Computational analysis of genome-scale growth-interaction data in *Saccharomyces cerevisiae*

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Benjamin James VanderSluis

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Chad L. Myers

August, 2014

Acknowledgements

In my years at the University of Minnesota, I have met many special people who have had a profound impact on me, both as a scientist and as an individual. I owe any success I might claim to their guidance, friendship, and support. First of all, I must thank Bob Hain, and the and the staff of Enet: Ray Muno, Hokan, and Paul Markfort. They hired me a sophomore undergraduate and gave me (in addition to a paycheck) a real education in computing. I am indebted to them for patience and their willingness to satisfy my curiosity. While I learned many technical lessons in my time there, perhaps the more lasting impact was the respect for learning and research that flowed freely in that environment. Their attitude and support opened the door for me to go to graduate school and pursue a career as a scientist.

I would also like to thank several collaborators from whom I have learned a great deal. The interdisciplinary nature of my work has introduced me to many biologists, who have taught me a great deal. Particularly, my friends from Toronto, Michael Costanzo, Charlie Boone, Anastasia Baryshnikova, Elena Kuzmin, and Matej Usaj. Working with Amy Caudy and David Hess was also a great pleasure and a valuable experience.

It has been a joy to work in such a stimulating and friendly environment over the past several years, and for that, I have to thank my current and former lab-mates, Jeremy Bellay, Yungil Kim, Raamesh Deshpande, Elizabeth Koch, Roman Briskine, Carles Pons, Colin Pesyna, Robert Schaefer, Stephanie DiPrima, Timothy Kunau, Justin Nelson, Scott Simpkins, Wen Wang, and Praveen Kumar. I am grateful to each of them, and I wish them all tremendous success.

Finally, I would like to thank my advisor Chad Myers. Chad has been my mentor, my advocate, and my friend since 2008. He has not only provided me with a model for how to do science well, but how to run a lab while engendering a shared excitement

for the work. I am very grateful for the opportunities Chad gave me, and the faith he put in me, and I am proud to have had an early contribution to what I know will be a lasting and successful lab.

Dedication

I have had an abundance of excellent role models to whom I owe everything, but none have had as positive or lasting an impact as my family. This dissertation is dedicated to my parents: Bill and Cheryl, and my two big brothers: Matt and Justin.

Abstract

In just two decades, advances in the experimental mapping and computational analysis of DNA sequences have resulted in complete reference genomes for thousands of different species. We therefore have a nearly complete "parts list" (that is, genes) for each of these organisms, but the task remains to discover the individual function of each of these genes, as well as characterize the organization and evolution of these individual genes into the many sub-systems at work inside the cell. Perturbation analysis is a crucial tool in identifying gene function and genetic relationships. In perturbation analysis, genes are selectively deleted or mutated, and any change in the resulting phenotype—for example, growth rate—can give an indication of gene function. We can then obtain a more complete functional map by systematically changing or combining genetic perturbations, and/or varying the environment under which we observe the phenotype. The focus of this dissertation is the development of computational methods to enable genome-scale perturbation analyses in yeast.

We begin the dissertation with a discussion of the first computational analysis of growth rate data for a comprehensive collection of deletion mutants in a wide variety of truly minimal environments. This analysis revealed how sources of nitrogen and carbon in the environment interact to determine growth rate, both in the context of wild-type strains, and in the context of individual single-mutants. We also discuss comparisons between experimental observation and *in silico* growth rate predictions which serve as a benchmark for current constraint-based modeling methods. Secondly, we discuss our efforts to map the complete genetic interaction network in yeast through a comprehensive set of double-mutant experiments. We explore the ability of genetic interactions and high-dimensional interaction profiles in to predict gene function, and describe both local and global properties of the genetic interaction network, which may reasonably be expected to be conserved to other organisms, such as humans. Lastly, we describe local properties of the genetic interaction network surrounding genes which have undergone ancient duplication. Using networks derived from both double- and triple-mutant experiments, we explore the consequences of duplication, divergence, and

retained common functionality, and speculate about the evolutionary process, and the constraints on that process which govern the fates of duplicate gene pairs.

Functional capabilities of genes are conserved across species to a surprising extent. Determining the functions of the remaining uncharacterized genes in yeast, will assist in the functional characterization of the thousands of remaining uncharacterized genes in human. Further, the mapping of the first complete eukaryotic genetic interaction network has direct impact on the study of complex, multi-genic phenotypes, including many human diseases. Meanwhile, the study of genetic interaction network structure, yields fundamental insights into the nature of cellular robustness, redundancy, and the evolutionary processes which give rise to them.

Contents

Acknowledgements **i**

Dedication **iii**

Abstract **iv**

List of Tables **xi**

List of Figures **xii**

1 Introduction **1**

1.1 *S. cerevisiae* as a model organism for computational biology 3

1.2 The impacts of whole-genome sequencing 5

1.3 Systematic perturbation collections for reverse genetic screens 8

1.4 High dimensional data and genetic networks 10

1.5 Gene duplication and models of evolutionary processes 12

1.6 Dissertation focus and organization 14

1.7 Remaining chapters 16

2 Broad metabolic sensitivity profiling of the yeast deletion collection **19**

2.1 Chapter Overview 19

2.2 Introduction 20

2.3 Results and Discussion 22

2.3.1 Prototrophic deletion set construction and profiling 22

2.3.2 Yeast wild-type growth suggests carbon/nitrogen interactions . . 22

2.3.3 Fitness determination of deletion mutants over the media conditions 27
2.3.4 Observations in galactose concur with previous auxotrophic studies 27
2.3.5 Liquid validation of mutant fitness measurements 28
2.3.6 Environmental sensitivities and genetic interaction degree 30
2.3.7 Mutant sensitivity profiles are predictive of gene function 31
2.3.8 Metabolic network models show modest ability to predict experimental data 31
2.3.9 Broad environmental surveys address incomplete gene annotations 39
2.3.10 Novel effects for genes with high fitness in standard conditions . 40
2.3.11 Novel phenotypes for uncharacterized ORFs 42
2.3.12 Clustering of metabolic conditions reveals carbon source as primary factor driving mutant profiles 42
2.3.13 Matrix factorization distinguishes carbon from nitrogen effects . 45
2.3.14 Environmental and genetic perturbations can provoke similar cellular states 46
2.4 Conclusions 49

3 Essential and non-essential genes in the complete genetic interaction network 52
3.1 Chapter Overview 52
3.2 Introduction 54
3.2.1 Defining and interpreting genetic interactions 54
3.2.2 Generating genetic interactions in yeast 55
3.3 Results and Discussion 59
3.3.1 Overview of interactions discovered 59
3.3.2 Assessment of experimental reproducibility 60
3.3.3 Correlation of reciprocal interactions 61
3.3.4 Replicate screening to estimate SGA precision and recall 62
3.3.5 Properties of the essential genetic interaction network 65
3.3.6 Variations in genetic interaction density 65
3.3.7 Information content of essential genetic profiles 65
3.3.8 Genetic interactions within and between functional modules . . . 67

3.3.9 The predictive power of essential profiles 72
3.3.10 The proteasome as an essential hub 74
3.3.11 Hierarchical structure in the genetic interaction network 78
3.4 Conclusions 84

4 Genetic interactions and the evolutionary trajectories of duplicate genes 86
4.1 Chapter Overview 86
4.2 Introduction 88
4.3 Results and Discussion 90
4.3.1 Genetic interactions are buffered subsequent to gene duplication 90
4.3.2 SGA data confirms synthetic lethality between duplicates 90
4.3.3 Redundancy between duplicates causes dissimilar profiles 92
4.3.4 Dosage duplicates are exceptions to the buffering model 94
4.3.5 Duplicates exhibit asymmetric genetic interaction patterns . . . 96
4.3.6 Dissecting divergent function through genetic interaction profiles 99
4.4 Conclusions 104

5 Complete functional profiles of paralogs revealed through trigenic genetic interactions 108
5.1 Chapter Overview 108
5.2 Introduction 109
5.3 Results and Discussion 111
5.3.1 Trigenic scoring model 111
5.3.2 Experimental Approach 114
5.3.3 Double-mutant query profile includes digenic interactions 119
5.3.4 Summary of trigenic interactions discovered 122
5.3.5 Functional validation of novel trigenic interactions 124
5.3.6 Different sub-types of trigenic interactions 127
5.3.7 Trigenic proportion as indicative of buffering capacity 132
5.3.8 Properties which correlate with trigenic proportion 132
5.3.9 Modeling evolutionary divergence 138
5.4 Conclusions 147

6 Conclusion and Discussion **149**

References **153**

A Appendix for Chapter 2 **183**

A.1 Construction of a prototrophic delection collection 183
A.2 Media preparation 183
A.3 Calculation of growth rate data 184
A.4 Definition and construction of a reference condition 184
A.5 Normalization of experimental rates against reference 184
A.6 Recovery of missing data 185
A.7 Transformation from normalized rates to z-scores 185
A.8 Spatial smoothing procedure 185
A.9 Choosing effect thresholds 186
A.10 Liquid growth confimation assay 188
A.11 Gene Ontology and KEGG annotations 188
A.12 Constraint-based metabolic modeling (FBA/MoMA) 188
A.13 Source signature decomposition via modified non-negative matrix factorization 190
A.14 Comparison to SGA data 190
A.15 Abbreviations 191

B Appendix for Chapter 3 **192**

B.1 Supplementary materials and methods 192
B.1.1 SGA array normalization 192
B.1.2 SGA gold standard definition 193
B.1.3 Protein-protein interaction data 193
B.1.4 Gene Ontology terms for functional prediction 193
B.1.5 Array gene function prediction via KNN 193
B.1.6 Hierarchical cluster filtering 194
B.2 Supplementary figures 195

C Appendix for Chapter 4 **197**
C.1 Ribosomal Duplicates 197
C.2 Sequence evolution rates support selection class distinction 198
C.3 Genetic interactions highlight the divergence of *GAS1* and *GAS2* 199
C.4 Self-reinforcing model of duplicate divergence. 199
C.5 Supplementary figures 201
C.6 Supplementary materials and methods 208
C.6.1 Definition of duplicates and singletons 208
C.6.2 Functionally related pairs 208
C.6.3 Significance of binomial proportions 208
C.6.4 Genetic interaction data and profile similarity calculations 209
C.6.5 Definition of dosage class 209
C.6.6 Ancestral proxies on the PPI network 210
C.6.7 Genetic interaction degree asymmetry 210
C.6.8 Chemical-genetic degree 210
C.6.9 Phylogenetic comparison for asymmetric pairs 211
C.6.10 Biological example profile similarity 211

D Appendix for Chapter 5 **212**
D.1 Supplementary methods 212
D.2 Supplemental figures and tables for Chapter 5 214

List of Tables

2.1 Sensitivity count for all conditions combinations 22
2.2 Overlap with previous galactose sensitivity screens 29
2.3 Performance of iMM904 when predicting slow-growth mutants 35
2.4 Performance of Sourceforge 5.35 when predicting slow-growth mutants . 36
2.5 Performance of iMM904 when predicting fast-growth mutants 37
2.6 Performance of Sourceforge 5.35 when predicting fast-growth mutants . 38
3.1 Strains surveyed by perturbation type 60
3.2 Genetic interaction counts by array and type 61
3.3 Replicate screening setup 62
4.1 Profile correlations for *SSO1* and *SSO2* 102
5.1 Trigenic interaction counts by dataset and type 124
5.2 Trigenic proportion correlations 134
A.1 FDR 20% thresholds for z-score data 187
B.1 Hierarchical clusters after filtereing 195
D.1 Trigenic proportion correlations 215

