is necessary.

Component	Amount per reaction (µl)	Final
Water	41	
T4 DNA Ligase Buffer (10X)	5	1X
Primer (100 µM) e.g. 'Vector for' or 'vector rev' in Table 2.1	3	6 µM
T4 Polynucleotide kinase (10 U μl^{-1})	1	10 U

■ **PAUSE POINT** The primers can be stored at -20°C for up to 1 year.

2 | Perform whole-plasmid PCR: Setup a PCR reaction with the following components.

Component	Amount per reaction (µl)	Final
Water	27	
Phusion HF buffer (5X)	10	1X
Phosphorylated Primers (6 µM) e.g. 'Vector for' or 'vector rev' in Table 2.1	5 of each	0.6 µM
dNTP mix (10 µM)	1	0.2 µM
Plasmid template (200 ng μl^{-1})	1	200 ng
Phusion polymerase (2 U μl^{-1})	1	2 U

3 | Run the samples in thermocycler with the following conditions:

Cycle number	Denature	Anneal	Extend
1	95°C, 2 min		
2-16	95°C, 30 s	55°C, 30 s	72°C, 1 min per kb

4 | When the PCR is finished, cool to room temperature (23°C) and add 1 µl DpnI restriction endonuclease to degrade the template plasmid. Incubate at 37°C for 1h.

5 | Run the PCR reaction on an agarose gel, visualize with UV/EtBr, and excise the appropriate fragment and purify. We utilize Zymoclean Gel DNA Recovery Kit (see REAGENTS) and follow the manufacturer's instructions.

? TROUBLESHOOTING

6 | Circularize the gel-purified fragment vector by performing a unimolecular blunt-ended ligation with the following components and incubating at room temperature for 1 h:

Component	**Amount per reaction (µl)**	**Final**
Water	4	
T4 DNA Ligase Buffer (10X)	1	1X
Gel purified PCR product (from step 5)	4	Varies
T4 DNA Ligase (400 U μl^{-1})	1	400 U

7 | Transform the ligation reaction into a cloning strain of *Escherichia coli* (*E. coli*) by mixing 100 µl of competent cells with 5 µl of the ligation reaction and incubating on ice for 15 min. Subsequently, heat-shock the mixture at 42°C for 45 s, cool for one min on ice, add 1 ml of LB broth (stored at room temperature), and then incubate it at 37°C for 1 h. Finally, spread 100 µl of cells onto LB-kanamycin plates, and let colonies grow at 37°C for 16 h.

8 | Pick two individual colonies and grow each in liquid culture (LB-kanamycin) for 16 h at 37°C.

9 | Isolate plasmid DNA using a ZR plasmid miniprep kit and Sanger sequencing. One or both plasmids usually have the appropriate sequence and can be used in subsequent steps.

10 | Prepare cassettes containing saturation mutants. Dissolve the forward and reverse oligonucleotides (e.g. 'Cas for' and 'Cas rev' in Table 2.1) in water to a final concentration of 100 µM. Combine 50 µl of forward and 50 µl of reverse oligonucleotides so that the final cassette concentration is 50 µM.

11 | Anneal cassettes by boiling followed by slow cooling: Boil 1 liter of water, float 100 µl of the cassettes in boiling water, remove the water from heat and allow the entire water bath to cool naturally to ambient temperature (~1 h).

12 | Dilute annealed cassettes to 0.5 µM in water.

13 | Digest the vector to generate cohesive ends complementary to the cassettes. Set up a BsaI digest as tabulated below and incubate at 50°C for 2 h.

Component	**Amount per reaction (µl)**	**Final**
Plasmid from step 9 (200 ng μl^{-1})	3	600 ng
NEB buffer 3 (10X)	5	1X

Bovine Serum Albumin (10 mg/ml – sterile filtered)	0.5	0.1 mg/ml
Water	36.5	
BsaI enzyme (10 U μl^{-1})	5	50 U

▲ **CRITICAL STEP** Sequential digestion with BsaI followed by a second enzyme that cuts between the BsaI sites (i.e. SphI in Figure 2.2) reduces undesired ligation products and improves library quality.

14 | Allow sample to cool to room temperature. Add 1 µl SphI enzyme and incubate at 37°C for 1 h.

15 | Column-purify the digested plasmid using a Zymoclean gel DNA recovery kit. The small BsaI and SphI fragments will not bind efficiently to silica columns, thereby reducing background ligation products.

16 | Setup a separate ligation reaction for each cassette containing the following reagents and incubate at room temperature for 1 h.

Component	**Amount per reaction (µl)**	**Final**
Digested plasmid from step 15 (~10 nM)	2	~ 1 nM
Annealed cassette from step 12 (0.5 µM)	2	100 nM
T4 DNA Ligase buffer (10X)	1	1X
Water	4	
T4 DNA Ligase (400 U μl^{-1})	1	400 U

17 | Place tubes from step 16 into an ice bath for 5 min.

18 | Add 100 µl of chemically competent *E. coli* to each tube. Incubate the tubes on ice for 15 min.

19 | Place the tubes in a 42°C water bath for 45 s, and then place them back on ice for 1 min. Add 1 ml of LB broth to each tube (maintained at room temperature) and incubate the tubes at 37°C for 1 h.

20 | To analyze transformation efficiency, plate 10 µl of the transformed *E. coli* onto selective plates (e.g LB+kanamycin).

▲ CRITICAL STEP The library size of a single randomized codon is 64. The probability of sampling each possible library member is related to the number of transformants in this step. A good rule of thumb is to have tenfold or greater coverage, meaning 640 or more total transformants and at least six colonies from the 10 µl that was plated. This procedure routinely produces 2,000 – 8,000 total transformants.

?TROUBLESHOOTING

21 | Inoculate the remaining 990 µl of recovery mixture from step 20 into a sterile flask containing 10 ml of selective liquid growth medium (e.g. LB+kanamycin) and grow the cultures overnight at 37°C on an orbital shaker at 180 r.p.m.

22 | After overnight growth of the cultures, prepare plasmid libraries. Libraries can be readily combined at this step. To prepare a library of ten different amino acid positions,

combine equal volumes of saturated culture for each position and prepare plasmid DNA from this combined culture using a ZR plasmid miniprep kit. We typically prepare a miniprep from 3 ml of culture and discard the remaining culture. We have found that growing cultures larger than 3 ml from bulk transformations is necessary for consistent yields in DNA preparations.

▲ **CRITICAL STEP** To assess the quality of the library, it is useful to prepare at least one library with a single randomized codon. Sanger sequencing of this sample should show incorporation of all four nucleotides at the randomized codon and homogeneous sequencing at all other positions.

Generating Libraries of Yeast ●TIMING ~1 week

23 | From a frozen stock, streak out the conditional yeast strain to be transformed at least 72 h before transformation. The media used must be permissive to your strain and will vary depending on the conditional strain used. Throughout this protocol we will provide example media that assume the use of a temperature sensitive strain[202]. This strain can be propagated on YPDA plates at 30°C.

24 | Allow individual colonies to grow to between 1 and 2 mm in diameter. Twenty hours before transformation, inoculate a single yeast colony into 3 ml of appropriate liquid medium (e.g., YPDA). Grow cultures overnight on an orbital rotator set at 180 r.p.m., at the appropriate permissive temperature for the conditional strain.

25 | When overnight cultures reach near-saturation, determine cell density by counting with a hemocytometer. Add 2.5 X 10^8 cells to a flask containing 50 ml of rich medium (e.g., 2x YPDA), which is sufficient for up to ten transformations. Incubate the cells on an orbital shaker at 180 r.p.m. for at least two cell doubling times, which is from 4 to 8 h depending on the strain.

26 | Prepare competent yeast from the cultures using the lithium acetate method[215, 216].

▲ CRITICAL STEP It is important to use freshly prepared competent cells in order to achieve efficient transformation.

27 | Add 1 µg of library plasmid DNA (from Step 22) to 360 µl of competent yeast and vortex briefly to mix.

▲ CRITICAL STEP Plasmid transformation into yeast should be performed to maximize independent transformed cells while minimizing the number of cells that acquire more than one plasmid[217].

▲ CRITICAL STEP In addition to the transformation of plasmids with point mutant libraries, you should also transform a negative control (vector without the gene of interest), as well as a positive control (vector with a wild type copy of the gene of interest). These controls enable you to monitor selection pressure in your experiment. When switched to selective conditions, the negative control strain should stop growing and the positive control should continue to grow robustly.

28 | Incubate the yeast-DNA mixture while rocking at room temperature for 30 min, and transfer it to a 42°C water bath for 30 min.

29 | Pellet the cells at 6,000*g* for 1 min at room temperature. Discard supernatant and re-suspend the cells in 1 ml of permissive medium (e.g., YPDA).

▲ CRITICAL STEP For the transformation of G418-resistant plasmids it is important to outgrow yeast under permissive conditions for at least 6 h at room temperature before exposure to G418 in Step 31.

30 | This step can be done overnight. Resuspend yeast transformation in 5 ml of medium lacking G418. Ampicillin can be added to a final concentration of 0.05 μg ml^{-1} at this stage and all subsequent yeast growth steps to hinder bacterial contamination. Grow at 25°C for 6-18 h to allow transformed cells to develop antibiotic resistance to G418.

31 | Spread 50 ul of each yeast transformation onto a plate containing G418. Incubate plates at 30°C for 48-72 h. A library of single codon variants for a 10-aa region contains 640 possible variants. Tenfold coverage or better is desired for sampling and represents 6,400 independent yeast transformants. Typical yeast plasmid transformations yield 20,000 – 100,000 independent transformants.

32 | Take the remaining yeast transformation from step 30 (~4.95 ml) and pellet at 3,000*g* for 5 min at 4°C. Aspirate supernatant and resuspend the pellet in 15 ml of permissive medium. Repeat for a total of five washes.

▲ **CRITICAL STEP** Extracellular plasmid will contribute noise in subsequent analyses and should be thoroughly washed away.

33 | Add the washed cells to 50 ml of sterile culture medium (e.g., YPDA) with G418 under otherwise permissive conditions.

Bulk Yeast Competitions ●TIMING ~1 week

34 | By using a cuvette (1 cm path length) in the Cary spectrophotometer (Cary, 50 UV), measure the optical density of the yeast cultures at 600 nm (OD_{600}) immediately after inoculation and record the measurement. Measure the OD_{600} periodically (e.g. every 12 h) to determine when the culture enters mid-logarithmic growth (usually between 12 and 48 h). When the culture enters mid-log phase (OD_{600}=0.4-1), dilute it as needed into fresh medium to maintain an OD_{600} between 0.1 and 1. Maintain cultures in log growth for a total of at least 48 h, targeting a final OD_{600} of 0.8.

? TROUBLESHOOTING

35 | Collect ~20 ml of cells of OD_{600} = 1.0 and place them in a 50-ml conical tube. This sample represents your yeast library before selection for mutational function. Adjust the

collected volume relative to the actual measured OD_{600}. For example, if the OD_{600} is 0.5, collect 40 ml of cells. Centrifuge the collected cells for 5 min at 3,000*g* at 4°C. Aspirate off the supernatant and wash with 25 ml of water. Centrifuge again, aspirate off supernatant, and then store pellet at -80°C.

36 | Pellet the remaining culture and resuspend in medium conditions that select for the function of the library gene. For example, if you are using a temperature sensitive strain[202], transfer the culture to the nonpermissive temperature. If you are using a shutoff strain[218] with your library constitutively expressed, transfer the culture to shutoff conditions. Record the OD_{600} every 2 h for the initial 12-h period and then every 8 h. Collect samples as described in Step 35 every 3-4 h for the first 12 h of growth and then every 8 h thereafter. Dilute samples to maintain the OD_{600} between 0.05 and 1. Continue growth experiments for about 20 generations of the wild type control. Growth of the negative control should stall. This time course has provided useful results for the analysis of essential yeast genes, including Hsp90, under multiple growth conditions in our laboratory. This time course could be adjusted to account for the desired fitness resolution (i.e., shorter time courses and fewer points would result in less-precise fitness measurements, but they could be used to distinguish null mutants from viable mutants), as well as for potential gene-specific and condition-specific effects.

▲ CRITICAL STEP Record all dilutions to accurately generate a growth curve for library, positive control, and negative control cells. During dilutions, always pass at least 1×10^7 cells to avoid population bottlenecks.

Preparation of DNA from yeast competitions ●TIMING ~1 week

37 | Remove the yeast pellets from -80°C freezer and resuspend them in 200 µl of P1 buffer containing RNAse (ZR plasmid miniprep kit).

38 | Add 5 µl of Zymolyase (150 units ml^{-1}) to each resuspended pellet and mix by pipetting. Incubate for 1.5 h at 37°C.

39 | Add 300 µl of P2 buffer to the suspension and invert ten times to mix. Incubate at room temperature for 5 min.

40 | Add 420 µl of P3 buffer. Invert ten times to mix.

41 | Centrifuge for 10 min at 18,000*g* at room temperature.

42 | Purify the DNA from the supernatant using a silica column.

43 | PCR amplify the DNA using primers specific to the library version of the gene (e.g., 'PCR1a for' and 'PCR1a rev', Table 2.1) and purify the resulting product on an agarose gel. Performing this step reduces sequencing of the conditional copy of the gene (i.e., the temperature-sensitive or shutoff version), which would otherwise be the dominant read

in the sequencing reaction. This can be accomplished using primers targeted to regions upstream and downstream of the coding region that are unique to the library plasmid.

Component	Amount per reaction (µl)	Final
Water	27	
Phusion HF buffer (5X)	10	1X
Primers (50 µM) e.g. 'PCR1a for' and 'PCR1a rev' in Table 2.1	0.5 of each	0.5 µM
dNTP mix (10 µM)	1	0.2 µM
Template DNA (from step 42)	10	Varies
Phusion polymerase (2 U μl^{-1})	1	2 U

Cycle number	Denature	Anneal	Extend
1	95°C, 2 min		
2 to ~21	95°C, 30 s	55°C, 30 s	72°C, 1 min per kb

▲ **CRITICAL STEP** To limit errors that contribute noise to subsequent fitness analyses, use a high-fidelity polymerase and minimizing PCR cycles. Typically, 18-22 cycles are sufficient to produce a strong PCR product at this stage, and care should be taken to avoid unnecessary PCR cycles throughout the rest of the protocol. To assess processing errors, include a control sample at this stage, consisting of a plasmid of homogeneous sequence. For example, use a plasmid encoding the wild type gene and perform the same PCR steps and manipulations. By using this control, we have found that the number of misreads from the entire processing procedure is compatible with reproducible fitness measurements that correlate with traditional fitness analyses of individual mutants[202].

44 | Perform PCR to add MmeI restriction site and 3' Illumina universal primer sequence (Figure 2.3) and purify the resulting product on a silica column.

Component	Amount per reaction (µl)	Final
Water	27	
Phusion HF buffer (5X)	10	1X
Primers (50 µM) e.g. 'PCR1b for' and PCR1b rev' in Table 2.1	0.5 of each	0.5 µM
dNTP mix (10 µM)	1	0.2 µM
Template (PCR product from step 43)	10	varies
Phusion polymerase (2 U μl^{-1})	1	2 U

Cycle number	Denature	Anneal	Extend
1	95°C, 2 min		
2 to ~11	95°C, 30 s	55°C, 30 s	72°C, 30 s

▲ **CRITICAL STEP** Typically 8-12 cycles are sufficient to produce a strong PCR product at this stage.

45 | Digest the PCR product from Step 44 with MmeI enzyme using the following setup. Incubate the PCR product at 37°C for 1 h, and then heat inactivate it at 80°C for 20 min.

Component	Amount per reaction (µl)	Final
PCR product (20 ng μl^{-1}, from step 44)	10	200 ng
NEB buffer 4 (10X)	2	1X
SAM (1 mM)	1	50 µM
Water	5	
MmeI enzyme (2 U μl^{-1})	2	4 U

▲ **CRITICAL STEP** Freshly prepare 1 mM SAM in water by diluting the concentrated stock solution. SAM is unstable in water and the 1 mM solution should be used immediately.

46 | Ligate adapters containing a binding site for 5' universal deep-sequencing primers and a barcode to the MmeI-digested DNA. Mix the components tabulated below and

incubate the mixture at room temperature for 30 min, and then heat-inactivate it at 65°C for 10 min.

Component	Amount per reaction (µl)	Final
MmeI digested DNA (10 ng μl^{-1}, from step 45)	15	150 ng
T4 DNA Ligase buffer (10X)	2	1X
Adapter (6 µM) e.g. 'Adapter for' and 'Adapter rev' in Table 2.1	2	600 nM
T4 DNA Ligase (400 U μl^{-1})	1	400 U

▲ **CRITICAL STEP** Use adapters whose overhangs are complementary to the overhangs from the MmeI digestion in Step 45. If you are planning to pool samples for sequencing, use bar codes that differ by at least 2 nucleotides from all other bar codes in order to minimize bar code switching from misreads.

47 | Separate the ligation reaction on an agarose gel, excise the ligated band, and purify it on a silica column.

▲ **CRITICAL STEP** At this step, the goal is to deplete adapter dimers from your sample. These dimers will readily separate from the product of interest. However, the MmeI digestion and adapter ligation reactions typically go to about 70% completion, which produces a complex banding pattern. However, neither the undigested nor unligated products PCR amplify in subsequent steps.

48 | Perform PCR on the gel-purified products (from Step 47) with Illumina universal primers. Separate the PCR product on an agarose gel, excise the appropriate band, and then column purify it. This sample is ready for deep sequencing.

Component	Amount per reaction (µl)	Final
Water	26	
Phusion HF buffer (5X)	10	1X
Illumina Universal primers (10 µM)	1 of each	0.2 µM
dNTP mix (10 µM)	1	0.2 µM
Template (from step 47)	10	varies
Phusion polymerase (2 U μl^{-1})	1	2 U

Cycle number	Denature	Anneal	Extend
1	95°C, 2 min		
2 to ~11	95°C, 30 s	55°C, 30 s	72°C, 30 s

▲ **CRITICAL STEP** PCR cycles should be minimized. This step typically requires 8-14 cycles of PCR. Samples representing different time-points can be pooled if they are distinctly bar coded and amplified with similar numbers of PCR cycles.

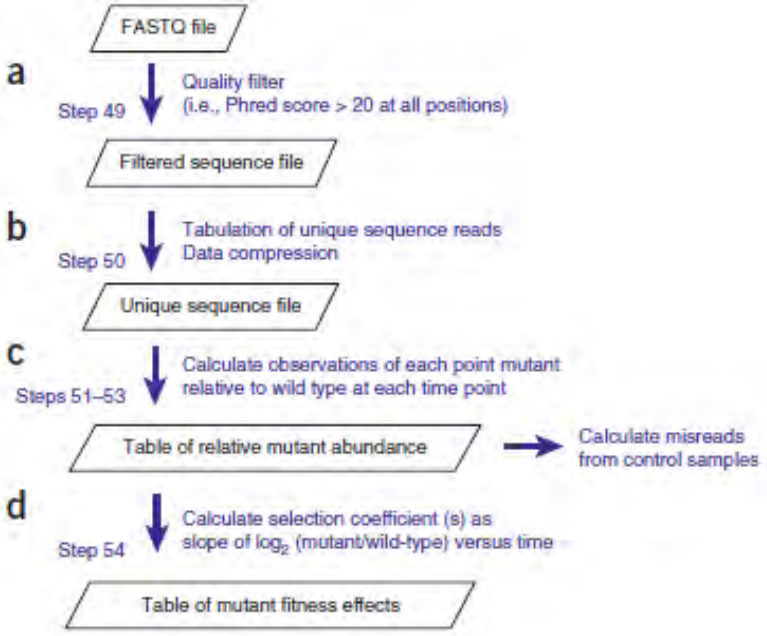
▲ **CRITICAL STEP** Misreads from deep-sequencing are highly variable and should be internally assessed in every run. This can be accomplished by generating a plasmid containing universal deep-sequencing primer sites, which can be utilized to generate a sequencing control with minimal PCR cycles (about eight) and hence minimal sequence heterogeneity. This sample should be mixed in with time point samples at about a 1/100 molar ratio in all analyses. This sample should ideally be distinct from all other samples in the sequencing reaction.

Analyzing the Sequencing Data: *Approximately 1 day*

▲**CRITICAL** All data analysis is performed as outlined (Figure 2.4) with custom programs (http://labs.umassmed.edu/Bolonlab). Knowledge of Linux, Perl, and deep-sequencing are required for the analysis steps.

Figure 2.4 Analysis pipeline for measuring fitness effects of mutations from deep-sequencing data. (a) Sequences that pass quality filtering at all positions in the read are stored in a sequence-only file. (b) The occurrence of each unique sequence read is summed, resulting in a substantial compression of file size. (c) At each time point, calculate the relative abundance of each point mutant in the library. For the control samples, calculate the misread rate per base. (d) Calculate the fitness of each point mutant on the basis of its change in relative abundance over time.

Figure 2.4


FASTQ file
a
Step 49
Quality filter
(i.e., Phred score > 20 at all positions)
Filtered sequence file
b
Step 50
Tabulation of unique sequence reads
Data compression
Unique sequence file
c
Steps 51–53
Calculate observations of each point mutant
relative to wild type at each time point
Table of relative mutant abundance
Calculate misreads
from control samples
d
Step 54
Calculate selection coefficient (s) as
slope of log2 (mutant/wild-type) versus time
Table of mutant fitness effects

49 | *Perform quality filtering*. By using the FASTQ file (the output file from sequence analysis) as input, check the quality score at all nucleotide positions for each read[219]. Define a threshold (we frequently use PHRED score of >20, which corresponds to >99% confidence). Create a new output file that contains sequences for which all base calls pass this threshold.

50 | Enumerate the unique sequence reads and how often they were observed. This serves to compress the data dramatically and speeds subsequent analyses.

51 | Generate an input mask file that describes the experiment, including the correspondence between bar code sequence and time-point, and the wild type sequence.

52 | Tabulate the number of reads of each possible single-codon variant at each time point. Notably, this step removes all sequences that contain apparent codon changes at two or more positions. This filtering step removes many misread events and improves signal-to-noise ratio.

▲ CRITICAL STEP Analyze the internal sequencing and processing controls (described in Steps 44 and 49). Sequencing and PCR/processing errors will appear as mutations in these samples. We have typically observed processing misread rates of ~2 in 1,000 base calls. Taking into account that ~90% of misreads will be filtered out as apparent double mutants in library samples, this translates to an effective noise per base called of ~2 in 10,000. With this misread rate, the vast majority of 36-base reads

(~0.9998^{36} = 99.2%) will be accurate over each base. Because the remaining misread noise is distributed over multiple mutants, the average signal-to-noise ratio for each mutant is ~100:1. The non-linear relationship between per-base misread rates and mutant noise makes it valuable to have low misread noise. If the processing per base misread rate is above 1 in 100, we typically perform a second sequencing analysis. By having an independent control for processing (including all PCR steps) and sequencing (without most PCR steps), it is possible to determine where problems occurred and go back to the appropriate step: re-doing processing and/or sequencing. Of note, misread errors are not random and can vary from run to run. Having internal controls should enable improved error handling in future work. In addition, misread errors are dependent on the sequencing platform being used. We have utilized Illumina sequencing in all of our analyses to date.

? TROUBLESHOOTING

53 | For each possible single-codon variant, calculate the mutant-to-wild type ratio at each time point. Of note, the abundance of wild type sequence reads in our plasmid libraries is typically between 1-4%, about tenfold higher than each point mutant because it is generated independently at each amino acid position. If each codon randomization is completely random, the wild type sequence would be present at 1.5% (1/64). This provides improved counting accuracy of the wild type sequence, which is used as the reference for calculating the relative abundance of all point mutants.

54 | Determine the slope of $\log_2$(mutant/wild type) versus time in wild type generations. This is a direct measure of fitness called the selection coefficient (s). For neutral mutations, s=0; whereas deleterious mutations have s<0 and beneficial mutants s>0. Of note, other groups have developed software for analyzing more complex mutant libraries that include multiple mutations[220, 221].

? TROUBLESHOOTING (Table 2.2)

Step	Problem	Possible reason	Solution
5	Multiple bands	Non-specific primer binding	Increase annealing temperature and/or identify appropriate band by running single-cut plasmid in adjacent lane.
20	Poor transformation efficiency	Ratio of cassette to vector, or mismatched overhangs.	With phosphorylated vector overhangs and non-phosphorylated cassette overhangs a molar ratio of 50 cassette to 1 vector works well. Occassionally (< 5%) BsaI may cut non-canonically – in this case make a new vector with the BsaI sites moved by 1 nucleotide.
34	Yeast with negative control plasmid do not halt growing in selective conditions.	Yeast strain contains another copy of the gene, or the gene is not essential.	Re-check or re-make conditional strain.
52	High noise level from sequencing misreads	Poor quality filtering and/or poor sequencing data.	Re-run the analysis with a more stringent quality cutoff or re-sequence.

Anticipated Results

Generating plasmid libraries of point mutants

The cassette ligation strategy generally produces 2,000-8,000 transformants. Background transformants (from ligations without any insert cassette) can vary depending on the overhangs left after BsaI digestion. The perfect match between the cassette and vector overhangs usually outcompetes background vector self-ligation. For this reason control transformants (from ligations performed without any insert) do not necessarily indicate a problem. Sanger sequencing of an individually randomized codon library is required to assess quality. If the Sanger chromatogram shows all four bases at randomized positions and homogeneous sequence before and after, then the library is appropriate for further use.

Generating libraries of yeast and bulk competitions

Plasmid transformations into yeast generally produce 20,000-100,000 independent transformants. Upon transfer to conditions that select for mutant function, growth of library cultures typically slow briefly compared to the growth of the positive control culture (because the average mutation in the library is deleterious relative to wild type). Growth of the negative control culture should plateau.

Preparation of DNA for sequencing

All samples should amplify by PCR, digest, and ligate to adapters with similar efficiency.

Analyzing the sequencing data

The quality of deep-sequencing data varies markedly from run-to-run. Internal sequencing and processing controls should be included in every sequencing sample. If necessary, quality filtering should be adjusted so that the internally determined miscall rate is well below the signal (abundance of mutants in the library). Within libraries, internal positive controls (i.e., silent mutations) should have near-neutral fitness effects ($s \approx 0$) and internal negative controls (i.e., stop codons) should have null-like fitness ($s \approx -1$). Because mutants with null-like fitness rapidly decrease in abundance, they can only be observed in early time-points. The switch to selective conditions during these early time points is not perfectly synchronized across all cells in the culture (i.e., in a shutoff experiment, variations in initial protein levels result in variation in shutoff timing in individual cells). This can result in apparent selection coefficients of null mutants that are >-1 (typically <-0.5, though). True nulls (i.e., internal stop codons) can be used to define a range of apparent fitness measurements that correspond to null fitness. Mutants that support yeast growth persist in the culture beyond the point where selection synchronization impacts fitness analysis.

Acknowledgements

This work was supported in part by grants from the National Institutes of Health (R01-GM083038) and the American Cancer Society (RSG-08-17301-GMC) to D.N.A.B.

Chapter III – Experimental Illumination of a Fitness Landscape

This chapter has been published previously as *Hietpas RT, Jensen JD, Bolon DNA. "Experimental illumination of a fitness landscape." Proc Natl Acad Sci U S A. 2011 May 10; 108(19):7896-901.*

The following chapter was a collaborative effort. I and Dr. Daniel N. A. Bolon designed the research and I performed the yeast growth competitions, DNA isolation, sequencing preparation, and binary competitions. I, Dr. Jeffrey D. Jensen and Dr. Daniel N. A. Bolon analyzed the data. I, Dr. Jeffrey D. Jensen, and Dr. Daniel N. A. Bolon prepared the manuscript.

Abstract

The genes of all organisms have been shaped by selective pressures. The relationship between gene sequence and fitness has tremendous implications for understanding both evolutionary processes and functional constraints on the encoded proteins. Here, we have exploited deep sequencing technology to experimentally determine the fitness of all possible individual point mutants under controlled conditions for a nine-amino acid region of Hsp90. Over the past five decades, limited glimpses into the relationship between gene sequence and function have sparked a long debate regarding the distribution, relative proportion and evolutionary significance of deleterious, neutral and advantageous mutations. Our systematic experimental measurement of fitness effects of Hsp90 mutants in yeast, evaluated in the light of existing population genetic theory, are remarkably consistent with a nearly neutral model of molecular evolution.

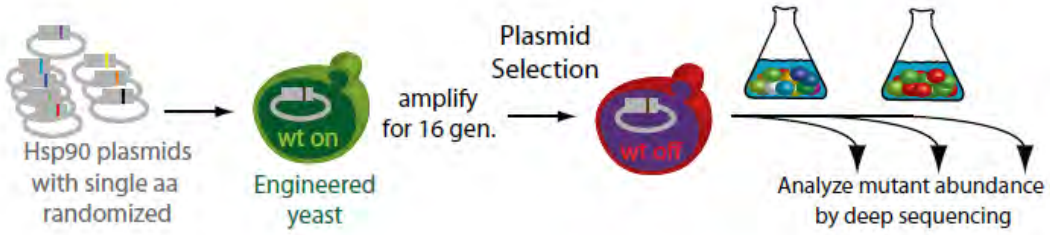
Introduction

The results of greater than 150 years of biological research has demonstrated that selection pressures shape the evolution of organisms[4]. The relationship between gene sequence and selective advantage/disadvantage provides the fundamental link between genotype and fitness. Until now, it had not been feasible to systematically measure this relationship because of the challenge of constructing and monitoring all possible genetic variants. Two classes of experiments have provided glimpses of the fitness landscape and inferences into the relationship between gene sequence and fitness: directed evolution[135, 152, 222] and microbial experimental evolution[207, 223]. In both of these approaches, the fitness landscape can only be inferred - either because the pool of starting mutations is unknown, or because mutational sampling is limited. Thus, the question remains: what does the fitness landscape look like for all possible point mutants?

Determining the fitness landscape of point mutations in a gene is conceptually simple: measure the fitness of organisms with each possible point mutation in a specific gene in an otherwise identical genetic background. To accomplish this in practice, there are two major technical challenges: generating high-quality systematic mutant libraries, and measuring fitness in high-throughput both accurately and with a large dynamic range. To address these challenges, we developed an approach that we call "extremely methodical and parallel investigation of randomized individual codons" (EMPIRIC) fitness (Figure 3.1).

Figure 3.1 EMPIRIC approach to experimentally determine fitness landscapes. Randomized individual codon libraries are introduced into a host cell whose only other copy of the gene is regulatable. The fitness of each individual codon mutation is determined by measuring its abundance in the mixed culture as a function of time under selective conditions.

Figure 3.1


Hsp90 plasmids
with single aa
randomized
wt on
Engineered
yeast
amplify
for 16 gen.
Plasmid
Selection
wt off
Analyze mutant abundance
by deep sequencing

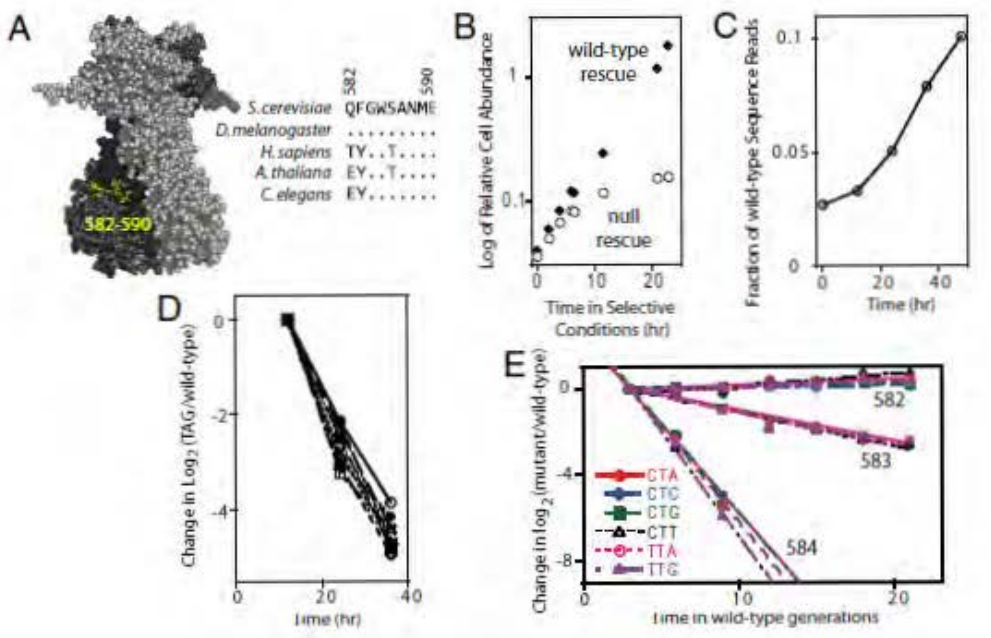
Results

We used the EMPIRIC approach to analyze yeast Hsp90, an essential chaperone in eukaryotes[123] required for the maturation of many kinases[200]. Of note, analyzing an essential gene maximizes the potential fitness range of mutants and the signal of the analysis. The amino acid sequence of Hsp90 is highly conserved among eukaryotes with 45% of the amino acids identical between the human and *S. cerevisiae* proteins. Based on the sequence and structure[171] of Hsp90, we focused on a nine amino acid region that contains a diversity of different amino acids with positions that vary in both their level of phylogenetic conservation among diversely related eukaryotes and their physical environment (solvent exposed and buried) in the structure of Hsp90 (Figure 3.2A). In addition, two solvent-exposed aromatic side chains (Phe583 and Trp585) were structurally intriguing for a chaperone based on their potential to bind to hydrophobic regions on binding partners. The randomization of this nine amino acid region resulted in the parallel analyses of 180 amino acid substitutions and >500 different codon variants – a task that would be daunting by traditional approaches.