List of Figures

2.1 Experimental overview 23
2.2 Average wild-type growth rates in all conditions 25
2.3 Linear model for growth rate suggests interactions 26
2.4 Galactose sensitivities overlap with previous studies 29
2.5 Predicting gene function with conditional sensitivity profiles 32
2.6 Assessment of constraint-based modeling predictions 34
2.7 Number of producible metabolites 39
2.8 Counting effects for under-characterized genes 41
2.9 *FMP32* shows specific proline/arginine sensitivity 43
2.10 Clustergram of conditions for high-variance mutants 44
2.11 Equivalence between genetic and environmental perturbation 48
3.1 Genetic interaction example 56
3.2 Genetic interaction interpretations 56
3.3 Correlation of AB - BA observations 63
3.4 SGA precision and recall of biological replicates 64
3.5 Genetic interaction network density 66
3.6 Information content of interactions 67
3.7 Genetic interactions and essential protein complexes 69
3.8 Profile similarity example 73
3.9 Predictive power of essential and non-essential profiles 75
3.10 Proteasome genetic interactions 76
3.11 Proteasome module-module genetic interactions 77
3.12 Defining hierarchical clusters 79
3.13 Map of cellular function 80

3.14 Interaction density at several levels of functional specificity 82
3.15 Genetic interaction locality 83
4.1 A buffering model for the genetic interactions of partially redundant genes 91
4.2 Distribution of genetic interactions supports buffering hypothesis 93
4.3 Genetic interaction similarity comparisons support selection distinction 96
4.4 Genetic interactions support asymmetric functional divergence. 100
4.5 Functional analysis of *CIK1*, *VIK1* 101
4.6 Updated model of asymmetric duplicate genetic interaction evolution . . 105
5.1 Replicate data from trigenic mini-array 119
5.2 Per query replicate correlations 120
5.3 Double-mutant query recovers digenic interactions 121
5.4 Trigenic degree distribution 125
5.5 Novel trigenic interactions predict gene function 126
5.6 Map of trigenic interaction sub-types 129
5.7 Complex relationship hypothesis for region A 131
5.8 Distribution of trigenic proportion 133
5.9 Divergence modeling examples 140
5.10 Divergence model confirms asymmetry 142
5.11 Hypothetical map of duplicate divergence space 144
5.12 Modeling the axis of divergence 146
B.1 Functional process coverage of included strains 196
C.1 Duplicate deletion has less impact on cell fitness 202
C.2 Proportion of Duplicates in "Dosage" class 203
C.3 Asymmetric pairs show retained shared functionality 204
C.4 *SSO1/SSO2* localization differences 205
C.5 Simplified model for duplicate sequence evolution 206
C.6 Asymmetric divergence model for duplicate genes 207
D.1 Pilot study degree scatterplot 214
D.2 Precision and digenic/trigenic overlap 215
D.3 Trigenic interaction predict protein-protein interactions 216

Chapter 1

Introduction

Data generation is no longer a constraining factor in modern molecular biology. In recent years, experimental methods and technologies have been scaled up to the extent that individual studies routinely contain millions of observations. The volume and diversity of data has become so vast that the central challenge now lies in making sense of it all. Computer scientists are responding to this challenge with new methods for integrating, summarizing, and interrogating these diverse data. In close collaboration with biologists, computer scientists not only work to address the technical and conceptual challenges raised by these new methodologies, but also to exploit the new opportunities they present. For example, computer science was integral to the success of early "shotgun" sequencing technology [1, 2]. By solving the genome fragment assembly problem, computer science empowered us to read an entire genome at once, instead of crawling along one chromosome at a time for fear of losing our place. The sequencing achievements of the late 1990's and early 2000's kicked off a revolution in biological investigation and experimental technology in which computer science would play a continually increasing role. New challenges in the post-genome era include the integration and comprehension of this wealth of data, as well as technical and statistical problems related to the reduced signal in the the data, and its sheer scale.

To exploit these opportunities, and to tackle the obstacles they present, we have a myriad of tools at our disposal. Perhaps most importantly, we have complete genome sequences. A complete listing of the genes in a given organism not only gives us a list of potential actors when making specific biological hypotheses, but gives us a framework

on which to organize all of our information and analyses. For a researcher such as me, who began his career after many complete genomes had become available, this is easy to take for granted. However, work on one particular project concerning *Cryptococcus neoformans* gave me a new appreciation of the benefits of having a complete genome in hand. While a *C.n.* reference sequence existed, it lacked the standardized nomenclature and robust functional annotation scheme used in more studied organisms such as *Saccharomyces cerevisiae* (yeast). This hindered analysis and comparisons with previous work and made communication of results much more difficult than I was accustomed to after working with yeast. Indeed, much of the analysis I performed was done through the lens of the yeast genome, leveraging the principle of evolutionary conservation of function to borrow as much knowledge as I could (including even gene names) from a better understood and annotated genome. Genes are the essential vocabulary of cellular operation, and until the advent of whole-genome sequencing, reasoning and communicating about cellular biology had to be done with an incomplete lexicon; the positive impact of this standardizing force cannot be overstated.

Another essential tool in modern biological investigation is perturbation analysis. While the study of mutants predates even the concept of the gene, the study of intentional mutation is much more recent, and the systematic perturbation of every gene in a genome in order to study each of their functions has only been possible for just under two decades [3, 4]. Enabled by the revolution in sequencing and the complete "parts list" it provided, systematic perturbation analysis has made an indispensable impact on the field of functional genomics. Not only has systematic perturbation analysis resulted in an explosion of gene-specific observations [5], but through new computational approaches, these observations have coalesced, engendering more systematic views into entire systems and organisms [6].

The study of biological networks is one such systematic view, and interest in the mapping and construction of these biological networks has increased dramatically in recent years. Network science abstracts a view of something, in this case cellular function, reducing it to a set of nodes and edges which can be interrogated computationally. For example, nodes can represent biological actors, such as genes or proteins, while edges might capture the relationships between them [7]. In many cases, such as protein-protein interaction networks, these relationships represent tangible mechanisms such

as the ability for one protein to physically bind to another, while in other cases edges represent something more conceptual, such as similarity with respect to experimental observations or the integration of many other heterogeneous data types [8, 9, 10]. Once the data has been cast as a network, computer scientists can apply a standard set of tools to study the architecture of the system. Sometimes network questions are simple, such as asking which nodes are connected after conversion to a network, while other questions are more complex, like asking how fragile the network is as a whole [11, 12], or what types of processes could give rise to a network with similar aggregate properties [13, 14], or which network sub-structures occur more often than expected [15, 16, 17].

Network-level characterization is one of the most important methods by which we address complex biological systems as a whole. Life is an emergent property of all of cellular components, which interact in complex ways, and must be studied in aggregate to achieve anything approaching true comprehension. Cellular metabolism is probably unrivaled in terms of large cellular processes we are able to model *in silico*, in large part due to its amenability to abstract network. Metabolism provides an excellent example of an instance where, after mapping the principle actors and their relationships, we can simulate emergent systems-level properties and check agreement with experimental observation. Mapping these networks is a huge challenge being addressed by modern computer scientists in close collaboration with experimental biologists, as is learning to integrate and interrogate networks to test hypotheses both broad and specific.

Paramount to the application of all of these tools is a model organism from which we can collect the vast amount of data required to test the resulting hypotheses. *S. cerevisiae* is such a model system and has played a key role in the development and application of these tools.

1.1 *S. cerevisiae* as a model organism for computational biology

S. cerevisiae has been used as a model organism for the study of biochemistry, genetics, and cell biology for decades. Owing to its many important roles in human history (both in and out of science), it is alternatively known as brewer's yeast, baker's yeast, or budding yeast. Yeast belongs to the fungal kingdom, which has a close evolutionary

relationship to animals. In its complex structure, functions, and organization, yeast resembles many other eukaryotes, including humans, despite their roughly one billion years of evolutionary divergence.

The *S. cerevisiae* genome contains roughly 12 million bases (Mb) of DNA that is organized into approximately 6,000 protein-coding genes [18]. Using gene count as a measure of complexity—albeit a notoriously unreliable one—this ranks well above the smallest observed free-living genomes, such as the bacterial genome of *Pelagibacter ubique* which has ∼1,354 genes (1.3 Mb) [19] and the genome of the eukaryotic parasite *Mycosporidium Nematocida parisii*, which has ∼2,600 genes (4 Mb) [20]. The *S.c.* genome is larger by half than the model prokaryote *Escherichia coli* which has ∼4,200 genes (4.6 Mb) [21]. Gene counts among eukaryotes vary widely; humans for example have between 20,000 and 40,000 genes. The largest number of genes observed in a eukaryote as of 2008 was about 60,000, belonging to a single-celled human-infecting parasite called *Trichomonas vaginalis* [22]. *S. cerevisiae* was chosen as a model organism for early sequencing efforts in part because of its position on the low end of this spectrum relative to other well-studied eukaryotes [23].

Yeast cells propagate vegetatively in either the haploid or diploid form and double themselves once every one and a half to eight hours. There is also a sexual cycle in which haploids of opposite mating types (so-called $MAT\mathbf{a}$ and $MAT\alpha$) can mate with one another to form a diploid cell. Diploid cells can undergo meiosis to produce four haploid spores, two with $MAT\mathbf{a}$ and two carrying $MAT\alpha$. The ability of yeast to persist vegetatively as either haploids or diploids allows us to grow isogenic colonies and therefore conveniently attribute colony-level properties such as growth rate to a single genotype. Meanwhile, their ability to reproduce sexually creates additional opportunities to construct those genotypes that we seek to test.

Another desirable feature found in yeast is the high frequency of homologous recombination, which is a mechanism commonly invoked to repair double-stranded breaks (DSBs) in DNA and also to produce novel genotypes during meiosis [24]. The DSB repair process uses the second chromosomal copy as a template when fixing the damaged region, and this process can be commandeered, allowing geneticists to replace or remove genomic DNA at a specific location in the genome [25] provided they provide a carefully crafted DNA sequence as an alternative template. In *S. cerevisiae*, the homologous

recombination pathway is frequently invoked even in the absence of a DSB provided the template is there. This feature allows yeast researchers to skip the laborious step of additionally causing a break at the precise region of the genome targeted for change. Recently, this technique, known as "gene targeting," has been made more efficient in organisms with lower rates of homologous recombination by the development of technologies which causes DSBs at the targeted locus, inducing homologous recombination at a higher frequency [26, 27, 28, 29].

One other facet of yeast's lifestyle is essential to mention here. In its role as a free-living single-celled organism, yeast must manufacture many of the essential raw materials needed to propagate. In fact, yeast metabolism is quite flexible; yeast can subsist in very simple environments and generate energy by either respiration or fermentation. This is in contrast to humans, which subsist by making comparatively few metabolites directly and obtaining the majority of its nutritional requirements through the consumption of other organisms. Yeast's humble place on the food-chain ensures that it has a metabolic toolbox that is well-equipped to teach us about basic cellular metabolism.

Yeast is one of many different species that have been domesticated by humans, and while the relationship does not quite predate the one with man's supposed best friend (dogs), evidence suggests that yeast domestication goes back several thousand years. The first brewing activity is thought to have taken place in Sumeria nearly 6,000 years ago [30]. It is thought that *S. cerevisiae* in particular came into widespread use for brewing some 1,000 years ago in the middle ages [31]. Since then, humans have cultivated many yeast strains for many different purposes, exploiting both yeast's respiratory and fermentative capacities in the bakery and brewery, respectively. Yeast ultimately found its way to the lab by 1940 [32]. Since then, perhaps no other species has contributed as much to modern molecular biology and genomics, and therefore to computational biology as well.

1.2 The impacts of whole-genome sequencing

The principle tool of the genomicist is a complete map of the genome itself. The first genetic map of *S. cerevisiae* was published in 1949 by Carl Lindegren [33], and covered

only four chromosomes. The very first complete DNA genome of any organism was not sequenced until 1977 by Frederick Sanger [34]. That organism, a bacteriophage, Phage Φ-X174, was quite simple with a genome of only around 6,000 bases (6Kb). It was not until 1996 that the technology had progressed enough to directly sequence and assemble a complete eukaryotic genome (yeast) [18]. In contrast to Sanger's phage, the yeast genome has 12.1Mb, over 2,000 times the amount of genetic material. A consortium of 94 labs in 19 countries worked in concert to release the first version of its complete sequence. This same period saw several other major sequencing milestones, including the first complete bacterial genome in 1995 (*H. influenzae* [1]), the first archaeal genome in 1996 (*M. jannaschii* [35]), and the first multi-cellular eukaryotic genome in 1998 (*C. elegans* [36]). By the time the first human genome draft was completed in 2000 [37, 38], it had become apparent that many genes and gene functions were conserved to a striking degree across vast stretches of the tree of life. Several experiments showing the ability to functionally complement human genes with the yeast homolog and vice versa highlighted the possibility of using yeast to study processes important for human physiology. In fact, gene function in one species so often mimics another that cases of interspecies functional complementation can be now be referred to as "routine" [6, 39].

Today, complete genome sequences exist (to varying degrees of completion) for thousands of species. These species include hundreds of eukaryotes, and among them, dozens of species closely related to the original model yeast. The explosion of sequence data gave rise to entirely new methods and analytical paradigms, ushering in the new field of genomics. This field concerned the investigation of many genes at once and required new computational methods to meet the challenge. In order to realize the promise of translating knowledge from the genes of one organism to another, bioinformaticists developed efficient methods for sequence alignment and ortholog recognition [40]. The study of non-genic regions also received a boon from the flood of sequence information as computational algorithms were refined. For example, transcription factor binding sites, which are less regularly structured than protein coding regions, could be discovered computationally by comparing the now large number of examples looking for subtly conserved patterns. Once these patterns were sufficiently well characterized, computational biologists could apply the models to other regions or other genomes to predict novel binding sites [41].

Beyond even the level of individual genes and the regions surrounding them, computational biologists have begun to formulate hypotheses about entire genomes. Computational methods for sequence analyses have established the evolutionary-genomic history of the entire yeast clade, revealing a great deal about the processes of gene duplication and loss, as well as genome-scale rearrangements [42, 43, 23]. These yeast sequences have also anchored even larger comparative studies, contributing much to what we know about molecular diversity, divergence, and speciation, on many different branches of the tree of life.

The collection and analysis of sequence data further spurred the development of other whole-genome technologies, which again, demanded novel computational methods. As one example, in 1997—a single year after the publication of the entire yeast genome—a complete, genome-wide expression profile was published using microarrays [44]. DNA microarrays can measure differential gene expression or detect known single-nucleotide polymorphisms (SNPs) for thousands of genes simultaneously [45]. Methods originally adapted to process and interpret microarrays, such as clustering in one or two dimensions [46, 47], or principal components analysis [48], are now routinely applied to many diverse data that are high-dimensional in nature [49, 50].

On a conceptual level, the lasting impacts of whole-genome sequencing on functional genomics have been profound. While proteins are not the sole participants of cellular function, they are responsible for most of the major structures and actions required for life. A complete catalog of coding genes, and therefore a listing of proteins, gives us a reasonable place to begin searching for the mutated gene responsible for any given phenotype of interest. It also encourages a more systems-oriented view of the cell, and provides a driving force behind systematic collaborative study. For example, the nomenclature for systematic gene names in *S. cerevisiae*; which encodes a chromosome number, arm identifier, gene number, and strand information; was conceived and refined in the 1990's as a product of the sequencing consortium [32]. As a result, even genes without a known phenotype, or those which might not be genes at all, could be identified, categorized, and annotated in a centralized and unambiguous way. Perhaps most fundamentally, a complete listing of every gene stands as a challenge, turning a classical genetics question on its head. Instead of asking "Which gene is responsible for phenotype X?" functional genomics asks "Can we find a function for each possible gene

Y?" Despite years of work, and much progress, this challenge remains relevant in even the most highly characterized organisms [51].

1.3 Systematic perturbation collections for reverse genetic screens

The idea of perturbing or deleting a specific gene and using the resulting observed phenotypes to infer something about the gene's function is known as "reverse genetics." It differs from "forward genetic" screens, which seek to isolate mutations responsible for causing an existing phenotype. Perhaps no single event has had such a lasting impact on reverse genetics (and therefore functional genomics) as the release of a near complete yeast deletion collection around the year 2000 [52, 3]. In a characteristically cooperative and coordinated fashion, the yeast genetics community constructed a set of ∼4,800 strains, each lacking a single specific gene. The design of the strains in the collection sought to facilitate experimentation and subsequent modification. As a result, many unique derivatives of the collection exist, each specialized for particular biochemical and genetic assays [5]. At the time of this writing, the deletion collection study published in 1999 (Winzeler *et al.*) has been cited over 2,800 times [52].

Other model organism communities have since followed suit in the construction of such collections. A collection for *E. coli* was released in 2006 [26], and a *Saccharomyces pombe* collection was released in 2010 [53]. The technical hurdle for producing such collections for each new organism is usually finding a mechanism for gene replacement that can be applied at genome-scale. Homologous recombination can be used to generate targeted deletions in mouse [54], or in *Caenorhabditis elegans* [28]. In *Drosophila melanogaster*, homologous recombination shows promise but cannot yet be applied at genome-scale, though large random mutation collections exist [55]. Meanwhile zinc-finger based mutation technology has produced made-to-order deletions in *Arabidopsis thaliana* a reality and may form the bases for a comprehensive deletion collection in the future [27]. New technologies such as CRISPR hold promise for the rapid production of such collections as well [28, 29].

These deletion collections have been successful at eliminating the rate-limiting step of mutant construction, making genome-wide screens a viable option for consideration

when planning any experiment. The have allowed researchers to use established assays at larger scale, measuring numerous phenotypes, or measuring phenotypes in different environments in search of additional sensitivities. In 2002, Giaever *et al.* not only characterized (qualitatively) the growth of each yeast mutant, but did so on several different types of media [3]. In the following years many studies leveraged the yeast deletion collection to detect changes in growth phenotype in different environments, many of them containing bioactive drugs with human therapeutic ends in mind [3, 56, 57, 58].

Among the important findings to emerge from this body of work was that for many genes, but not all, there was at least one environment in which that gene was essential for growth. This partially accounted for the surprising observation that in an organism as supposedly efficient as yeast, only 20% of genes were required for growth in basal conditions [3]. These experiments also provided the first unbiased measurements of gene pleiotropy. While the idea of "one gene for one function" was by this time already obsolete, it was unknown how many genes were pleiotropic or to how many different phenotypes those genes might contribute [58]. To this day, the list of genes with no apparent defect in any environment in yeast remains substantial, though it is much shortened. Between 2011 and 2013 the number of ORFs listed as "Uncharacterized" in the Saccharomyces Genome Database dropped from 897 to 846. At the time of this writing, 722 ORFs still remain uncharacterized [59]. Meanwhile, so much perturbation data exists that, instead of classifying gene function by the chemical to which it is sensitive, it may soon be possible to classify unknown bioactive compounds by the set of gene deletion mutants that sensitive to the chemical [60, 61]. Modeling the action of known bioactive compounds in terms of their chemical structure and binding properties has proven to be a significant computational challenge. Ultimately, our interest in bioactive compounds extends beyond their specific targets to include downstream effects as well as even a correct prediction of which gene product a putative drug candidate targets will not inform us as to the consequences of that gene's inactivation. Regulatory responses, alternate pathways, and potential genetic interactions with secondary targets all need to be considered, and perturbation studies can yield much of this information. Computational modeling of these chemical properties will no doubt continue to improve, meanwhile, perturbation-based analysis of these compounds allows us to begin facing

this complex challenge immediately.

1.4 High dimensional data and genetic networks

As diverse phenotypic data became available for a large number of genes, dimensionality reduction and data integration became a new computational focus. Methods that leverage high-dimensional data frequently work by finding coherent patterns across profiles, often reducing the dimensionality to a single measurement of profile "similarity" (or distance). The similarity of high-dimensional profiles has proven quite effective at predicting known functional relationships between genes, and is therefore suitable for making novel functional predictions for poorly characterized genes. Related methods such as clustering in one or two dimensions [46, 47, 49], and principal components analysis [48], provided alternative approaches to detecting common patterns, and applying them to inform us about the relationships between sets of genes or experimental environments.

Another related approach is to build networks of pairwise similarity relationships. Networks provide an important tool in conceptualizing the organization of cellular operation [7]. For example, gene-gene expression similarity measured across high-dimensional microarray experiments can be used to generate coexpression networks, which have been hugely influential in determining how genes respond (and respond together) in different cell types or different environmental conditions [62]. When methods were needed to aggregate the data from many heterogeneous experiments and present it in a useful format, the network again proved a useful construct [9]. For example, functional linkage networks integrate pairwise gene data from many different assays and summarizes them in one single network. Most often, these network are constructed using a Bayesian framework which is trained to predict a known—if incomplete—functional standard, such as annotation to a common term in the Gene Ontology [8, 63, 64]. These networks can be used to make predictions about particular gene functions and relationships and in the process can be adapted to different prediction settings by marshaling the data sets included, accounting for their relative usefulness or reliability given the task at hand. For example, if a given dataset (such as gene-gene co-expression to salt-related stress) boosts prediction of general functional relationships but is shown to provide no

additional information to the specific task of predicting known protein-protein interactions, it can be withheld to create a functional linkage network more suitable to the prediction of novel protein-protein interactions.

In addition to computationally derived networks, there are a number of important experimental technologies that measure specific types of gene-gene relationships and give rise to their own networks. The most prominent of these technologies measure direct physical interactions between proteins [65], interactions between proteins and DNA, or logical relationships such as genetic interactions [66]. These individual networks have demonstrated a number of interesting properties. For example, they are often scale-free, or close to it, with a large majority of genes having very few connections, and a minority of "hubs" having many connections [7]. These hubs tend to be centrally located in the network and represent the most influential genes. Indeed, hub status in one type of network frequently predicts status in another, suggesting that networks of complementary types often portray the same underlying functional organization. For example, essential genes are more likely to be hubs in both the protein-protein interaction network and the genetic interaction network [11, 67], [Chapter 3]. Jeong *et al* suggested that their central location in these cellular networks is what makes these genes essential, while others contend that the observed relationship is the result of their increased degree influencing their probability of participating in certain essential interactions [68]. Still others contend that a minority of highly connected protein complexes, many highly enriched for essential genes, are causing the correlation, and that it is not a general phenomenon [69].