Our analysis method monitors plasmid abundance that we expect to parallel with cell growth such that selective pressure begins to impact plasmid abundance at about the same time that it impacts cell growth. When cells with null rescue plasmids were switched to nonpermissive conditions, growth began to retard noticeably after 8 h and was stably slowed after 12 h (Figure 3.2B). At this time we also observed the effects of selective pressure on the relative abundance of our point mutant plasmid library as

Figure 3.2 Hsp90 region analyzed and application of selection pressure to point mutants of Hsp90 in yeast. (A) Positions 592-600 are highlighted in yellow in the dimeric structure of *S. cerevisiae* Hsp90. (B) Growth of an Hsp90 temperature sensitive yeast strain at 36ºC is rescued with a wild type Hsp90 plasmid. (C&D) Deep sequencing analysis of a library of single-codon mutants of Hsp90 from amino acids 582-590 grown in mixed culture. (C) Relative abundance of wild type sequence as a function of time in selective conditions where the only other copy of Hsp90 is inactivated. (D) The ratio of TAG stop codons to wild type codons decreases steeply over time in selective conditions. (E) Observed fitness of leucine synonyms at positions 582, 583, and 584.

Figure 3.2


A
582-590
582
590
S.cerevisiae QFGWSANME
D.melanogaster .....
H.sapiens TY..T....
A.thaliana EY..T....
C.elegans EY.....
B
wild-type rescue
null rescue
Log of Relative Cell Abundance
Time in Selective Conditions (hr)
C
Fraction of wild-type Sequence Reads
Time (hr)
D
Change in Log2 (TAG/wild-type)
Time (hr)
E
Change in log2 (mutant/wild-type)
582
583
584
CTA
CTC
CTG
CTT
TTA
TTG
Time in wild-type generations

monitored by deep sequencing (Figure 3.2C and D). Starting at 12 h, the relative abundance of wild type sequence reads starts to increase consistent with the wild type sequence having better fitness relative to the average point mutant (Figure 3.2C). Because we generate our libraries with mixtures of all four nucleotides at each codon position, stop codons are included in our library and provide an internal monitor of selection pressure. Stop codons at all of the positions that we analyzed rapidly decrease in relative abundance starting at 12 h in selective conditions (Figure 3.2D) consistent with the known requirement of sequences C-terminal to this region for Hsp90 function[188]. From these results, we conclude that our deep sequencing approach is an effective means to monitor selective pressure.

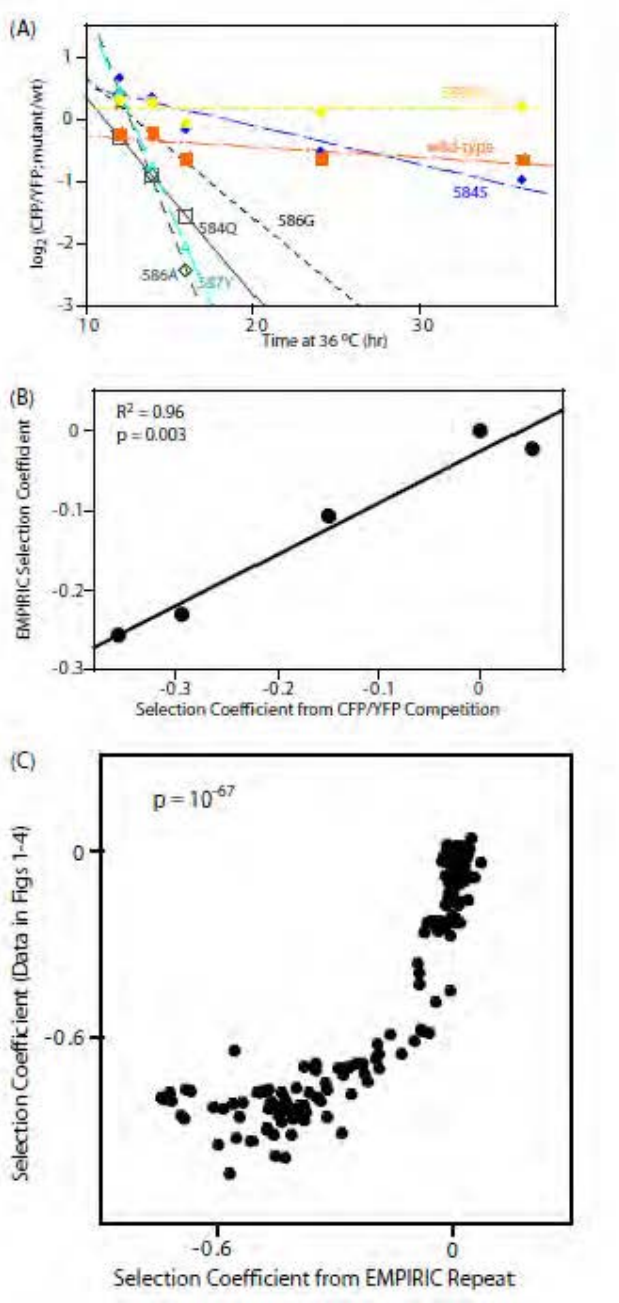
The inclusion of wild type sequences in our libraries serves as an internal benchmark to calculate the competitive fitness of each mutant. Under our experimental conditions, the doubling time of a homogeneous culture of yeast harboring wild type Hsp90 plasmid was 4 h (Figure 3.2B). By measuring the change in the ratio of a mutant to wild type sequence reads as a function of this wild type generation time (Figure 3.2E), we calculate the relative fitness of the mutant as a selection coefficients (s)[99]. Because fitness is related to the change in abundances as a function of time, it does not require equal abundance of each variant at the beginning of the experiment. Thus, biases in the mutational process (i.e., from oligonucleotide synthesis) did not preclude the analysis of fitness of any mutants. The selection coefficients represent the difference in fitness between the mutant and wild type. For yeast, fitness is proportional to the inverse of the

doubling time, and by definition wild type fitness is 1. Thus, a selection coefficient of zero (no change in mutant to wt ratio over time) means that a mutant is as fit as wild type, a negative selection coefficient means that a mutant is less fit than wild type (-1 if a mutant does not support any proliferation), and a positive selection coefficient means that a mutant is more fit than wild type.

We calculated selection coefficients from our EMPIRIC fitness measurements for each codon mutant in our library (Table A1). For six mutants, we compared the EMPIRIC measured fitness effects to those measured by traditional two strain competition using strains with different colored fluorescent proteins (Figure S3.1). EMPIRIC and the bistrain competitions both parse wt-like and null-like mutants similarly, and for strains that persist in the cultures and that are therefore monitored with higher signal, we observe a strong positive correlation (R^2= 0.92, P=0.003 for two-tailed Student's *t* test) between EMPIRIC and biculture fitness measurements. Of note, one advantage of EMPIRIC measurements is that all mutants experience identical environmental conditions because they are physically located in the same flask compared with bistrain fitness competitions where each mutant is grown in separate flasks. To examine the reproducibility of EMPIRIC measurements, we repeated the EMPIRIC experiment (Figure S3.1) and observe a strong correlation (R^2=0.82, P=10^{-67} two-tailed Student's *t* test).

Figure S3.1 Validation of EMPIRIC measurements. A strain with wt Hsp90 labeled with YFP was grown in competition with strains labeled with CFP under identical growth conditions to those in the EMPIRIC analyses. CFP strains included the following Hsp90 genes: wild type, G584S, G584Q, S586A, S586G, A587Y, and E590W (Panel A). For Hsp90 mutants that persist in the CFP/YFP competition and whose fitness could be accurately determined by this method the correlation with EMPIRIC fitness values (Panel B) is very strong ($R^2 = 0.96$) and is statistically significant ($p = 0.003$). The EMPIRIC growth competition and deep sequencing were repeated (Panel C). Fitness measurements from the repeat experiment are based on three time points (corresponding to 0, 3, and 6 generations of selection pressure). There is a clear correlation between the fitness effects measured in these independent experiments (linear relationship has an R^2 of 0.82 - fit not shown and a p-value of 10^{-67} two-tailed student T test). The relationship is also distinctly non-linear, likely indicating slight differences, such as media composition or temperature in the environmental conditions of the repeat experiment. The agreement between these experimental replicates indicates that EMPIRIC measurements are reproducible.

Figure S3.1

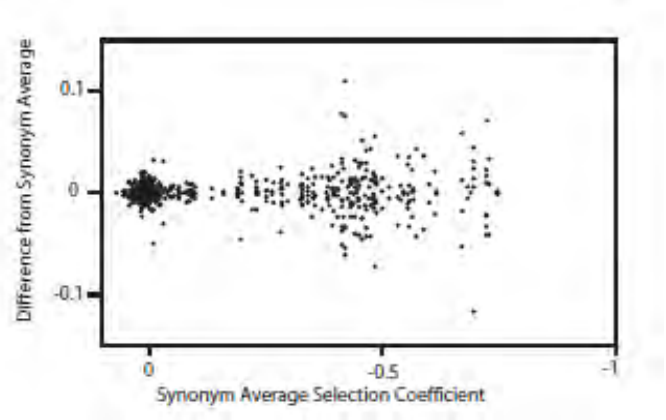

(A)
log2 (CFP/YFP: mutant/wt)
Time at 36 °C (hr)
wild-type
5845
586G
584Q
586A
587Y
(B)
R² = 0.96
p = 0.003
EMPIRIC Selection Coefficient
Selection Coefficient from CFP/YFP Competition
(C)
p = 10-67
Selection Coefficient (Data in Figs 1-4)
Selection Coefficient from EMPIRIC Repeat

The average selection coefficient for all codons for all codons from our EMPIRIC analyses was s=-0.42 ($P=10^{-170}$ using two-tailed Student's t test compared with null hypothesis that the mean s=0), indicating that the average mutation in this region is of deleterious effect. For comparison, the average selection coefficient of all stop codons was -0.75. The fitness of stop codons is greater than a true null (s=-1), indicating that the relative plasmid abundance of these nonsense mutations falls off rapidly with time in selective conditions, but that slow plasmid replication may persist in some fraction of these cells.

We examined selection coefficient differences between synonymous codons, which code for identical protein sequences. Synonymous codon substitutions among homologous genes are widely used in population genetic analysis as a measure of the neutral mutation rate[224] with the underlying assumption that these substitutions do not impact fitness and their dynamics are governed by genetic drift. This assumption is imperfect because species have distinct codon preferences within coding regions, indicating that selective pressure may in fact distinguish between synonymous codons[225]. Our experimental data enabled us to analyze the variation in fitness between synonymous substitutions. We observe increased fitness variability between synonymous substitutions where the average synonym fitness was null-like (Figure S3.2), which is likely caused by sampling noise (due to the low abundance of these codons in the competing culture). To minimize this noise, we calculated the variability among selection coefficients for

Figure S3.2 Variability in experimental fitness among synonymous codons as a function of the fitness of the synonym average. Amino acid substitutions with poor fitness decrease rapidly in abundance and are therefore sequenced with reduced frequency, increasing the noise in their measurement.

Figure S3.2


Difference from Synonym Average
0.1
0
-0.1
0
-0.5
-1
Synonym Average Selection Coefficient

synonymous codons with high fitness (s>-0.05, n=151) as a root mean square deviation (rmsd=0.018 compared to the synonym mean). For comparison we observe an rmsd of 0.35 when all possible substitutions including those that result in an amino acid change are considered. Thus, synonymous substitutions caused fitness changes that pale in magnitude to amino acid changing substitutions consistent with the expectation of neutrality at synonymous sites[226].

We averaged the observed EMPIRIC selection coefficient of synonymous codons to generate fitness profiles of each amino acid at each position (Figure 3.3A and Supporting Figure S3.3). The fitness of the hydrophobic amino acids exhibited less variability within a position than polar amino acids (rmsd of selection coefficients for VILMFYW of 0.15 compared to 0.26 for KRHDENQST). From a fitness perspective, the specific geometry of hydrophobic amino acids had a smaller impact compared to the varied physical properties of polar amino acids. To compare geometrical sensitivity among amino acids with similar physical properties, we compared selection coefficients at each position between the following pairs: D/E, K/R, N/Q, V/I, L/M, W/Y. The three polar pairs differ more than the three hydrophobic pairs (selection coefficient rmsd of 0.23 and 0.10 respectively), indicating that the fitness of polar amino acids is more sensitive to geometry than the fitness of hydrophobic amino acids. The relative insensitivity of amino acid substitutions between hydrophobic amino acids is consistent with the finding that hydrophobic cores of proteins can be efficiently repacked with different hydrophobic sequences[227-229]. All polar amino acids can form hydrogen bonds

Figure 3.3 Amino acid profile in phylogenetic alignment poorly predicts EMPIRIC fitness profile. (A) Heat map representation of the EMPIRIC fitness profile with the wild type amino acids outlined in red. (B) Information content logos generated from amino acids with wt-like EMPIRIC fitness (B) and a phylogenetic alignment of 448 Hsp90 protein sequences (C). (D) The dominant genetic code is optimized for single-base substitutions between codons with wt-like fitness compared to randomly simulated codes (+2.4σ). (E) Distribution of tolerated and phylogenetically observed amino acids expressed as an entropy where zero corresponds to a frozen position and 3 corresponds to unrestrained positions. (F) Relationship between tolerated amino acid profile from EMPIRIC fitness measurements and phylogenetic alignment. Linear regression indicates a very weak correlation with R^2 of 0.15. (G) EMPIRIC fitness analyzed as a function of amino acid prevalence in the phylogenetic alignment. Most amino acids observed in the phylogenetic alignment are well tolerated when made in the yeast homologue.

Figure 3.3

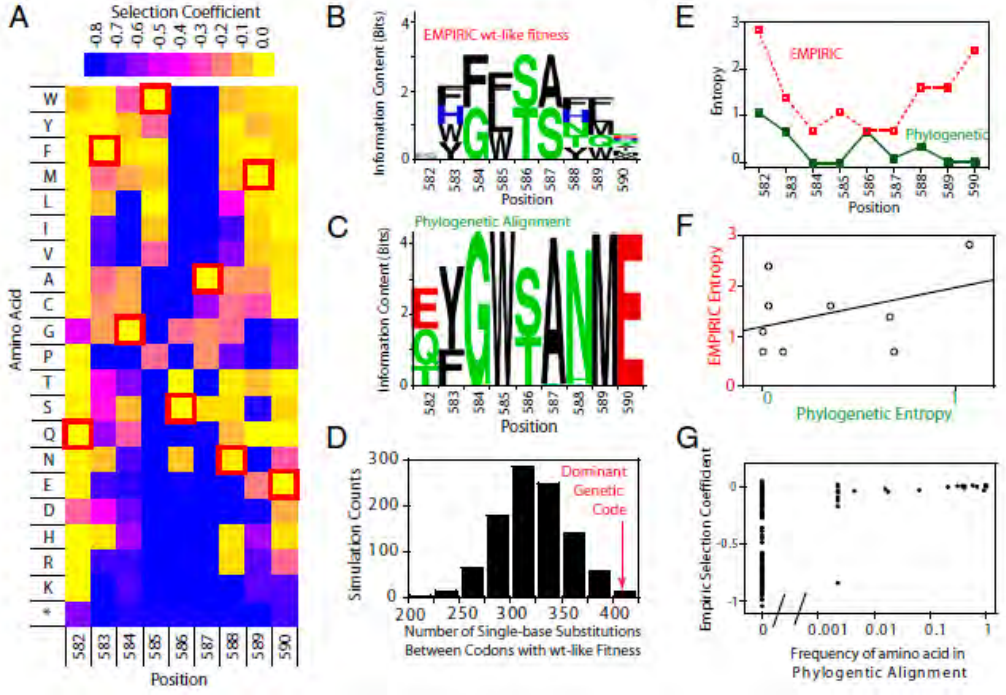
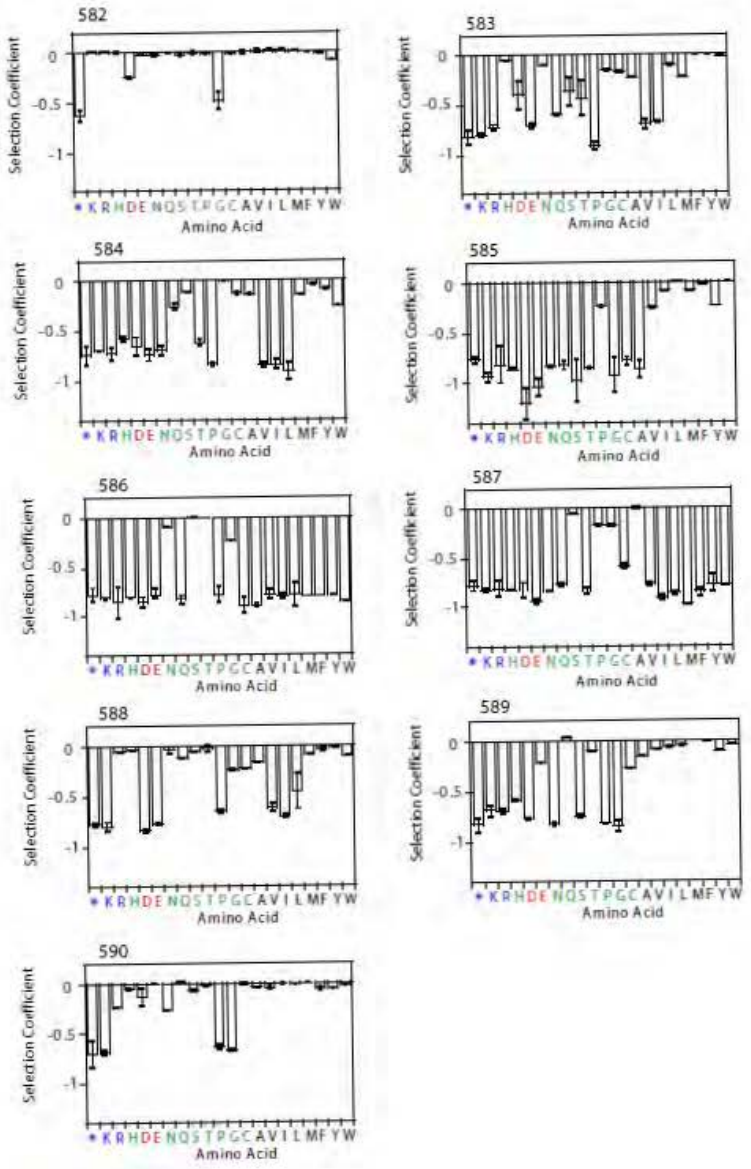

A
Selection Coefficient
-0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0
Amino Acid
W Y F M L I V A C G P T S Q N E D H R K *
582 583 584 585 586 587 588 589 590
Position
B
EMPIRIC wt-like fitness
Information Content (Bits)
Position
C
Phylogenetic Alignment
Information Content (Bits)
Position
D
Simulation Counts
Dominant Genetic Code
Number of Single-base Substitutions Between Codons with wt-like Fitness
E
EMPIRIC
Phylogenetic
Entropy
Position
F
EMPIRIC Entropy
Phylogenetic Entropy
G
Empiric Selection Coefficient
Frequency of amino acid in Phylogentic Alignment

Figure S3.3 Selection coefficients measured for each amino acid substitution at positions 582-590 of Hsp90. The wild type sequence is: Q582, F583, G584, W585, S586, A587, N588, M589, E590.

Figure S3.3

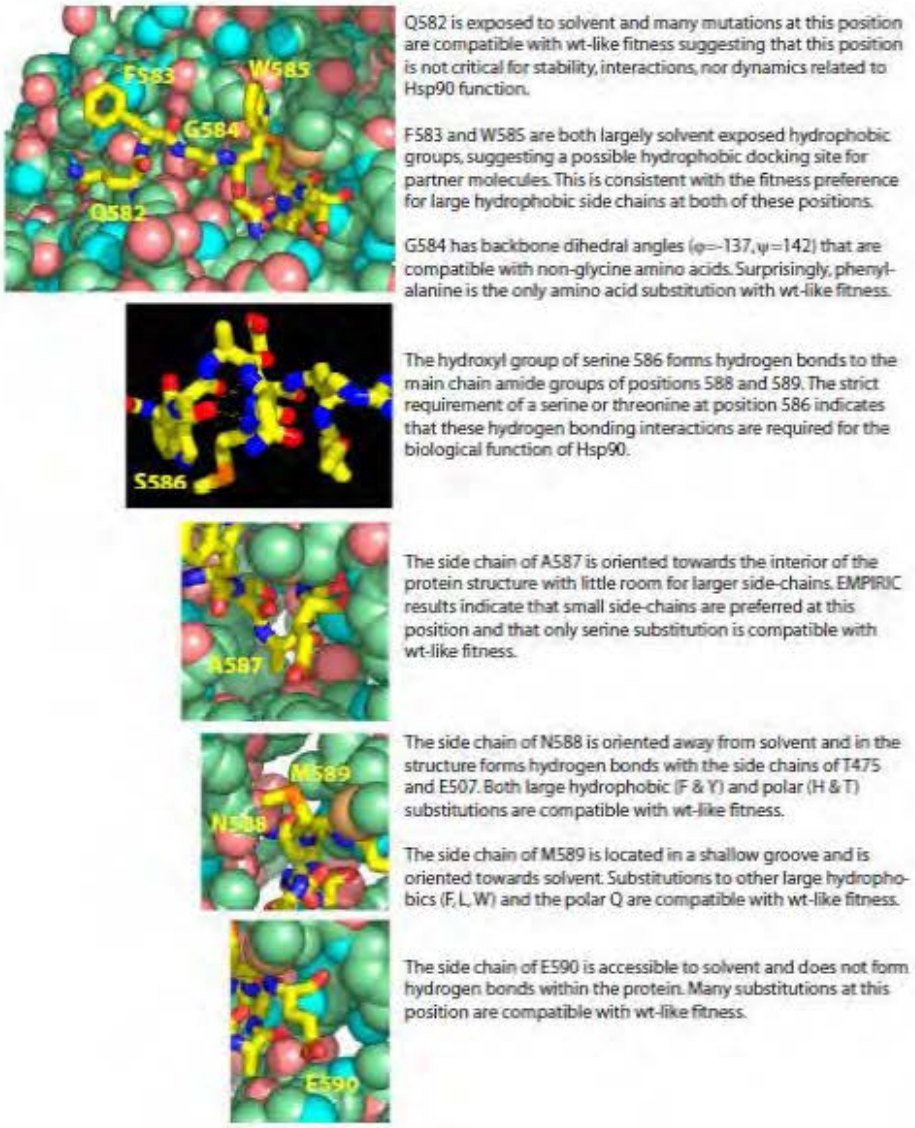

582
583
584
585
586
587
588
589
590
Selection Coefficient
0
-0.5
-1
* K R H D E N Q S T P G C A V I L M F Y W
Amino Acid

whose thermodynamic energy varies sharply with distance and angle[230], providing a rationale for the greater variability of the fitness of polar amino acids.

Based on the distribution of observed fitness effects, we classified mutants as wt-like if they had a selection coefficient within 5% of wild type or better (s>-0.05). We chose this cutoff value because it is three times the rmsd between synonyms and, thus, represents a 99% confidence interval. We generated a logo of amino acids with wt-like fitness to analyze the patterns for underlying physical requirements for fitness at each position (Figure 3.3B). Two of the nine positions analyzed exhibited a clear and consistent physical requirement for wt-like fitness: large hydrophobic side chains for position 585 and a gamma-hydroxyl group for position 586. The physical properties for the preferred amino acids at these positions enable mechanistic predictions. The fitness preference for large hydrophobic amino acids of varied geometry (tryptophan, leucine, phenylalanine) at position 585, which is located on the surface of the Hsp90 structure, is consistent with involvement in loose contacts with hydrophobic partner molecules. The preference for only serine or threonine at position 586 indicates that the hydroxyl group common to both of these amino acids is important for function. In the structure of Hsp90, this hydroxyl group forms hydrogen bonds to two main-chain amide groups (Figure S3.4). Although many hydrogen bonds are not important for protein function[231], our fitness measurements indicate that the hydrogen bonds formed at position 586 are critical for the function of Hsp90. Indeed, although tremendous strides have been made in understanding the relationship between protein structure and stability[228, 232], the ability to

Figure S3.4 Structural images of the amino acids in yeast Hsp90 analyzed by EMPIRIC.

Figure S3.4



Q582 is exposed to solvent and many mutations at this position are compatible with wt-like fitness suggesting that this position is not critical for stability, interactions, nor dynamics related to Hsp90 function.

F583 and W585 are both largely solvent exposed hydrophobic groups, suggesting a possible hydrophobic docking site for partner molecules. This is consistent with the fitness preference for large hydrophobic side chains at both of these positions.

G584 has backbone dihedral angles (φ=-137, ψ=142) that are compatible with non-glycine amino acids. Surprisingly, phenylalanine is the only amino acid substitution with wt-like fitness.

The hydroxyl group of serine 586 forms hydrogen bonds to the main chain amide groups of positions 588 and 589. The strict requirement of a serine or threonine at position 586 indicates that these hydrogen bonding interactions are required for the biological function of Hsp90.

The side chain of A587 is oriented towards the interior of the protein structure with little room for larger side-chains. EMPIRIC results indicate that small side-chains are preferred at this position and that only serine substitution is compatible with wt-like fitness.

The side chain of N588 is oriented away from solvent and in the structure forms hydrogen bonds with the side chains of T475 and E507. Both large hydrophobic (F & Y) and polar (H & T) substitutions are compatible with wt-like fitness.

The side chain of M589 is located in a shallow groove and is oriented towards solvent. Substitutions to other large hydrophobics (F, L, W) and the polar Q are compatible with wt-like fitness.

The side chain of E590 is accessible to solvent and does not form hydrogen bonds within the protein. Many substitutions at this position are compatible with wt-like fitness.

predict from structure the most important stabilizing contacts remains an unmet challenge. EMPIRIC fitness measurements provide a high-throughput approach to identify these important interactions experimentally and, hence, a route to develop and train predictive algorithms with improved accuracy.

Most of the other positions analyzed exhibit a preference for amino acids with varied physical properties. For example, at position 584, both glycine (the wild type amino acid) and phenylalanine result in wt-like fitness. These amino acids differ dramatically in their physical properties: phenylalanine is large and hydrophobic, and glycine is the smallest amino acid and imparts flexibility on the protein main-chain. Despite their disparate physical properties, these two amino acids are clearly distinguished in fitness from all others. This type of physical plasticity illustrates the degenerate relationship between physics and biology: biology is governed by physical interactions, but biological requirements can have multiple physical solutions. The observed absence of phenylalanine at position 584 in a broad phylogenetic alignment (Figure 3.3C) is consistent with the genetic code requiring two base substitutions to make this amino acid transition and the deleterious fitness effects of any of the single base substitutions.

Indeed, the fitness landscape combined with the genetic code may have broad impacts on evolutionary processes. The EMPIRIC approach provides a long-sought route (via larger datasets) to accurately examine the influence of the genetic code on evolution.

For example, it makes it possible to determine whether the dominant genetic code is optimized for sampling evolutionarily neutral/favorable mutations. To address this issue in our dataset of the fitness effects of >500 codon replacements, we counted the number of single-base substitutions that result in transitions between two codons with wt-like fitness for the dominant genetic code and for 1,000 randomly simulated genetic codes (Figure 3.3E). We find that the genetic code is highly optimized (+2.4 σ) to favor single-base substitutions between codons with wt-like fitness compared with randomly generated codes as predicted from theoretical considerations of amino acid similarity[233]. Thus, the genetic code generally permits single-base substitution pathways between codons with wt-like fitness.

To assess the EMPIRIC fitness profile against the evolutionary record, we compared our experimental results against the Hsp90 species tree (Figure 3.3C). For almost every position, the amino acid entropy is higher for EMPIRIC fitness (Figure 3.3E), indicating that more amino acid substitutions are compatible with high fitness in yeast Hsp90 than are observed in the phylogenetic alignment of Hsp90. Indeed, the relative amino acid entropy from the phylogenetic alignment was a poor predictor of the EMPIRIC entropy (Figure 3.3F). The number and distribution of substitutions in the phylogenetic alignment did not accurately indicate the number of amino acids that would be compatible with high fitness experimentally. Many factors could contribute to this observation, including distinct fitness profiles under environmental conditions experienced in natural selection, and fitness differences beyond our ability to differentiate

resulting in meaningful selection pressures in nature. Importantly, the genes in the phylogenetic alignment vary widely in their codon usage and hence their nucleotide sequence (Figure S3.5), indicating that mutational sampling occurred in this region and were subject to distinct evolutionary pressures in different organisms owing to varying selection intensities and/or effective population sizes.

Although the absence of a substitution in the phylogenetic alignment was a poor prognosticator of fitness effects, we find that all 17 amino acid substitutions observed at least twice in the phylogenetic alignment had wt-like ($s>-0.05$) experimental fitness (Figure 3.3G). We do note that five amino acid substitutions that were observed only once in the phylogenetic alignment fall below this fitness cutoff, but only one is null-like ($s<-0.5$). In contrast, of the amino acid substitutions absent from the phylogenetic alignment, only 20% had wt-like experimental fitness (Figure S3.6). Thus, the presence of an amino acid in a phylogenetic alignment was predictive that the corresponding point mutation in the yeast protein will be biochemically functional and evolutionarily nearly neutral. These observations indicate the important role of drift in the fixation of equivalent substitutions, and highlight the dominant role of purifying selection in suppressing deleterious fixations [24].

Figure S3.5 Nucleotide conservation (Panel A) among Hsp90 genes from an evolutionarily broad distribution of eukaryotes compared to protein alignment (Panel B) from the same region (amino acids 582-590). The variation observed at the nucleotide level indicates that mutational sampling has occurred within this dataset of Hsp90 sequences. Protein alignment from six species (*C. posadasii, A. irradians, C. japonica, E. gracilis, B. rapa,* and *S. cerevisiae*) separated by similar evolutionary distance (Panel C). The similarity of the amino acid profiles in Panels B and C indicate that the parental set of sequences is representative of the diversity of Hsp90 sequences in natural populations.

Figure S3.5

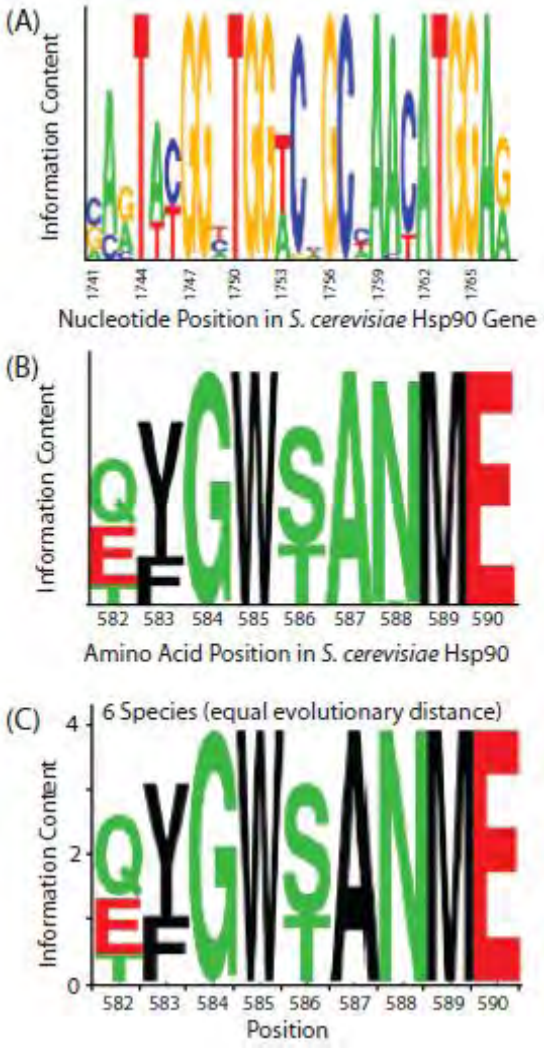
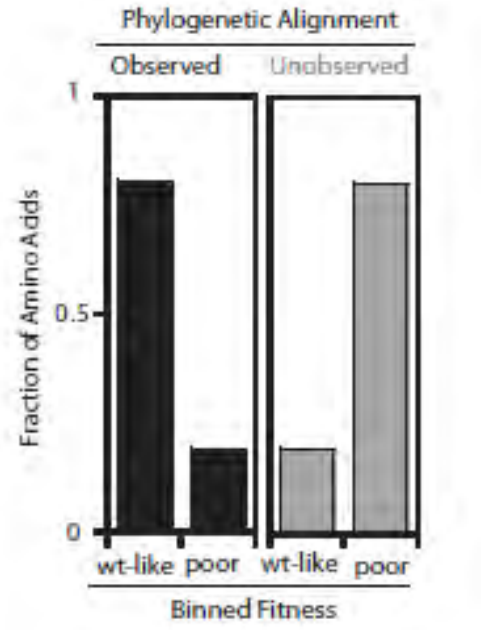

(A)
Information Content
1741
1744
1747
1750
1753
1756
1759
1762
1765
Nucleotide Position in S. cerevisiae Hsp90 Gene
(B)
Information Content
582 583 584 585 586 587 588 589 590
Amino Acid Position in S. cerevisiae Hsp90
(C)
6 Species (equal evolutionary distance)
Information Content
4
2
0
582 583 584 585 586 587 588 589 590
Position

Figure S3.6 EMPIRIC fitness of amino acids both observed and unobserved in a wide phylogenetic alignment. Fraction of amino acids with EMPIRIC fitness similar to wild type ($s>-0.05$) and poorly fit ($s<-0.05$) parsed by their observation or lack thereof in the phylogenetic alignment.