Genetic interactions, by definition, measure the effects of multiple perturbations at once in search of surprising phenotypes, and give rise to their own network structure. For example, if two individual mutations show no apparent phenotype, yet prove to be fatal in combination, they are said to have a "synthetic lethal" genetic interaction. The yeast deletion collection, along with recent technological advances in robotics have enabled the systematic construction of millions of double-mutant strains in an effort to map the entire genetic interaction network. The scale of these experiments has increased so rapidly that the number of genetic interactions in the BioGRID database has already eclipsed the number of physical interactions by nearly a factor of two [70]. The information gained from genetic interaction mapping efforts is directly applicable

to the study of complex phenotypes. Consider, for example the so-called "missing heritability problem," which refers to the fact that the effects we see as a result of individual genes fail to account for the total heritability of many diseases and other complex phenotypes [71, 72, 73]. Because individual genes are seldom responsible for observed phenotypes, mapping the genetic architecture by which genes jointly affect these complex phenotypes is essential to addressing current problems such as missing heritabiltiy and complex genetic disorders in humans.

The exploration of fitness consequences in both gene-environment and gene-gene space continues and the two can be related to one another with high-dimensional observations. For example, it was shown that correlation between the sensitivities of a drug condition and a gene deletion, when measured across the same set of mutant backgrounds, could indicate a targeting relationship between the drug and the deleted gene [57, 66]. This is one example where the use of profiles can overcome the noise and functional ambiguity of individual interactions. Lots of information, each bit of which provides no real mechanistic information, can aggregate to provide real biological insight. Mapping of the first comprehensive digenic network is nearly complete, however there is no reason to suspect that genetic redundancies do not involve more than two genes. And so, despite the incomprehensible number of possible triple-mutant combinations in yeast ($\sim 3.6 \times 10^{10}$), targeted explorations of the trigenic genetic interaction space have begun.

1.5 Gene duplication and models of evolutionary processes

Gene duplication was recognized early on as a primary source of raw genetic material and therefore key to the evolutionary process. This case was put forward most completely by Sumusu Ohno in his 1970 work Evolution by Gene Duplication [74]. His argument was that a second copy of a gene relieves selection pressure on the first and one or both of them are then able to mutate, perhaps dividing their common responsibilities. Alternatively, one may assume the bulk of the ancestral function while the other specializes or becomes useful in some novel capacity. This has been a very active area of research since Ohno's work but his central assertions are still accepted. Yeast provides a platform on which we can come to understand these fundamental evolutionary forces.

In addition to driving the evolutionary history of a genome, duplication is thought to contribute to the observed robustness against null mutations in many organisms. For example, if extant duplicate genes have retained common functionality, it may explain the observed scarcity of essential duplicates, in yeast.

Many models, not all of which are mutually exclusive, have been proposed and refined to explain the process by which duplicate genes might diverge [75]. One model, known as duplication-mutation-complementation (DMC) is perhaps closest to Ohno's original idea. In this model duplication results in two fully functional redundant gene copies, each of which then accumulates complementary mutations until they have completely partitioned ancestral function in a process called sub-functionalization. This process has been applied *in silico* to network models, and has been shown to result in network structures resembling the real protein-protein interaction network [13]. Another model, called escape from adaptive conflict (EAC) proposes a slightly different model of sub-functionalization. In the EAC model, genes perform multiple functions which are not discrete, but are subtly different. In cases such as this a gene cannot converge to optimal performance for one function, without a corresponding sacrifice in the performance of the other. The two functions are said to be in "adaptive-conflict" and a duplication event can resolve this conflict, by allowing two distinct gene copies to specialize. One example of this hypothesis was demonstrated by Voordeckers *et al.* who compared the kinetic properties of the *MALS* gene family in several closely related yeasts [76]. Through computational reconstruction of key pre-duplication versions of the enzymes, and *in vitro* reaction velocity measurements on different substrates, coupled with sequence and structural analysis, the authors show that the evolution of an enzyme preferential to isomaltose-like sugars came at the expense of its ability to metabolize maltose. Thus, these two similar functions were in adaptive-conflict, and multiple post-duplication copies were able to specialize for both of these competing roles, each becoming more efficient than the ancestor for one particular task.

Previous work has shown that an ancestor of *S. cerevisiae* survived a duplication of its entire genome [77, 42]. This whole-genome duplication (WGD) event is thought to have taken place approximately 100 million years ago, and was followed by a period of rapid reorganization and genome reduction. Several hundred duplicate pairs survived the yeast WGD event, and together with many pairs from small-scale duplication events,

make up the nearly 33% of the yeast genome [78]. Whole-genome duplication events are traumatic, and we generally detect very few survived instances in most extant organisms. On the other hand, small-scale duplication (SSD) events, in which a small number of genes are duplicated, are common, and commonly survived. Sequence analyses have revealed that most species retain a significant level of duplicated genes. To discover the reason such a large amount of potentially redundant genetic material is retained in such a wide variety of organisms, we must integrate the diverse functional data we have for duplicated genes, as well as construct computational models of the processes by which they are created or destroyed. Yeast is well equipped to answer many questions about duplications and functional specialization through experiment, thanks in part to its interesting evolutionary history. However, it is the creation of evolutionarily relevant computational models, and their ability to make testable predictions in a wide range of organisms, which will ultimately lead to a better understanding of these processes.

1.6 Dissertation focus and organization

The central challenge in modern computational biology lies in the integration and interpretation of the vast amount of data now available. Computer scientists are responding to these challenges, and in the process, shifting the focus of study from particular genes to networks and network structures, and eventually, to the whole organism. This dissertation concerns three central issues, which are not mutually exclusive, and have arisen as a natural consequence of the new era of functional genomics and systems biology. These three issues can be characterized as the problems of *scale*, *signal*, and *systems*, and the focus of my work has been to address these issues in the specific domain of mutant growth-rate analysis in yeast.

The *scale* problem refers to the amount of data from which we must extract specific biological information. The number of data points in even rudimentary studies is expanding with the advance of experimental technology, and shows no sign of slowing. To accommodate this trend, we need efficient methods to process and summarize large amounts of data. We need to formulate hypotheses about how entire distributions might behave and test them accordingly. We also need to understand that these distributions seldom conform to well understood theoretical models, so we must employ empirical and

non-parametric methods to limit our assumptions. Growing along with the number of data points in any given study is their dimensionality. Frequently, we must make sense of thousands of different measurements for each of thousands of genes. Here we must develop special methods to visualize the data, before we even begin to test hypotheses. Dimensionality reduction is often the first step, but care needs to be taken to prevent the loss of useful information. Furthermore, we have to distinguish which dimensions carry useful information and discover their relationships to one another, which we can do by applying techniques from the field of machine learning.

The *signal* problem is the price we pay for the scale of modern experimental technology. As a general rule, individual observations made in high throughput are not as reliable as those made in comparatively smaller experiments, and we must adjust our thinking accordingly. Much of the increased noise characteristic of genome-scale assays comes in the form of systematic effects which can introduce spurious structures in complex data. We can address these effects in our analyses by using computational techniques which target known or unknown non-biological effects for removal. We can also apply computational methods much earlier in the experimental design phase to mitigate these effects. Often, reduced signal is only a minor problem because large experiments can be exploratory in nature, narrowing the search space for a particular phenomenon. For example, high throughput experiments followed by small-scale confirmation to eliminate false-positives, can be a cost effective strategy to search for drug-target or protein-protein interactions, which are rare among millions possibilities. Some of the same methods we employ to deal with the scale of the data (for example, reducing high-dimensional profile pairs to similarity scores) also help us overcome increased noise by lessening the reliance on any single observation. And, just as the pooling observations in a single data set can increase functional signal, so too can integrating data from many diverse experiments.

The third issue is that of *systems*. This refers to the opportunity we have, as a result of scale, to treat biology more holistically. We can now begin to assemble all of the information collected with single-gene reductionism in mind, and compile it together to build complex models involving many genes at once. For example, genome-wide data may be used to infer the organization of an entire cellular subsystem, such as the metabolic or regulatory network. Alternatively, it may be used to study how a common

process affects genes in aggregate, such as mechanisms of evolution and functional divergence. Or it may be used to inform notions of higher level organization, such as the distribution of pathway lengths or branching factors, or the hierarchical organization of modular subsystems. Each of these ends requires observations be made at scale and each are reasonably tolerable to lower levels of signal. Taken together, these three interrelated issues form the core set of challenges to modern biological investigation, and provide the most interesting applications to today's computational biologists, functional genomicists, and/or bioinformaticists.

1.7 Remaining chapters

This dissertation's remaining chapters each touch on issues raised in this section. Here I present a brief outline of the remaining chapters and discuss how each of them relates to particular issues regarding the study of mutant growth rates in *S. cerevisiae*.

Chapter 2 covers the study of single-mutant growth rates on a wide variety of media. The scale of the study encompassed the entire non-essential genome and 28 different, metabolically relevant environments. This was the first study of its kind conducted on prototrophic mutants and uncovered many novel condition-specific sensitivities. Many mutants showed novel effects in more than one environment, an indication of pleiotropy. These novel sensitivities will help shorten the list of uncharacterized genes in the model eukaryote, and therefore aid in cross-species research in matters of basic metabolic function. Further, the scale of the project enabled us to answer questions concerning the absolute number of genes with a such specific metabolic growth signatures. These included genes with no direct involvement in metabolism such as transporters or transcription factors. The relationship of the environments to one another, as well as the number of mutants tested, allowed us to ascertain that carbon sources have a much greater impact on the internal state of the yeast cells than do nitrogen sources. The signal problem was addressed heavily in the preliminary data processing procedures, by the application of methods that were robust to noise. The amount of available data allowed us to measure false discovery characteristics, and disentangle robust carbon- and nitrogen-based effects. I also demonstrate the use of high dimensional profile similarity and guilt-by-association to predict gene function in spite of experimental noise. I also

was able to leverage existing models of the complete metabolic system and simulate the effects of systematic perturbations. Comparisons between these simulations and our observations identified those environmental conditions which were not yet well suited to simulation. Finally, I integrated the prototrophic growth data with the genetic interaction network, finding correspondence between the number of sensitivities in our study and node degree in the network. An integrated similarity score further allowed us to search for correspondences between environmental and genetic perturbation, and I speculated as to the biological underpinnings of such a correspondence.

Chapter 3 explores large-scale genetic interactions derived from double-mutant fitness experiments. I give an overview of genetic interactions and use them as a tool to compare and contrast the placement of essential and non-essential genes in yeast's genetic architecture. Here again, use of high-dimensional profiles as predictors of knockout consequence proves useful in determining gene function, and considerable attention is paid to demonstrating the usefulness of genetic interactions despite the absence of mechanistic information and increased noise. Computational analysis of genetic interactions in related species has shown that properties associated with genetic interactions can be conserved even when specific individual interactions are not [79]. I also explore the genetic interaction network as a whole, demonstrating systematic properties which might reasonably be expected to be conserved, such as its modular structure. The tendency of genes to cooperate in physical modules has been demonstrated experimentally by the characterization of protein complexes [65], we and others have used computational techniques to reveal a more logical modular structure, such as that found in gene expression patterns [46, 47], or in genetic interaction networks [66]. I also examine the organization of these modular structures in relation to one another. For example, much previous work has uncovered hierarchical structure in the transcription factor network, where the expression of specific groups of genes are controlled by transcription factors, who are themselves controlled by transcription factors from a higher level [80, 81]. I show that genetic interactions also have a hierarchical organizational structure, and that this structure is useful for characterizing different relationships between cellular processes. These hierarchical relationships are also helpful for distinguishing different classes of genetic interactions, and will prove useful in their interpretation in yeast and their mapping in other species.

Chapter 4 again examines double-mutant genetic interactions, but focuses on the specific case of duplicated genes. Many pairs of duplicated genes retain some measure of functional overlap, and this property causes specific aberrations in the genetic interaction network. The genetic interaction network is particularly well suited to the study of functional robustness and the interplay between gene duplication and genetic interaction network structure can give us valuable insights into the mechanisms of evolutionary divergence. In particular, I set out to explain an apparent contradiction whereby duplicate genes show reduced genetic interaction profile similarity, despite their obvious signs of redundant function. I exploit the large number of duplicate pairs and their comprehensive genetic interaction profiles to explore the nature and extent of this functional redundancy, and I integrate diverse functional data to define classes of duplicate pairs that I found to have different network properties. I also demonstrate the asymmetric nature of duplicate evolution and use a computational model to establish which biological assumptions are necessary to explain the observed asymmetry.

Chapter 5 follows predictions made in Chapter 4 to their logical conclusion with another study of redundancy and divergence in duplicate gene evolution. Here I explore triple-mutant genetic interactions, or trigenic interactions, which relate the retained common function of duplicate pairs not apparent in the pairwise genetic interaction network. I introduce a novel model for scoring trigenic interactions and show how extra care must be taken in the design stages of trigenic experiments as the previously acceptable level of experimental noise compounds itself. I also demonstrate several different sub-classes of trigenic interactions, as well as the variation in the number of trigenic interactions relative to digenic interactions. By integration of diverse genomic data I discover physiological indicators of retained functional redundancy, and use these indicators to inform an updated model of the process of gene duplication and divergence. This model, which can make predictions about gene loss, functional asymmetry, and retained common functionality, can generalized to the gene duplication in any organism.

Chapter 6 concludes the dissertation with reflection and suggested directions for future work.

Chapter 2

Broad metabolic sensitivity profiling of the yeast deletion collection

2.1 Chapter Overview

This chapter covers an in-depth analysis of single-mutant data gathered for a prototrophic derivative of the entire deletion collection. Genome-wide sensitivity screens in yeast have been immensely popular following the construction of a collection of deletion mutants of non-essential genes. The complete collection was grown in environments consisting of one of four possible carbon sources paired with one of seven nitrogen sources, for a total of 28 different well-defined metabolic environments. The relative contributions to mutants' fitness of each carbon and nitrogen source were determined using multivariate statistical methods. The mutant profiling recovered known and novel genes specific to the processing of nutrients and accurately predicted functional relationships, especially for metabolic functions. A benchmark of genome-scale metabolic network modeling is also given to demonstrate the present level of agreement between current *in silico* predictions and hitherto unavailable experimental data. These data address a fundamental deficiency in our understanding of the model eukaryote *Saccharomyces cerevisiae* and its response to the most basic of environments. I demonstrate

utility in characterizing genes of unknown function and illustrate how these data can be integrated with other whole-genome screens to interpret similarities between seemingly diverse perturbation types.

This chapter has been adapted from a previously published study entitled "Broad metabolic sensitivity profiling of a prototrophic yeast deletion collection" [82], on which I was a co-first author. The article version was published in 2014 in Genome Biology.

The statistical analysis in the paper was carried out by me. Specifically, I performed the analysis of wild-type growth rates, the assessment of functional predictive power, overlap with previous experiments and all analysis regarding data collected for validation experiments. Additionally, I performed all cluster analysis, the multi-variate profile decomposition, and conceived and interpreted the data from the flux balance analysis. I also made all the figures and was principally responsible for the text of the paper.

The study was originally conceived by David Hess, Olga G Troyanskaya, my advisor Chad L Myers, and Amy A Caudy. Drs. Caudy and Hess are experimental biologists specializing in yeast metabolism and performed the experiments, they also helped write the paper with special attention to any section requiring their metabolic expertise. Corey Nislow provided us with access to lab equipment. Balázs Szappanos and Balázs Papp are experts in flux balance analysis and provided crucial insight and guidance over the course of the project and in the writing of the paper. Colin Pesyna and Tahin Syed worked on the scoring pipeline which converted raw plate images into informative growth scores. Elias W Krumholz generated flux balance analysis data and helped to quantify the agreement between contraint-based model predictions and our experimental observations.

2.2 Introduction

Large-scale gene deletion screens have become common in *Saccharomyces cerevisiae* due to efforts in the yeast community to assemble a near complete collection of non-essential single-mutant strains [3]. The subsequent refinement of mating-based high-throughput strain construction techniques such as Synthetic Genetic Array (SGA) analysis [83] has further driven the creation of customized yeast deletion arrays. While quantitative single-mutant fitness assays have been performed [50], they are generally limited to

a single growth medium. A few notable exceptions have begun to explore this space [84, 57, 56, 58], but the conditions of interest are often chosen with human therapeutic ends in mind and are limited to known drugs or small molecules of unknown biological effect. A decade and a half after the sequencing of the best-studied eukaryote, a systematic exploration of mutant growth across basic nutrient environments is conspicuously absent. These data would be valuable for metabolic researchers and computational biologists that attempt to model the metabolic network of the cell using methodologies such as flux balance analysis (FBA) [85] because the defined growth conditions are amenable to modeling.

Yeast strain collections used in previous high-throughput assays (that is, the deletion collection) are auxotrophic [3], and therefore unable to survive in minimal media unless provided additional nutrients. This requirement reflects the historical use of auxotrophic markers for genetic selection. The resulting requirement for nutrient supplementation precludes systematic testing of the yeast deletion collection on specific combinations of carbon and nitrogen sources because the auxotrophic nutrient supplements can also be used as carbon and nitrogen sources. Previous work has shown not only that nutrient supplementation can have different physiological consequences from genetic complementation [86] but also that auxotrophies can alter the expression of many other genes [87].

To address this deficiency in genome-scale data on growth in other, defined media, we constructed a prototrophic version of the yeast deletion collection and then screened this collection of 4,772 mutants against 28 defined minimal media conditions. These 28 conditions were formed by using all pairwise combinations of four carbon sources and seven nitrogen sources (Table 2.1, Fig. 2.1). These screens of the prototrophic collection revealed numerous interactions between carbon and nitrogen sources with respect to wild-type growth rate, underscoring the need to perform growth experiments in a combinatorial fashion. Mutant data revealed condition-specific sensitivities across all conditions, including many effects for uncharacterized genes and mutants that are healthy under standard laboratory conditions. We show that the data have power to predict functional relationships between genes and are otherwise validated via a separate liquid assay as well as through comparison with previous studies involving galactose. We also present a method for distinguishing carbon and nitrogen effects from their

combined profiles and additionally provide a benchmark of current constraint-based modeling techniques and their ability to predict our experimental data.

2.3 Results and Discussion

2.3.1 Prototrophic deletion set construction and profiling

Briefly, a *MATα* strain carrying the SGA marker [88, 89] was crossed to the *MAT***a** yeast deletion set [3], selected for diploids, and sporulated. Prototrophic haploids were selected using the SGA approach [88]. The final genotype of these 4,772 strains is *MAT***a** *yfgΔ0::KanMX can1Δ::STE2pr-SpHIS5 his3Δ1 lyp1Δ0*. These strains were then pinned out onto plates containing one of four different carbon sources along with one of seven nitrogen sources. All 28 carbon:nitrogen combinations were included to produce a broad set of well-defined metabolic conditions. The plates were imaged in time course in order to estimate growth rates from measurements of colony size (Fig. 2.1; see Appendix A.1 for details).

2.3.2 Yeast wild-type growth suggests carbon/nitrogen interactions

The mean growth rate of all wild-type replicates was calculated in each condition, which revealed extensive variation across the profiled conditions (Fig. 2.2; see Appendix A.2). As expected, wild-type yeast grow substantially faster on glucose or galactose than on glycerol or ribose. Similarly, urea is a consistently poor nitrogen source with glutamine

Fast / Slow	Ammonium	Proline	Glutamate	Glutamine	Arginine	Urea	Allantoin
Glucose	41.5 / 41*	186 / 417	133 / 354	169 / 286	173 / 920	135 / 219	95 / 284
Galactose	132 / 461	276 / 658	400 / 906	270 / 877	452 / 530	154 / 216	124 / 545
Ribose	312 / 345	981 / 462	306 / 412	291 / 192	437 / 46	388 / 345	379 / 492
Glycerol	NA	NA	NA	NA	NA	NA	NA

Table 2.1: Each condition is comprised of one carbon and one nitrogen source. For each combination the number of mutants which had a significant z-scores (FDR 20%) are reported here. Faster than expected growth (positive z-score) and slower than expected growth (negative z-score) are reported separately (left and right respectively). *Glucose:Ammonium numbers represent the mean count over six replicates. Growth on glycerol was too noisy to confidently determine z-scores and so sensitivity counts are not presented.