Figure S3.6

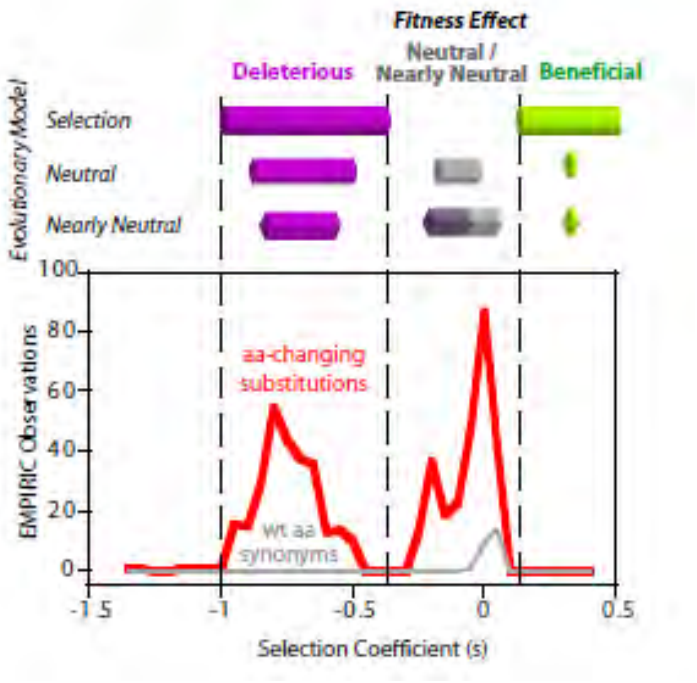

Phylogenetic Alignment
Observed
Unobserved
1
0.5
0
Fraction of Amino Adds
wt-like
poor
wt-like
poor
Binned Fitness

Discussion

The overall distribution of EMPIRIC fitness is bimodal (Figure 3.4) with a clustering of amino acids with fitness similar to wild type (s≈0), and a broader distribution of mutations of deleterious effect. Based on this distribution, we classified the first mode as 'nearly neutral', and the second as 'deleterious'. Evaluating this directly measured distribution of fitness effects for >500 codon variants against the rich field of predictions from population genetics is of tremendous interest. Indeed, understanding this underlying distribution of selection coefficients has been a central focus of evolutionary biology over the past five decades[107]. Contrary to recent inference made in Drosophila favoring models of frequent recurrent and strongly positive selection[107], but similar to inferences from genome-wide analyses of polymorphisms from *S. cerevisiae* and *S. paradoxus*[234], our direct observations in yeast are remarkably consistent with a nearly neutral model of molecular evolution[19], in which a large proportion of new mutations are strongly deleterious and are eliminated via purifying selection, whereas the great majority of remaining mutations are nearly neutral, with dynamics largely dictated by genetic drift (Figure 3.4). Importantly, these initial results pertain to a conserved region of a highly conserved gene under a single growth condition. Examining and comparing the distribution of fitness effects for regions of variable levels of conservation, and under variable growth conditions, will be of extreme interest and should be a subject of future investigation.

Figure 3.4 Distribution of fitness effects of mutations from population genetic models and EMPIRIC measurement. For each model, the relative abundance of each type of fitness effect is illustrated by the length of the bar segment. The experimentally measured selection coefficients were binned in 0.05 increments.

Figure 3.4


Fitness Effect
Deleterious
Neutral / Nearly Neutral
Beneficial
Evolutionary Model
Selection
Neutral
Nearly Neutral
EMPIRIC Observations
100
80
60
40
20
0
aa-changing substitutions
wt aa synonyms
-1.5
-1
-0.5
0
0.5
Selection Coefficient (s)

As with the first techniques of protein electrophoresis allowing biologists to glimpse the extensive protein-level variation[80, 235] spurring the development of the neutral[18, 236] and nearly neutral theories of molecular evolution[19] - as well as the introduction of DNA sequencing technology allowing for inference to be drawn from nucleotide-level variation[237] - the EMPIRIC technique provides another layer of understanding, enabling direct measure of the distribution of selection coefficients by considering each possible point mutation at each site. In doing so, the EMPIRIC approach exposes a broad range of long-standing questions in population genetics to experimental examination including the effects of environmental conditions and genetic background on fitness landscapes.

Specific Materials and Methods

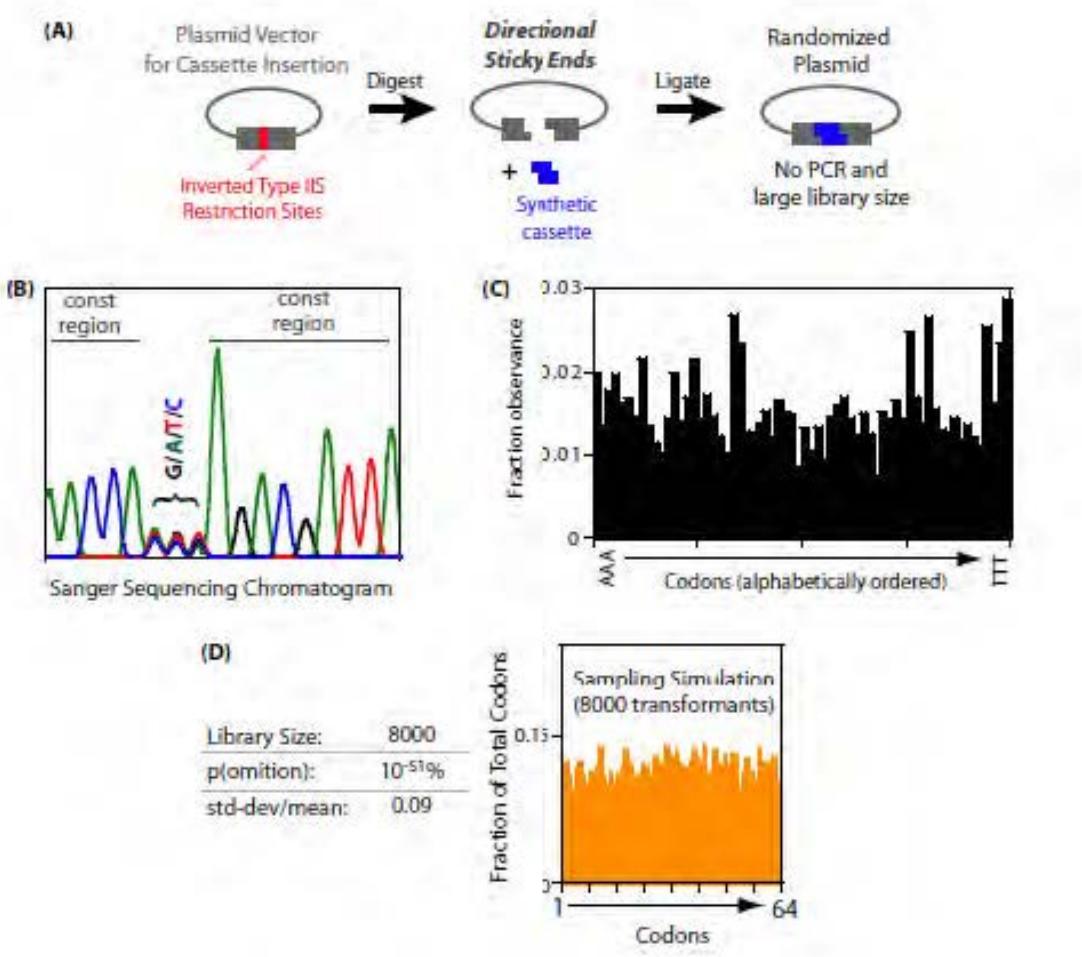
Library construction

In EMPIRIC fitness measurements, high-quality single codon substitution libraries that avoid multiple mutations are important for two reasons: enabling fitness changes to be directly attributed to distinct mutations, and providing a library size that can be accurately monitored in high-throughput. To generate these point mutant libraries, we optimized a cassette ligation strategy to rapidly generate plasmids containing single codons fully randomized to generate all 64 possibilities (Figure S3.7).

We constructed a plasmid with a self-encoded removable fragment (SERF) composed of inverted BsaI restriction sites, such that treatment with this enzyme results in directional sticky ends and removal of the BsaI sites. To reduce background ligation products, an SphI site, introduced between the BsaI sites was also digested in these vectors. We introduced a silent HpaII restriction site adjacent to the randomized region in order to facilitate adapter ligation required for deep sequencing. We generated the SERF vector by PCR from a yeast 417 shuttle plasmid containing a KanMX marker and the HSP82 (systematic name of yeast Hsp90) gene driven by a GPD promoter that expresses Hsp90 to endogenous levels[67]. The HSP82 region of the SERF vector was fully sequenced to ensure the fidelity of the entire gene. Annealed oligonucleotide cassettes with a single codon randomized as NNN were ligated into the vector and transformed

Figure S3.7 Mutagenesis strategy. (A) Vectors are prepared with inverted BsaI sites, such that digestion removes the enzyme binding site and provides directional sticky ends for the efficient ligation of an oligonucleotide cassette. (B) Sanger sequencing chromatogram of an experimentally generated single-codon randomized library. The homogeneous peaks in the constant regions demonstrates that the procedure does not result in detectable levels of undesired ligation products such as vectors lacking insert. (C) Solexa sequencing indicates that all 64 codons are well-sampled for the randomized position. (D) simulated sampling of randomized codons demonstrates that the experimentally achievable sampling of 8,000 independent transformants results in small codon variability from sampling.

Figure S3.7


(A)
Plasmid Vector for Cassette Insertion
Digest
Directional Sticky Ends
Ligate
Randomized Plasmid
Inverted Type IIS Restriction Sites
+
Synthetic cassette
No PCR and large library size
(B)
const region
const region
G/A/T/C
Sanger Sequencing Chromatogram
(C)
Fraction observance
0.03
0.02
0.01
0
AAA
TTT
Codons (alphabetically ordered)
(D)
Library Size: 8000
p(omition): 10^-51%
std-dev/mean: 0.09
Sampling Simulation (8000 transformants)
Fraction of Total Codons
0.15
1
64
Codons

into *Excherichia coli*. Transformants were grown in mixed liquid culture from which plasmid DNA was isolated.

Growth Competition

We used iG170D *S. cerevisiae* cells[179] engineered with a temperature-sensitive chromosomal copy of Hsp90. This yeast strain grows robustly at the permissive temperature of 25°C, rapidly slows growing at the non-permissive temperature of 36°C. Growth at the non-permissive temperature is rescued with a plasmid bearing wild type Hsp90 (Figure 3.2B). Plasmid libraries for each randomized position were transformed into using the lithium acetate method. Transformants were grown in mixed liquid culture with G418 selection at 25°C. After growing to saturation (2 d), cultures were outgrown overnight at 25°C and an equal number of cells for each randomized position combined into a single culture. This culture was then heated in a water bath to 39°C for 15 min to rapidly inactivate G170D Hsp90 and subsequently grown at 36°C. These cultures were diluted every 8 h to maintain a culture density less than $<10^7$ cells/ml. Samples for analysis corresponding to $\approx 2\times10^8$ cells were harvested at different time points. All yeast growth was performed in synthetic dextrose medium with 200 μg/ml G418 and 50 μg/ml ampicillin. Growth rates were determined for a strain with a wild type Hsp90 rescue plasmid under identical conditions (doubling time of 4 h at 36°C).

To validate our EMPIRIC approach, we experimentally determined the fitness effects of six point mutants through binary competition of strains fluorescently labeled

with either CFP or YFP as described. Briefly, the CFP and YFP genes were chromosomally integrated into the iG170D parental yeast strain used in the EMPIRIC measurements. Plasmids containing either the wild type yeast Hsp90 gene or a panel of six mutants was introduced into each strain. Individually, CFP-labeled strains with either wild type or one of the six mutants were grown in competition with a YFP-labeled strain containing wild type Hsp90 under experimental conditions identical to the EMPIRIC experiment, and fluorescent measurements were made as a function of time.

DNA preparation and sequencing

Yeast pellets were lysed with Zymolyase and total DNA was purified using a silica column. A region containing all of the randomized codons was PCR amplified with primers that added a 3' Illumina sequencing primer binding site. After purifying the PCR product on a silica column, a sticky end was created adjacent to the randomized region by digestion with the enzyme HpaII. This sticky end was ligated to an oligonucleotide cassette that included a three base barcode with a Hamming distance[238] of two between any two codes (used to distinguish each time-point sample) and a 5' Illumina sequencing primer binding site. The ligation reactions for each time-point were column purified, combined and amplified in a single reaction with Illumina genomic sequencing primers. This PCR product was separated on an agarose gel and purified prior to 36-base Illumina sequencing.

Data analysis

Illumina sequencing resulted in fastq file from which 2.6×10^7 reads were used for time-dependent analysis based upon stringent accuracy requirements (greater than 99% confidence across all 36 bases). The occurrence of each point mutant at each time-point was tabulated. Ten of the randomized codon sequences resulted in the formation of internal HpaII sites and were removed from further analysis. The ratio of each single-codon mutation relative to the wild type sequence was calculated for each time point on a $\log_2$ scale. Selection coefficients (s) for each mutation were determined as the slope of this ratio to time in wt generations. Selection coefficients for all stop codons were determined from the 12, 24, and 36 h time-points. Selection coefficients of mutants within three standard deviations of the stop codon mean ($s<-0.50$) were considered null-like and analyzed in the same manner. For all other mutations, selection coefficients were determined from the 12, 24, 36, 48, 60, 72, and 84 h time-points. To check for systematic influences of codon bias on fitness, we calculated the fitness difference between a codon and the average for all synonymous codonsand compared this difference to the relative abundance of the codon in highly expressed yeast genes. For this analysis, we chose the 13 genes with the highest experimentally observed expression in *S. cerevisiae*[239]. We averaged over all synonymous codons to calculate amino acid fitness and used the standard deviation to estimate noise in our system. Amino acids were considered wt-like if their amino acid selection coefficient was greater than -0.05. Fitness logos of wt-like amino acids were generated by creating sequences with an equal number of each wt-like amino acid and the program weblogo[240]. A similar logo was produced for the 448

sequences obtained using BLASTP with the full-length yeast Hsp82 protein that aligned fully within this region.

Simulations of alternate genetic codes

Genetic codes were chosen randomly for the twenty amino acids plus stop codons with the requirement that each of these twenty one possible classes be encoded by at least one codon. The EMPIRIC fitness measurements were then searched using these codes for all single-base substitutions between codons with wt-like fitness. The simulation was run for 1000 iterations and compared to the dominant biological code.

Acknowledgements

Thank you to P. Zamore for illuminating the potential of deep sequencing, O. Rando, A. Wong and J. Bowie for thoughtful discussion, and T. Pederson for comments on the manuscript. This work was supported by grants from the National Institutes of Health (R01-GM083038-01A) and the American Cancer Society (RSG-08-17301-GMC) to D.N.A.B.; and grants from the National Science Foundation and an award from the Worcester Foundation to J.D.J.

Chapter IV – Shifting Fitness Landscapes in Response to Altered Environments

This work has been submitted, reviewed, and resubmitted to *Evolution* as *Hietpas RT*, Bank C*, Jensen JD[¥], Bolon DNA[¥]. "Shifting fitness landscapes in response to altered environments"*

This work was a highly collaborative effort between the Daniel N. A. Bolon lab and the Jeffrey D. Jensen lab. I performed the yeast growth competitions, DNA preparation and sequencing, initial sequence analysis, binary competition and qPCR assay, and full experimental repeat. Dr. Claudia Bank performed rigorous mathematical analyses including synonym normalization, FGM fit and analysis, cost of adaptation analysis, and gamma distribution. I, Dr. Claudia Bank, Dr. Jeffrey D. Jensen, and Dr. Daniel N. A. Bolon analyzed the data and prepared the manuscript.

Abstract

The role of adaptation in molecular evolution has been contentious for decades. Here, we shed light on the adaptive potential in *Saccharomyces cerevisiae* by presenting systematic fitness measurements for all possible point mutations in a region of Hsp90 under four environmental conditions. Under elevated salinity, we observe numerous beneficial mutations with growth advantages up to 7% relative to the wild type. All of these beneficial mutations were observed to be associated with high costs of adaptation. We thus demonstrate that an essential protein can harbor adaptive potential upon an environmental challenge, and report a remarkable fit of the data to a version of Fisher's geometric model that focuses on the fitness trade-offs between mutations in different environments.

Introduction

As multiple whole genome projects come to fruition, the presence of whole genome data within and between populations and species has spurred the development of a large class of test statistics aimed at describing the distribution of fitness effects (DFE), or some aspect of the distribution, using polymorphism and divergence data[107, 108, 171, 241-244]. Yet, the abundance of empirical, theoretical, and computational study has led to contentious findings. Turning to perhaps the best-studied organism in population genetics, *Drosophila melanogaster*, estimates are far from consistent, and reconciling results with one another is often challenging. For example, considering polymorphism data, Li & Stephan[171] and Jensen et al[243] estimate the mean beneficial selection coefficient (s) at 0.002, whereas Macpherson *et al.*[107] estimate at 0.01, and Andolfatto[108] at 0.00001.

Considering other avenues for illuminating the DFE apart from statistical inference, we come to the rich field of experimental evolution. These studies have often come in the form of mutation accumulation experiments, with most results recapitulating the basic expectations of Timofeeff-Ressovsky[245] and Muller[246] that most mutations which affect phenotype must be strongly deleterious owing to billions of years of gradual improvements. However, one of the main limitations of such experiments derives from the structure of the experiments themselves. Observations are limited to considering only mutations that happen to spontaneously occur, and are generally focused upon characterizing the rate of accumulation of deleterious variants. In recent years, panels of

tens to hundreds of spontaneous or engineered[247] mutants have been investigated for their impacts on fitness under variable conditions in both bacteria[111, 248-250] and viruses[251, 252]. These studies have provided important insights into the environmental dependence of mutant fitness effects. However, precise measurement of fitness effects for large panels of mutants in an otherwise identical genetic background and under distinct environmental conditions remains an arduous challenge.

In order to overcome many of these limitations, Hietpas *et al.*[202] recently proposed a methodology coined 'extremely methodical and parallel investigation of randomized individual codons' (EMPIRIC), which generates high quality systematic mutant libraries and measures fitness with a large dynamic range. The EMPIRIC approach enables the investigation of all possible point mutations (and their effect relative to wild type) for a given region. This approach has the advantage of mutation accumulation experiments, inasmuch as the DFE of new mutations may be directly observed, rather than inferred from the distribution of segregating and fixed mutations, but has the additional benefit of allowing for a systematic exploration of the full mutational landscape for regions of genes. We developed EMPIRIC to provide precise and reproducible measurements of individual mutations including performing experiments with yeast rapidly expanded from a single colony to provide a homogeneous genetic background, maintaining large populations throughout to minimize stochastic fluctuations in mutant frequencies, and including the wild type sequence in our competitions to provide a direct reference. Thus our approach avoids pitfalls common to previous bulk competitions[253]. Given the

experimental procedure, a given identical mutation (on an identical genetic background) may be readily compared across multiple environments. This benefit however comes at the cost of two distinct trade-offs: 1) we here focus on a specific genomic region, rather than the whole-genome search associated with mutation accumulation experiments, and 2) the experimentally controlled environment is not necessarily related to the complex and variable environmental pressures experienced by natural populations.

We focused our analyses of environmental conditions on two parameters (temperature and salinity) expected to affect the biophysical and biochemical properties of multiple proteins and hence to place distinct pressures on the Hsp90 chaperone system. As its name indicates, heat shock protein 90 is a chaperone that is up regulated in response to elevated temperature[190]. Heat-induced expression of Hsp90 in yeast is required for efficient growth at temperatures above 37ºC[123]. Genome-wide analyses of mRNA abundance indicate that Hsp90 is transiently up regulated upon heat shock at 37ºC with a maximum 8-fold response after 10 min, but in steady state growth at this temperature it is not significantly up regulated[254]. In addition, Hsp90 regulates the global transcriptional response to elevated temperature by binding to Heat Shock Factor 1 (HSF1). Under normal conditions, Hsp90 binds to HSF1 keeping it in an inactive state. Elevated temperature causes Hsp90 to release HSF1 that triggers the main transcriptional response to heat shock[199].

In contrast to Hsp90's central role in the response to heat stress, it plays a more modest role in the response to osmotic stress. For example, elevated salinity (0.7 or 1 M sodium chloride) does not cause a statistically significant change in Hsp90 mRNA levels in yeast over either short or long time periods[254-256]. While Hsp90 is not up regulated in response to elevated salinity, its basal function is required for robust growth under hyperosmotic conditions[257]. Indeed Hsp90 and its co-chaperone Cdc37 are both required for activation of the high osmolarity glycerol pathway[258, 259] that yeast require for growth under conditions of elevated salinity[260]. While Hsp90 function is intimately linked to the yeast response to elevated temperature, it plays a more indirect role in the yeast response to elevated salinity.

In the initial study[202], considering a strongly conserved region of a strongly conserved gene (Hsp90, an essential chaperone in eukaryotes[123] required for the maturation of many kinases[200]), results were quite clear – with a strong bimodal distribution containing mutations nearly equivalent to wild type, mutations that were strongly deleterious, and no observed beneficial mutations. We made the case that this observation was remarkably consistent with Ohta's[19] expectation under the nearly neutral model. However, this first proof-of-principle study may in some ways be regarded as the most likely scenario for replicating the predictions of the nearly neutral model, in that - given the gentle growth conditions and the evolutionary importance of this region - it might be considered unlikely to identify mutations more fit than wild type for the reasons argued by Muller[246].

One valuable approach for interpreting unique data of this type is within the context of Fisher's geometric model (FGM)[12] – a widely utilized framework for interpreting adaptation to novel environments – which yields expectations of the proportion of beneficial to deleterious mutations, as well as the mutational steps sizes and distances characterizing adaptive walks[261, 262]. The comparison of evolutionary models with experimental measures has a rich tradition (e.g. examining the underlying causes of epistasis[263]). For interpreting our data, the FGM provides an intuitive framework for quantifying the cost of adaptation. The model itself is straightforward (see also Figure 1 in Orr 1998[262]). For a given environment, the fitness of an individual is characterized by its position in an *n*-dimensional phenotype space. Fitness is assumed to decrease radially from a single phenotypic optimum. Hence, the Euclidean distance to the optimum determines the fitness of an individual. Random vectors originating from the current phenotype represent new mutations. Those mutations that decrease the distance to the optimum are considered beneficial and can hence contribute to adaptation, whereas those that increase the distance to the optimum are considered deleterious. The better the current phenotype is adapted to the environment, the closer it is to the phenotypic optimum. Thus, for a population near optimum, fewer mutations are expected to be beneficial, because, by solely geometrical arguments, the probability that a randomly occurring mutation will decrease the distance to the optimum becomes lower as the optimum nears. Different environments correspond to differently located optima in phenotype space. The position of each of these optima (relative to one another, and the current phenotype) determines the number and magnitude of beneficial mutations and the

expected cost of adaptation between different environments. Of note, the FGM inherently proposes (potentially high) costs of adaptation: as soon as the phenotypic optimum is relocated upon a change in environment, subsets of the phenotype space arise in which mutations are beneficial in one environment while deleterious in the other.

Continuing to exploit the EMPIRIC high throughput approach, we examine changes in the DFE for a region of the Hsp90 gene in associated with novel selective pressures – here in the form of variations in temperature and salinity. We uniquely address two general and long-standing points of both theoretical and empirical interest: i) the applicability of Fisher's geometric model[12] for populations facing adaptive challenges, and ii) a characterization of the relative cost of adaptation.

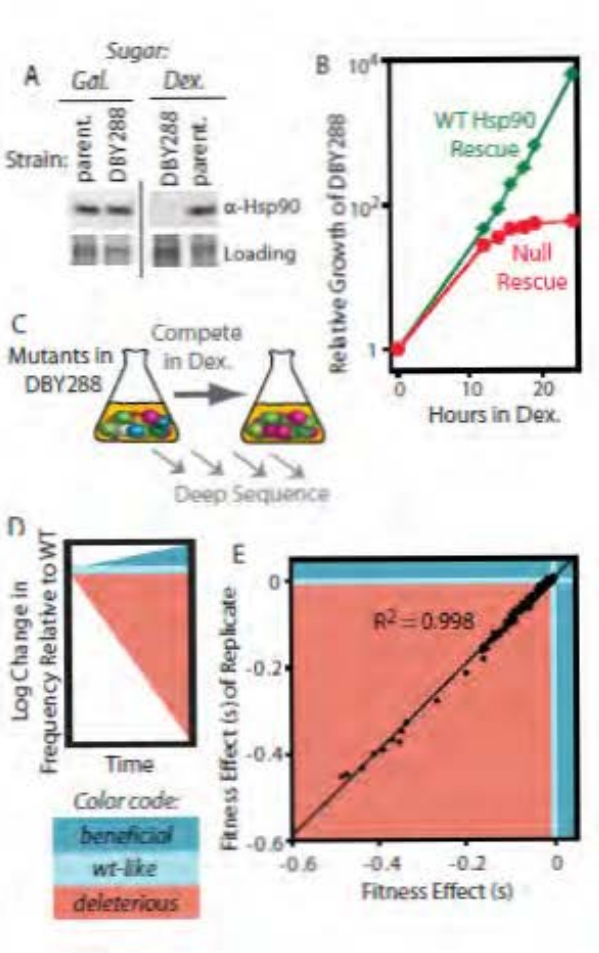
Results and Discussion

To facilitate fitness analyses of Hsp90 mutants under varied temperature and salinity, we created the DBY288 shutoff strain. In this strain, the only chromosomal copy of Hsp90 is driven by a strictly galactose-dependent promoter. In galactose (Gal) medium this strain expresses Hsp90 to a level similar to the parental strain, but in dextrose (Dex) medium Hsp90 expression is fully shutoff (Figure 4.1A). Transformation of this strain with a plasmid that constitutively expresses Hsp90 rescues growth in dextrose compared to a control plasmid lacking Hsp90 (Figure 4.1B). This conditional strain allows libraries of Hsp90 mutants to be amplified in Gal medium, and then competed in Dex medium where growth depends on the function of the library Hsp90 variants. Importantly, these bulk competitions can be performed under conditions of varied temperature and salinity.

We analyzed mutants in a region encompassing amino acids 582-590 of Hsp90. This region forms a hydrophobic patch on the surface of the Hsp90 structure including Phe583 and Trp585 that forms a putative substrate binding site[80], and that we speculate has the potential to impact client maturation as a function of distinct environments. We transformed a systematic library of mutations including all possible point mutations in this region into the DBY288 strain, amplified the resulting yeast library in Gal medium, and performed a bulk competition in Dex medium (Figure 4.1C) under four different conditions (30ºC, 36ºC, 30ºC+S, 36ºC+S, where +S indicates elevated salinity). For cells harboring wild type Hsp90, both elevated temperature and salinity reduced growth rate

Figure 4.1 EMPIRIC fitness analyses in a shutoff strain. The yeast strain DBY288 was engineered to express Hsp90 at endogenous levels when grown in galactose medium, and shutoff in dextrose medium, confirmed by Western analysis in (A). Growth of DBY288 yeast in dextrose was rescued with a plasmid that constitutively expresses Hsp90 (B). To measure fitness effects of mutants, systematic point mutant libraries were competed in bulk with deep sequencing used toreadout the abundance of each mutant (C). The time-dependent change in mutant frequency provided a direct examination of relative growth (D). This approach provided precise and reproducible measurements of fitness effects (E).

Figure 4.1


A
Sugar:
Gal.
Dex.
Strain:
parent.
DBY288
DBY288
parent.
α-Hsp90
Loading
B
WT Hsp90 Rescue
Null Rescue
Relative Growth of DBY288
Hours in Dex.
C
Mutants in DBY288
Compete in Dex.
Deep Sequence
D
Log Change in Frequency Relative to WT
Time
Color code:
beneficial
wt-like
deleterious
E
R² = 0.998
Fitness Effect (s) of Replicate
Fitness Effect (s)

with the combined 36ºC+S condition causing the greatest reduction (Supporting Figure S4.1), indicating a greater joint stress.

The relative abundance of each mutant in bulk competition was determined by focused deep sequencing of samples extracted from the culture over time (Figure 4.1C). The change over time in relative abundance of an amino acid substitution relative to the wild type amino acid (Figure 4.1D) yields a direct readout of the relative fitness effect of each point mutant in the library. Importantly, each point mutant is analyzed with precise control over genetic background and environmental sampling relative to all other mutants because the library is transformed into the same batch of yeast, and the bulk competitions are performed in the same flask where rapid mixing ensures that all mutants experience identical conditions. The ability to standardize genetic background and environmental conditions results in measurements of fitness effects that are highly reproducible (Figure 4.1E). The precision of these measurements enables unique systematic exploration of the distribution of fitness effects that we have analyzed with regard to expectations from Fisher's geometric model (FGM).

Experimental competitions under each environmental condition were managed to provide robust fitness measurements (Figure 4.2). In order to limit stochastic fluctuations, population sizes were maintained in gross excess to the diversity of our libraries at all steps. In addition, cells were analyzed rapidly once subject to selective conditions in order to limit the influence of potential secondary mutations. With these safeguards in

Figure S4.1 Growth of wild type Hsp90 in the four investigated environmental conditions. Growth of yeast harboring wild type Hsp90 were monitored by OD_{600} under experimental propagation identical to those used in the bulk competitions. The apparent doubling times under each condition are indicated in parentheses.

Figure S4.1

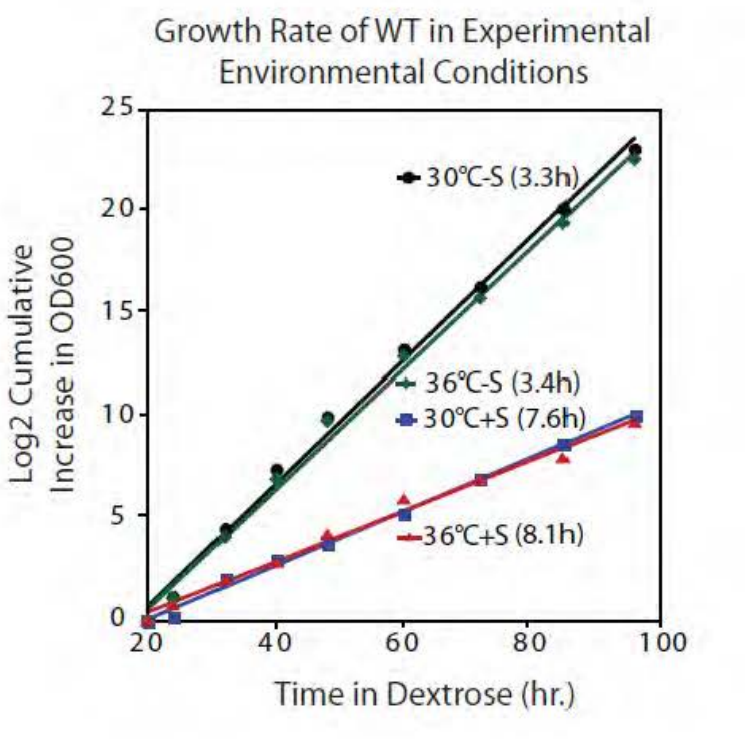
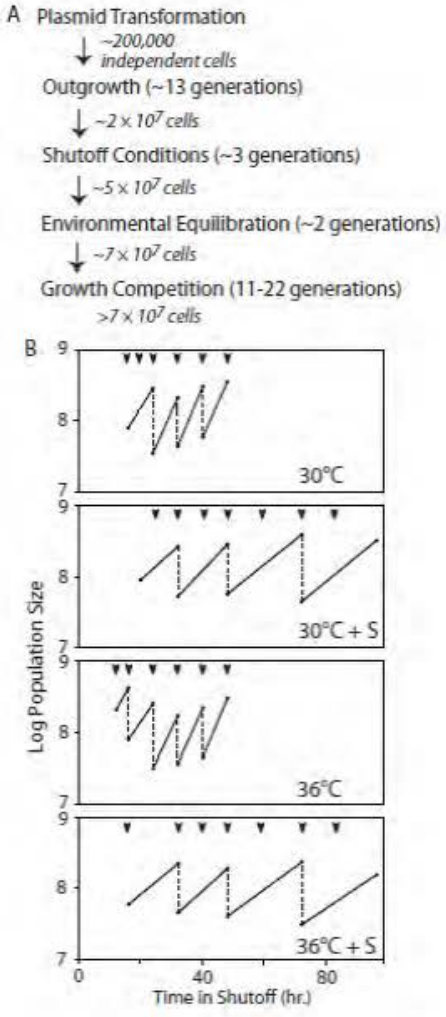

Growth Rate of WT in Experimental Environmental Conditions
Log2 Cumulative Increase in OD600
Time in Dextrose (hr.)
30℃-S (3.3h)
36℃-S (3.4h)
30℃+S (7.6h)
36℃+S (8.1h)
0
5
10
15
20
25
20
40
60
80
100

Figure 4.2 Population management during bulk competition experiments. (A) Complete experimental outline from initial transformation. The number of independent transformants and minimum population sizes during all phases of growth were in gross excess to the diversity of our library containing 567 mutants. (B) Trajectories indicating population sizes during the selection phase in each environmental condition. Dashed lines indicate dilutions, arrowheads indicate time points harvested and used for calculation of fitness.

Figure 4.2


A
Plasmid Transformation
~200,000 independent cells
Outgrowth (~13 generations)
~2 × 10^7 cells
Shutoff Conditions (~3 generations)
~5 × 10^7 cells
Environmental Equilibration (~2 generations)
~7 × 10^7 cells
Growth Competition (11-22 generations)
>7 × 10^7 cells
B
Log Population Size
30°C
30°C + S
36°C
36°C + S
9
8
7
0
40
80
Time in Shutoff (hr.)

place we analyzed the fitness effects of mutations under conditions of elevated temperature and/or elevated salinity (Figure 4.3, Supporting Table A2). Compared to the standard condition (0 mutations with s>0.01; s_{max}=0.007), we observe a dramatic increase in the number and magnitude of beneficial mutations in Hsp90 under conditions of elevated salinity (in 30ºC+S: 38 mutations with s>0.01,s_{max}=0.083; in 36ºC+S: 24 mutations with s>0.01; s_{max}=0.044).

As recent studies have observed that the adaptation of yeast to new conditions can be dramatically influenced by standing mutations[264], and mutations to multiple adaptive mutational pathways have been reported for yeast in elevated salinity[265], we took additional steps to ensure that the observed beneficial fitness effects (Figure 4.3) were caused by the specifically induced amino acid changes in Hsp90. Firstly, we analyzed synonymous substitutions underlying the identified beneficial amino acid, as most were encoded by multiple codons (e.g. N588P is encoded by four separate nucleotide variants). During transformation, each codon variant should enter a distinct pool of cells. If adaptation were primarily driven by secondary mutations within the genomes of these pools of cells, we would expect to observe highly variable fitness measurements among codons of the same amino acid. In contrast, if adaptation is primarily due to the Hsp90 amino acid substitution, then synonymous substitutions should have a narrow distribution. Among the beneficial amino acids that we observe, the distribution of fitness effects for synonymous codons is indeed narrow (Supporting Figure S4.2), indicating that

Figure 4.3 Fitness of mutants in different environments. (A) Histogram of mutant fitness in 30ºC+S. In order to exemplify the costs of adaptation to elevated salinity, the fitness effects of the beneficial (dark blue area) and wild type like (light blue area) mutations identified in 30ºC+S are indicated as dark and light blue bars in the corresponding histograms in 30ºC (B), 36ºC (C), and 36ºC+S (D). The selection coefficients of three independently validated beneficial mutations in 30ºC+S and their corresponding selection coefficients in the other environments are indicated by black triangles. (Two overlapping triangles in (D) are marked with a star). (E) Proportions of beneficial, wild type like, deleterious and strongly deleterious mutations in the different environments.

Figure 4.3

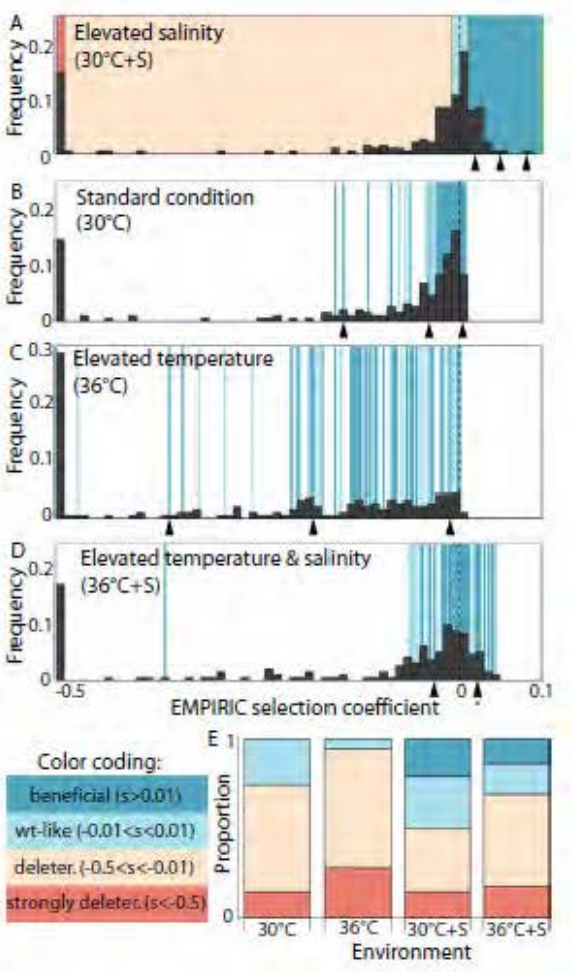
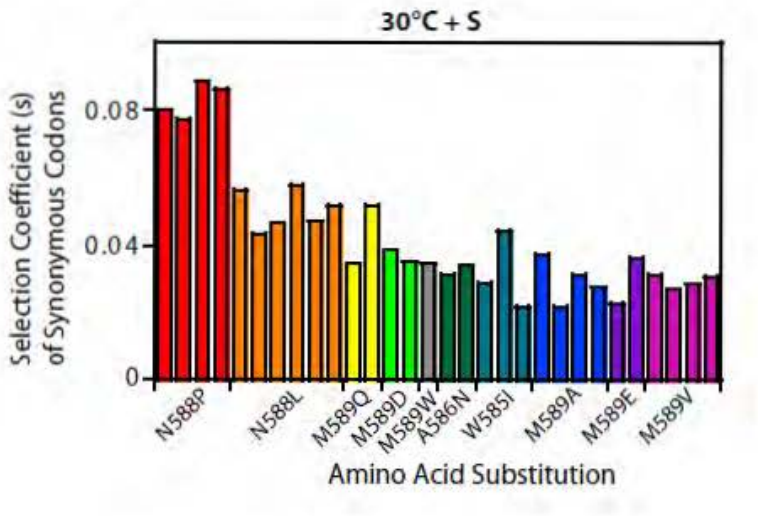

A
Elevated salinity
(30°C+S)
B
Standard condition
(30°C)
C
Elevated temperature
(36°C)
D
Elevated temperature & salinity
(36°C+S)
Frequency
EMPIRIC selection coefficient
Color coding:
beneficial (s>0.01)
wt-like (-0.01<s<0.01)
deleter. (-0.5<s<-0.01)
strongly deleter. (s<-0.5)
E
Proportion
30°C
36°C
30°C+S
36°C+S
Environment

Figure S4.2 Fitness effects of synonymous substitutions for the ten amino acid substitutions with the greatest benefit in elevated salinity. Independently measured synonymous mutations were consistently beneficial, indicating that the amino acid sequence was a dominant cause of the observed fitness benefit. Only one of the top ten substitutions was due to an amino acid change encoded by a single codon (M589W). These analyses indicate that the majority of observed beneficial amino acid substitutions are caused by the protein sequence.

Figure S4.2


30°C + S
Selection Coefficient (s) of Synonymous Codons
0.08
0.04
0
N588P
N588L
M589Q
M589D
M589W
A586N
W585I
M589A
M589E
M589V
Amino Acid Substitution

the induced mutations in Hsp90 are the primary determinant underlying the experimental observation.

Secondly, we performed a follow-up 30ºC+S study with two experimental replicates using independent transformations. In these experiments we observe a similar number of beneficial mutations and a strong correlation between each independent replicate (Supporting Figure S4.3). Thirdly, we developed a qPCR based binary competition assay (Supporting Figure S4.4) that we used as an alternative strategy to measure fitness effects of six individual mutations (Supporting Figure S4.5). The fitness effects observed by binary competition correlate well with those measured by EMPIRIC bulk competitions. From this collection of experiments, we conclude that our fitness measures are primarily owing to the induced amino acid changes.

We studied the distribution of fitness effects along two environmental axes comprising elevated temperature and elevated salinity. These are associated with different expectations concerning the potential for adaptation and the distance to the optimum in the FGM. As a heat shock protein, Hsp90 is per definition suited to cope with elevated temperature, and its function becomes even more essential upon heat stress as discussed above. Hence, we hypothesize that the endogenous sequence is close to the phenotypic optimum under high temperature and beneficial mutations rare. In contrast, we hypothesize that elevated salinity may pose a novel environmental challenge that is

Figure S4.3 Correlation between biological replicates at elevated salinity (30°C+S). To examine potential influences of variations among populations of transformed cells in promoting adaptation to elevated salinity, additional bulk competitions were performed starting with two independent transformations. We observed a strong correlation for all mutants that persist in the bulk culture ($s>-0.2$), including mutations with an adaptive benefit relative to wild type under this condition.

Figure S4.3

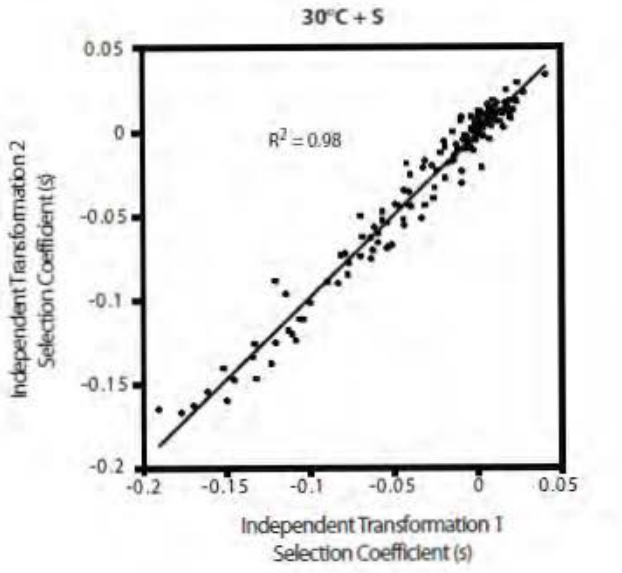

30°C + S
0.05
0
-0.05
-0.1
-0.15
-0.2
$R^2 = 0.98$
Independent Transformation 2
Selection Coefficient (s)
-0.2
-0.15
-0.1
-0.05
0
0.05
Independent Transformation 1
Selection Coefficient (s)

Figure S4.4 Schematic of qPCR based analyses of selection coefficients**.** (A) On the left is the parental plasmid used in bulk competition (Plasmid A), and on the right is a nearly identical construct (Plasmid B) with a 50 base pair insert in a non-functional region of the plasmid. Distinct primers were designed in order to distinguish these plasmids by qPCR. (B) Plasmid A and Plasmid B were amplified with both the match and mismatch primers to determine the dynamic measurement range and specificity of each primer. (C) Representative standard curve generated with both match and mismatch template/primer sets indicate that PCR efficiency is robust for match pairs, but that mismatch pairs require many more cycles in order to generate measureable PCR product. (D) Plasmid abundance was calculated for both the match and mismatch template/primer pairs with an input of approximately 0.1 ng of template. Match plasmids were readily detected with minimal noise from mismatch plasmids. (E) Analyses of a binary competition between yeast harboring either wt or N588P Hsp90. These two strains of yeast were competed as in the bulk experiments with samples isolated at different time points under selection. Lysates from these samples were amplified with plasmid specific primers to determine the abundance of each plasmid in the lysate. (F) The change in mutant to wild type abundance over time in wild type generations was analyzed to determine the selection coefficient as in the bulk competitions.

Figure S4.4

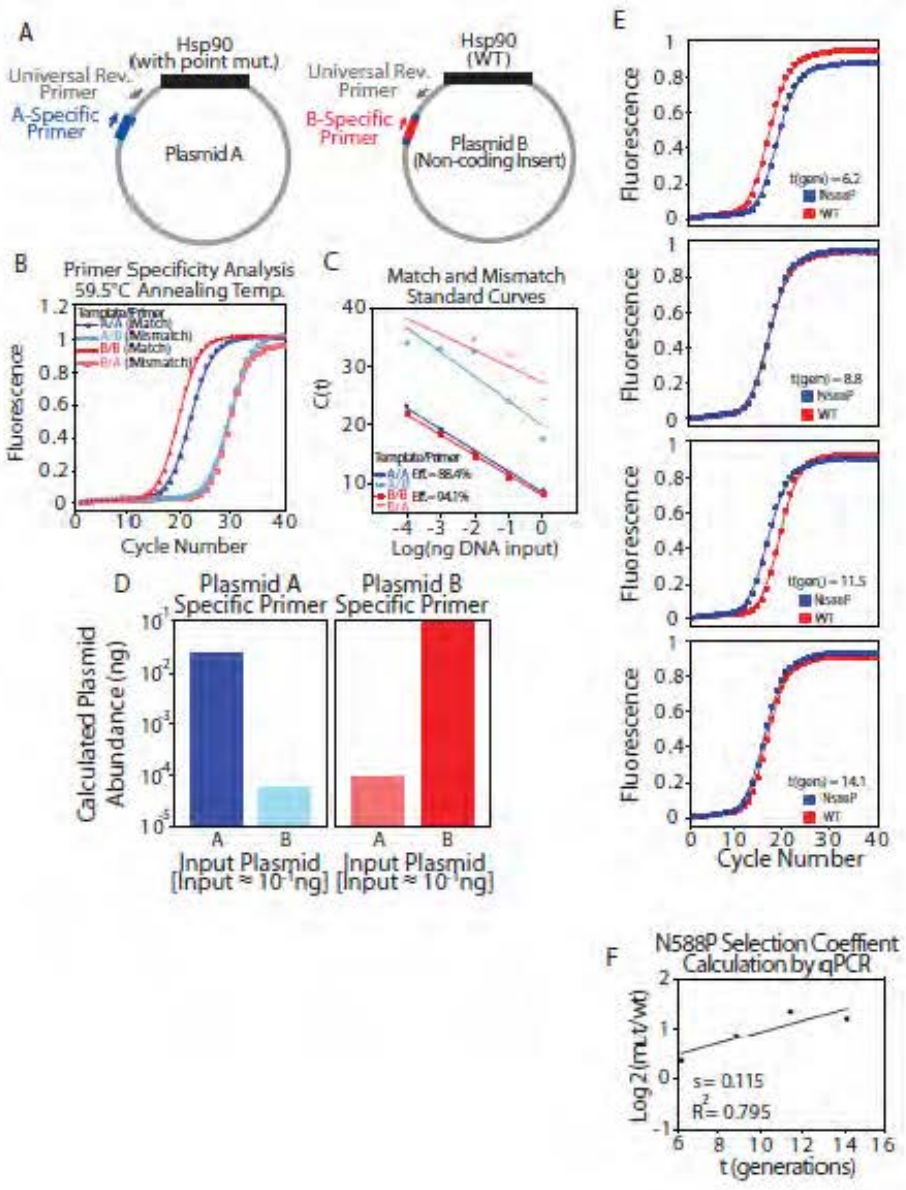
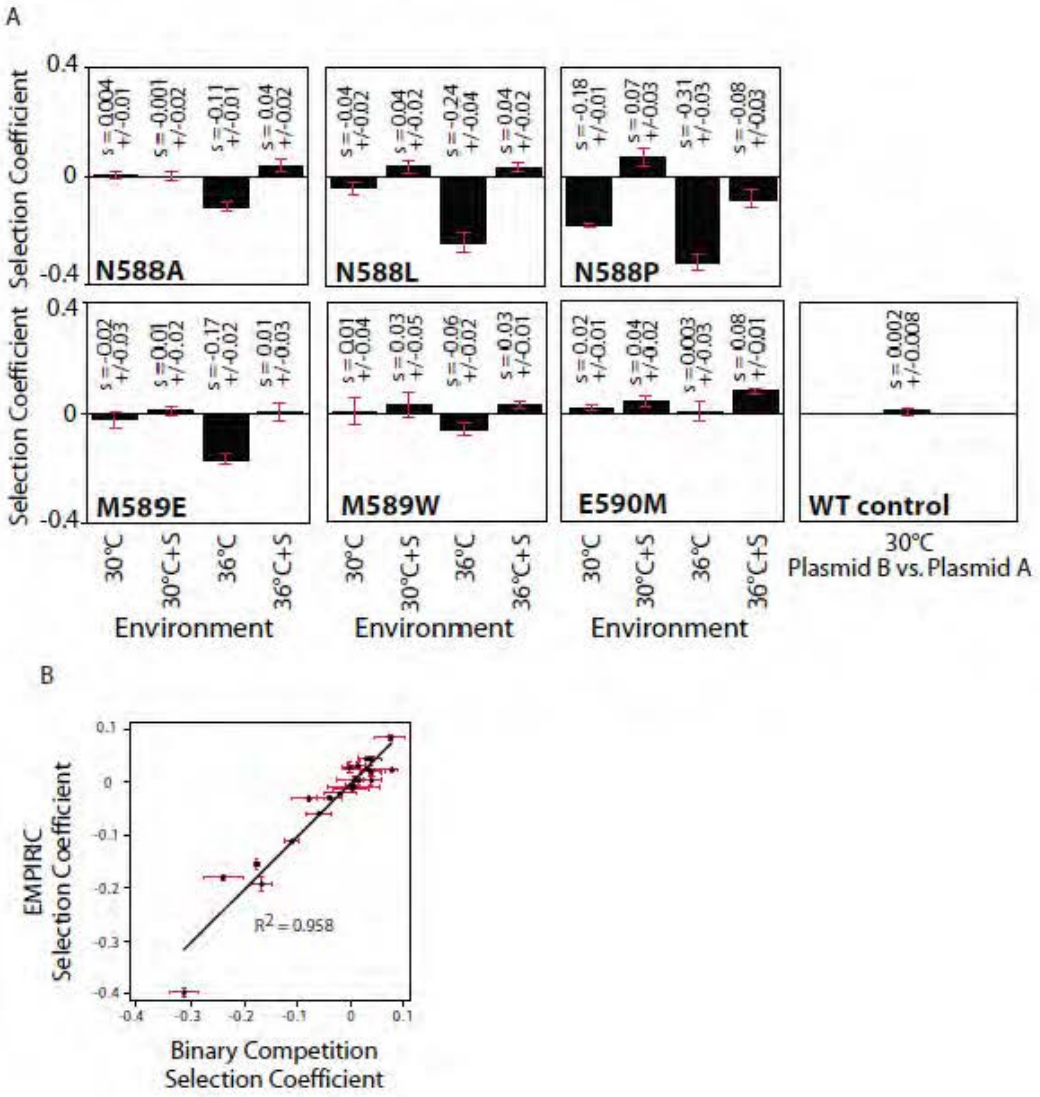

A
Hsp90 (with point mut.)
Universal Rev. Primer
A-Specific Primer
Plasmid A
Hsp90 (WT)
Universal Rev. Primer
B-Specific Primer
Plasmid B (Non-coding Insert)
B
Primer Specificity Analysis
59.5°C Annealing Temp.
Fluorescence
Cycle Number
C
Match and Mismatch Standard Curves
C(t)
Log(ng DNA input)
D
Plasmid A Specific Primer
Plasmid B Specific Primer
Calculated Plasmid Abundance (ng)
Input Plasmid [Input ≈ 10⁻¹ ng]
E
Fluorescence
Cycle Number
F
N588P Selection Coeffient Calculation by qPCR
Log 2(mut/wt)
s = 0.115
R² = 0.795
t (generations)

Figure S4.5 Fitness measurements made by binary competition correlate with EMPIRIC results. (A) Binary competitions were performed between wild type Hsp90 and six different point mutants in four distinct environmental conditions and fitness effects of the mutants determined using a qPCR approach. Bar graphs represent the average selection coefficient from three independent qPCR analyses with error bars representing the standard deviation. Competition between two wild type coding sequences results in indistinguishable growth, indicating that the non-coding changes made to distinguish plasmids by qPCR do not perturb fitness. (B) Correlation between fitness measurements made by binary competition and EMPIRIC.

Figure S4.5


A
Selection Coefficient
0.4
0
-0.4
N588A
s = 0.004 +/-0.01
s = -0.001 +/-0.02
s = -0.11 +/-0.01
s = 0.04 +/-0.02
N588L
s = -0.04 +/-0.02
s = 0.04 +/-0.02
s = -0.24 +/-0.04
s = 0.04 +/-0.02
N588P
s = -0.18 +/-0.01
s = 0.07 +/-0.03
s = -0.31 +/-0.03
s = -0.08 +/-0.03
M589E
s = -0.02 +/-0.03
s = 0.01 +/-0.02
s = -0.17 +/-0.02
s = 0.01 +/-0.03
M589W
s = 0.01 +/-0.04
s = 0.03 +/-0.05
s = -0.06 +/-0.02
s = 0.03 +/-0.01
E590M
s = 0.02 +/-0.01
s = 0.04 +/-0.02
s = 0.003 +/-0.03
s = 0.08 +/-0.01
WT control
s = 0.002 +/-0.008
30°C
30°C+S
36°C
36°C+S
Environment
30°C
Plasmid B vs. Plasmid A
B
EMPIRIC
Selection Coefficient
0.1
0
-0.1
-0.2
-0.3
-0.4
R2 = 0.958
Binary Competition
Selection Coefficient

not directly associated with the function of Hsp90. Therefore, we expect a relocation of the phenotypic optimum that yields increased potential for adaptation.

The above expectations are clearly supported by our results: both under standard conditions (30ºC) and under elevated temperature (36ºC) we find no beneficial mutations, indicating that the current phenotype is close to the phenotypic optimum (cf. Figure 4.3B,C). In contrast, we find numerous beneficial mutations under elevated salinity (30ºC+S) and under combined elevated temperature and salinity (36ºC+S, cf. Figure 4.3A,D). As hypothesized, the proportion of deleterious mutations grows with increased heat stress in the 36ºC and 36ºC+S environments as compared with 30ºC (Figure 4.3E). In addition, Figure 4.3E shows that the proportion of beneficial mutations under high salinity is reduced upon increased heat stress, indicating that the current phenotype is closer to the phenotypic optimum in 36ºC+S than in 30ºC+S – which is supported by the estimated distances to the optimum under the assumption that the DFE follows a shifted gamma distribution. We further observe high costs of adaptation. In particular, mutations that were found to be beneficial in 30ºC+S (represented by dark blue background in the histogram in Figure 4.3A) are very likely to be deleterious in the low-salinity environments (as indicated by dark blue bars in Figure 4.3B,C).

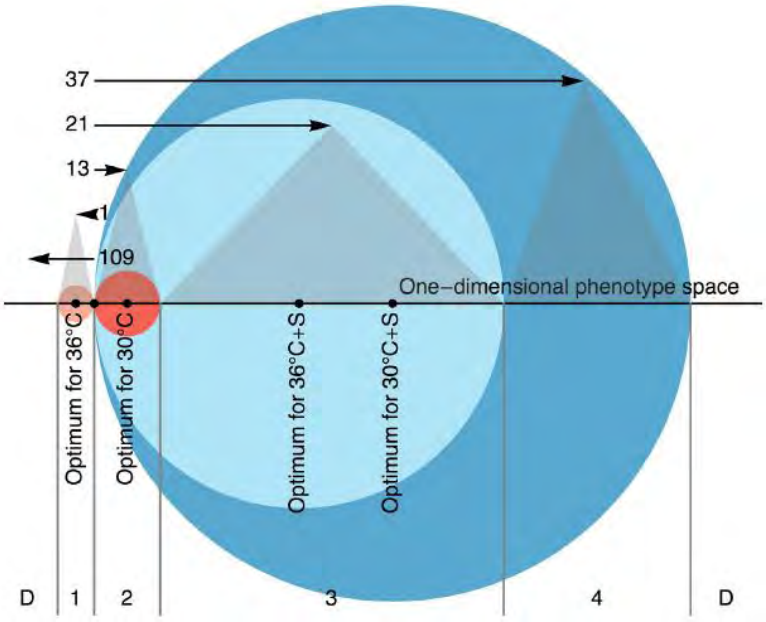
We were interested in the question of whether the observed costs of adaptation are corresponding with the predictions of an FGM that is extended to more than one environment. Roughly, this idea has been sketched in Martin and Lenormand 2006b,

Figure 7 (but see also Figure 4.4): if different phenotypic optima characterizing the fitness peaks for different environments are drawn in a phenotype space according to the FGM, subsets of the phenotype space arise in which mutations are deleterious in one but beneficial in the other environment. Hence, as soon as two environments have differently located optima, costs of adaptation are predicted. Moreover, this is necessarily true for every choice of two environments that represent different distances between the current phenotype and the optimum (e.g., standard environment vs. adaptive challenge). Here, we do not fit explicit distributions to the DFEs for individual environments. Instead, we use a simplified FGM that reduces the information for each mutant to the sign of its selection coefficient – which indicates its rough location in phenotype space as compared with the current phenotype and the optimum. By comparing this location between environments we can test whether there is a configuration of the FGM (i.e., an arrangement of the current phenotype and the four different optima) that is in accordance with our observed costs of adaptation (which is determined by the sign effect of a mutation in all four environments).

In order to determine the effective number of dimensions of the phenotype space, we utilize the results of Martin and Lenormand[266]. Our data for the standard environment yielded n_e=1.08, which is in concordance with previously published genome-wide values for *S. cerevisiae*[266]. In addition, we estimated the distances to the optimum under the assumption that the DFE (neglecting the deleterious mode with s<=-0.5) follows a shifted gamma distribution as suggested by Martin & Lenormand 2006b (further detailed in

Figure 4.4 Graphical representation of the fit of the FGM. The horizontal black solid line represents the one-dimensional phenotype space, whereas the vertical solid black line indicates the position of the current phenotype. For each environment, a colored circle (evocative of the original two-dimensional picture of the FGM (cf. Figure 1, Orr 1998[262])) is drawn tangential to the current phenotype with its radius corresponding to the distance to the phenotypic optimum (indicated as indexed black dot) of the respective environment. The interval corresponding to the projection of each circle onto the one-dimensional phenotype space represents the area into which mutations have to fall according to the FGM in order to decrease the distance to the optimum and hence to be beneficial in a particular environment. 181 of 189 analyzed mutations are in agreement with a one-dimensional FGM, in which the distance of the current phenotype to the optimum for each environment in phenotype space is assumed to rank according to the selection coefficient of the most beneficial mutation (here, the radius of the circles is drawn proportional to the selection coefficient of the best mutant). This yields 5 categories of mutations that are characterized by the overlap of the projected circles (cf. also Figure 4.5): (D) Deleterious in all environments, (1) $s>0$ in 36ºC only, (2) $s<0$ in 36ºC only, (3) $s<0$ in 30ºC and 36ºC; $s>0$ in 30ºC+S and 36ºC+S, (4) $s>0$ in 36ºC+S only. Arrows represent exemplary mutations for each of these categories, indexed with the respective observed number of such mutations per category. Gray triangles indicate the projection of these categories onto the respective interval in phenotype space (see also Figure 4.5).

Figure 4.4

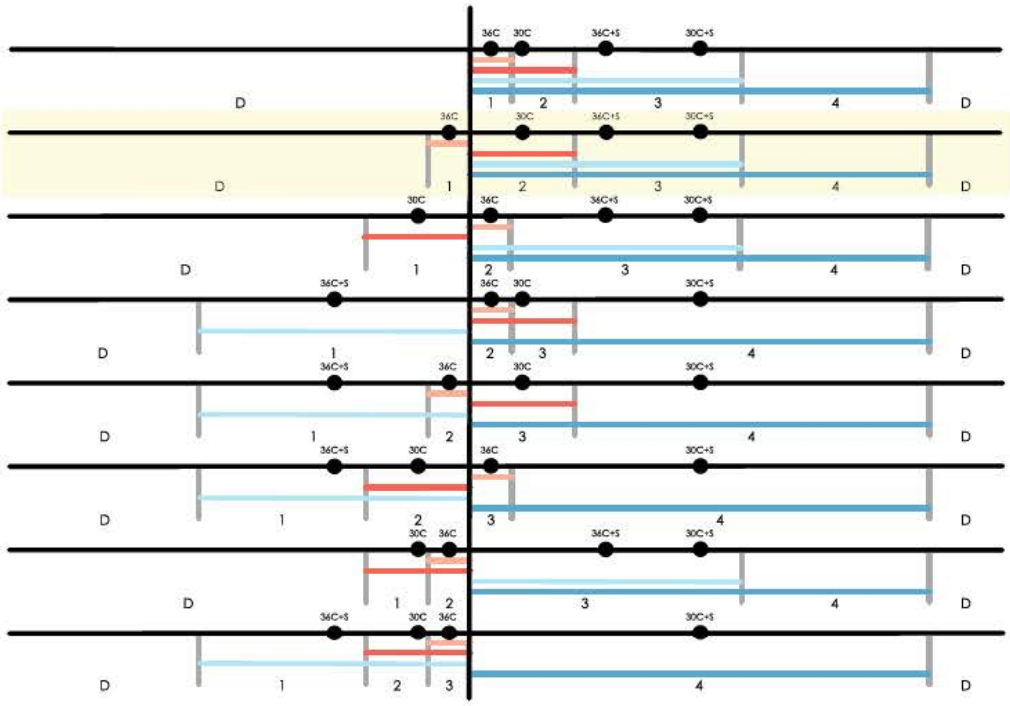

37
21
13
1
109
One−dimensional phenotype space
Optimum for 36°C
Optimum for 30°C
Optimum for 36°C+S
Optimum for 30°C+S
D
1
2
3
4
D

Methods). Mutations were categorized according to the sign of their selection coefficient in each of the four environments, resulting in 16 possible categories of mutations – however, any possible geometric arrangement according to a one-dimensional FGM as defined for our purpose contains only 5 categories of mutations, which are determined by the ranking of the distances to the optimum (which we fixed according to the estimated distances) and the arrangement of the optima in phenotype space (cf. Figure 4.4). This resulted in 8 possible models that are consistent with the FGM (cf. Figure 4.5). A particular FGM arrangement harbors a chosen mutation if it includes the mutation's category.

Despite the restriction on the dimensionality (i.e., on the number of mutational categories allowed), our best fit pursuant to our simplified version of the FGM (graphically represented in Figure 4.4) harbors 95.8% of all mutations (99.5% if the sign of mutations that were previously categorized as wt-like are neglected). In addition, we compared the 8 models that are compatible with an arrangement according to our definition of the FGM (cf. Figure 4.5) with fits to all 4,368 possible subsets containing 5 out of 16 mutational categories (i.e., all alternative models with the same complexity, but not compatible with an FGM-like geometrical arrangement). We find that our best fit ranks second among all models, where the overall best model harbors 97.4% of all observed mutations. Furthermore, all eight models that are in concordance with the FGM are within the best 5% of all models, harboring between 79.9% and 95.8% of mutations.

Figure 4.5 All possible realizations of the FGM in a one-dimensional phenotype space. Horizontal black solid lines represent the phenotype space, whereas the vertical solid black line indicates the position of the current phenotype. For each environment, an interval (representing the one-dimensional version of what is commonly drawn as a circle in the two-dimensional FGM) is drawn as a colored line that spreads from the current phenotype along the phenotype space to twice the distance between the current phenotype and the phenotypic optimum (indicated as indexed black dot) for that environment. This interval represents the area in which mutations have to fall according to a one-dimensional FGM in order to decrease the distance to the optimum and hence to be beneficial. All 8 combinatory possibilities to realize this model are shown, with the second one (highlighted in yellow) representing the best fit that is elaborated in Figure 4.4. Every realization contains 5 categories of mutations (D, 1-4) that are characterized by the overlap of colored intervals.

Figure 4.5



Conclusions

Many of the greatest accomplishments of the genomic era come from the empirical evaluation of the fundamental theoretical models of evolution proposed nearly a century ago by the founders of the field – Fisher, Haldane, and Wright (for an excellent overview of the early field, see Crow 1987[267]). It is in this vein which we have evaluated Fisher's geometric model. We have experimentally observed many of Fisher's expectations regarding adaptive step sizes as they relate to the distance from an optimum state.

Our observations suggest a number of noteworthy implications. First, we observe a striking number of beneficial mutations in a small region of an essential protein. This demonstrates that genomic regions under high constraint harbor hitherto unrecognized potential for adaptation upon environmental change. Interpreting these identified beneficial mutations in light of the known biology of Hsp90 suggests that biochemical context well predicts adaptive response, implying an important role of regulators in dictating adaptive potential. Although this region of Hsp90 is strongly conserved in eukaryotes[202], some of the salt beneficial mutations that we observed experimentally (e.g., S586T and N588H) are also found in nature. Second, the simple framework of the FGM is sufficient to explain important aspects of our data. In particular, the observed costs of adaptation and the number of shared beneficial mutations between the two high-salinity environments are remarkably consistent with a one-dimensional FGM. And while it is far from conclusive, the repeated observation that beneficial mutations in one

environment tend to be mildly to strongly deleterious in all other environments, ought to serve as a note of caution against recent arguments[234, 268, 269] for the pervasive role of standing variation in adaptation. Third, we observe that the potential for adaptation is reduced in the combined high-temperature-and-salinity as compared with the high-salinity environment. Combined with the experimental observation that this joint environment also results in the greatest reduction in growth rate, this result is shown to be consistent with the expectation of the FGM. Further, this observation echoes the notion of Haldane[81], who suggested the difficulty inherent in simultaneous selection for multiple traits.

Specific Materials and Methods

Plasmid library construction

Saturative single codon substitution libraries of amino acids 582-590 were generated in plasmid p417GPD that constitutively expresses Hsp90 as previously described[202].

Yeast transformation and selection

Constitutively expressed libraries of Hsp90 mutants were introduced into a shutoff strain, amplified in galactose medium, and then competed in dextrose medium. These studies used the DBY288 yeast strain (can1-100 ade2-1 his3-11,15 leu2-3,12 trp1-1 ura3-1 hsp82::leu2 hsc82::leu2 ho::pgals-hsp82-his3) where expression of the only Hsp90 gene in these cells strictly depends on galactose. A single colony of DBY288 was picked from a synthetic raffinose/galactose (Gal) plate and inoculated into 25 ml 2xYPA Gal medium (20 g yeast extract, 40 g peptone, and 0.2 g adenine hemisulphate per liter with 1% (w/v) raffinose and 1% galactose) and grown at 30ºC on an orbital shaker to late log phase. The culture density was calculated by hemocytometry and 10^8 cells were inoculated into 50 ml of fresh 2xYPA Gal medium. The culture was grown for 5 h at 30ºC with agitation, harvested by centrifugation at 3,000*g* for 5 min and transformed by the standard lithium acetate protocol[215, 216] with plasmid (either mutant libraries, a positive control with wild type Hsp90, or negative control lacking Hsp90). The generation time of cells harboring wild type Hsp90 was determined by following the

change in optical density over time (Supporting Figure S4.1).To examine measurement precision, replicate competition experiments were performed under the 30ºC condition with cultures split after transformation (Figure 4.1E).

After heat shock at 42ºC for 30 min, the cells were pelleted at 3,000*g* for 5 min and washed twice with 500 µl Gal medium (1.7 g yeast nitrogen base without amino acids, 5 g ammonium sulfate, 0.1 g aspartic acid, 0.02 g arginine, 0.03 g valine, 0.1 g glutamic acid, 0.4 g serine, 0.2 g threonine, 0.03g isoleucine, 0.05g phenylalanine, 0.03g tyrosine, 0.04g adenine hemisulfate, 0.02g methionine, 0.1g leucine, 0.03g lysine, 0.01g uracil per liter with 1% raffinose and 1% galactose) and recovered in 5 ml Gal medium for 16 h. The cells were pelleted by centrifugation at 3,000*g* for 5 min and inoculated into 50 ml Gal medium with 200 µg/ml G418. The culture was then allowed to grow at 30ºC on an orbital shaker to near-saturation (about 48 h). 20 ml of this culture was washed with fresh Gal medium and the pellet was inoculated into 100 ml synthetic Gal medium containing 100 µg/ml ampicillin. This culture was grown for 12 h in log phase, diluting when necessary. The log phase cells were then inoculated to an OD_{600} of ~0.1 in 150 ml of synthetic Dextrose (Dex) medium (identical composition to Gal medium except with 2% dextrose as the sugar source) with 100 µg/ml ampicillin that were grown at 30ºC on an orbital shaker. After 8 h, the culture was split into four different environmental conditions (30ºC, 36ºC, 30ºC+S, 36ºC+S; where S represents the addition of 0.5 M sodium chloride). The culture was then grown for 12-20 generations in log phase

(diluting when needed). Samples were reserved at different time points throughout the experiment by pelleting 10^9 cells and storing the pellets at -80ºC.