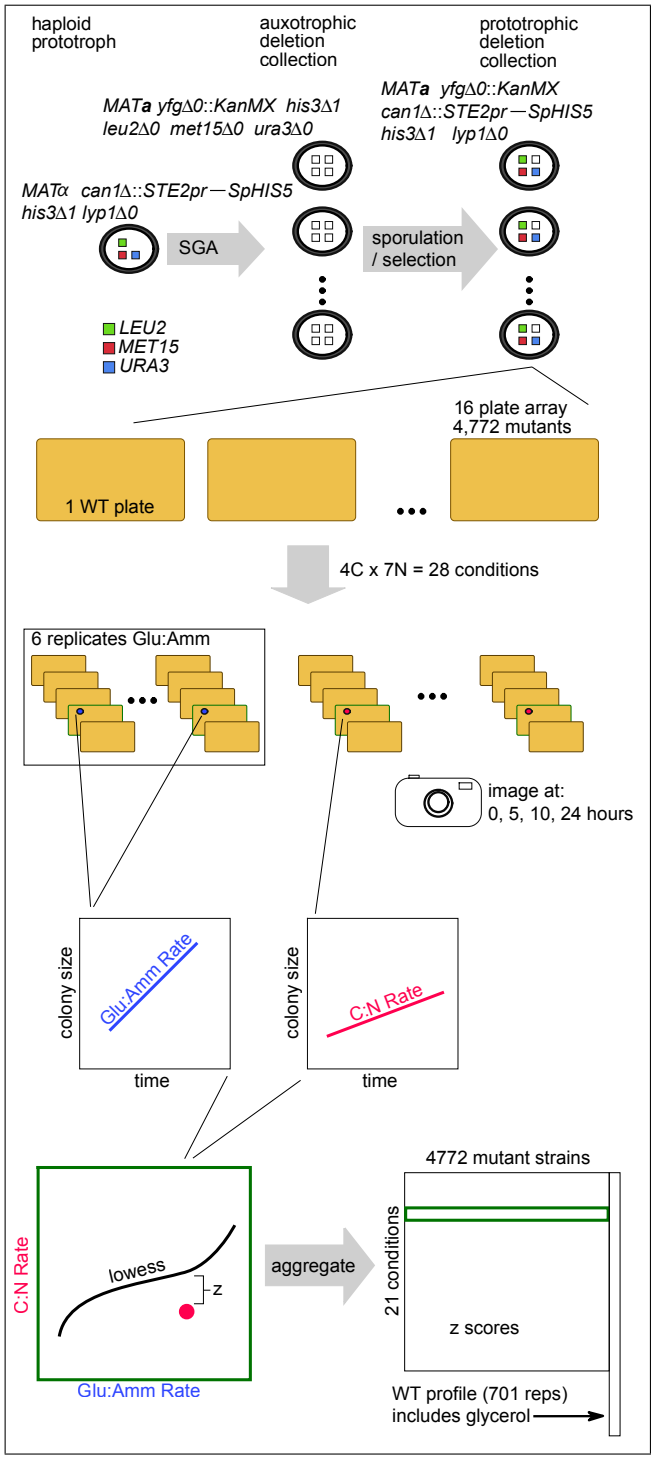

Figure 2.1: Experimental overview. A custom prototrophic strain is mated to the entire deletion collection and haploids are selected via SGA. The resulting prototrophic deletion collection is plated out onto 28 distinct metabolic media, and time course growth rate data are extracted from plate images. Growth rates are normalized to a glucose:ammonia reference (constructed from six replicates) and z-scores are calculated for each deletion, in each condition (except glycerol). WT, wild-type

and ammonium generally preferred. To systematically examine the interactions between carbon and nitrogen sources over our entire dataset, a linear model was fit to the logarithm of wild-type growth rates under the assumption that independent contributions to growth rate would combine multiplicatively (a multiplicative model fit better than simple alternatives such as an additive formulation). Indeed, the model suggests that pairs of nitrogen and carbon sources commonly interact to produce a wild-type growth rate phenotype that is different from what might be predicted assuming independent contributions, evidenced by the fact that the majority of the interaction terms in the linear model were significant (Fig. 2.3). For example, consider the apparent increase in growth rate observed under ribose:glutamate when compared to glucose:glutamate (Fig. 2.2), observable as a positive interaction between ribose and glutamate (Fig. 2.3). When paired with glucose, glutamate is the nitrogen source that yields the fourth fastest growth rate. However, when paired with a much poorer carbon source (for example, ribose or glycerol) glutamate becomes the nitrogen source that yields the fastest growth rate. This interaction is likely caused by the ability of the cell to utilize glutamate not only as a source of nitrogen, but as a secondary carbon source in the presence of a poor primary carbon source. When glutamate is de-aminated for use as a nitrogen source, alpha-ketoglutarate is produced and can be subsequently utilized for energy production via the tricarboxylic acid cycle. This dual role is not specific to glutamate. For example, glutamine is utilized in a similar manner, though the ratio of "free" carbon skeletons per nitrogen produced is less efficient (1:2 as opposed to 1:1). Despite the fact that many of the nitrogen sources share this property, we hereafter continue to refer to them simply as "nitrogen sources" for simplicity. Our results show that the wild-type growth rate can be predicted from independent contributions of carbon and nitrogen sources in only 3 of our 28 conditions (Fig. 2.3). Significant interaction terms in all but three conditions signify the complex interdependencies throughout the metabolic network, thus underscoring the importance of testing each pair of sources systematically.

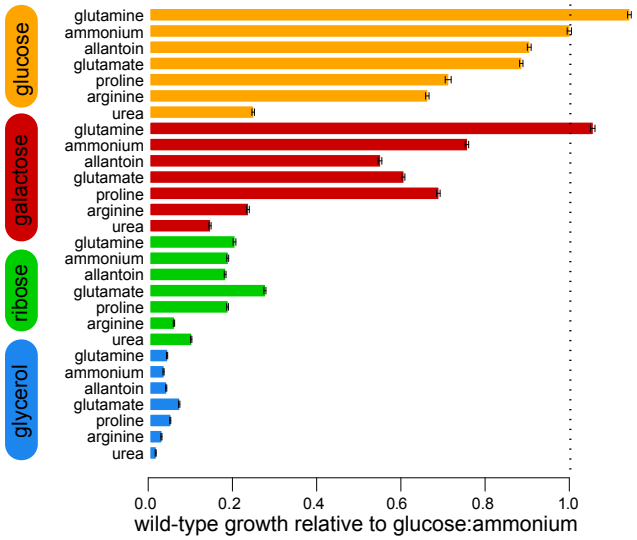

Figure 2.2:
Average wild-type growth rates in all conditions. Conditions are grouped and colored by carbon source. Nitrogen sources are ordered by growth rate when paired with glucose, and all values are relative to the glucose:ammonium rate. Error bars represent standard error with 701 wild-type replicates.

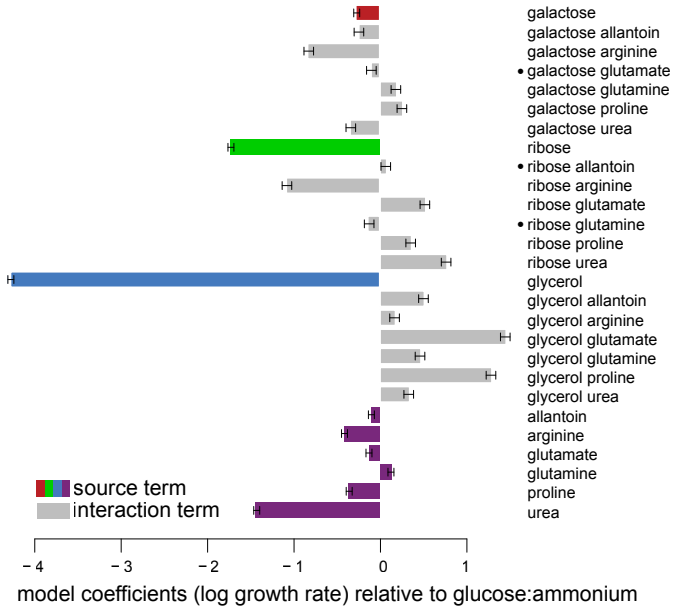

Figure 2.3: An additive linear model fit to log-transformed growth values. Terms for individual carbon and nitrogen sources are colored, interaction terms are gray. All but the three terms marked with (•) are significant ($p < 0.01$), error bars represent standard error.

2.3.3 Fitness determination of deletion mutants over the media conditions

In an effort to identify mutant growth defects specific to particular conditions, we derived a model designed to score growth rate for each deletion strain in a given condition relative to its growth under a reference condition (glucose:ammonium). First, the growth rate data were normalized from each experimental condition with respect to the glucose:ammonium reference (see Appendix A.4). This controlled for the growth rate differences observable in wild-type cells across the different conditions (Fig. 2.2) and enabled us to focus on more subtle effects due only to the genetic perturbation. A modified z-score was then calculated for each mutant strain (see Appendix A.9). This measure is negative if the strain grew slower in the test condition than would be expected due to the nutrient environment alone, and positive if the strain grew faster than expected. The distribution of growth rates in the 701 wild-type replicates was used to assess the statistical significance of mutant effects in each condition and estimate a false discovery rate (FDR) for any gene-environment interactions (see Appendix A.7). Table 2.1 shows the number of deletions that grew slower or faster than expected at an FDR threshold of 20%, and thresholds can be found in Table A.1. While the large number of wild-type replicates allowed for confidence in the small differences in reference strain growth between various nitrogen sources when paired with glycerol, the mutant data on glycerol proved to be too noisy due to extremely slow growth to call mutant effects. Therefore, no growth rate (z-score) data are presented for mutant strains on glycerol.

2.3.4 Observations in galactose concur with previous auxotrophic studies

To build additional confidence in our high-throughput dataset, we compared lists of mutants deficient for growth under galactose to data from several previous studies which had tested the auxotrophic deletion collection in a variety of experimental conditions. Giaever *et al.* [3]; Kuepfer *et al.* [90]; and Dudley *et al.* [58] each included a condition in which galactose is the major source of carbon, and the overlap between the deletions that we call as effects in our galactose conditions and sensitivities collected from these

three experiments is highly significant (Fig. 2.4; Table 2.2). We define a galactose-sensitive gene for this purpose as having a significant fitness defect in at least four of our seven galactose conditions and we obtain a list of 565 such genes (using FDR 20%; Table A.1). This list covers approximately 50% of the sensitive genes identified in each of the three previous auxotrophic screens (Giaever n=23, $p < 1 \times 10^{-11}$; Kuepfer n=120, $p < 2 \times 10^{-16}$; Dudley n=16, $p < 1 \times 10^{-6}$; hypergeometric; Fig. 2.4; Table 2.2). Additionally, we discover 385 mutants sensitive under galactose not revealed in any of these previous studies. For comparison, the overlap between two of the previous genome-wide studies (Giaever *et al.* and Kuepfer *et al.*) was only 15 genes, 12 of which are recovered in this study (Fig. 2.4). We suggest two primary reasons for the increased number of galactose sensitive mutants discovered in our study. The first is that 47% of these new galactose sensitive genes did not have a phenotype when the standard laboratory nitrogen source (ammonium) was used. Thus, the testing of a wide-range of nitrogen sources revealed additional galactose sensitive mutants. The second reason is that previous studies used more stringent thresholds for galactose phenotypes. Smaller quantitative measurements of fitness defects across multiple galactose:nitrogen source combinations allow for increased sensitivity in detecting galactose phenotypes compared with other studies.

Another possible explanation for differences between our galactose results and those from the Dudley study is the absence of antimycin A in our media. Antimycin A inhibits energy production from respiratory pathways and forces the strains to ferment galactose. In our experiments, yeast had access to oxygen and could perform both respiration and fermentation with galactose as carbon source, which is the natural metabolism of galactose by *S. cerevisiae* [91].