DNA preparation, sequencing, and analysis

Yeast lysis, DNA preparation and sequencing was performed as described[270]. Sequencing was performed by the UMass deep sequencing core facility, and generated ~30 million reads of 99% confidence at each read position as judged by PHRED scoring[271, 272]. The relative abundance of each mutant relative to wild type was calculated at each sampled time point. The slope of the logarithm of relative mutant abundance versus time in generations was used as a direct measure of relative fitness. To account for sequencing noise, an outlier detection based on the boxplot rule was performed for each mutant's trajectory – hence, data points outside the range spanned by the 50% confidence interval extended by 1.5 times the interquartile range on each side were excluded from the linear regression. In order to obtain normalized selection coefficients for each data set such that wild type fitness represents s=0, we selected all mutants that result in wild type synonyms as a reference set and calculated its mean and standard deviation. To account for potential outliers of the distribution of synonyms, we neglected those mutations further than two standard deviations away from the mean, and defined the resulting new mean as the normalization constant representing s=0. Hence, each selection coefficient on the nucleotide level is calculated as the slope of its absolute read numbers minus the normalization constant. The selection coefficient of each amino acid is thereupon obtained as a weighted average of the selection coefficients of all synonymous codons.

Weights are assigned after outlier detection (according to the boxplot rule) on the time-point level to account for the effect of low read numbers in the initial library. For Figure 4.2, mutations were categorized as indistinguishable from wild type ("wt-like") if $|s|<0.01$, beneficial if $s>0.01$, deleterious if $-0.01>s>-0.5$ and strongly deleterious if $s<-0.5$.The threshold for the wt-like category represents a strongly conservative choice to assure that beneficial mutations are truly advantageous – however, we observed no qualitative differences in the results when this threshold is set to $|s|<0.005$.

Reproducibility of fitness effects in a bulk competition replicate

To further investigate potential fitness contributions from background mutations in pools of transformed cells and to vet the reproducibility of bulk fitness measurements, we performed a subsequent full experimental replicate under 30ºC+S conditions that included separate transformations. DBY288 cells were transformed and selected as in the original experiment at 30ºC+S. Mutants with fitness effects $s>-0.2$ were compared between full experimental replicates (Supporting Figure S4.3) and exhibited a high level of reproducibility ($R^2=0.98$).

Confirmation of mutant fitness effects by binary competition

To confirm the fitness measurement generated by the EMPIRIC approach, we developed an independent qPCR based assay to measure the fitness of a subset of mutants by binary competition (Supporting Figure S4.4 and S4.5). The binary competition assays competed cells bearing a single point mutant against cells bearing wild type plasmid. A

50 base pair region was inserted into a non-coding region of the wild type plasmid in order to distinguish mutant from wild type. This insertion did not alter the growth property of the host strain, but did enable quantification of the relative abundance of wild type and mutant cells in binary competitions. Wild type and point mutant plasmid were mixed at a 1:1 molar ratio and co-transformed into DBY288 cells. Growth, selection and lysis procedures were identical to the EMPIRIC experiment. For qPCR analysis, a common reverse primer and a wild type or mutant specific forward primer was used produce a 300 base pair amplicon. The qPCR reactions consisted of the following: 1X SYBR Green I gel stain, 500 nM each forward and reverse primer, 50 µM each dNTP, 1X Phusion HF buffer, 0.5 mM additional magnesium chloride, 0.5 µl Phusion DNA polymerase, in a final volume of 50 µl. PCR conditions were as follows: 94ºC for 2 min; 40 cycles of 94ºC for 30 s, 59.5ºC for 30 s, 72ºC for 30 s. Standard curves were generated by analyzing dilution series (1 to 10^{-4} ng) of wild type and mutant plasmids with both primer sets in triplicate. Experimental samples contained 1-2 µl of lysate as template with equal volumes amplified with each primer set. Selection coefficient measurements were repeated in triplicate in order to assess measurement precision.

Correspondence with the fitness trade-offs predicted by the FGM

Given four environments and a straightforward distinction of two classes of mutations (beneficial if $s>0$, deleterious if $s<0$) in each environment, there are $2^4=16$ possible categories of mutations, ranging from “deleterious in all environments” to “beneficial in all environments”. From Martin and Lenormand[273], Eq. 4a, one can

determine the effective number of phenotypic dimensions in the FGM as $n_e=2E(s)^2/V(s)$, where E(s) and V(s) are the mean and the variance of the distribution of deleterious (but not lethal) mutation effects in a population that is close to the optimum (in our experiment corresponding to all mutations with $-0.5<s<0$ from 30ºC). We obtain $n_e\approx1.08$, clearly supporting a one-dimensional phenotype space. Despite being restricted to a small region of a single protein, our data is in agreement with whole-genome estimates from mutation accumulation experiments[266]. Even though the theory was developed for populations close to the optimum, we obtain similar numbers for the high salinity environments (30ºC+S: $n_e=0.80$; 36ºC+S: $n_e=1.40$), whereas the result from the high-temperature environment would indicate a higher complexity of its phenotype space (36ºC: $n_e=2.79$).

We estimated the distance to the optimum in each environment by fitting a shifted gamma distribution following Martin & Lenormand 2006b (see Equation 5), by neglecting the strongly deleterious mode of the distribution. Its location parameter s_0 determines the distance to the optimum. We obtain $s_0=0.007$ for 30ºC,$s_0=0.002$ for 36ºC, $s_0=0.087$ for 30ºC+S, and $s_0=0.045$ for 36ºC+S. Notably, the same ranking and similar proportions are obtained if the mean, median, or the maximum of all beneficial mutations are taken as reference for the distance to the optimum. For combinatorial reasons, there exist eight different geometric arrangements of the 4 optima in phenotype space that each contain 5 categories of mutations (cf. Figure 4.5). We identify the best fit in accordance with the FGM (shown in Figure 4.4) to harbor 181/189 (=95.8%) of all mutations. The

189 amino acid substitutions represent all twenty amino acids plus stop codons at nine positions. Fitness measurements of amino acid substitutions were averaged over synonymous substitutions resulting in independent measures of the wild type amino acid at each position. Classifications according to the best fit are shown in the last column of Supporting Table A2, labeled as indicated in Figure 4.4 (additionally, the label "I" represents incongruous mutations). Note that all but one of the incongruous mutations can be classified if the sign of wt-like mutations is neglected (hence, incongruity is likely explained by the limits of accuracy of the experiment).

ACKNOWLEDGEMENTS

This work was supported in part by grants from the National Institutes of Health (R01-GM083038) and the American Cancer Society (RSG-08-17301-GMC) to D.N.A.B., and by grants from the Swiss National Science Foundation and a European Research Council (ERC) Starting Grant to J.D.J. We would like to thank vital-IT and the Swiss Institute of Bioinformatics (SIB) for computational resources. The authors declare no conflict of interest.

Chapter V – Experimental Characterization of Intragenic Epistatic Effects

This work is in preparation for submission as *Hietpas RT, Bank C, Jensen JD, Bolon DNA. "Experimental characterization of intragenic epistatic effects"*

This work was a collaborative effort. I designed and performed yeast growth competitions, DNA isolation and preparation, initial data analysis, and structure-function analysis of epistatic effects. Dr. Claudia Bank applied her Bayesian MCMC approach to the initial sequencing data and applied rigorous mathematical analyses to give the data statistically meaningful cutoffs. I, Dr. Claudia Bank, Dr. Jeffrey D. Jensen, and Dr. Daniel N. A. Bolon prepared the manuscript.

Abstract

Mutations are the source of evolutionary variation, and the interactions of multiple mutations can have profound effects on fitness and evolutionary trajectories. We have recently described the distribution of fitness effects of all single mutations for a nine amino acid region of yeast Hsp90 (Hsp82) implicated in substrate binding. Here, we present a distribution of intragenic epistatic effects within this region in seven Hsp90 point mutant backgrounds to gauge the frequency and magnitude of epistatic effects. We find negative epistasis between substitutions common, and positive epistasis to be relatively rare. Structural analyses indicate a correlation between local residue environment and the predominant type of epistasis. Negative epistasis was mainly associated with mutations at solvent inaccessible positions. In contrast, all observations of positive epistasis involved at least one mutation at a solvent exposed position, and commonly also involved a second mutation at a solvent inaccessible position. These observations suggest that the interplay between mutations that impact main chain conformation and the biophysical properties of solvent facing side chains frequently have complex fitness effects.

Introduction

Mutation is the source of evolutionary variation, and over immense timescales, the cumulative effects of mutations have given rise to an enormous diversity of life. New mutations may be grouped into three general categories based upon their effect on organismal fitness; deleterious, neutral, and beneficial. Previous experimental studies indicate that the majority of new mutations are slightly to strongly deleterious with a vast minority conferring a fitness benefit[202, 252], consistent with population genetic theory[18]. Single nucleotide substitutions are the most common form of new mutation, and are frequently observed in human disease[274]. While the simultaneous occurrence of two or more new mutations within a single gene is extremely rare, over many replicative events multiple mutations can accumulate in the same gene with important consequences. The potential interdependence of mutations can have a tremendous impact on evolutionary trajectories. This was first recognized by Bateson[275], who coined the term epistasis in what was one of the first joint considerations of Darwinian evolution with Mendelian genetics.

The interdependence or epistasis of mutations can be assessed by comparing the fitness effects of a combined mutant with the fitness effects of each individual mutant. Here, we consider mutations independent if the fitness of the combined mutant equals the product of the fitness of each individual mutant. Combinations of mutations that deviate from this rule are considered interdependent or epistatic. The interdependent fitness effects of epistatic mutations are directional and can result in combined mutants with

fitness that is increased (referred to as positive epistasis) or decreased (negative epistasis) relative to independence. Epistasis can have a dramatic impact on the accumulation of mutations within populations including the probability of fixation[276-278]. While epistasis is central to evolution, for most experimental systems it is challenging to investigate because of combinatorial complexity[162].

Compensatory mutations represent a form of epistasis with many biologically and medically important ramifications. Compensatory mutations rescue fitness defects of primary mutations, but are of little fitness consequence in the parental background. For example, many studies demonstrate that compensatory mutations play an important role in the adaptation of microbes and viruses to antibiotic or antiviral treatment[149-151, 279, 280]. In many of these cases, the primary drug resistance mutation was found to be deleterious in the absence of drug, but a secondary mutation that had a neutral fitness effect in the parental genotype increased the fitness of the primary mutation, promoting the persistence of the primary mutation to in the absence of drug treatment. A recent meta-analysis indicates that 83% of all compensatory mutations occur within the same gene as the primary mutation, which emphasizes the relevance of intragenic epistasis in evolution[281].

Intragenic epistasis has a rich history of investigation in the framework of protein structure-function relationships. Double mutant cycles compare the biochemical properties of combined mutants to individual mutants and have proven a powerful

approach to investigate the interdependence of mutations on protein stability or activity[153]. Double mutant cycles have been utilized to investigate a variety of intra-protein interactions including functional residues[154], long-range structural interactions[155], exposed and buried salt bridges[156, 157], and hydrogen bond networks[158] as well as protein-protein interactions[159]. While double mutant cycles provide valuable biophysical and biochemical insights[282], measuring the biochemical properties of many protein variants can be laborious and the connections between biochemical function and fitness complex[283-285].

Epistasis has also been studied *in vivo*. For example, in yeast the effects of specific mutations on fitness can be rapidly analyzed in the background of thousands of individual gene knockouts using the epistatic miniarray profile (E-MAP) approach[286], or synthetic genetic analysis (SGA)[101, 287]. While epistasis mapping by these approaches has been extremely useful for detecting physiological connections between gene products, it is not well suited to investigate intragenic epistasis, or comprehensively screen point mutants.

We previously developed an approach that we term EMPIRIC to quantify the fitness effects of all possible point mutations in a gene or region of a gene[202, 212] and used this approach to comprehensively delineate the distribution of fitness effects for a nine amino acid region of yeast Hsp90 (also known as Hsp82). Hsp90 is a homodimeric protein chaperone that plays an essential role in stress responses, kinase activation, and

hormone receptor maturation[177, 288, 289]. To successfully perform these functions, Hsp90 binds to numerous co-chaperones during the process of substrate maturation. By 2-hybrid and SGA analysis, Hsp90 has been found to interact with ~3% of the yeast proteome. The sheer number and transient nature of these interactions makes the elucidation of mechanistic detail difficult by standard biochemical methods[176, 177].

The region of Hsp90 that is the focus of this work (amino acids 582-590) has many hallmarks of a putative substrate binding interface[171, 202, 290] including two positions with solvent exposed aromatic residues (F583 and W585). In our previous work, we observed that yeast fitness requires large hydrophobic amino acids at both of these positions[202, 285], indicating that they provide a critical hydrophobic docking site. We have also observed that a buried intra-molecular hydrogen bond mediated by S586 is critical for yeast fitness indicating that the main-chain conformation of this region is important for function. Together, these observations lead us to investigate epistasis in this region in order to understand how the surface properties and main-chain conformational preferences contribute to fitness. To probe relationships between protein structure and epistasis in this region of Hsp90 we used EMPIRIC to systematically determine fitness effects of point mutants in the background of seven local anchor mutations (Figure 5.1A).

We chose anchor mutations that based on structural inspection were likely to alter either the exterior composition or the main-chain conformation of this region of Hsp90, but that were well tolerated in the parental background (Table 5.1). We hypothesized that

Figure 5.1 Experimental setup. (A) Structure of yeast Hsp90 illustrating the region (amino acids 582-590) investigated in yellow. Solvent exposed surface area calculations by AREAIMOL to the right of the structure (B) Independent fitness is calculated by multiplying the calculated fitness of each point mutant constituent (left). Library competition of double mutant libraries was performed to measure observed fitness.

Figure 5.1

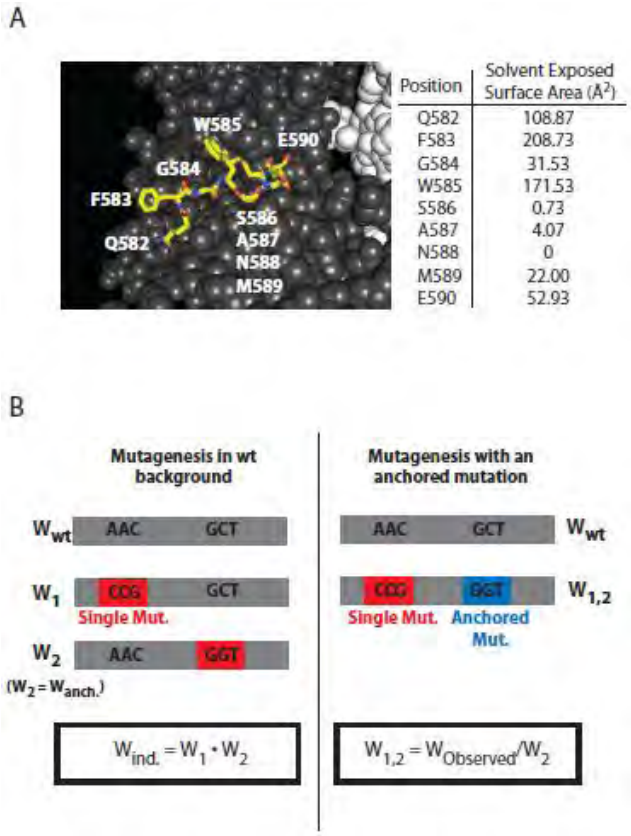

A
W585
E590
G584
F583
S586
Q582
A587
N588
M589
Position
Solvent Exposed Surface Area (Å²)
Q582 108.87
F583 208.73
G584 31.53
W585 171.53
S586 0.73
A587 4.07
N588 0
M589 22.00
E590 52.93
B
Mutagenesis in wt background
Mutagenesis with an anchored mutation
W_wt AAC GCT
W_1 CCG GCT
Single Mut.
W_2 AAC GGT
(W_2 = W_anch.)
W_ind. = W_1 · W_2
AAC GCT W_wt
CCG GCT W_1,2
Single Mut. Anchored Mut.
W_1,2 = W_Observed / W_2

Table 5.1

Anchor Mutation (codon)	Fitness Confidence Interval 2.5%	Median W	97.5%
F583N (AAC)	0.971	0.981	0.991
G584F (TTC)	0.977	0.989	1.003
G584S (TCC)	0.962	0.975	0.988
W585L (TTG)	0.967	0.976	0.986
S586G (GGT)	0.964	0.974	0.983
A587G (GGT)	0.983	0.989	0.994
N588F (TTC)	0.996	1.006	1.015

these non-conservative, but well tolerated substitutions would sensitize Hsp90 to secondary mutations and thus provide extensive sampling of the interplay between mutations impacting protein conformation and/or exterior composition. We sampled anchor mutations at both solvent accessible as well as core solvent inaccessible positions because core positions in proteins tend to have a dominant impact on structure and dynamics, while positions on the solvent accessible surface tend to play a primary role in mediating intermolecular interactions[230, 291, 292]. There are exceptions to this general trend because the impact of a mutation on protein stability, dynamics and function depends on detailed atomic interactions that are not perfectly captured by surface/core classification. For example, glycine mutations typically increase protein flexibility at any position because the lack of heavy atoms in the glycine side chain provides greater access to main chain conformations than any other amino acid. To broadly span potential mutant structural effects, we chose anchor mutations (Figure 5.1A) at two solvent exposed positions (F583N and W585L), two mutations at a glycine position (G584F and G584S), and three mutations at solvent inaccessible positions (S586G, A587G, and N588F).

To precisely investigate epistasis it is critical to control the genetic background to avoid fitness differences introduced by unintended background mutations. Controlling the genetic background is a challenge for approaches that involve isolating potentially fitness defective variants as strong selection pressure can rapidly select for secondary mutations. However, the EMPIRIC approach is ideally suited to address this challenge because all mutants are introduced into the same batch of competent yeast rapidly expanded from a

single colony. In addition, cells carrying mutations are initially expanded with the co-expression of a second copy of wt Hsp90 reducing potential selection pressure. Rapid and stringent shutoff of this second copy then enables measurement of the fitness supported by each mutant as the sole Hsp90 expressed in yeast, and the introduced Hsp90 mutations should be the predominant contributor to fitness differences in these experiments.

The datasets we obtained in this work provide direct measures of the fitness effects of 2,866 double amino acid substitutions including the magnitude and frequency of epistatic interactions between substitutions. To calculate the epistasis of mutant fitness effects, we compared direct measures of the fitness effects of double mutants to the predicted independent fitness effect of each individual mutant. We find epistatic interactions are common, and the majority of epistasis is negative in direction. Interpretation of these results in light of the structure of Hsp90 indicates a complex interplay between mutations that impact conformation and exterior facing composition.

Results and Discussion

Experimental quantification of fitness effects

To ensure continuity between the single site and double mutant data for comparison, we examined the reproducibility of fitness measurements and the overall distribution of fitness effects in our datasets. To investigate reproducibility, we included single site substitutions at the anchor position in our double mutant libraries. We compared the fitness effects of these single site substitutions with our previous measurements of all single site substitutions in this region (Figure S5.1). The strong correlation (R^2=0.93) between these full experimental replicates indicate that the estimates of fitness effects from bulk competitions are very reproducible, consistent with the strict control of both genetic background and environmental conditions in these experiments.

Having established the reproducibility of our measures of fitness effects, we next investigated the distribution of fitness effects (DFE) of the engineered double mutants. Similar to our previous observations of single mutants, we also observe a bi-modal DFE for double mutants with peaks at null-like and wt-like fitness (Figure 5.2A). We compared the frequency of fitness effects between the wild type background and anchor mutant background (Figure 5.2B). In the wild type background, the frequency of both null-like and intermediate fitness effects was lower than in the anchor mutant background, but of higher frequency in the wt-like category. These results suggest the

Figure S5.1 Correlation of selection coefficients. Selection coefficients from a previous single site selection experiment was correlated with the abridged single site library incorporated into the double mutant libraries.

Figure S5.1

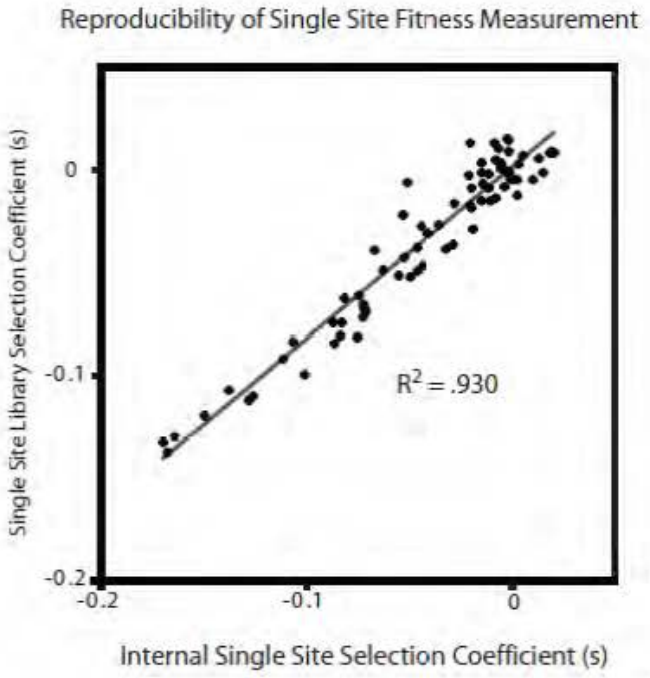
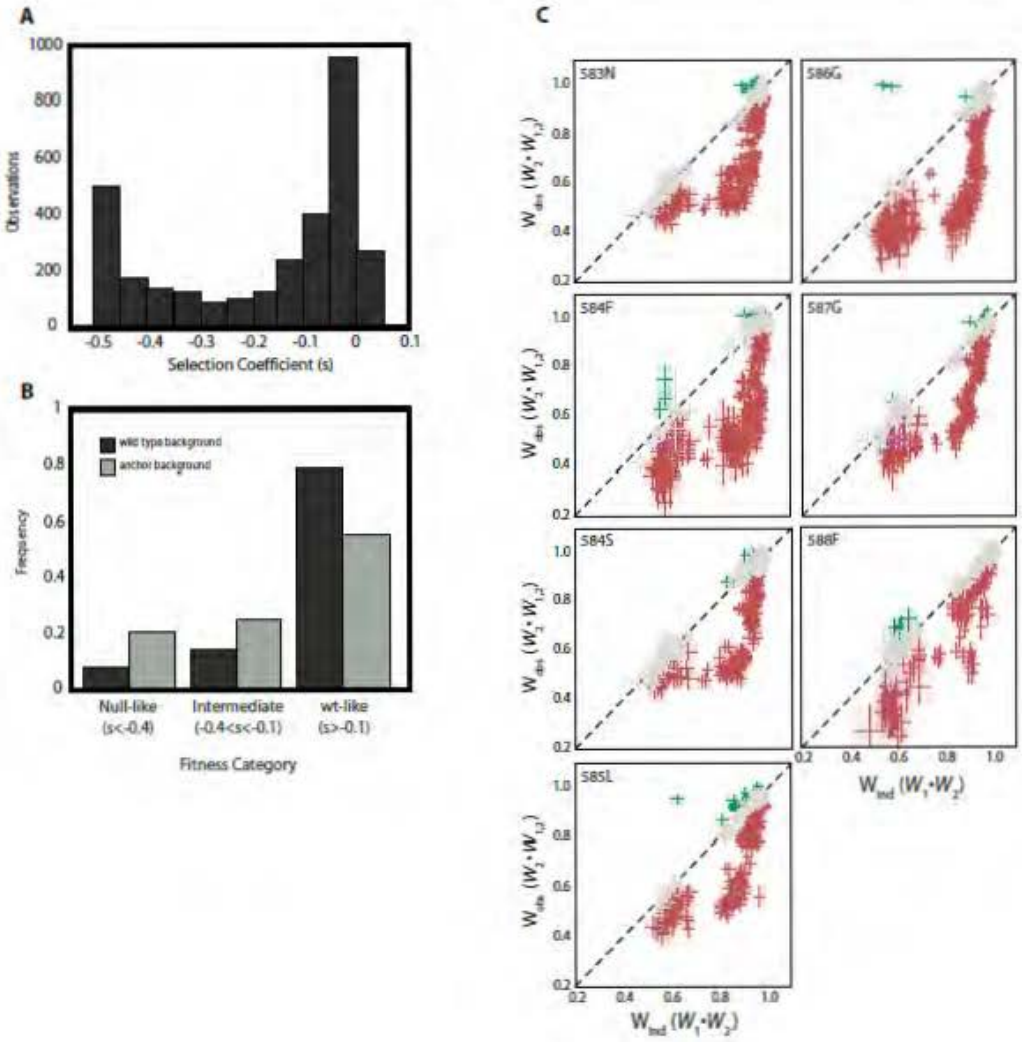

Reproducibility of Single Site Fitness Measurement
Single Site Library Selection Coefficient (s)
0
-0.1
-0.2
$R^2 = .930$
-0.2
-0.1
0
Internal Single Site Selection Coefficient (s)

Figure 5.2 Distributions of fitness and epistatic effects. (A) The distribution of fitness effects of all protein coding double mutants. (B) The frequency of fitness effects between single site and double mutant libraries. (C) The distribution of epistatic effects for all protein coding double codon substitutions based on analysis of 95% confidence intervals. Green points indicate positive epistasis, red points indicate negative epistasis, and gray points indicate independence.

Figure 5.2


A
Observations
Selection Coefficient (s)
B
wild type background
anchor background
Frequency
Null-like (s<-0.4)
Intermediate (-0.4<s<-0.1)
wt-like (s>-0.1)
Fitness Category
C
S83N
S86G
S84F
S87G
S84S
S88F
S85L
W_ind (W1•W2)

mutants chosen to occupy the anchor position do, in fact, sensitize the genetic background to epistatic effects.

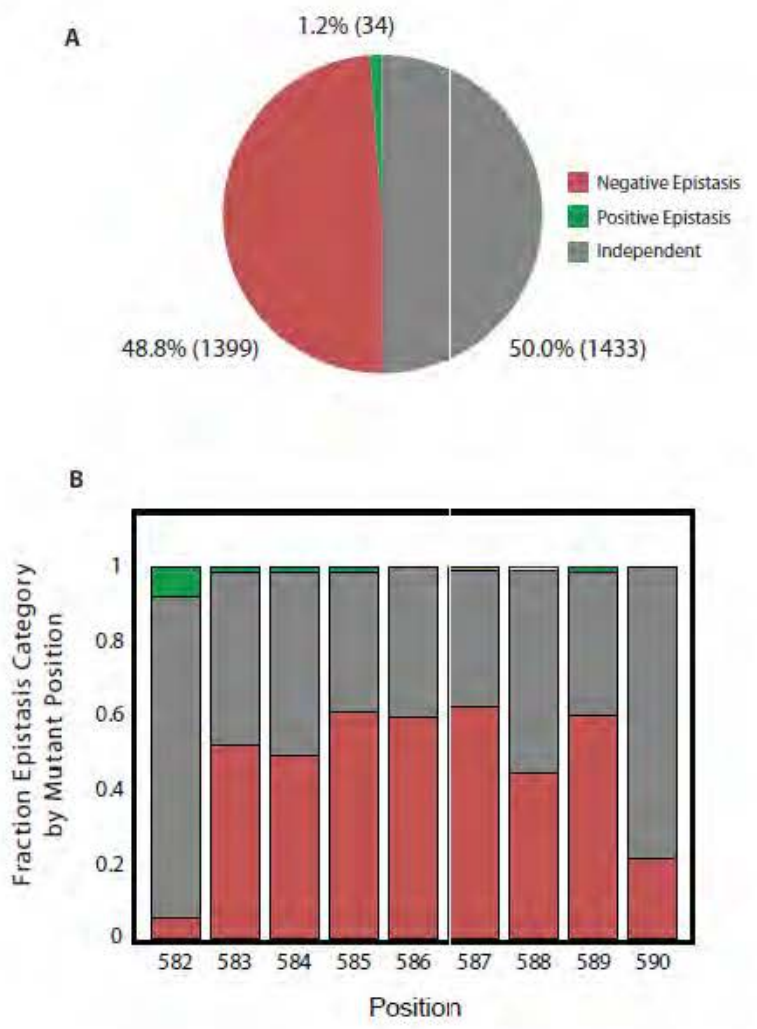
The distribution of epistatic effects

To categorize epistasis at each measured protein coding double mutant, the 95% confidence interval (CI) of fitness was estimated by the MCMC (Figure 5.2C). We observe epistatic interactions to be common, with negative epistasis occurring much more frequently than positive epistasis. Whereas negatively epistatic and independent interactions explained 98.8% of all double mutant data, positive epistasis was observed in only 1.2% of all double codon substitutions (Figure 5.3A). In order to evaluate commonalities between the type of mutant interaction and the biochemical properties of each position, the epistatic category (positive, negative or independent) was separated by single mutant position and plotted as a fraction of total interactions at each position (Figure 5.3B).

We find that of all interactions containing mutations at position 582, nearly 8% were positively epistatic whereas less than 6% were negatively epistatic. Mutations in position 590 also produced fewer negatively epistatic events (21% of all interactions), but no positively epistatic interactions. Negatively epistatic and independent mutations, on the other hand, were more common at positions other than 582 and 590 (45% at position 588 (lowest) and 62% at position 587 (highest)). The largest fraction of negative epistasis occurred in the context of position 585-587 mutations, and this is likely due to crosstalk

Figure 5.3 Frequency of epistasis. (A) The epistatic category (positive/negative/additive) of double mutants as a fraction of all protein coding double mutants. Values within parentheses indicate the number of observations. (B) The fraction of each epistatic category by mutant position.

Figure 5.3


A
1.2% (34)
Negative Epistasis
Positive Epistasis
Independent
48.8% (1399)
50.0% (1433)
B
Fraction Epistasis Category by Mutant Position
1
0.8
0.6
0.4
0.2
0
582
583
584
585
586
587
588
589
590
Position

between hydrogen bonding contributing to local destabilization at position 586 and 587, and exposed hydrophobic composition at position 585.

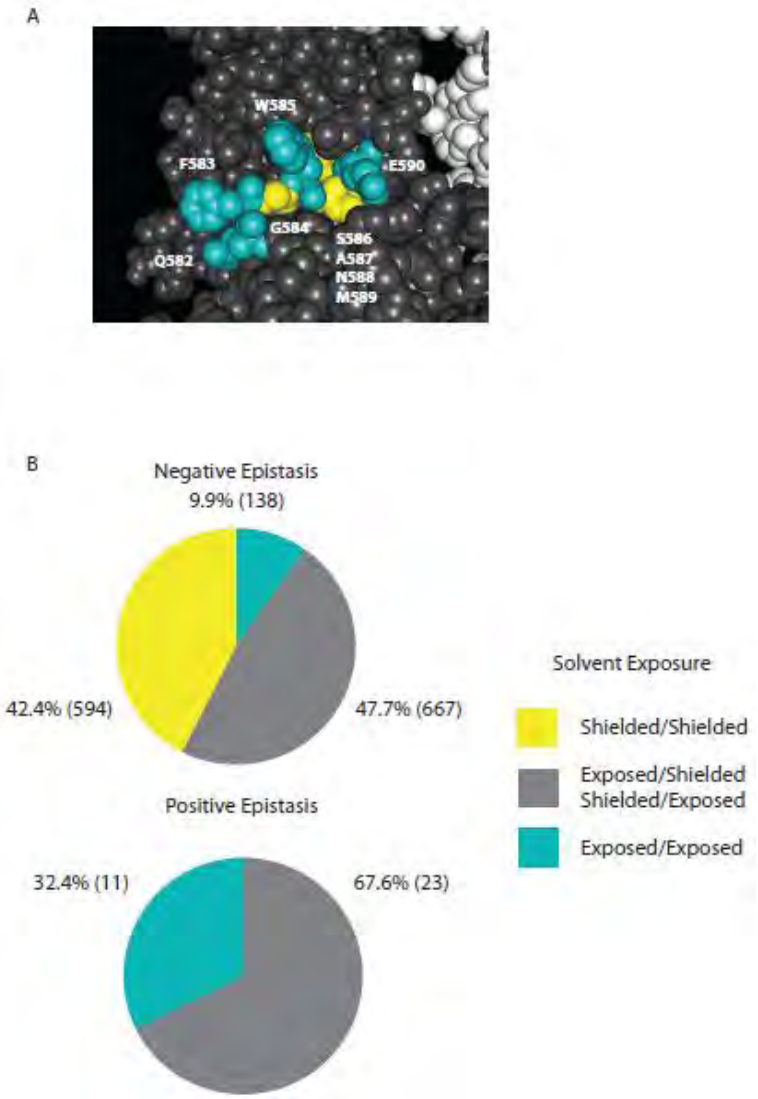
Biochemical bases of mutant interactions

We further probed mutant interactions in terms of the biochemical properties of this region in light of a solved crystal structure of Hsp90[171]. Based on this structure, amino acids 582-590 form a loop with two bulky hydrophobic residues projecting into solvent as well as several other residues buried and solvent shielded in the structure (Figure 5.4A). To probe the structural features of a position and their association with a particular category of epistasis, we classified each position as either solvent exposed or solvent shielded based on solvent exposed surface area estimation by AREAIMOL (CCP4 Software Suite) (Figure 5.1A), and grouped by epistatic category.

Based on the categorization of solvent exposure, we discovered several associations between epistasis and solvent exposed surface area. Double mutants categorized as negatively epistatic displayed a clear prominence of core mutations with 90.1% of all negative epistasis involving at least one core residue (Figure 5.4B, top). Previous biochemical analyses of stability indicate that solvent shielded mutations make evolutionarily conserved energetic contributions to global folding/unfolding[293, 294]. Biophysical analysis of mutants within this region indicate single mutants have very little effect on global folding[285] but more general studies of protein stability indicate mutations are associated with non-additive destabilizing energies[295]. A likely explanation for the

Figure 5.4 Biochemical bases for dependent interactions. (A) Solvent exposed/shielded assignment for interrogated positions on an Hsp90 monomer (gray). Cyan spheres indicate residues with significant solvent exposure and yellow spheres indicate solvent shielded positions. (B) Epistatic interactions classified by fraction exposed/shielded composition.

Figure 5.4


A
W585
F583
E590
G584
S586
A587
N588
M589
Q582
B
Negative Epistasis
9.9% (138)
42.4% (594)
47.7% (667)
Positive Epistasis
32.4% (11)
67.6% (23)
Solvent Exposure
Shielded/Shielded
Exposed/Shielded
Shielded/Exposed
Exposed/Exposed

abundance of core mutations contributing to negative epistasis is that the position of surface residues plays an important role for the proper function of this loop, and local destabilization of the solvent shielded residues in this region alters surface residue positioning, and therefore fitness.

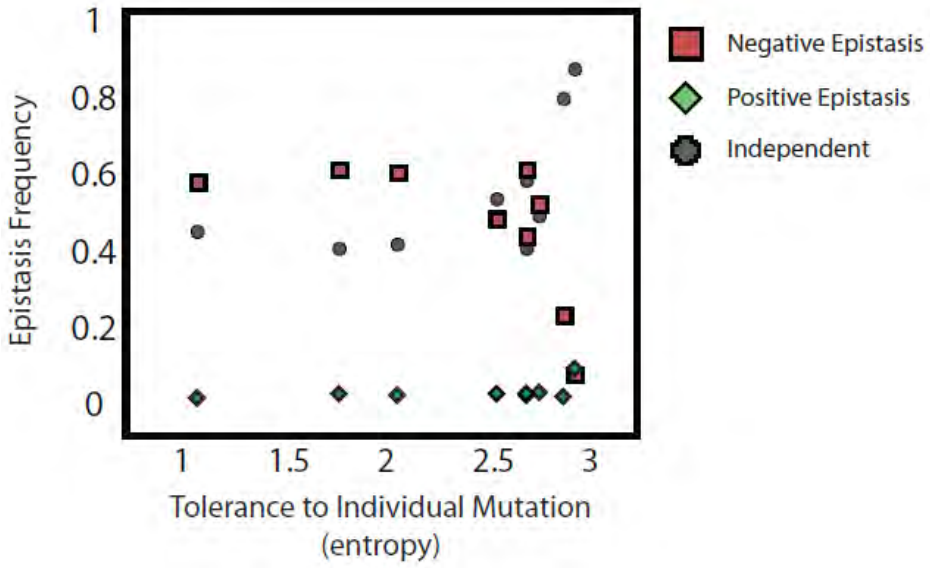
To further investigate epistatic interactions in terms of protein structure, we applied the same analysis to double mutants exhibiting positive epistasis. Based on the previous finding analysis for negative epistasis, we found no mutant combinations involved in positive epistasis could be attributed to two solvent shielded interactions, and therefore all observations of positive epistasis included at least one solvent exposed position. However, the largest fraction of double mutants in the positive epistasis category included a combination of one exposed and one shielded position. Within the same logical framework as negative epistasis, we hypothesize that the fitness defect introduced by altering the local stability of the region with a core mutation can be alleviated by a compensatory surface mutation, or vice versa. This region of Hsp90 has previously been characterized as a putative protein binding interface due to characteristic hydrophobic residues projecting into solvent[171, 202], so the overarching explanation for fitness effects within this region is based on crosstalk between the hydrophobic character of residues solvent exposed positions and the orientation of these residues dictated by solvent shielded residue packing.

Epistasis and sensitivity to mutation

Having previously described the mutational sensitivity of this region in terms of positional entropy – and finding that this region of interest contains positions of both low and high robustness - we examined how tolerance to mutation in the parental background correlated with epistasis. Using an entropy term, we quantified tolerance to mutation (with 0 representing a frozen position that cannot mutate and 3 representing a position that tolerates all 20 amino acids) against the fraction of each epistasis category per position (Figure 5.5). Negative epistasis is negatively associated with mutational promiscuity, whereas independent mutant combinations are positively associated with increased promiscuity. Positive epistasis appears slightly positively correlated with mutational robustness, but the limited number of observations may obfuscate this finding. These results appear to fit mechanistically into a broader evolution context: a position within a protein with stringent biochemical requirements is less able to participate in epistatic suppression of deleterious phenotypes than less stringent positions. Based on this framework, elucidating the biochemical requirements of a protein through means of experimentally determined DFEs may lend significant insight into potential sites of compensatory secondary mutations.

Figure 5.5 Mutational sensitivity predicts predominant epistasis category. The mutational sensitivity of each position (enumerated as entropy) correlated with the fraction of each epistatic category as a fraction of all mutations occurring at each position.

Figure 5.5


Epistasis Frequency
1
0.8
0.6
0.4
0.2
0
1
1.5
2
2.5
3
Tolerance to Individual Mutation
(entropy)
Negative Epistasis
Positive Epistasis
Independent

Conclusions

In this study we experimentally determined the distribution of epistatic effects between a large panel of closely linked double amino acid substitutions, which allowed us to draw both biochemical and evolutionary conclusions about patterns of intragenic epistasis. Our experiments indicate that epistasis between these residues of Hsp90 correlates with the biochemical properties of the region - including residue burial, solvent exposed surface area, and intraprotein interactions. These findings also lend experimental evidence of frequent negative intragenic epistasis which may not necessarily be observed in natural populations due to strong negative selection. Interestingly, among our observed interactions that were positively epistatic - the great majority may be attributed to surface-core double mutants, possibly owing to compensatory effects.

The interdependence of surface and core mutations in the context of protein evolution also has interesting connections to the work presented here. Previous analysis of the rate of evolution between surface and core residues indicates the conservation of a surface position dictates the conservation of closely associate core residues[296]. This coupling effect is especially visible in the association of positive epistasis most frequently occurring in the context of one solvent exposed and one solvent shielded mutation. Additionally, studies of protein interaction surfaces indicate more interactive protein surfaces are more highly conserved[297]. These results in combination with our previous analysis of mutational sensitivity may also indicate this region is a coevolving sector which is generally refractory to substitution, but possesses adaptive potential based on the

specific cohort of interacting proteins[298]. Taken together, this highly conserved loop of Hsp90 likely participates in a significant fraction of Hsp90s interaction network, explaining the crosstalk between surface and core residues, and further lending evidence to the hypothesis that the amino acid 582-590 region is a protein binding interface.

Specific Materials and Methods

Anchored library generation

Seven Hsp90 point mutations (F583N, G584F, G584S, W585L, S586G, A587G, and N588F) previously observed to have wt-like (growth rate within 2.6% of wt) fitness under the conditions utilized in these studies were chosen as anchors. Within each of these seven anchored Hsp90 backgrounds, systematic site saturation mutagenesis was used to introduce second point mutations throughout the amino acid 582-590 region. The amino acid position fixed as the anchor was chosen to act as a single site library control to determine if the fitness effects of individual mutations were reproducible. In addition, one position was held constant to provide an internal estimate of misreads due to processing and sequencing of samples. Mutagenesis was carried out as previously described[270].

Yeast transformation and selection conditions

Yeast manipulations and growth competitions were performed as previously described[285]. Briefly, mutants were encoded on 417GPD, a plasmid/promoter system that closely matches the endogenous expression of Hsp90[285, 299]. Each anchored library was separately transformed into the DBY288 strain of *S. cerevisiae* (can1-100 ade2-1 his3-11,15 leu2-3,12 trp1-1 ura3-1 hsp82::leu2 hsc82::leu2 ho::pgals-hsp82-his3). Transformed cells were amplified in medium containing galactose (SRGal -H +G418; per liter: 1.7g yeast nitrogen base without amino acids, 5g ammonium sulfate, 0.1g aspartic acid, 0.02g arginine, 0.03g valine, 0.1g glutamic acid, 0.4g serine, 0.2g threonine, 0.03g

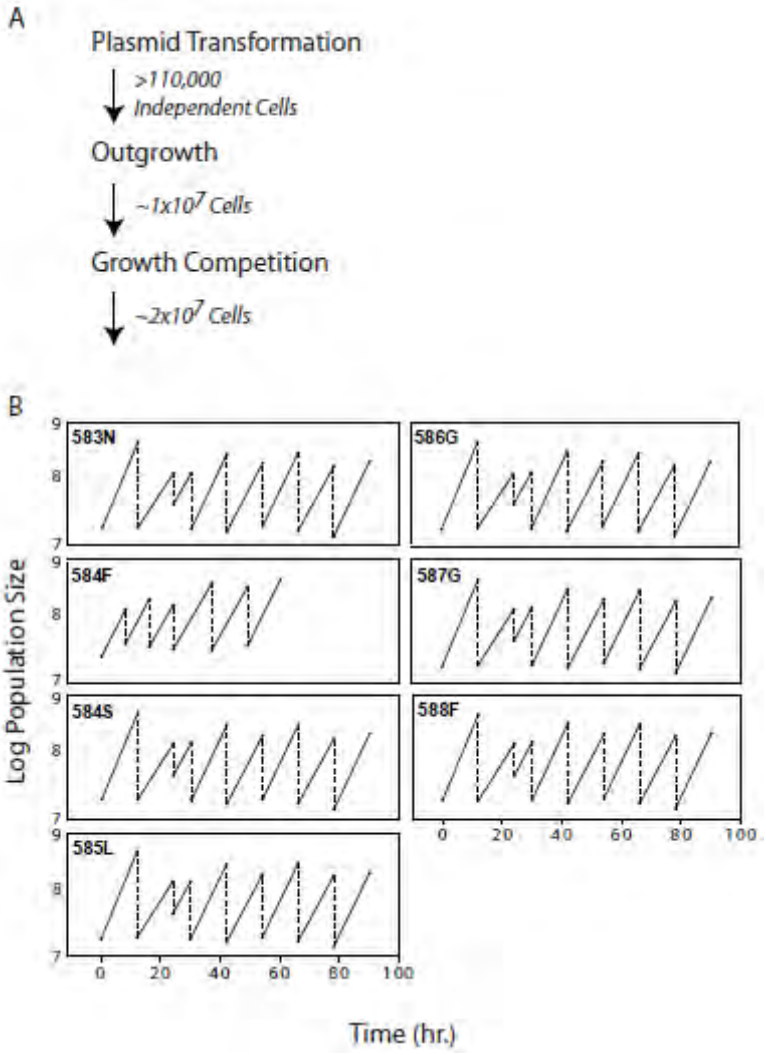
isoleucine, 0.05g phenylalanine, 0.03g tyrosine, 0.04g adenine hemisulfate, 0.02g methionine, 0.1g leucine, 0.03g lysine, 0.01g uracil 200mg G418, with 1% w/v raffinose and 1% galactose) such that wt Hsp82 protein was co-expressed along with each mutant. Transformation of the library yielded on average 110,000 individual isolates. Following the amplification of transformants, cells were diluted into fresh SRGal –H +G418 medium and grown to mid log phase. Selection was initiated by transferring cells to shutoff conditions consisting of synthetic dextrose medium (SD –H +G418; identical to SRGal –H +G418 media but with 2% dextrose in place of raffinose and galactose). Yeast cells were diluted periodically (with minimum population sizes in gross excess to library diversity) to maintain log phase growth and samples isolated at different time points in shutoff conditions (Figure S5.2).

Sequencing and data analysis

Lysis, sample preparation, and sequencing were performed as previously described[270]. In brief, plasmid DNA was harvested from yeast pellets, and the mutated region was selectively amplified by PCR and prepared with Illumina primer binding sites and a barcode used to distinguish each time point. Sequencing was performed on the Illumina HiSeq platform at the UMass Medical School core sequencing facility. Sequencing produced 21.6 million reads of 99% read confidence per base and the relative abundance of each mutant at each time point extracted as previously described[270].

Figure S5.2 Population management. (A) Preparation of cells for library selection. Population size was maintained at $>10^7$ cells to ensure complete library sampling. (B) Growth and dilution scheme for experimental populations. Solid lines indicate growth and dashed lines indicate dilutions. 584F competition was performed separately from other competitions, therefore dilution and time in selection differs.

Figure S5.2


A
Plasmid Transformation
>110,000
Independent Cells
Outgrowth
~1x10^7 Cells
Growth Competition
~2x10^7 Cells
B
583N
586G
584F
587G
584S
588F
585L
Log Population Size
Time (hr.)

Data processing

In order to correct for sequencing errors, we performed an outlier correction for each individual mutant. Outliers are identified based on the residuals of a linear regression for each mutant's trajectory, based on a modified z-score. We use a Bayesian Monte Carlo Markov Chain (MCMC) modeling approach to obtain confidence intervals for the selection coefficients. The model is based on the assumption that, starting from an initial population size, each mutant grows exponentially in the bulk competition. At each time point, samples are drawn according to a multinomial distribution, with parameters represented by the overall read number (N) and the expected fraction of individuals at that time point, given the proposed growth rates (p). We implemented a Metropolis-Hastings algorithm in R to compute the stationary distribution of the MCMC. A burn-in phase of 1,000,000 accepted parameter combinations and a subsequent estimation phase of 10,000,000 accepted values ensured convergence of the MCMC for the high number of parameters that are estimated simultaneously. Sub-sampling of every 1,000th parameter combination resulted in the data sets used for further analysis. The R package "coda" was used to ensure convergence of the MCMC. A corrected mean (i.e., discarding the contribution of mutants outside two standard deviations on both sides of the mean) of the growth rates of all mutations synonymous to the sequence that is similar to the original wild type except for the anchored mutation is used as normalization constant that determines r=1 (s=0) for each data set, and all growth rates are rescaled accordingly.

Classification of mutations

A combination of mutations was classified as positively epistatic if the lower limit of the combined 95% confidence interval (CI) of the product of the double mutant's growth rate and the anchored mutant's growth rate in the wt background was larger than the upper limit of the combined 95% CI of the product of the individual mutant's growth rates in the wt background. A combination of mutations was classified as negatively epistatic if the upper limit of the combined 95% CI of the product of the double mutant's growth rate and the anchored mutant's growth rate in the wt background was smaller than the upper limit of the combined 95% CI of the product of the individual mutant's growth rates in the wt background. Combined CIs were obtained by multiplying the individual lower and upper limits.

Acknowledgements

The computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics. Financial support for this work was provided by grant R01-GM083038 from the National Institutes of Health to D.N.A.B., and by grants from the Swiss National Science Foundation and the European Research Council to J.D.J.

Chapter VI - General Discussion

Summary

The basis and necessity of the work presented within this dissertation is generally focused on the accurate measurement of the distribution of fitness effects of new mutations, and the applicability of these measurements to long-standing models in population genetics and molecular evolution. In Chapter II, I present a novel methodology for accurate and high-throughput measurement relative fitness in the yeast *Saccharomyces cerevisiae* coined EMPIRIC. Outlined in detail is the methodology necessary to examine any essential gene of interest in yeast, from library generation to data analysis. Although this dissertation is both 'yeast-centric' and focused on a relatively small region of the Hsp90 protein, it should be reiterated the EMPIRIC methodology can be much more broadly applied to model systems including cancer cells, viruses, and bacterium to study processes including drug resistance, structure-function relationships, rational design/directed evolution, and others. This method can also be scaled and modified to suit the throughput and signal-to-noise necessities of a specific system, as well as being linked to other technologies such as cell surface display and FACS analysis.

As described in chapters III and IV, I presented the distribution of fitness effects of new mutations in the amino acid 582-590 region of yeast Hsp90 in the context of evolutionarily relevant variables including optimal growth conditions, reduced expression level, thermal stress, and osmotic stress. Most directly interpretable from these results is the pervasive nature of bimodal DFEs first predicted by Ohta and Kimura. Regardless of

environmental or expression level perturbations, we consistently observe mutations distributed as either strongly deleterious or near-neutral with a very limited number of mutations conferring intermediate fitness effects. Although the relative proportion and mean fitness of these maxima change in response to perturbation, the bimodal distribution is always retained. Results from Roscoe *et. al* [212] indicate this finding is not simply an artifact of this region of Hsp90, but may be broadly applicable across the proteome, although studies of several proteins with diverse structures and functions are needed to validate these findings.

The frequency and magnitude of beneficial mutations is also of central importance to our current understanding of evolutionary biology. However, experimental characterization of beneficial mutations has lagged behind theoretical predictions because beneficial mutations are relatively rare[300]. Utilizing the EMPIRIC method, we can generate an unbiased measurement of not only deleterious and near-neutral mutations, but beneficial mutations as well (if they exist in the system). In Chapter III, the initial biomodal fitness landscape under thermal stress conditions did not contain any significantly beneficial mutations, but this is not necessarily an unexpected outcome as the role of Hsp90 at elevated temperature is well characterized, and the protein sequence is likely optimized for this condition. However, we did find that phylogenetic conservation may be a poor predictor of mutational promiscuity, and the current genetic code is highly optimized to sample single mutations of wild type-like fitness.

Although our initial study did not identify any beneficial mutations, in Chapter IV when environmental conditions were perturbed from optimum to one of three other conditions, several beneficial mutations were isolated under elevated salinity with a maximum fitness benefit of ~8%, which we find to be consistent with many of Fisher's expectations of adaptive evolution. The observed relationships of fitness between conditions was also remarkably consistent with a one dimensional Fisher geometry which encompassed 95.8% of all mutations and 99.5% if the sign of wild type-like mutations was ignored. Beneficial mutations isolated in one condition also paid a fitness cost in alternate conditions (cost of adaptation) indicating environmental specialization of these beneficial mutations. Interestingly, the finding that growth rate was most retarded in conditions of combined environmental perturbation while producing fewer adaptive mutations than osmotic stress alone appears to support Haldane's predictions of genetic interference during selection for multiple traits simultaneously[81].

Epistasis is a vitally important feature of adaptive evolution, especially in the context of pathogenesis and drug resistance. Although studied in the context of interaction networks and biophysical interactions within proteins, a comprehensive distribution of intragenic epistatic effects may allow the EMPIRIC technique to be repurposed for use in elucidating functional domains of proteins with no solved structure. In chapter V of this dissertation I present a distribution of intragenic epistatic effects in the context of seven genetic backgrounds. We find epistatic interactions are common, and the majority of epistasis is negative in directionality. However, we also isolate a number

of positively epistatic interactions and find negative epistasis to be highly correlated with mutations in solvent shielded positions whereas positive epistasis is correlated with surface positioning, consistent with the prediction that this solvent exposed hydrophobic loop is a docking site for protein-protein interaction.

Future Directions

Although the work presented here is a comprehensive interrogation of a biochemically intriguing region of Hsp90, there are still avenues which have been left unexplored. Hsp90 is a highly interactive protein, and the region encompassing amino acids 582-590 is of obvious interest as a putative docking site. We and others[179] have uncovered several interesting features of this loop which are indicative of a protein interaction surface, but these studies do not present direct evidence for the nature and number of interactions taking place in conjunction with this region. Due to the numerous and transient nature of potential interactions, many standard analytical techniques such as crystallography, NMR, and various binding studies are not ideal. Utilizing the EMPIRIC technique, it may now be possible to study specific interactions in much finer detail by combining our genetic approaches with biochemical analyses such as molecular dynamics simulations or potentially FACS analysis if a suitably small set of putative docking partners can be identified

In the format presented here, EMPIRIC relies on the ability to link mutations to organismal fitness. In the case of classic Hsp90 specific assays for steroid hormone

receptor maturation, Hsp90 function relies on clients which are not endogenous to yeast, and therefore are not directly linked to fitness. However, through simple genetic manipulations placing an essential metabolic or chemotherapeutic resistance gene downstream from a steroid hormone receptor response element, it may be possible to directly link the ability of Hsp90 to mature these specific clients to growth rate. Mutations which cause either partial or wholly defective interactions with a specific client would present a phenotype proportional to the defect. By studying the underlying structural and biophysical requirements of this and other regions in the context of a specific function of Hsp90, it may be possible to biochemically define the interaction surface of binding to steroid hormone receptors.

As well as steroid hormone receptors, Hsp90 interacts with a number of signal transduction molecules including a repertoire of kinases. A second well characterized assay for Hsp90 function is for the activation of the tyrosine kinase v-src. However, yeast do not possess endogenous tyrosine kinases, and when v-src is activated in yeast cells by interaction with Hsp90, a fitness defect caused by an undetermined toxic mechanism is observed. Since active v-src introduces a fitness cost, it presents an interesting opportunity to studying the role of Hsp90 mutations in the activation of kinases in general. Instead of the typical fitness readout generated by EMPIRIC in which mutants of wild type fitness eventually dominate the culture, in a kinase activation study, mutations of wild type fitness would be depleted from culture whereas mutants defective for interaction with v-src would eventually dominate culture. With the inverse nature of

selection in this system, the readout is complicated by deciphering whether a mutant of low fitness is due to a defect in kinase activation or unfit for a myriad of other reasons. By comparison of a fitness landscape without v-src present (of which we already have a number of examples) to a competition with v-src present, a Venn diagram approach could be used to parse general fitness defects from mutations capable of activating kinases, and therefore have the potential to illuminate the mechanism of interaction.

Extension of the EMPIRIC Methodology

The general EMPIRIC methodology can be logically extended not only to other regions of yeast Hsp90, but to a vast number of other proteins to answer questions in evolutionary biology, protein biochemistry, and drug design. The most obvious and straightforward extension of the work presented here is the systematic analysis of the distribution of fitness effects for an ensemble of structurally and functionally diverse proteins. Although we observe a bimodal distribution of fitness effects for this region of Hsp90, it is only now becoming clear how applicable these findings are to the entire yeast proteome. As previously mentioned, Roscoe *et al.* have determined the DFE for the entire ubiquitin gene, and found a bimodal distribution for new mutations[212]. However, both Hsp90 and ubiquitin are highly conserved proteins with pleiotropic roles in the cell. Simple extensions to further test the bimodal DFE hypothesis include monofunctional metabolic enzymes, structural and cytoskeletal proteins, membrane proteins, and a myriad of other functional classes. Perhaps following exhaustive studies of numerous proteins, we can begin to generate a database of DFEs for use in evolutionary biology, as

well as a library of known mutations and their fitness effects which could be broadly applied to biochemical questions.

Comprehensive mutational analyses may also be of great value to structural biology, especially in the context of difficult-to-study systems such as membrane proteins. Historically, membrane proteins have been extremely difficult to characterize because of their unique native amphipathic microenvironment. Methods such as crystallography and NMR have resulted in relatively few solved structures (as compared to water soluble proteins) because lipid conditions of the membrane must be accurately mimicked to produce refractive crystals with biologically relevant structures. High-throughput mutational fitness analyses combined with structural modeling and biophysical analyses may be able to fill the current gap in structural knowledge by better defining the unique residue properties required for both external and internal structures.

Characterization of chemotherapeutic drug resistance is a rational synthesis of evolutionary biology, biochemistry, and drug design. Drug resistance is a commonly isolated phenomenon in populations of viruses, bacteria, yeast, protists, and cancer cells undergoing chemotherapeutic therapies due to strong selective pressure for mutations conferring a fitness benefit in the presence of drug. Practically, the ability of cells to adapt to pharmaceutical treatment is a significant economic and medical problem, yet few solutions exist to predict resistance mutations before they occur in a population. In the EMPIRIC technique, we have the potential to screen all possible point mutations within a

pharmaceutically relevant region of a protein under conditions of drug selection. By choosing drug concentrations near the half maximal inhibitory concentrations, even small fitness advantages conferred by a mutation may be detectible and further studied to predict potential routes of drug resistance. Additionally, discovering drug resistance mutations before they become clinically relevant may allow for the rational design of better drugs with difficult-to-transverse adaptive pathways.

Broader Impact

The need for the comprehensive understanding of the mechanistic detail of evolution has never been greater in the face of challenges such as sustainable energy, world food supply, and climate change. Evolutionary thought is no longer limited to biology, but has become an integral component in fields as diverse as sociology and anthropology, computer science, and criminal justice[301]. We find ourselves at a monumental time in human history where, for the first time, it is possible to analyze the genetic blueprint of every organism on earth to solve problems which seemed intractable a generation ago. My hope is that the work presented in this dissertation can act as both a tool to understand evolutionary biology as well as an experimental starting point for understanding the full distribution of mutational effects on fitness.

The process of evolution is continuous and dynamic as all organisms attempt to adapt to new challenges and changing environments. Although the evolutionary biology literature is vast, the experimental evidence for many evolutionary processes remains

relatively scarce due to our difficulty in adapting logically sound and controlled experimental systems to a ~4 billion year old ongoing trial. As we and others develop methodologies to understand the mechanisms of evolution, we may potentially not only be able to rewind the tape of life to discover our origins, but have the ability to look forward. With a more sound understanding of specific genotype to phenotype relationships, the ability to make accurate predictions about human health will allow for targeted preventative medicine and disease treatment to not only extend the quality and quantity of human life, but also address many of the economic and social costs associated with disease.

More broadly, many of the findings presented here can be conceptually applied to the numerous challenges facing the human race. By uncovering adaptive mutations in altered environmental conditions, it may be possible to design better crops to feed a growing population in the face of climate change. Likewise, understanding the fitness consequences of epistasis may allow for improved design of enzymatic and whole organism systems to generate sustainable energy sources such as biofuels, hydrogen cells, and solar conversion systems. While I certainly do not claim to have solved the woes of humankind, my hope is the results herein presented may one day play a small role in addressing some of our species gravest challenges.

Bibliography

1. Kirk, G. S., Raven, J. E. & Schofield, M. 514 (Cambridge University Press, Cambridge, UK, 1983).
2. Cicero, M. T. De Natura Deorum: With an English Translation by H. Rackham, M.A. (Havard University Press, Cambridge, MA, USA, 1933).
3. Needham, J. & Ronan, C. A. The Shorter Science and Civilization in China: An Abridgement by Colin A. Ronan (Cambridge University Press), 1995).
4. Darwin, C. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life (Down, Bromley, Kent), 1859).
5. Wallace, A. R. (Natural History Museum: Wallace Letters Online), 2013).
6. Tschermak, E. Concerning Artificial Crossing in Pissum Sativum. Genetics 35(5), 42-47 (1950).
7. de Vries, H. Concerning the Law of Segregation of Hybrids. Genetics 35(5), 30-32 (1950).
8. Correns, C. G. Mendel's Law Concerning the Behavior of Progeny of Varietal Hybrids. Genetics 35(5), 33-41 (1950).
9. Mendel, G. in Meetings of the Brünn Natural History Society (ed. Bateson, W.) 39 (Brünn, Czech Republic, 1865).
10. Castle, W. E. & Little, C. C. Reversion in guinea-pigs and its explanation: Experimental studies of the inheritance of color in mice. (1913).
11. Fisher, R. The Correlation Between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh 52, 399-433 (1918).
12. Fisher, R. The Genetical Theory of Natural Selection (Oxford University Press, 1930).
13. Skipper, R. The persistence of the R.A. Fisher-Sewall Wright controversy. Biology and Philosophy 17, 341-367 (2002).
14. Wright, S. The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution. Proceedings of the Sixth International Congress of Genetics, 356-366 (1932).
15. Dobzhansky, T. Genetics and the Origin of Species (Columbia University Press, New York, 1941).
16. Ford, E. B. Polymorphism and Taxonomy. Heredity 9, 255-264 (1954).
17. Mayr, E. Systematics and the Origin of Species (Columbia University Press, 1942).
18. Kimura, M. Evolutionary rate at the molecular level. Nature 217, 624-6 (1968).
19. Ohta, T. Slightly deleterious mutant substitutions in evolution. Nature 246, 96-8 (1973).
20. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171, 737-8 (1953).
21. Franklin, R. E. & Gosling, R. G. Molecular configuration in sodium thymonucleate. Nature 171, 740-1 (1953).

22. Haas, J. W. The Reverend Dr William Henry Dallinger, F.R.S. (1839-1909). Notes Rec R Soc Lond 54, 53-65 (2000).
23. Huey, R. B. & Rosenzweig, F. in Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments (eds. Garland, T. & Rose, M. R.) 752 (University of California Press, 2009).
24. Rose, M. R. & Charlesworth, B. Genetics of life history in Drosophila melanogaster. II. Exploratory selection experiments. Genetics 97, 187-96 (1981).
25. Rose, M. R. & Graves, J. L., Jr. What evolutionary biology can do for gerontology. J Gerontol 44, B27-9 (1989).
26. Rose, M. R. Genetics of increased lifespan in Drosophila. Bioessays 11, 132-5 (1989).
27. Rose, M. R., Vu, L. N., Park, S. U. & Graves, J. L., Jr. Selection on stress resistance increases longevity in Drosophila melanogaster. Exp Gerontol 27, 241-50 (1992).
28. Rose, M. R., Matos, M. & Passananti, H. B. Methuselah Flies: A Case Study in the Evolution of Aging (World Scientific Publishing Company Inc., 2004).
29. 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, 7899-906 (2008).
30. Khan, A. I., Dinh, D. M., Schneider, D., Lenski, R. E. & Cooper, T. F. Negative epistasis between beneficial mutations in an evolving bacterial population. Science 332, 1193-6 (2011).
31. Wielgoss, S. et al. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. Proc Natl Acad Sci U S A 110, 222-7 (2013).
32. Blount, Z. D., Barrick, J. E., Davidson, C. J. & Lenski, R. E. Genomic analysis of a key innovation in an experimental Escherichia coli population. Nature 489, 513-8 (2012).
33. Seresht, A. K. et al. Long-term adaptation of Saccharomyces cerevisiae to the burden of recombinant insulin production. Biotechnol Bioeng (2013).
34. Chen, X., Gao, B., Shi, W. & Li, Y. [Expression and secretion of human interferon alpha A in yeast Kluyveromyces lactis]. Yi Chuan Xue Bao 19, 284-8 (1992).
35. Fleer, R. et al. High-level secretion of correctly processed recombinant human interleukin-1 beta in Kluyveromyces lactis. Gene 107, 285-95 (1991).
36. Iwata, T. et al. Efficient secretion of human lysozyme from the yeast, Kluyveromyces lactis. Biotechnol Lett 26, 1803-8 (2004).
37. Zhu, D. X. et al. Purification and characterization of the biologically active human truncated macrophage colony-stimulating factor expressed in Saccharomyces cerevisiae. Biol Chem Hoppe Seyler 374, 903-8 (1993).
38. Fleer, R. et al. Stable multicopy vectors for high-level secretion of recombinant human serum albumin by Kluyveromyces yeasts. Biotechnology (N Y) 9, 968-75 (1991).

39. Belter, J. G., Carey, H. V. & Garland, T., Jr. Effects of voluntary exercise and genetic selection for high activity levels on HSP72 expression in house mice. J Appl Physiol 96, 1270-6 (2004).
40. Wallace, I. J. et al. Functional significance of genetic variation underlying limb bone diaphyseal structure. Am J Phys Anthropol 143, 21-30 (2010).
41. Thomson, S. L., Garland Jr, T., Swallow, J. G. & Carter, P. A. Response of Sod-2 enzyme activity to selection for high voluntary wheel running. Heredity (Edinb) 88, 52-61 (2002).
42. Dlugosz, E. M., Chappell, M. A., McGillivray, D. G., Syme, D. A. & Garland, T., Jr. Locomotor trade-offs in mice selectively bred for high voluntary wheel running. J Exp Biol 212, 2612-8 (2009).
43. Rhodes, J. S. & Garland, T. Differential sensitivity to acute administration of Ritalin, apomorphine, SCH 23390, but not raclopride in mice selectively bred for hyperactive wheel-running behavior. Psychopharmacology (Berl) 167, 242-50 (2003).
44. Keeney, B. K. et al. Differential response to a selective cannabinoid receptor antagonist (SR141716: rimonabant) in female mice from lines selectively bred for high voluntary wheel-running behaviour. Behav Pharmacol 19, 812-20 (2008).
45. Hartmann, J. et al. Fine mapping of "mini-muscle," a recessive mutation causing reduced hindlimb muscle mass in mice. J Hered 99, 679-87 (2008).
46. Mullis, K. et al. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 51 Pt 1, 263-73 (1986).
47. Bartlett, J. M. & Stirling, D. A short history of the polymerase chain reaction. Methods Mol Biol 226, 3-6 (2003).
48. Wells, J. A., Vasser, M. & Powers, D. B. Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites. Gene 34, 315-23 (1985).
49. Reidhaar-Olson, J. F. & Sauer, R. T. Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. Science 241, 53-7 (1988).
50. Leung, D. W., Chen, E. & Goeddel, D. V. A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. Technique 1, 11-15 (1989).
51. Cadwell, R. C. & Joyce, G. F. Mutagenic PCR. PCR Methods Appl 3, S136-40 (1994).
52. Stemmer, W. P. Rapid evolution of a protein in vitro by DNA shuffling. Nature 370, 389-91 (1994).
53. Hartley, J. L., Temple, G. F. & Brasch, M. A. DNA cloning using in vitro site-specific recombination. Genome Res 10, 1788-95 (2000).
54. Smith, G. P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science 228, 1315-7 (1985).

55. Mattheakis, L. C., Bhatt, R. R. & Dower, W. J. An in vitro polysome display system for identifying ligands from very large peptide libraries. Proc Natl Acad Sci U S A 91, 9022-6 (1994).
56. Oliphant, A. R., Brandl, C. J. & Struhl, K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. Mol Cell Biol 9, 2944-9 (1989).
57. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide libraries. Nat Biotechnol 15, 553-7 (1997).
58. Yang, X., Kathuria, S. V., Vadrevu, R. & Matthews, C. R. Betaalpha-hairpin clamps brace betaalphabeta modules and can make substantive contributions to the stability of TIM barrel proteins. PLoS One 4, e7179 (2009).
59. Ibarra-Molero, B., Zitzewitz, J. A. & Matthews, C. R. Salt-bridges can stabilize but do not accelerate the folding of the homodimeric coiled-coil peptide GCN4-p1. J Mol Biol 336, 989-96 (2004).
60. Serrano, L., Sancho, J., Hirshberg, M. & Fersht, A. R. Alpha-helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. J Mol Biol 227, 544-59 (1992).
61. Shim, J. Y. Understanding functional residues of the cannabinoid CB1. Curr Top Med Chem 10, 779-98 (2010).
62. Volpato, J. P. & Pelletier, J. N. Mutational 'hot-spots' in mammalian, bacterial and protozoal dihydrofolate reductases associated with antifolate resistance: sequence and structural comparison. Drug Resist Updat 12, 28-41 (2009).
63. Kerscher, S., Kashani-Poor, N., Zwicker, K., Zickermann, V. & Brandt, U. Exploring the catalytic core of complex I by Yarrowia lipolytica yeast genetics. J Bioenerg Biomembr 33, 187-96 (2001).
64. Chillakuri, C. R., Sheppard, D., Lea, S. M. & Handford, P. A. Notch receptor-ligand binding and activation: insights from molecular studies. Semin Cell Dev Biol 23, 421-8 (2012).
65. Campbell, K. S. & Purdy, A. K. Structure/function of human killer cell immunoglobulin-like receptors: lessons from polymorphisms, evolution, crystal structures and mutations. Immunology 132, 315-25 (2011).
66. Rodriguez-Martin, T. et al. Tau phosphorylation affects its axonal transport and degradation. Neurobiol Aging (2013).
67. Oh, Y. et al. Mitotic exit kinase Dbf2 directly phosphorylates chitin synthase Chs2 to regulate cytokinesis in budding yeast. Mol Biol Cell 23, 2445-56 (2012).
68. Moreira, I. S., Fernandes, P. A. & Ramos, M. J. Hot spots--a review of the protein-protein interface determinant amino-acid residues. Proteins 68, 803-12 (2007).
69. Kouadio, J. L., Horn, J. R., Pal, G. & Kossiakoff, A. A. Shotgun alanine scanning shows that growth hormone can bind productively to its receptor through a drastically minimized interface. J Biol Chem 280, 25524-32 (2005).