2.3.5 Liquid validation of mutant fitness measurements

We independently validated our single-mutant fitness measurements by measuring the growth rate of 40 mutants in a liquid growth assay performed across 20 of the experimental conditions (excluding ribose:arginine and all glycerol pairings, see Appendix A.10). The overall correlation between wild-type strain growth rates from these two different approaches was 0.65 ($p < 0.003$; Pearson), suggesting general agreement between growth rates determined on solid and liquid media. We then adjusted the liquid

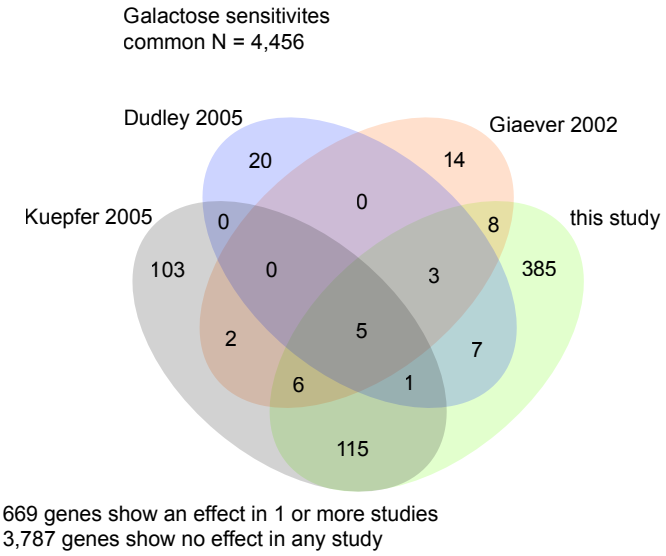

Figure 2.4: Overlap between mutants sensitive on galactose from several different studies. For this study, galactose sensitivity is defined as a significant z-score in four or more of our seven galactose conditions. In each case, N denotes the total number of genes the studies have in common. Venn diagram image is used with permission under the terms of the Creative Commons Attribution-Share Alike 3.0 Unported license [92].

Study	genes	gal hits	common genes	common hits	n1	n2	p-value
Giaever 2002	4743	42	4554	23	544	41	9.73×10^{-12}
Kuepfer 2005	4856	311	4585	135	552	255	$< 1.11 \times 10^{-16}$
Dudley 2005	4992	41	4463	16	530	36	9.39×10^{-7}

Table 2.2: Overlap with previous whole-genome galactose sinsitivity screens. Comparison with three previous studies is shown, where hits in our study are comprised of genes sensitive in at least 4 out of 7 galactose conditions. "genes" denotes the total number of genes in the relevent external study, "n1" denotes the number of our "hits" which were tested in the external study, and "n2" denotes the reverse. These values are used to compute significance of the overlap according to the hypergeometric cummulative distribution (p-value column).

growth scores, controlling for the wild-type rate in the given condition and the relevant mutant rate in glucose:ammonium so they would reflect condition-specific effects, similar to our modified z-score derived from the agar experiment. The Spearman rank correlation between the adjusted liquid growth score and our agar z-score (for 40 mutant strains $\times$ 19 conditions) was 0.34 ($p < 2.2 \times 10^{-16}$). Further excluding glucose conditions (which are generally sparser in the z-score data as a consequence of our use of glucose:ammonium as a reference) increases this correlation to 0.38. Thus, we conclude that there is reasonable agreement between the high-throughput measures and a lower-throughput liquid growth assay, including for condition-specific effects.

2.3.6 Number of environmental sensitivities is correlated with single-mutant fitness and genetic interaction degree

We compared our growth measurements with other quantitative phenotypes measured on the auxotrophic deletion collection. For example, genetic interaction mapping efforts have measured the single-mutant fitness of all deletion strains from the auxotrophic background on minimal complete media [50, 66] and found a correlation between the magnitude of the fitness defect and the number of genetic interactions for each single mutant (genetic interaction degree). The prevailing explanation for this correlation is that genes that display a fitness defect represent the subset that are playing an active role under the condition tested, are additionally not completely buffered by other genes, and/or contribute to a wider variety of cellular processes. We observe a similar correlation between the single-mutant fitness defect (as previously measured on minimal complete media [50]) and the number of significant condition-specific sensitivities in our study ($r = 0.33, p < 5 \times 10^{-100}$; Pearson). Additionally, there is a partial correlation between the number of genetic interactions a gene has and the number of environments with which it interacts, even after controlling for single-mutant fitness defect ($r = 0.18, p < 5 \times 10^{-31}$; Pearson). This echoes a previously observed correlation between genetic interaction degree and sensitivities in more complex chemical environments ($r = 0.4, p < 1 \times 10^{-5}$) [56, 66]. These results confirm that our study is uncovering more effects for genes known to be pleiotropic or central under a variety of environmental backgrounds [58]. These findings also suggest that hubs are conserved across different network types, with many of the same genes conferring robustness to genetic, chemical,

and environmental perturbations.

2.3.7 Mutant sensitivity profiles are predictive of gene function

Previous genetic interaction studies have shown that high profile similarity for mutant sensitivity across varied environmental conditions or diverse genetic backgrounds (for example, genetic interaction profiles) is highly predictive of similar gene function [58, 66, 57]. We applied an analogous logic to our data to see if similar environmental sensitivity profiles would also be predictive of similar function. Using co-annotation to an informative set of Gene Ontology (GO) terms [93, 94] as our standard for functional similarity, we ranked all pairs of genes by their profile similarity (Pearson) and evaluated these rankings with respect to known functional relationships. We measured a precision of approximately 35% at a recall of 1,000 gene pairs (2-fold over a random baseline of 17%; Fig. 2.5). Additionally, when we restrict our predictions to those genes with a known involvement in metabolism (663) we see a much higher precision (precision $\sim$ 65% at recall = 100), though a similar performance over the increased background rate (1.7-fold over 38%). The higher performance for metabolism related predictions is likely due to the direct relevance of the environmental conditions chosen to the study of basic metabolism. Thus, we have demonstrated an ability to predict general gene function using the guilt-by-association principle, and the diverse yet minimal environments chosen for this assay are well-suited to reveal sensitivities in the metabolic network of this newly created prototrophic collection.

2.3.8 Metabolic network models show modest ability to predict experimental data

The prototroph growth data on minimal media presented here are uniquely suited to bring experimental data to bear on theoretical predictions of constraint-based analysis of metabolic networks. Constraint-based modeling is a widely used approach to study the metabolic capacity of genome-scale biochemical networks in steady state without requiring detailed enzyme kinetic parameters [85]. FBA is the most popular constraint-based approach to computationally predict the phenotypes under environmental and genetic perturbations and has been shown to successfully predict gene essentiality, and

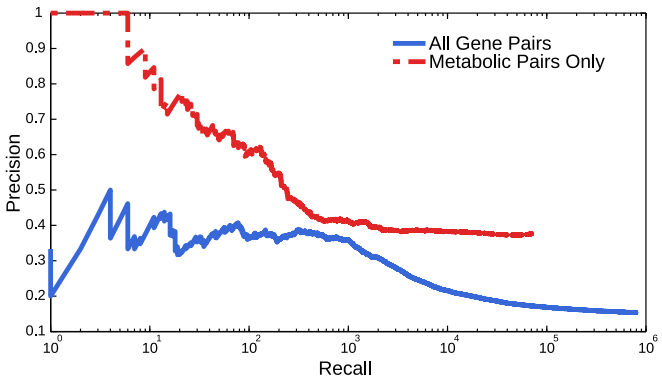

Figure 2.5: Precision-Recall analysis assessing the ability of gene-gene similarity to predict co-annotation to an informative term in the Gene Ontology. Results for all gene pairs are shown in blue, and results for a subset of metabolism related genes (included in iMM904 model) is shown in red.

to a lesser extent, condition-specific essential status in yeast [90, 95]. We used our sensitivity data to evaluate the ability of constraint-based models to predict subtler quantitative sensitivities in a condition-specific manner. We predicted biomass yield, a proxy for growth, in all conditions using two versions of the yeast metabolic network reconstruction: the more recent Sourceforge Yeast Consensus Reconstruction v5.35 (hereafter Yeast5) [96], and iMM904 [97]. Additionally, we applied two alternative algorithms to predict mutant phenotypes, namely standard FBA [98] and minimization of metabolic adjustment (MoMA) [99]. Predicted biomass production fluxes were normalized with respect to every mutant's predicted biomass production in glucose:ammonium and the wild-type prediction in each condition to make scores analogous to our experimental z-scores. The prediction of z-scores as opposed to raw growth rates was chosen to assess the adaptability of each model's performance in the face of varied environments, an admittedly more difficult scenario than predicting global or condition-specific essentiality. Though the output of the models is quantitative, many conditions predict only a few discrete levels of resulting biomass production and therefore yield identical predictions for the majority of mutants. The mode of the output accounted for between 39% and 95% of the predictions, so we assessed model performance by comparing the predicted set of slow mutants (below the mode biomass production) to our set of significant z-scores in each condition. Three metrics were collected to assess the performance of

each model-method combination: average precision (across all 20 predicted conditions), average recall, and the number of conditions in which precision exceeded random expectation (at $p < 0.05$ hypergeometric; Fig. 2.6; Tables 2.3–2.4). Results for positive z-score prediction (above the mode biomass) are also available in Tables 2.5–2.6 (see Appendix A.12).

Prediction of condition-specific slow growth proved consistently above random expectation (Fig. 2.6), though values of precision are much lower than those previously reported in predicting qualitative essentiality (>90% [95]). One key difference between our study and Snitkin *et al.* [95] (as with Dudley *et al.* [58] in the section on galactose sensitivity above) is the latter's inclusion of antimycin A in the media, which inhibits energy production from respiration, whereas our strains could naturally respire and ferment. Our results show an advantage for the more recent Yeast5 model over the iMM904 model, as well as a slight advantage for standard FBA over MoMA. The Yeast5 model was able to perform above random expectation in 14 out of 20 conditions with a mean precision of 25% and a mean recall of 18% (Fig. 2.6; Table 2.4). Recall scores for MoMA were generally higher than for FBA owing to a much smaller fraction of the predictions equal to the mode, though this was generally associated with a loss of precision. Galactose conditions appear to be well captured by the two models, and consistently perform above random. By contrast, all three conditions for which no model-method achieved significance involved glucose (glucose:allantoin, glucose:glutamine, glucose:urea). Thus, while the overall performance demonstrates an above-random ability of these models to predict quantitative and condition-specific perturbation effects, their modest precision and recall scores (< 50%) suggest substantial room for improvement.

An examination of false positives (predicted sensitive by the model but not observed in the data) and false negatives (observed sensitive, not predicted) showed some functional coherency. Specifically, Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment of false positives in many conditions revealed connections to central carbon metabolism (for example, the tricarboxylic acid cycle), and half of the conditions showed enrichment for the KEGG sulfur metabolism pathway in the model for false positives. This suggests potential pathways that may need attention for the development of improved models.