70. Simonsen, S. M. et al. Alanine scanning mutagenesis of the prototypic cyclotide reveals a cluster of residues essential for bioactivity. J Biol Chem 283, 9805-13 (2008).
71. Arguello, J. M., Whitis, J. & Lingrel, J. B. Alanine scanning mutagenesis of oxygen-containing amino acids in the transmembrane region of the Na,K-ATPase. Arch Biochem Biophys 367, 341-7 (1999).
72. Traxlmayr, M. W. & Obinger, C. Directed evolution of proteins for increased stability and expression using yeast display. Arch Biochem Biophys 526, 174-80 (2012).
73. Gershenson, A. & Arnold, F. H. Enzyme stabilization by directed evolution. Genet Eng (N Y) 22, 55-76 (2000).
74. Akbulut, N., Tuzlakoglu Ozturk, M., Pijning, T., Issever Ozturk, S. & Gumusel, F. Improved activity and thermostability of Bacillus pumilus lipase by directed evolution. J Biotechnol 164, 123-9 (2013).
75. Wang, J. et al. Enhanced activity of Rhizomucor miehei lipase by directed evolution with simultaneous evolution of the propeptide. Appl Microbiol Biotechnol 96, 443-50 (2012).
76. Sideri, A., Goyal, A., Di Nardo, G., Tsotsou, G. E. & Gilardi, G. Hydroxylation of non-substituted polycyclic aromatic hydrocarbons by cytochrome P450 BM3 engineered by directed evolution. J Inorg Biochem 120, 1-7 (2013).
77. Joo, H., Arisawa, A., Lin, Z. & Arnold, F. H. A high-throughput digital imaging screen for the discovery and directed evolution of oxygenases. Chem Biol 6, 699-706 (1999).
78. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94, 441-8 (1975).
79. Liu, L. et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012, 251364 (2012).
80. Quail, M. A. et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13, 341 (2012).
81. Haldane, J. B. S. A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. 838-844 (1927).
82. Gallet, R., Cooper, T. F., Elena, S. F. & Lenormand, T. Measuring selection coefficients below 10(-3): method, questions, and prospects. Genetics 190, 175-86 (2012).
83. Schlager, G. & Dickie, M. M. Natural mutation rates in the house mouse. Estimates for five specific loci and dominant mutations. Mutat Res 11, 89-96 (1971).
84. Moses, A. M. & Davidson, A. R. In vitro evolution goes deep. Proc Natl Acad Sci U S A 108, 8071-2 (2011).
85. O'Bleness, M., Searles, V. B., Varki, A., Gagneux, P. & Sikela, J. M. Evolution of genetic and genomic features unique to the human lineage. Nat Rev Genet 13, 853-66 (2012).

86. Langergraber, K. E. et al. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proc Natl Acad Sci U S A 109, 15716-21 (2012).
87. Linnen, C. R., Kingsley, E. P., Jensen, J. D. & Hoekstra, H. E. On the origin and spread of an adaptive allele in deer mice. Science 325, 1095-8 (2009).
88. Hoekstra, H. E., Hirschmann, R. J., Bundey, R. A., Insel, P. A. & Crossland, J. P. A single amino acid mutation contributes to adaptive beach mouse color pattern. Science 313, 101-4 (2006).
89. Nielsen, R. & Yang, Z. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol Biol Evol 20, 1231-9 (2003).
90. Sanjuan, R., Moya, A. & Elena, S. F. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci U S A 101, 8396-401 (2004).
91. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. Genetics 156, 297-304 (2000).
92. Seaborg, D. M. Was Wright right? The canonical genetic code is an empirical example of an adaptive peak in nature; deviant genetic codes evolved using adaptive bridges. J Mol Evol 71, 87-99 (2010).
93. Jones, S. (Core Coursework, Block II, UMass Medical School GSBS, 2013).
94. Montelone, B. A. Yeast Mating Type (ed. Sciences, E. o. L.) (MacMillan Publishing Group, 2002).
95. Hampsey, M. A review of phenotypes in Saccharomyces cerevisiae. Yeast 13, 1099-133 (1997).
96. Wloch, D. M., Szafraniec, K., Borts, R. H. & Korona, R. Direct estimate of the mutation rate and the distribution of fitness effects in the yeast Saccharomyces cerevisiae. Genetics 159, 441-52 (2001).
97. Baryshnikova, A. et al. Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. Nat Methods 7, 1017-24 (2010).
98. DeLuna, A. et al. Exposing the fitness contribution of duplicated genes. Nat Genet 40, 676-81 (2008).
99. Hegreness, M., Shoresh, N., Hartl, D. & Kishony, R. An equivalence principle for the incorporation of favorable mutations in asexual populations. Science 311, 1615-7 (2006).
100. Jarosz, D. F. & Lindquist, S. Hsp90 and environmental stress transform the adaptive value of natural genetic variation. Science 330, 1820-4 (2010).
101. Tong, A. H. et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 294, 2364-8 (2001).
102. St Onge, R. P. et al. Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. Nat Genet 39, 199-206 (2007).
103. Roguev, A., Wiren, M., Weissman, J. S. & Krogan, N. J. High-throughput genetic interaction mapping in the fission yeast Schizosaccharomyces pombe. Nat Methods 4, 861-6 (2007).

104. Babu, M. et al. Systems-level approaches for identifying and analyzing genetic interaction networks in Escherichia coli and extensions to other prokaryotes. Mol Biosyst 5, 1439-55 (2009).
105. Hamilton, M. Population Genetics (Wiley-Blackwell, 2009).
106. Gillespie, J. H. A simple stochastic gene substitution model. Theor Popul Biol 23, 202-15 (1983).
107. Macpherson, J. M., Sella, G., Davis, J. C. & Petrov, D. A. Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila. Genetics 177, 2083-99 (2007).
108. Andolfatto, P. Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome. Genome Res 17, 1755-62 (2007).
109. Ratner, V. A., Bubenshchikova, E. V. & Vasil'eva, L. A. [Prolongation of MGE 412 transposition induction after gamma-irradiation in an isogenic line of Drosophila melanogaster]]. Genetika 37, 485-93 (2001).
110. Altenburg, E. & Muller, H. J. The Genetic Basis of Truncate Wing,-an Inconstant and Modifiable Character in Drosophila. Genetics 5, 1-59 (1920).
111. MacLean, R. C. & Buckling, A. The distribution of fitness effects of beneficial mutations in Pseudomonas aeruginosa. PLoS Genet 5, e1000406 (2009).
112. Costanzo, M. S., Brown, K. M. & Hartl, D. L. Fitness trade-offs in the evolution of dihydrofolate reductase and drug resistance in Plasmodium falciparum. PLoS One 6, e19636 (2011).
113. Zhou, Q., Liao, L. J. & Huang, H. J. [Impacts of HIV-1 resistance mutations associated with nucleoside reverse transcriptase inhibitors on viral fitness]. Bing Du Xue Bao 28, 291-6 (2012).
114. McCubrey, J. A. et al. Ras/Raf/MEK/ERK and PI3K/PTEN/Akt/mTOR cascade inhibitors: how mutations can result in therapy resistance and how to overcome resistance. Oncotarget 3, 1068-111 (2012).
115. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. Science 188, 107-16 (1975).
116. Manceau, M., Domingues, V. S., Mallarino, R. & Hoekstra, H. E. The developmental role of Agouti in color pattern evolution. Science 331, 1062-5 (2011).
117. Liebermeister, W., Klipp, E., Schuster, S. & Heinrich, R. A theory of optimal differential gene expression. Biosystems 76, 261-78 (2004).
118. Elena, S. F. & Lenski, R. E. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat Rev Genet 4, 457-69 (2003).
119. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420, 186-9 (2002).
120. Cooke, J., Nowak, M. A., Boerlijst, M. & Maynard-Smith, J. Evolutionary origins and maintenance of redundant gene expression during metazoan development. Trends Genet 13, 360-4 (1997).
121. Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. Nature 436, 588-92 (2005).

122. Bershtein, S., Mu, W., Serohijos, A. W., Zhou, J. & Shakhnovich, E. I. Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness. Mol Cell 49, 133-44 (2013).
123. Borkovich, K. A., Farrelly, F. W., Finkelstein, D. B., Taulien, J. & Lindquist, S. hsp82 is an essential protein that is required in higher concentrations for growth of cells at higher temperatures. Mol Cell Biol 9, 3919-30 (1989).
124. Wagner, A. Birth and death of duplicated genes in completely sequenced eukaryotes. Trends Genet 17, 237-9 (2001).
125. Hartl, D. L. & Clark, A. G. Principles of Population Genetics (Sinauer Associates, Inc., 2006).
126. Mumberg, D., Muller, R. & Funk, M. Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. Gene 156, 119-22 (1995).
127. Zaret, K. S. & Sherman, F. Mutationally altered 3' ends of yeast CYC1 mRNA affect transcript stability and translational efficiency. J Mol Biol 177, 107-35 (1984).
128. Alberts, B. et al. Molecular Biology of the Cell (Garland Science, 2007).
129. Mercer, T. R. & Mattick, J. S. Structure and function of long noncoding RNAs in epigenetic regulation. Nat Struct Mol Biol 20, 300-7 (2013).
130. Bergman, Y. & Cedar, H. DNA methylation dynamics in health and disease. Nat Struct Mol Biol 20, 274-81 (2013).
131. de Visser, J. A., Cooper, T. F. & Elena, S. F. The causes of epistasis. Proc Biol Sci 278, 3617-24 (2011).
132. Guilmatre, A. & Sharp, A. J. Parent of origin effects. Clin Genet 81, 201-9 (2012).
133. Weinreich, D. M., Watson, R. A. & Chao, L. Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. Evolution 59, 1165-74 (2005).
134. Gould, S. J. Wonderful Life: The Burgess Shale and the Nature of History (W. W. Norton & Co., 1989).
135. Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. Nature 445, 383-6 (2007).
136. Zuckerkandl, E. & Pauling, L. Evolutionary Divergence and Convergence in Proteins. Evolving Genes and Proteins (1965).
137. Malcolm, B. A., Wilson, K. P., Matthews, B. W., Kirsch, J. F. & Wilson, A. C. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. Nature 345, 86-9 (1990).
138. Steiner, C. C., Weber, J. N. & Hoekstra, H. E. Adaptive variation in beach mice produced by two interacting pigmentation genes. PLoS Biol 5, e219 (2007).
139. Nagel, R. L. Epistasis and the genetics of human diseases. C R Biol 328, 606-15 (2005).
140. Maisnier-Patin, S. et al. Genomic buffering mitigates the effects of deleterious mutations in bacteria. Nat Genet 37, 1376-9 (2005).

141. Sanjuan, R., Moya, A. & Elena, S. F. The contribution of epistasis to the architecture of fitness in an RNA virus. Proc Natl Acad Sci U S A 101, 15376-9 (2004).
142. Salathe, P. & Ebert, D. The effects of parasitism and inbreeding on the competitive ability in Daphnia magna: evidence for synergistic epistasis. J Evol Biol 16, 976-85 (2003).
143. de Visser, J. A., Hoekstra, R. F. & van den Ende, H. An experimental test for synergistic epistasis and its application in Chlamydomonas. Genetics 145, 815-9 (1997).
144. Mukai, T. The Genetic Structure of Natural Populations of DROSOPHILA MELANOGASTER. VII Synergistic Interaction of Spontaneous Mutant Polygenes Controlling Viability. Genetics 61, 749-61 (1969).
145. Gerke, J., Lorenz, K. & Cohen, B. Genetic interactions between transcription factors cause natural variation in yeast. Science 323, 498-501 (2009).
146. Paulander, W., Maisnier-Patin, S. & Andersson, D. I. Multiple mechanisms to ameliorate the fitness burden of mupirocin resistance in Salmonella typhimurium. Mol Microbiol 64, 1038-48 (2007).
147. Kafri, R., Bar-Even, A. & Pilpel, Y. Transcription control reprogramming in genetic backup circuits. Nat Genet 37, 295-9 (2005).
148. Costanzo, M. et al. The genetic landscape of a cell. Science 327, 425-31 (2010).
149. Wu, N. C. et al. Systematic identification of H274Y compensatory mutations in influenza A virus neuraminidase by high-throughput screening. J Virol 87, 1193-9 (2013).
150. Abed, Y., Pizzorno, A., Bouhy, X. & Boivin, G. Role of permissive neuraminidase mutations in influenza A/Brisbane/59/2007-like (H1N1) viruses. PLoS Pathog 7, e1002431 (2011).
151. Bloom, J. D., Gong, L. I. & Baltimore, D. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. Science 328, 1272-5 (2010).
152. Lunzer, M., Golding, G. B. & Dean, A. M. Pervasive cryptic epistasis in molecular evolution. PLoS Genet 6, e1001162 (2010).
153. Carter, P. J., Winter, G., Wilkinson, A. J. & Fersht, A. R. The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus). Cell 38, 835-40 (1984).
154. Wall-Lacelle, S., Hossain, M. I., Sauve, R., Blunck, R. & Parent, L. Double mutant cycle analysis identified a critical leucine residue in the IIS4S5 linker for the activation of the Ca(V)2.3 calcium channel. J Biol Chem 286, 27197-205 (2011).
155. Istomin, A. Y., Gromiha, M. M., Vorov, O. K., Jacobs, D. J. & Livesay, D. R. New insight into long-range nonadditivity within protein double-mutant cycles. Proteins 70, 915-24 (2008).
156. Luisi, D. L. et al. Surface salt bridges, double-mutant cycles, and protein stability: an experimental and computational analysis of the interaction of the Asp 23 side chain with the N-terminus of the N-terminal domain of the ribosomal protein l9. Biochemistry 42, 7050-60 (2003).

157. Vaughan, C. K., Harryson, P., Buckle, A. M. & Fersht, A. R. A structural double-mutant cycle: estimating the strength of a buried salt bridge in barnase. Acta Crystallogr D Biol Crystallogr 58, 591-600 (2002).
158. Jang, D. S. et al. Structural double-mutant cycle analysis of a hydrogen bond network in ketosteroid isomerase from Pseudomonas putida biotype B. Biochem J 382, 967-73 (2004).
159. Schreiber, G. & Fersht, A. R. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. J Mol Biol 248, 478-86 (1995).
160. 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, 929-32 (2006).
161. Ortlund, E. A., Bridgham, J. T., Redinbo, M. R. & Thornton, J. W. Crystal structure of an ancient protein: evolution by conformational epistasis. Science 317, 1544-8 (2007).
162. Lehner, B. Molecular mechanisms of epistasis within and between genes. Trends Genet 27, 323-31 (2011).
163. 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).
164. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. Proc Natl Acad Sci U S A 103, 5869-74 (2006).
165. Sarkar, I., Hauber, I., Hauber, J. & Buchholz, F. HIV-1 proviral DNA excision using an evolved recombinase. Science 316, 1912-5 (2007).
166. Reetz, M. T. et al. Expanding the substrate scope of enzymes: combining mutations obtained by CASTing. Chemistry 12, 6031-8 (2006).
167. 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, 111-4 (2006).
168. Crevel, G., Bates, H., Huikeshoven, H. & Cotterill, S. The Drosophila Dpit47 protein is a nuclear Hsp90 co-chaperone that interacts with DNA polymerase alpha. J Cell Sci 114, 2015-25 (2001).
169. McDaniel, D. et al. Basal-level expression of the yeast HSP82 gene requires a heat shock regulatory element. Mol Cell Biol 9, 4789-98 (1989).
170. Sorger, P. K. & Nelson, H. C. Trimerization of a yeast transcriptional activator via a coiled-coil motif. Cell 59, 807-13 (1989).
171. Ali, M. M. et al. Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. Nature 440, 1013-7 (2006).
172. Panaretou, B. et al. ATP binding and hydrolysis are essential to the function of the Hsp90 molecular chaperone in vivo. Embo J 17, 4829-36 (1998).
173. Obermann, W. M., Sondermann, H., Russo, A. A., Pavletich, N. P. & Hartl, F. U. In vivo function of Hsp90 is dependent on ATP binding and ATP hydrolysis. J Cell Biol 143, 901-10 (1998).

174. Wandinger, S. K., Richter, K. & Buchner, J. The Hsp90 chaperone machinery. J Biol Chem 283, 18473-7 (2008).
175. da Silva, V. C. & Ramos, C. H. The network interaction of the human cytosolic 90 kDa heat shock protein Hsp90: A target for cancer therapeutics. J Proteomics 75, 2790-802 (2012).
176. Millson, S. H. et al. A two-hybrid screen of the yeast proteome for Hsp90 interactors uncovers a novel Hsp90 chaperone requirement in the activity of a stress-activated mitogen-activated protein kinase, Slt2p (Mpk1p). Eukaryot Cell 4, 849-60 (2005).
177. Zhao, R. et al. Navigating the chaperone network: an integrative map of physical and genetic interactions mediated by the hsp90 chaperone. Cell 120, 715-27 (2005).
178. Jakob, U., Lilie, H., Meyer, I. & Buchner, J. Transient interaction of Hsp90 with early unfolding intermediates of citrate synthase. Implications for heat shock in vivo. J Biol Chem 270, 7288-94 (1995).
179. Nathan, D. F. & Lindquist, S. Mutational analysis of Hsp90 function: interactions with a steroid receptor and a protein kinase. Mol Cell Biol 15, 3917-25 (1995).
180. Wayne, N. & Bolon, D. N. Charge-rich regions modulate the anti-aggregation activity of Hsp90. J Mol Biol 401, 931-9 (2010).
181. Youker, R. T., Walsh, P., Beilharz, T., Lithgow, T. & Brodsky, J. L. Distinct roles for the Hsp40 and Hsp90 molecular chaperones during cystic fibrosis transmembrane conductance regulator degradation in yeast. Mol Biol Cell 15, 4787-97 (2004).
182. Norby, J. G. Coupled assay of Na+,K+-ATPase activity. Methods Enzymol 156, 116-9 (1988).
183. Johnson, J. L. Evolution and function of diverse Hsp90 homologs and cochaperone proteins. Biochim Biophys Acta 1823, 607-13 (2012).
184. Retzlaff, M. et al. Asymmetric activation of the hsp90 dimer by its cochaperone aha1. Mol Cell 37, 344-54 (2010).
185. Fontana, J. et al. Domain mapping studies reveal that the M domain of hsp90 serves as a molecular scaffold to regulate Akt-dependent phosphorylation of endothelial nitric oxide synthase and NO release. Circ Res 90, 866-73 (2002).
186. Meyer, P. et al. Structural and functional analysis of the middle segment of hsp90: implications for ATP hydrolysis and client protein and cochaperone interactions. Mol Cell 11, 647-58 (2003).
187. Richter, K., Muschler, P., Hainzl, O. & Buchner, J. Coordinated ATP hydrolysis by the Hsp90 dimer. J Biol Chem 276, 33689-96 (2001).
188. Wayne, N. & Bolon, D. N. Dimerization of Hsp90 is required for in vivo function. Design and analysis of monomers and dimers. J Biol Chem 282, 35386-95 (2007).
189. Young, J. C., Obermann, W. M. & Hartl, F. U. Specific binding of tetratricopeptide repeat proteins to the C-terminal 12-kDa domain of hsp90. J Biol Chem 273, 18007-10 (1998).

190. McKenzie, S. L., Henikoff, S. & Meselson, M. Localization of RNA from heat-induced polysomes at puff sites in Drosophila melanogaster. Proc Natl Acad Sci U S A 72, 1117-21 (1975).
191. Rutherford, S. L. & Lindquist, S. Hsp90 as a capacitor for morphological evolution. Nature 396, 336-42 (1998).
192. Queitsch, C., Sangster, T. A. & Lindquist, S. Hsp90 as a capacitor of phenotypic variation. Nature 417, 618-24 (2002).
193. Casanueva, M. O., Burga, A. & Lehner, B. Fitness trade-offs and environmentally induced mutation buffering in isogenic C. elegans. Science 335, 82-5 (2011).
194. Yeyati, P. L., Bancewicz, R. M., Maule, J. & van Heyningen, V. Hsp90 selectively modulates phenotype in vertebrate development. PLoS Genet 3, e43 (2007).
195. Chen, G., Bradford, W. D., Seidel, C. W. & Li, R. Hsp90 stress potentiates rapid cellular adaptation through induction of aneuploidy. Nature 482, 246-50 (2012).
196. Pavelka, N. et al. Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. Nature 468, 321-5 (2010).
197. Selmecki, A. M., Dulmage, K., Cowen, L. E., Anderson, J. B. & Berman, J. Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance. PLoS Genet 5, e1000705 (2009).
198. Davies, A. E. & Kaplan, K. B. Hsp90-Sgt1 and Skp1 target human Mis12 complexes to ensure efficient formation of kinetochore-microtubule binding sites. J Cell Biol 189, 261-74 (2010).
199. Stemmann, O., Zou, H., Gerber, S. A., Gygi, S. P. & Kirschner, M. W. Dual inhibition of sister chromatid separation at metaphase. Cell 107, 715-26 (2001).
200. Whitesell, L. & Lin, N. U. HSP90 as a platform for the assembly of more effective cancer chemotherapy. Biochim Biophys Acta 1823, 756-66 (2012).
201. Trepel, J., Mollapour, M., Giaccone, G. & Neckers, L. Targeting the dynamic HSP90 complex in cancer. Nat Rev Cancer 10, 537-49 (2010).
202. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. Experimental illumination of a fitness landscape. Proc Natl Acad Sci U S A 108, 7896-901 (2011).
203. Pool, J. E., Hellmann, I., Jensen, J. D. & Nielsen, R. Population genetic inference from genomic sequence variation. Genome Res 20, 291-300 (2010).
204. Jensen, J. D., Wong, A. & Aquadro, C. F. Approaches for identifying targets of positive selection. Trends Genet 23, 568-77 (2007).
205. Zou, J., Guo, Y., Guettouche, T., Smith, D. F. & Voellmy, R. Repression of heat shock transcription factor HSF1 activation by HSP90 (HSP90 complex) that forms a stress-sensitive complex with HSF1. Cell 94, 471-80 (1998).
206. Giaever, G. et al. Functional profiling of the Saccharomyces cerevisiae genome. Nature 418, 387-91 (2002).
207. Lenski, R. E. Quantifying fitness and gene stability in microorganisms. Biotechnology 15, 173-92 (1991).
208. Cunningham, B. C. & Wells, J. A. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. Science 244, 1081-5 (1989).

209. Fowler, D. M. et al. High-resolution mapping of protein sequence-function relationships. Nat Methods 7, 741-6 (2010).
210. Pitt, J. N. & Ferre-D'Amare, A. R. Rapid construction of empirical RNA fitness landscapes. Science 330, 376-9 (2010).
211. Ernst, A. et al. Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. Mol Biosyst 6, 1782-90 (2010).
212. Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D. & Bolon, D. N. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. J Mol Biol 425, 1363-77 (2013).
213. Tsalik, E. L. & Gartenberg, M. R. Curing Saccharomyces cerevisiae of the 2 micron plasmid by targeted DNA damage. Yeast 14, 847-52 (1998).
214. Guthrie, C. & Fink, G. R. Guide to Yeast Genetics and Molecular and Cell Biology. Methods Enzymol 350 (2002).
215. Gietz, R. D., Schiestl, R. H., Willems, A. R. & Woods, R. A. Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/PEG procedure. Yeast 11, 355-60 (1995).
216. Gietz, R. D. & Schiestl, R. H. Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nat Protoc 2, 38-41 (2007).
217. Scanlon, T. C., Gray, E. C. & Griswold, K. E. Quantifying and resolving multiple vector transformants in S. cerevisiae plasmid libraries. BMC Biotechnol 9, 95 (2009).
218. Johnston, M. & Davis, R. W. Sequences that regulate the divergent GAL1-GAL10 promoter in Saccharomyces cerevisiae. Mol Cell Biol 4, 1440-8 (1984).
219. Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38, 1767-71 (2009).
220. Fowler, D. M., Araya, C. L., Gerard, W. & Fields, S. Enrich: software for analysis of protein function by enrichment and depletion of variants. Bioinformatics 27, 3430-1 (2011).
221. Pitt, J. N., Rajapakse, I. & Ferre-D'Amare, A. R. SEWAL: an open-source platform for next-generation sequence analysis and visualization. Nucleic Acids Res 38, 7908-15 (2010).
222. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 10, 866-76 (2009).
223. Cooper, T. F., Rozen, D. E. & Lenski, R. E. Parallel changes in gene expression after 20,000 generations of evolution in Escherichiacoli. Proc Natl Acad Sci U S A 100, 1072-7 (2003).
224. Nielsen, R. Molecular signatures of natural selection. Annu Rev Genet 39, 197-218 (2005).
225. Singh, N. D., Bauer DuMont, V. L., Hubisz, M. J., Nielsen, R. & Aquadro, C. F. Patterns of mutation and selection at synonymous sites in Drosophila. Mol Biol Evol 24, 2687-97 (2007).
226. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351, 652-4 (1991).

227. Lim, W. A. & Sauer, R. T. Alternative packing arrangements in the hydrophobic core of lambda repressor. Nature 339, 31-6 (1989).
228. Dahiyat, B. I. & Mayo, S. L. De novo protein design: fully automated sequence selection. Science 278, 82-7 (1997).
229. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. Deciphering the message in protein sequences: tolerance to amino acid substitutions. Science 247, 1306-10 (1990).
230. Marshall, S. A. & Mayo, S. L. Achieving stability and conformational specificity in designed proteins via binary patterning. J Mol Biol 305, 619-31 (2001).
231. Whitesell, L., Mimnaugh, E. G., De Costa, B., Myers, C. E. & Neckers, L. M. Inhibition of heat shock protein HSP90-pp60v-src heteroprotein complex formation by benzoquinone ansamycins: essential role for stress proteins in oncogenic transformation. Proc Natl Acad Sci U S A 91, 8324-8 (1994).
232. Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. Science 302, 1364-8 (2003).
233. Di Giulio, M. The extension reached by the minimization of the polarity distances during the evolution of the genetic code. J Mol Evol 29, 288-93 (1989).
234. Hernandez, R. D. et al. Classic selective sweeps were rare in recent human evolution. Science 331, 920-4 (2011).
235. Lewontin, R. C. & Hubby, J. L. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics 54, 595-609 (1966).
236. King, J. L. & Jukes, T. H. Non-Darwinian evolution. Science 164, 788-98 (1969).
237. Kreitman, M. Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature 304, 412-7 (1983).
238. Hamming, R. W. Error detecting and error correcting codes. The Bell System Technical Journal 2, 147-160 (1950).
239. Ghaemmaghami, S. et al. Global analysis of protein expression in yeast. Nature 425, 737-41 (2003).
240. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. Genome Res 14, 1188-90 (2004).
241. Eyre-Walker, A., Woolfit, M. & Phelps, T. The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173, 891-900 (2006).
242. Keightley, P. D. & Eyre-Walker, A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177, 2251-61 (2007).
243. Jensen, J. D., Thornton, K. R. & Andolfatto, P. An approximate bayesian estimator suggests strong, recurrent selective sweeps in Drosophila. PLoS Genet 4, e1000198 (2008).
244. Boyko, A. R. et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4, e1000083 (2008).
245. Timofeeff-Ressovsky, N. W. Mutations and geographical variation (ed. Huxley, J. S.) (Clarendon Press, Oxford, 1940).

246. Muller, H. J. Evidence of the precision of genetic adaptation. Harvey Lecture Series 43, 165-229 (1950).
247. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. Nat Rev Genet 8, 610-8 (2007).
248. Schenk, M. F., Szendro, I. G., Krug, J. & de Visser, J. A. Quantifying the adaptive potential of an antibiotic resistance enzyme. PLoS Genet 8, e1002783 (2012).
249. Kassen, R. & Bataillon, T. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat Genet 38, 484-8 (2006).
250. Trindade, S., Sousa, A. & Gordo, I. Antibiotic resistance and stress in the light of Fisher's model. Evolution 66, 3815-24 (2012).
251. Lalic, J., Cuevas, J. M. & Elena, S. F. Effect of host species on the distribution of mutational fitness effects for an RNA virus. PLoS Genet 7, e1002378 (2011).
252. Vale, P. F., Choisy, M., Froissart, R., Sanjuan, R. & Gandon, S. The distribution of mutational fitness effects of phage phiX174 on different hosts. Evolution 66, 3495-507 (2012).
253. Manna, F., Gallet, R., Martin, G. & Lenormand, T. The high-throughput yeast deletion fitness data and the theories of dominance. J Evol Biol 25, 892-903 (2012).
254. Gasch, A. P. et al. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11, 4241-57 (2000).
255. Causton, H. C. et al. Remodeling of yeast genome expression in response to environmental changes. Mol Biol Cell 12, 323-37 (2001).
256. Berry, D. B. & Gasch, A. P. Stress-activated genomic expression changes serve a preparative role for impending stress in yeast. Mol Biol Cell 19, 4580-7 (2008).
257. Yang, X. X. et al. The molecular chaperone Hsp90 is required for high osmotic stress response in Saccharomyces cerevisiae. FEMS Yeast Res 6, 195-204 (2006).
258. Hawle, P. et al. Cdc37p is required for stress-induced high-osmolarity glycerol and protein kinase C mitogen-activated protein kinase pathway functionality by interaction with Hog1p and Slt2p (Mpk1p). Eukaryot Cell 6, 521-32 (2007).
259. Yang, X. X. et al. Cdc37p is involved in osmoadaptation and controls high osmolarity-induced cross-talk via the MAP kinase Kss1p. FEMS Yeast Res 7, 796-807 (2007).
260. Hohmann, S. Osmotic stress signaling and osmoadaptation in yeasts. Microbiol Mol Biol Rev 66, 300-72 (2002).
261. Orr, H. A. A minimum on the mean number of steps taken in adaptive walks. J Theor Biol 220, 241-7 (2003).
262. Orr, H. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. Evolution 52, 935-949 (1998).
263. Martin, G., Elena, S. F. & Lenormand, T. Distributions of epistasis in microbes fit predictions from a fitness landscape model. Nat Genet 39, 555-60 (2007).
264. Gresham, D. et al. The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. PLoS Genet 4, e1000303 (2008).

265. Dhar, R., Sagesser, R., Weikert, C., Yuan, J. & Wagner, A. Adaptation of Saccharomyces cerevisiae to saline stress through laboratory evolution. J Evol Biol 24, 1135-53 (2011).
266. Martin, G. & Lenormand, T. A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. Evolution 60, 893-907 (2006).
267. Crow, J. F. Anecdotal, historical and critical commentaries on genetics twenty-five years ago in genetics: motoo kimura and molecular evolution. Genetics 116, 183-4 (1987).
268. Hermisson, J. & Pennings, P. S. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169, 2335-52 (2005).
269. Karasov, T., Messer, P. W. & Petrov, D. A. Evidence that adaptation in Drosophila is not limited by mutation at single sites. PLoS Genet 6, e1000924 (2010).
270. Hietpas, R., Roscoe, B., Jiang, L. & Bolon, D. N. Fitness analyses of all possible point mutations for regions of genes in yeast. Nat Protoc 7, 1382-96 (2012).
271. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8, 186-94 (1998).
272. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8, 175-85 (1998).
273. Martin, G. & Lenormand, T. The distribution of beneficial and fixed mutation fitness effects close to an optimum. Genetics 179, 907-16 (2008).
274. Bertram, J. S. The molecular biology of cancer. Mol Aspects Med 21, 167-223 (2000).
275. Bateson, W. Mendel's Principles of Heredity (Cambridge University Press, Cambridge, 1909).
276. Kermany, A. R. & Lessard, S. Effect of epistasis and linkage on fixation probability in three-locus models: an ancestral recombination-selection graph approach. Theor Popul Biol 82, 131-45 (2012).
277. de Oliveira, V. M., da Silva, J. K. & Campos, P. R. Epistasis and the selective advantage of sex and recombination. Phys Rev E Stat Nonlin Soft Matter Phys 78, 031905 (2008).
278. Campos, P. R. Fixation of beneficial mutations in the presence of epistatic interactions. Bull Math Biol 66, 473-86 (2004).
279. Martinez, J. P. et al. Fitness ranking of individual mutants drives patterns of epistatic interactions in HIV-1. PLoS One 6, e18375 (2011).
280. Hall, B. G. Experimental evolution of a new enzymatic function. II. Evolution of multiple functions for ebg enzyme in E. coli. Genetics 89, 453-65 (1978).
281. Poon, A., Davis, B. H. & Chao, L. The coupon collector and the suppressor mutation: estimating the number of compensatory mutations by maximum likelihood. Genetics 170, 1323-32 (2005).
282. Horovitz, A. Double-mutant cycles: a powerful tool for analyzing protein structure and function. Fold Des 1, R121-6 (1996).

283. Gout, J. F., Kahn, D. & Duret, L. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet 6, e1000944 (2010).
284. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A 102, 14338-43 (2005).
285. Jiang, L., Mishra, P., Hietpas, R. T., Zeldovich, K. B. & Bolon, D. N. Latent Effects of Hsp90 Mutants Revealed at Reduced Expression Levels. PLoS Genet, In Press (2013).
286. Schuldiner, M. et al. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. Cell 123, 507-19 (2005).
287. van Opijnen, T., Bodi, K. L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods 6, 767-72 (2009).
288. Pratt, W. B. & Toft, D. O. Regulation of signaling protein function and trafficking by the hsp90/hsp70-based chaperone machinery. Exp Biol Med (Maywood) 228, 111-33 (2003).
289. Nemoto, T., Ohara-Nemoto, Y., Ota, M., Takagi, T. & Yokoyama, K. Mechanism of dimer formation of the 90-kDa heat-shock protein. Eur J Biochem 233, 1-8 (1995).
290. Harris, S. F., Shiau, A. K. & Agard, D. A. The crystal structure of the carboxy-terminal dimerization domain of htpG, the Escherichia coli Hsp90, reveals a potential substrate binding site. Structure 12, 1087-97 (2004).
291. Cordes, M. H., Davidson, A. R. & Sauer, R. T. Sequence space, folding and protein design. Curr Opin Struct Biol 6, 3-10 (1996).
292. Dill, K. A. Dominant forces in protein folding. Biochemistry 29, 7133-55 (1990).
293. Gupta, R., Capalash, N. & Sharma, P. Restriction endonucleases: natural and directed evolution. Appl Microbiol Biotechnol 94, 583-99 (2012).
294. Das, P., Kapoor, D., Halloran, K. T., Zhou, R. & Matthews, C. R. Interplay between drying and stability of a TIM barrel protein: a combined simulation-experimental study. J Am Chem Soc 135, 1882-90 (2013).
295. Ohmae, E., Iriyama, K., Ichihara, S. & Gekko, K. Nonadditive effects of double mutations at the flexible loops, glycine-67 and glycine-121, of Escherichia coli dihydrofolate reductase on its stability and function. J Biochem 123, 33-41 (1998).
296. Toth-Petroczy, A. & Tawfik, D. S. Slow protein evolutionary rates are dictated by surface-core association. Proc Natl Acad Sci U S A 108, 11151-6 (2011).
297. Kim, P. M., Lu, L. J., Xia, Y. & Gerstein, M. B. Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314, 1938-41 (2006).
298. McLaughlin, R. N., Jr., Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. Nature 491, 138-42 (2012).

299. Picard, D. et al. Reduced levels of hsp90 compromise steroid receptor action in vivo. Nature 348, 166-8 (1990).
300. Orr, H. A. The population genetics of beneficial mutations. Philos Trans R Soc Lond B Biol Sci 365, 1195-201 (2010).
301. Losos, J. B. et al. Evolutionary biology for the 21st century. PLoS Biol 11, e1001466 (2013).

Appendix

Table A1

Selection Coefficients (Chapter III)

position	codon	aa	S
582	aaa	K	0.016527
582	aac	N	-0.02585
582	aag	K	0.015981
582	aat	N	-0.0237
582	aca	T	0.002578
582	acc	T	0.007921
582	acg	T	-0.03388
582	act	T	0.005527
582	aga	R	0.016136
582	agc	S	-0.03042
582	agg	R	0.017307
582	agt	S	-0.01906
582	ata	I	0.016769
582	atc	I	0.020264
582	atg	M	0.007341
582	att	I	0.026021
582	caa	Q	0
582	cac	H	0.006102
582	cag	Q	0.002206
582	cat	H	-0.00059
582	cca	P	-0.01214
582	ccc	P	-0.01938
582	ccg	P	-0.02238
582	cct	P	-0.00629
582	cga	R	0.011154
582	cgc	R	0.007826
582	cgg	R	0.009909

582	cgt	R	0.010772
582	cta	L	0.016014
582	ctc	L	0.009362
582	ctg	L	0.007467
582	ctt	L	0.032177
582	gaa	E	-0.01724
582	gac	D	-0.25208
582	gag	E	-0.02075
582	gat	D	-0.23387
582	gca	A	-0.00851
582	gcc	A	-0.00569
582	gcg	A	-0.00384
582	gct	A	0.005481
582	gga	G	-0.25513
582	ggc	G	-0.58161
582	ggg	G	-0.53116
582	ggt	G	-0.57639
582	gta	V	0.004786
582	gtc	V	0.004832
582	gtg	V	-0.00666
582	gtt	V	0.023479
582	taa	*	-0.55396
582	tac	Y	-0.01498
582	tag	*	-0.64089
582	tat	Y	-0.01341
582	tca	S	-0.02617
582	tcc	S	-0.02193
582	tcg	S	-0.02363
582	tct	S	-0.01079
582	tga	*	-0.67755
582	tgc	C	-0.01605
582	tgg	W	-0.07939
582	tgt	C	-0.01577
582	tta	L	0.019493
582	ttc	F	-0.0041
582	ttg	L	0.011481
582	ttt	F	-0.00016
583	aaa	K	-0.81673
583	aac	N	-0.09716
583	aag	K	-0.77478

583	aat	N	-0.09208
583	aca	T	-0.51948
583	acg	T	-0.51231
583	act	T	-0.24972
583	aga	R	-0.7196
583	agc	S	-0.24609
583	agg	R	-0.7362
583	agt	S	-0.53996
583	ata	I	-0.69055
583	atc	I	-0.64685
583	atg	M	-0.22337
583	att	I	-0.67122
583	caa	Q	-0.59968
583	cac	H	-0.04363
583	cag	Q	-0.58396
583	cat	H	-0.04097
583	cca	P	-0.87157
583	cct	P	-0.95312
583	cga	R	-0.73081
583	cgc	R	-0.66025
583	cgg	R	-0.77167
583	cgt	R	-0.68778
583	cta	L	-0.09377
583	ctc	L	-0.10196
583	ctg	L	-0.10086
583	ctt	L	-0.10108
583	gaa	E	-0.72999
583	gac	D	-0.25137
583	gag	E	-0.67106
583	gat	D	-0.52989
583	gca	A	-0.21489
583	gcg	A	-0.22781
583	gct	A	-0.23131
583	gga	G	-0.14935
583	ggc	G	-0.15042
583	ggg	G	-0.14884
583	ggt	G	-0.13864
583	gta	V	-0.66609
583	gtc	V	-0.71339
583	gtg	V	-0.66352

583	gtt	V	-0.74809
583	taa	*	-0.81577
583	tac	Y	-0.00044
583	tag	*	-0.73482
583	tat	Y	-0.00383
583	tca	S	-0.52919
583	tcg	S	-0.26846
583	tct	S	-0.21761
583	tga	*	-0.89047
583	tgc	C	-0.17124
583	tgg	W	-0.0195
583	tgt	C	-0.16453
583	tta	L	-0.09611
583	ttc	F	0.000216
583	ttg	L	-0.09333
583	ttt	F	0
584	aaa	K	-0.69371
584	aac	N	-0.74702
584	aag	K	-0.7005
584	aat	N	-0.64762
584	aca	T	-0.63215
584	acc	T	-0.52621
584	acg	T	-0.673
584	act	T	-0.65463
584	aga	R	-0.68609
584	agc	S	-0.10233
584	agg	R	-0.71931
584	agt	S	-0.10345
584	ata	I	-0.96976
584	atc	I	-0.76513
584	atg	M	-0.13989
584	att	I	-0.79547
584	caa	Q	-0.28341
584	cac	H	-0.60216
584	cag	Q	-0.2286
584	cat	H	-0.54864
584	cca	P	-0.84962
584	ccc	P	-0.81987
584	ccg	P	-0.86799
584	cct	P	-0.8213

584	cga	R	-0.72292
584	cgc	R	-0.71147
584	cgg	R	-0.71034
584	cgt	R	-0.78193
584	cta	L	-0.82848
584	ctc	L	-0.83588
584	ctg	L	-0.90347
584	ctt	L	-0.98797
584	gaa	E	-0.80147
584	gac	D	-0.73838
584	gag	E	-0.67962
584	gat	D	-0.56308
584	gca	A	-0.13226
584	gcc	A	-0.13923
584	gcg	A	-0.14539
584	gct	A	-0.13269
584	gga	G	0.009799
584	ggc	G	0.002484
584	ggg	G	0.008195
584	ggt	G	0
584	gta	V	-0.8082
584	gtc	V	-0.79434
584	gtg	V	-0.90572
584	gtt	V	-0.86284
584	taa	*	-0.76372
584	tac	Y	-0.08638
584	tag	*	-0.8213
584	tat	Y	-0.07867
584	tca	S	-0.10402
584	tcc	S	-0.11136
584	tcg	S	-0.12124
584	tct	S	-0.10176
584	tga	*	-0.6444
584	tgc	C	-0.1382
584	tgg	W	-0.26103
584	tgt	C	-0.11224
584	tta	L	-0.90505
584	ttc	F	-0.04159
584	ttg	L	-0.9939
584	ttt	F	-0.02812

585	aaa	K	-0.89047
585	aac	N	-0.82278
585	aag	K	-0.97773
585	aat	N	-0.83997
585	aca	T	-0.83649
585	acc	T	-0.92897
585	acg	T	-0.77505
585	act	T	-0.85564
585	aga	R	-0.83249
585	agc	S	-0.99046
585	agg	R	-0.72024
585	agt	S	-0.95356
585	ata	I	-0.08586
585	atc	I	-0.08818
585	atg	M	-0.08
585	att	I	-0.07816
585	caa	Q	-0.8668
585	cac	H	-0.86811
585	cag	Q	-0.77924
585	cat	H	-0.84513
585	cca	P	-0.22681
585	ccc	P	-0.22136
585	ccg	P	-0.2291
585	cct	P	-0.24077
585	cga	R	-0.8028
585	cgc	R	-0.77235
585	cgg	R	-0.7469
585	cgt	R	-0.99128
585	cta	L	0.013211
585	ctc	L	0.013232
585	ctg	L	0.001746
585	ctt	L	0.017239
585	gaa	E	-1.13188
585	gac	D	-1.04756
585	gag	E	-0.94413
585	gat	D	-1.35184
585	gca	A	-0.85257
585	gcc	A	-0.8287
585	gcg	A	-0.83873
585	gct	A	-0.95586

585	gga	G	-0.9997
585	ggc	G	-0.73895
585	ggg	G	-0.86272
585	ggt	G	-1.09846
585	gta	V	-0.25486
585	gtc	V	-0.26284
585	gtg	V	-0.24048
585	gtt	V	-0.2336
585	taa	*	-0.77193
585	tac	Y	-0.23984
585	tag	*	-0.78755
585	tat	Y	-0.23315
585	tca	S	-0.88104
585	tcc	S	-1.3347
585	tcg	S	-0.95422
585	tct	S	-0.76723
585	tga	*	-0.72938
585	tgc	C	-0.82415
585	tgg	W	0
585	tgt	C	-0.73701
585	tta	L	0.01514
585	ttc	F	-0.03212
585	ttg	L	0.00426
585	ttt	F	-0.015
586	aaa	K	-0.7898
586	aac	N	-0.09011
586	aag	K	-0.83259
586	aat	N	-0.09134
586	aca	T	0.000581
586	acc	T	0.000281
586	acg	T	-0.00116
586	act	T	0.003344
586	aga	R	-0.79745
586	agc	S	0.014337
586	agg	R	-0.81726
586	agt	S	0.011962
586	ata	I	-0.69895
586	atc	I	-0.86955
586	atg	M	-0.79818
586	att	I	-0.83024

586	caa	Q	-0.87863
586	cac	H	-0.79763
586	cag	Q	-0.7834
586	cat	H	-0.81918
586	cca	P	-0.77746
586	ccc	P	-0.86479
586	cct	P	-0.68799
586	cga	R	-0.78222
586	cgc	R	-0.87814
586	cgg	R	-1.15212
586	cgt	R	-0.6917
586	cta	L	-0.81453
586	ctc	L	-0.81315
586	ctg	L	-0.6917
586	ctt	L	-0.72148
586	gaa	E	-0.82415
586	gac	D	-0.80038
586	gag	E	-0.71412
586	gat	D	-0.91339
586	gca	A	-0.82626
586	gcc	A	-0.87496
586	gcg	A	-0.96485
586	gct	A	-0.91611
586	gga	G	-0.24056
586	ggc	G	-0.22504
586	ggg	G	-0.22493
586	ggt	G	-0.22832
586	gta	V	-0.78475
586	gtc	V	-0.85837
586	gtg	V	-0.62253
586	gtt	V	-0.83385
586	taa	*	-0.82626
586	tac	Y	-0.80471
586	tag	*	-0.78919
586	tat	Y	-0.78012
586	tca	S	0.012029
586	tcc	S	0.005105
586	tcg	S	0.003483
586	tct	S	0
586	tga	*	-0.69371

586	tgc	C	-0.96902
586	tgg	W	-0.86099
586	tgt	C	-0.81918
586	tta	L	-0.72381
586	ttc	F	-0.80471
586	ttg	L	-0.90328
586	ttt	F	-0.80038
587	aaa	K	-0.83952
587	aac	N	-0.83172
587	aag	K	-0.79518
587	aat	N	-0.83935
587	aca	T	-0.79298
587	acc	T	-0.92466
587	acg	T	-0.75822
587	act	T	-0.87962
587	aga	R	-0.78658
587	agc	S	-0.04622
587	agg	R	-0.81541
587	agt	S	-0.04786
587	ata	I	-0.95662
587	atc	I	-0.85326
587	atg	M	-0.98696
587	att	I	-0.93768
587	caa	Q	-0.74588
587	cac	H	-0.83084
587	cag	Q	-0.78076
587	cat	H	-0.813
587	cca	P	-0.18369
587	ccc	P	-0.17111
587	ccg	P	-0.17494
587	cct	P	-0.16462
587	cga	R	-0.8182
587	cgc	R	-0.84603
587	cgg	R	-0.83689
587	cgt	R	-0.71784
587	cta	L	-0.83616
587	ctc	L	-0.92175
587	ctg	L	-0.86472
587	ctt	L	-0.8514
587	gaa	E	-0.9686

587	gac	D	-0.91339
587	gag	E	-0.91656
587	gat	D	-0.73947
587	gca	A	0.006095
587	gcc	A	0.007192
587	gcg	A	0.006495
587	gct	A	0
587	gga	G	-0.17382
587	ggc	G	-0.17755
587	ggg	G	-0.17993
587	ggt	G	-0.16436
587	gta	V	-0.7358
587	gtc	V	-0.75085
587	gtg	V	-0.82454
587	gtt	V	-0.74979
587	taa	*	-0.76087
587	tac	Y	-0.67388
587	tag	*	-0.73242
587	tat	Y	-0.86644
587	tca	S	-0.04501
587	tcc	S	-0.05302
587	tcg	S	-0.05025
587	tct	S	-0.04941
587	tga	*	-0.83112
587	tgc	C	-0.55655
587	tgg	W	-0.77813
587	tgt	C	-0.60626
587	tta	L	-0.85571
587	ttc	F	-0.80578
587	ttg	L	-0.88832
587	ttt	F	-0.91278
588	aaa	K	-0.75883
588	aac	N	-0.06339
588	aag	K	-0.83822
588	aat	N	0
588	aca	T	-0.02776
588	acc	T	-0.0345
588	acg	T	-0.03348
588	act	T	0.004146
588	aga	R	-0.0513

588	agc	S	-0.05777
588	agg	R	-0.04979
588	agt	S	-0.05516
588	ata	I	-0.68141
588	atc	I	-0.70699
588	atg	M	-0.08482
588	att	I	-0.71914
588	caa	Q	-0.10879
588	cac	H	-0.03807
588	cag	Q	-0.10843
588	cat	H	-0.03537
588	cca	P	-0.61983
588	ccc	P	-0.66869
588	ccg	P	-0.66519
588	cct	P	-0.66424
588	cga	R	-0.05157
588	cgc	R	-0.05189
588	cgg	R	-0.04849
588	cgt	R	-0.05181
588	cta	L	-0.52268
588	ctc	L	-0.24561
588	ctg	L	-0.53137
588	ctt	L	-0.57403
588	gaa	E	-0.7724
588	gac	D	-0.85634
588	gag	E	-0.78413
588	gat	D	-0.8234
588	gca	A	-0.17524
588	gcc	A	-0.16631
588	gcg	A	-0.1536
588	gct	A	-0.16241
588	gga	G	-0.22752
588	ggc	G	-0.22963
588	ggg	G	-0.23603
588	ggt	G	-0.24083
588	gta	V	-0.6404
588	gtc	V	-0.61361
588	gtg	V	-0.62575
588	gtt	V	-0.57294
588	taa	*	-0.77014

588	tac	Y	-0.02022
588	tag	*	-0.7834
588	tat	Y	-0.02091
588	tca	S	-0.04641
588	tcc	S	-0.05029
588	tcg	S	-0.04917
588	tct	S	-0.04982
588	tga	*	-0.78919
588	tgc	C	-0.2257
588	tgg	W	-0.10194
588	tgt	C	-0.22056
588	tta	L	-0.56057
588	ttc	F	-0.05277
588	ttg	L	-0.27224
588	ttt	F	-0.02435
589	aaa	K	-0.7401
589	aac	N	-0.79699
589	aag	K	-0.62704
589	aat	N	-0.84147
589	aca	T	-0.09829
589	acc	T	-0.09767
589	acg	T	-0.09553
589	act	T	-0.09798
589	aga	R	-0.70711
589	agc	S	-0.72577
589	agg	R	-0.67937
589	agt	S	-0.71537
589	ata	I	-0.06457
589	atc	I	-0.06356
589	atg	M	0
589	att	I	-0.05861
589	caa	Q	0.039382
589	cac	H	-0.58101
589	cag	Q	0.042361
589	cat	H	-0.58616
589	cca	P	-0.77746
589	ccc	P	-0.85779
589	cct	P	-0.81347
589	cga	R	-0.71654
589	cgc	R	-0.66757

589	cgg	R	-0.67932
589	cgt	R	-0.66122
589	cta	L	-0.0463
589	ctc	L	-0.05522
589	ctg	L	-0.0509
589	ctt	L	-0.04348
589	gaa	E	-0.2064
589	gac	D	-0.76983
589	gag	E	-0.20381
589	gat	D	-0.76087
589	gca	A	-0.16378
589	gcc	A	-0.15879
589	gcg	A	-0.15023
589	gct	A	-0.15761
589	gga	G	-0.79379
589	ggc	G	-0.83659
589	ggg	G	-0.85038
589	ggt	G	-0.90029
589	gta	V	-0.08179
589	gtc	V	-0.0885
589	gtg	V	-0.0856
589	gtt	V	-0.07914
589	taa	*	-0.78144
589	tac	Y	-0.10563
589	tag	*	-0.80038
589	tat	Y	-0.09661
589	tca	S	-0.71132
589	tcc	S	-0.80567
589	tcg	S	-0.77342
589	tct	S	-0.73653
589	tga	*	-0.89346
589	tgc	C	-0.28105
589	tgg	W	-0.04708
589	tgt	C	-0.26688
589	tta	L	-0.03549
589	ttc	F	-0.00598
589	ttg	L	-0.0494
589	ttt	F	-0.00481
590	aaa	K	-0.66504
590	aac	N	-0.25564

590	aag	K	-0.72148
590	aat	N	-0.2598
590	aca	T	-0.00857
590	acc	T	-0.01629
590	acg	T	-0.01916
590	act	T	-0.01987
590	aga	R	-0.22836
590	agc	S	-0.06668
590	agg	R	-0.21841
590	agt	S	-0.07495
590	ata	I	-0.00098
590	atc	I	0.000312
590	atg	M	-0.00083
590	att	I	-0.00555
590	caa	Q	0.014282
590	cac	H	-0.04995
590	cag	Q	0.013613
590	cat	H	-0.04494
590	cca	P	-0.66445
590	ccc	P	-0.68154
590	ccg	P	-0.56887
590	cct	P	-0.66532
590	cga	R	-0.24954
590	cgc	R	-0.23201
590	cgg	R	-0.2221
590	cgt	R	-0.23542
590	cta	L	-0.01024
590	ctc	L	-0.02276
590	ctg	L	-0.00215
590	ctt	L	-0.00841
590	gaa	E	0
590	gac	D	-0.20557
590	gag	E	0.006411
590	gat	D	-0.04747
590	gca	A	-0.03945
590	gcc	A	-0.03183
590	gcg	A	-0.0429
590	gct	A	-0.03589
590	gga	G	-0.69263
590	ggc	G	-0.6902

590	ggg	G	-0.67493
590	ggt	G	-0.67509
590	gta	V	-0.03294
590	gtc	V	-0.0409
590	gtg	V	-0.03146
590	gtt	V	-0.05756
590	taa	*	-0.58736
590	tac	Y	-0.05403
590	tag	*	-0.6917
590	tat	Y	-0.05514
590	tca	S	-0.05612
590	tcc	S	-0.06711
590	tcg	S	-0.06673
590	tct	S	-0.05594
590	tga	*	-0.83682
590	tgc	C	-0.00906
590	tgg	W	-0.02363
590	tgt	C	-0.00276
590	tta	L	-0.00395
590	ttc	F	-0.04645
590	ttg	L	-0.00542
590	ttt	F	-0.0638

Table A2

Selection coefficients and FGM category of mutations in all conditions

position	aa	s_30C	s_36C	s_30C+S	s_36C+S	FGM category
582	*	-0.5	-0.5	-0.5	-0.5	D
582	A	-0.01224	0.002158	-0.01189	-0.01171	1
582	C	-0.01346	-0.06306	-0.0067	-0.01485	D
582	D	-0.03163	-0.14915	-0.00351	-0.04671	D
582	E	-0.00773	-0.02429	0.004305	0.015753	3
582	F	-0.0343	-0.09032	-0.02799	-0.06276	D
582	G	-0.03472	-0.17725	-0.00496	-0.06023	D
582	H	-0.01131	-0.05402	0.012183	-0.00591	4
582	I	-0.01175	-0.01703	-0.02156	-0.02468	D
582	K	0.005045	-0.00722	-0.02138	-0.00602	I
582	L	-0.01245	-0.02148	-0.02174	-0.02323	D
582	M	0.006005	-0.01395	0.004018	-0.0097	I
582	N	-0.00998	-0.06597	0.015005	-0.00282	4
582	P	-0.01327	-0.03998	-0.00048	-0.00888	D
582	Q	-0.00739	-0.00472	0.005578	0.019578	3
582	R	-0.00366	-0.0072	-0.01586	0.000861	I
582	S	-0.01939	-0.0459	-0.0098	-0.03863	D
582	T	-0.01196	-0.02037	-0.01044	-0.02317	D
582	V	-0.02009	-0.02226	-0.02361	-0.02881	D
582	W	-0.0155	-0.12535	0.001925	-0.01439	4
582	Y	-0.02348	-0.08402	0.007272	-0.03828	4
583	*	-0.5	-0.5	-0.5	-0.5	D
583	A	-0.03445	-0.14785	0.00512	-0.00776	4
583	C	-0.00737	-0.12409	-0.00723	0.011352	I
583	D	-0.1045	-0.19192	-0.01802	-0.06916	D
583	E	-0.07922	-0.271	-0.05278	-0.09632	D
583	F	-0.00479	-0.00876	0.008448	0.006692	3
583	G	-0.02339	-0.10958	0.005854	0.01809	3
583	H	0.006401	-0.0447	0.001032	0.032445	2
583	I	-0.04781	-0.24383	-0.02832	-0.07326	D
583	K	-0.09798	-0.36777	-0.02842	-0.05797	D
583	L	-0.01362	-0.12029	-0.00758	-0.00813	D
583	M	-0.03347	-0.17263	-0.00184	-0.01787	D
583	N	-0.02047	-0.09933	-0.01253	0.022973	I

583	P	-0.16459	-0.5	-0.11419	-0.23745	D
583	Q	-0.04549	-0.21656	-0.01285	-0.0317	D
583	R	-0.04532	-0.27048	-0.02038	-0.03953	D
583	S	-0.04099	-0.17936	-0.01714	-0.02221	D
583	T	-0.03943	-0.18604	-0.00468	-0.00828	D
583	V	-0.04901	-0.25081	-0.03725	-0.07944	D
583	W	-0.0056	-0.02273	0.006744	-0.01678	4
583	Y	-0.00396	0.001385	0.008319	0.007033	I
584	*	-0.5	-0.5	-0.5	-0.5	D
584	A	-0.01569	-0.08038	-0.00351	-0.00173	D
584	C	-0.02555	-0.10933	-0.01966	-0.02548	D
584	D	-0.04778	-0.22323	-0.02931	-0.04788	D
584	E	-0.08672	-0.28291	-0.05731	-0.14805	D
584	F	-0.02057	-0.11275	0.022149	-0.03657	4
584	G	-0.0008	-0.00171	0.005769	0.003098	3
584	H	-0.04469	-0.19955	-0.0168	-0.05146	D
584	I	-0.40934	-0.5	-0.44691	-0.5	D
584	K	-0.12717	-0.46682	-0.09926	-0.22589	D
584	L	-0.15349	-0.5	-0.13585	-0.29897	D
584	M	-0.01989	-0.11223	0.011797	-0.01805	4
584	N	-0.06476	-0.27284	-0.04498	-0.06866	D
584	P	-0.5	-0.5	-0.5	-0.5	D
584	Q	-0.05498	-0.18927	-0.01796	-0.04944	D
584	R	-0.11436	-0.43446	-0.06013	-0.18604	D
584	S	-0.03338	-0.10196	-0.00914	-0.01447	D
584	T	-0.05158	-0.23785	-0.02319	-0.07404	D
584	V	-0.13056	-0.4962	-0.10887	-0.23322	D
584	W	-0.04189	-0.22588	-0.04067	-0.06969	D
584	Y	-0.01797	-0.13258	-0.00999	-0.04737	D
585	*	-0.5	-0.5	-0.5	-0.5	D
585	A	-0.12386	-0.39814	-0.02101	-0.11712	D
585	C	-0.07678	-0.25592	-0.00862	-0.04977	D
585	D	-0.5	-0.5	-0.5	-0.5	D
585	E	-0.5	-0.5	-0.43898	-0.5	D
585	F	-0.02822	-0.10478	0.025971	0.001511	3
585	G	-0.46856	-0.5	-0.48139	-0.5	D
585	H	-0.06628	-0.33793	-0.02354	-0.10342	D
585	I	-0.11381	-0.13018	0.031944	0.02463	3
585	K	-0.46016	-0.5	-0.29476	-0.5	D
585	L	-0.04135	-0.05761	0.009901	0.016114	3

585	M	-0.01486	-0.07793	-0.01275	-0.00509	D
585	N	-0.11334	-0.48898	-0.08054	-0.18963	D
585	P	-0.08839	-0.18131	0.000677	-0.01422	4
585	Q	-0.12019	-0.49628	-0.02969	-0.18102	D
585	R	-0.22237	-0.5	-0.05943	-0.23534	D
585	S	-0.24551	-0.5	-0.15025	-0.33315	D
585	T	-0.16838	-0.49767	-0.03675	-0.16009	D
585	V	-0.08506	-0.19802	0.016245	0.002716	3
585	W	-0.00739	-0.01156	0.00778	0.021743	3
585	Y	-0.03525	-0.19191	-0.01083	-0.0463	D
586	*	-0.5	-0.5	-0.5	-0.5	D
586	A	-0.13546	-0.5	-0.11889	-0.29349	D
586	C	-0.103	-0.5	-0.08055	-0.23223	D
586	D	-0.5	-0.5	-0.5	-0.5	D
586	E	-0.5	-0.5	-0.5	-0.5	D
586	F	-0.5	-0.5	-0.5	-0.5	D
586	G	-0.02149	-0.17712	0.007943	-0.04805	4
586	H	-0.5	-0.5	-0.5	-0.5	D
586	I	-0.5	-0.5	-0.5	-0.5	D
586	K	-0.43018	-0.5	-0.5	-0.5	D
586	L	-0.5	-0.5	-0.5	-0.5	D
586	M	-0.5	-0.5	-0.5	-0.5	D
586	N	-0.01215	-0.12683	0.03306	-0.01264	4
586	P	-0.40758	-0.5	-0.39387	-0.5	D
586	Q	-0.5	-0.5	-0.5	-0.5	D
586	R	-0.5	-0.5	-0.5	-0.5	D
586	S	0.000803	-0.00275	0.004988	0.005485	2
586	T	0.00387	-0.05442	0.022022	0.001847	2
586	V	-0.5	-0.5	-0.5	-0.5	D
586	W	-0.5	-0.5	-0.5	-0.5	D
586	Y	-0.5	-0.5	-0.5	-0.5	D
587	*	-0.5	-0.5	-0.5	-0.5	D
587	A	-0.00225	-0.00104	-0.00124	0.005667	I
587	C	-0.02741	-0.1974	-0.00763	-0.05592	D
587	D	-0.5	-0.5	-0.5	-0.5	D
587	E	-0.5	-0.5	-0.5	-0.5	D
587	F	-0.204	-0.5	-0.23256	-0.39318	D
587	G	-0.01432	-0.11698	0.012992	-0.01257	4
587	H	-0.05952	-0.32964	-0.02841	-0.07717	D
587	I	-0.12082	-0.5	-0.09411	-0.19517	D

587	K	-0.23686	-0.5	-0.10807	-0.31209	D
587	L	-0.31213	-0.5	-0.04055	-0.29641	D
587	M	-0.14799	-0.5	-0.07629	-0.20424	D
587	N	-0.08463	-0.5	-0.03882	-0.08112	D
587	P	-0.01876	-0.13006	0.022302	-0.01042	4
587	Q	-0.05812	-0.39084	-0.04508	-0.07634	D
587	R	-0.06255	-0.40272	-0.02621	-0.06901	D
587	S	-0.00999	-0.07911	0.008239	-0.01484	4
587	T	-0.09014	-0.5	-0.07723	-0.14694	D
587	V	-0.04077	-0.33561	-0.02589	-0.06097	D
587	W	-0.18531	-0.5	-0.05501	-0.22646	D
587	Y	-0.14569	-0.5	-0.11385	-0.26346	D
588	*	-0.5	-0.5	-0.5	-0.5	D
588	A	-0.00701	-0.11209	0.02637	0.002435	3
588	C	-0.00179	-0.12155	0.007346	-0.00539	4
588	D	-0.22553	-0.5	-0.20239	-0.4372	D
588	E	-0.16553	-0.5	-0.0954	-0.38992	D
588	F	-0.00272	-0.01264	-0.00892	-0.00211	D
588	G	-0.01665	-0.15547	0.01556	-0.01142	4
588	H	-0.00332	-0.06087	0.010716	-0.01964	4
588	I	-0.0623	-0.34296	0.029177	-0.04672	4
588	K	-0.07579	-0.47384	0.004617	-0.05863	4
588	L	-0.03816	-0.18017	0.05063	0.021839	3
588	M	0.006973	-0.0218	0.022694	0.040277	2
588	N	-0.02279	-0.04396	0.000969	0.011879	3
588	P	-0.14498	-0.35976	0.082948	-0.03179	4
588	Q	-0.00681	-0.0832	0.023413	-0.00329	4
588	R	-0.00741	-0.06371	0.002788	-0.02277	4
588	S	-0.00624	-0.07277	0.009548	-0.01479	4
588	T	-0.00359	-0.03982	0.008354	0.011724	3
588	V	-0.02076	-0.20769	0.012371	-0.02192	4
588	W	-0.00197	-0.07215	-0.00423	-0.06317	D
588	Y	0.0063	-0.01585	0.005165	0.000399	2
589	*	-0.5	-0.5	-0.5	-0.5	D
589	A	-0.02586	-0.1827	0.0298	-0.00955	4
589	C	-0.02379	-0.19999	0.016351	-0.04681	4
589	D	-0.15511	-0.5	0.037236	-0.365	4
589	E	-0.02115	-0.16865	0.029732	0.005088	3
589	F	-0.00803	-0.04994	0.006575	-0.00612	4
589	G	-0.14055	-0.5	-0.15195	-0.3078	D

589	H	-0.04067	-0.20662	-0.00102	-0.03243	D
589	I	-0.0016	-0.08708	-0.00133	-0.01906	D
589	K	-0.06801	-0.29016	-0.0087	-0.04032	D
589	L	-0.01099	-0.08445	0.0087	-0.04459	4
589	M	-0.00739	-0.01156	0.00778	0.021743	3
589	N	-0.08671	-0.5	-0.01677	-0.15049	D
589	P	-0.5	-0.5	-0.5	-0.5	D
589	Q	-0.02637	-0.13296	0.043434	-0.01376	4
589	R	-0.05018	-0.27866	-0.01553	-0.05628	D
589	S	-0.04063	-0.32254	0.006106	-0.05327	4
589	T	-0.01098	-0.136	0.029013	-0.00935	4
589	V	-0.00227	-0.12403	0.029669	-0.01013	4
589	W	-0.0099	-0.05895	0.034708	0.044117	3
589	Y	-0.0124	-0.09473	0.012688	-0.00265	4
590	*	-0.5	-0.5	-0.5	-0.5	D
590	A	0.004616	-0.08104	0.01581	0.009161	2
590	C	0.004772	-0.02259	0.005824	0.015702	2
590	D	-0.00243	-0.18476	0.00742	-0.02805	4
590	E	-0.00629	-0.00745	0.001658	0.003966	3
590	F	-0.00559	-0.07867	0.019151	0.007374	3
590	G	-0.04763	-0.32026	-0.0309	-0.09234	D
590	H	-0.00095	-0.06659	0.025592	0.011292	3
590	I	0.001602	-0.01165	-0.01081	0.003191	I
590	K	-0.06483	-0.3473	-0.04435	-0.09504	D
590	L	0.002429	-0.02131	0.017009	0.030591	2
590	M	0.003429	-0.01216	0.019073	0.022234	2
590	N	-0.01975	-0.20025	-0.01042	-0.04115	D
590	P	-0.5	-0.5	-0.5	-0.5	D
590	Q	0.005478	-0.00254	0.01198	0.021362	2
590	R	-0.02277	-0.184	0.004154	-0.02012	4
590	S	-0.0103	-0.10759	0.004801	-0.01189	4
590	T	0.005986	-0.03725	0.008033	0.02313	2
590	V	0.005363	-0.03341	0.02496	0.022353	2
590	W	-0.00011	-0.03523	0.022599	0.036124	3
590	Y	0.003962	-0.06013	0.020846	0.020657	2