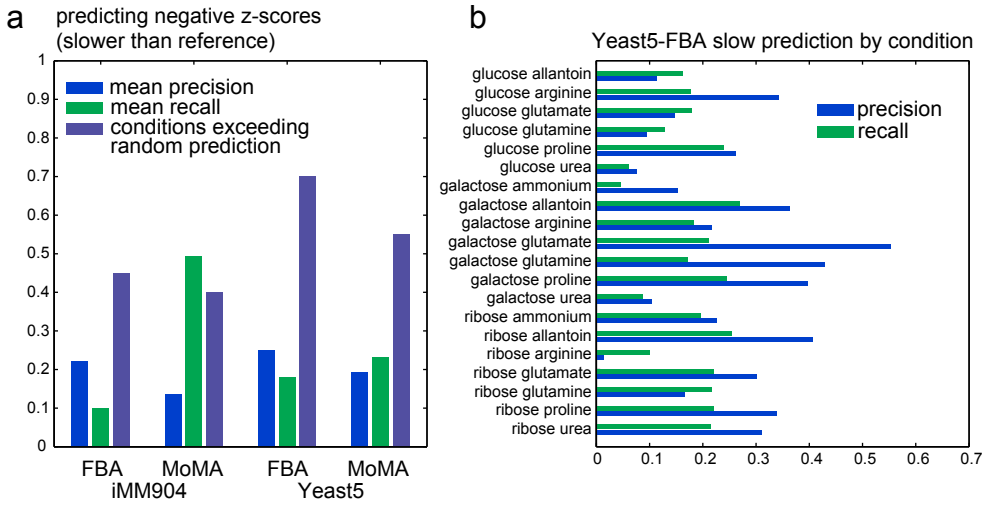

Figure 2.6:
(a) Assessment of constraint-based modeling predictions for slow growth. Precision and recall (blue and green) were calculated for each model in each of 20 conditions with glycerol:* and glucose:ammonium excluded—means are shown here. The fraction of conditions in which predicted model mutants overlap significantly with significant z-score effects is shown in purple. (See Tables 2.3, 2.4)
(b) Precision and recall scores as in **(a)** for each individual condition using the Sourceforge version 5.35 and standard FBA. (See Table 2.4)

We also attempted to leverage existing metabolic models to demonstrate the widespread metabolic consequences of these common auxotrophies. To accomplish this, we ran the models again using prototrophic and auxotrophic versions of the network on glucose:ammonium and characterized each metabolite as either: i) produced in the auxotroph and the prototroph; ii) produced in the prototroph only; or iii) included in the model but not produced in an optimal solution (see Appendix A.12). The simulations show that a significant proportion of producible metabolites (18% in iMM904 and 7% in Yeast5; Fig. 2.7) are unavailable in the auxotrophic network. This means that consequences of using auxotrophic strains, even under supplementation for their specific deficiencies, may have a broader impact than expected. It is our hope that the collection and accompanying growth data presented here will prove invaluable to the metabolic modeling community as it continues to refine the structure of its models as well as their underlying biological assumptions.

condition/model	imm904FBA-P	imm904FBA-R	imm904MOMA-P	imm904MOMA-R
glucose allantoin	0.073	0.059	0.093	0.49
glucose arginine	0.25	0.119	0.207	0.5
glucose glutamate	0.154	0.039	**0.104**	0.647
glucose glutamine	0.061	0.079	0.076	0.5
glucose proline	0.118	0.029	**0.15**	0.75
glucose urea	0.053	0.065	0.056	0.548
galactose ammonium	**0.333**	0.057	0.13	0.149
galactose allantoin	**0.318**	0.135	**0.185**	0.538
galactose arginine	**0.295**	0.217	0.114	0.422
galactose glutamate	**0.438**	0.048	**0.29**	0.32
galactose glutamine	**0.554**	0.221	0.23	0.436
galactose proline	**0.333**	0.068	**0.189**	0.602
galactose urea	**0.205**	0.191	0.073	0.468
ribose ammonium	0.154	0.051	0.094	0.114
ribose allantoin	**0.293**	0.109	0.177	0.482
ribose arginine	0.016	0.1	**0.026**	0.8
ribose glutamate	0.143	0.023	**0.164**	0.591
ribose glutamine	**0.179**	0.222	0.074	0.467
ribose proline	0.294	0.054	**0.168**	0.591
ribose urea	0.171	0.106	0.097	0.439
mean	0.22175	0.0996	0.13485	0.4927

Table 2.3: Performance of iMM904 when predicting slow-growth mutants. In each condition the table gives the precision (P) and recall (R) for *in silico* mutant preditions with below-the-mode flux using significant experimental z-scores as the standard for true positives. Precision scores for prediction scenarios which performed better than random expectation (hyper-geometric $p < 0.05$) are highlighted in **bold**.

condition/model	yeast535FBA-P	yeast535FBA-R	yeast535MOMA-P	yeast535MOMA-R
glucose allantoin	0.114	0.163	0.108	0.143
glucose arginine	**0.343**	0.178	0.278	0.078
glucose glutamate	**0.148**	0.18	0.138	0.18
glucose glutamine	0.094	0.128	0.075	0.103
glucose proline	**0.262**	0.239	**0.221**	0.254
glucose urea	0.077	0.061	0.058	0.182
galactose ammonium	0.154	0.046	**0.214**	0.172
galactose allantoin	**0.364**	0.269	**0.26**	0.183
galactose arginine	**0.217**	0.183	0.22	0.11
galactose glutamate	**0.554**	0.212	**0.472**	0.233
galactose glutamine	**0.429**	0.171	**0.317**	0.143
galactose proline	**0.397**	0.245	**0.302**	0.284
galactose urea	0.105	0.087	**0.095**	0.609
ribose ammonium	**0.227**	0.197	**0.149**	0.632
ribose allantoin	**0.406**	0.255	**0.3**	0.191
ribose arginine	0.015	0.1	**0.071**	0.3
ribose glutamate	**0.302**	0.221	0.19	0.128
ribose glutamine	**0.167**	0.217	0.088	0.109
ribose proline	**0.339**	0.22	**0.21**	0.516
ribose urea	**0.311**	0.215	0.074	0.092
mean	0.25125	0.17935	0.192	0.2321

Table 2.4: Performance of Sourceforge 5.35 when predicting slow-growth mutants. In each condition the table gives the precision (P) and recall (R) for *in silico* mutant predictions with below-the-mode flux using significant experimental z-scores as the standard for true positives. Precision scores for prediction scenarios which performed better than random expectation (hyper-geometric $p < 0.05$) are highlighted in **bold**.

condition/model	imm904FBA-P	imm904FBA-R	imm904MOMA-P	imm904MOMA-R
glucose allantoin	0	0	0.019	0.118
glucose arginine	**0.333**	0.069	0.055	0.138
glucose glutamate	0	0	0.02	0.059
glucose glutamine	0	0	0.016	0.091
glucose proline	0	0	0	0
glucose urea	0	0	0.028	0.105
galactose ammonium	0.167	0.034	0.059	0.138
galactose allantoin	0	0	0.027	0.111
galactose arginine	0.125	0.018	0.087	0.109
galactose glutamate	0	0	0.103	0.355
galactose glutamine	0	0	0.054	0.182
galactose proline	0	0	0.083	0.098
galactose urea	0	0	0.039	0.167
ribose ammonium	0	0	0.02	0.069
ribose allantoin	0.125	0.02	0.104	0.157
ribose arginine	0.333	0.027	0.068	0.068
ribose glutamate	0.118	0.047	0.091	0.116
ribose glutamine	0.333	0.03	0.042	0.121
ribose proline	0.25	0.035	0.143	0.049
ribose urea	0	0	0.064	0.086
mean	0.0892	0.014	0.0561	0.11685

Table 2.5: Performance of iMM904 when predicting fast-growth mutants. In each condition the table gives the precision (P) and recall (R) for *in silico* mutant preditions with above-the-mode flux using significant experimental z-scores as the standard for true positives. Precision scores for prediction scenarios which performed better than random expectation (hyper-geometric $p < 0.05$) are highlighted in **bold**.

condition/model	yeast535FBA-P	yeast535FBA-R	yeast535MOMA-P	yeast535MOMA-R
glucose allantoin	0.087	0.118	0.028	0.529
glucose arginine	0.136	0.107	0.042	0.536
glucose glutamate	0.042	0.053	0.025	0.421
glucose glutamine	0.094	0.125	0.041	0.583
glucose proline	0.04	0.04	0.032	0.4
glucose urea	0.018	0.05	0.031	0.45
galactose ammonium	0.125	0.033	0.068	0.167
galactose allantoin	0.105	0.133	0.028	0.6
galactose arginine	0.13	0.055	0.091	0.582
galactose glutamate	**0.25**	0.136	0.088	0.475
galactose glutamine	**0.156**	0.143	0.064	0.6
galactose proline	0.148	0.095	0.071	0.5
galactose urea	0.021	0.062	0.031	0.188
ribose ammonium	0.067	0.032	**0.116**	0.258
ribose allantoin	0.125	0.061	0.069	0.449
ribose arginine	0.13	0.041	0.12	0.575
ribose glutamate	0.125	0.068	0.066	0.5
ribose glutamine	0.115	0.091	0.042	0.424
ribose proline	0.321	0.064	0.198	0.234
ribose urea	0.053	0.038	0.081	0.472
mean	0.1144	0.07725	0.0666	0.44715

Table 2.6: Performance of Sourceforge 5.35 when predicting fast-growth mutants. In each condition the table gives the precision (P) and recall (R) for *in silico* mutant predictions with above-the-mode flux using significant experimental z-scores as the standard for true positives. Precision scores for prediction scenarios which performed better than random expectation (hyper-geometric $p < 0.05$) are highlighted in **bold**.

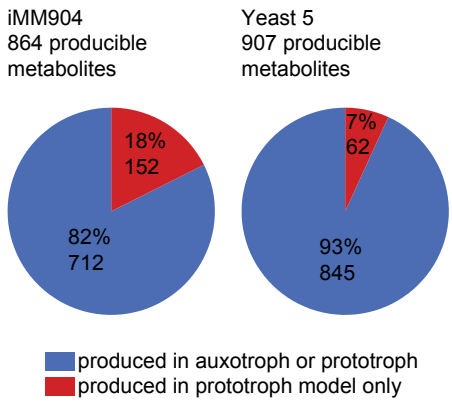

Figure 2.7:
Number of producible metabolites for iMM905 and Yeast5 metabolic models. For each model the total number of producible metabolites was counted based on simulation in glucose:ammonium (see Appendix A.12). The procedure was repeated for a model in which reactions involving auxotrophic marker genes (*HIS3, URA3, LEU2,* and *MET15*) were disabled. The chart shows the proportion of metabolites that the auxotrophic model fails to produce in red.

2.3.9 Broad environmental surveys address incomplete gene annotations

A primary motivation for measuring fitness across diverse environments is the discovery of novel phenotypes for mutants that have near wild-type fitness under previously tested conditions. The existence of such mutants in a eukaryotic genome with approximately 6,000 genes is driven by two main factors. The first is genetic redundancy, whereby genes are performing vital functions within the cell, but their importance is not captured by single-mutant fitness because other genes are present that buffer the loss of function. This occurs at both the level of individual genes buffering one another (e.g. duplicate genes [78, 100]) and at the level of larger network structures (for example, parallel pathways). These buffered functions are rapidly being mapped by genetic interaction studies that delete multiple genes simultaneously [88, 83, 101, 84, 66, 102]. The remaining contributing factor is environmental robustness, whereby a gene presumably has an important function under some evolutionarily relevant circumstance that is not reflected in a laboratory environment (for example, nutrients/media, temperature, stress etc.). Thus, an important motivation for complete pairwise coverage of

basic metabolic conditions is the detection of novel fitness defects for genes that become necessary only as the condition space is more broadly surveyed. Interestingly, of the 729 remaining uncharacterized mutants in the auxotrophic collection for which we have single-mutant fitness measurements in synthetic complete media, a significant fraction of them (609) have a fitness greater than 99% of wild-type (hypergeometric $p < 7 \times 10^{-66}$) [59]. Despite the ever-increasing availability of high-throughput genomic data for these genes, the task of eliminating this set has seen only marginal success since 2007 [51]. It is possible that these genes (many of which only have orthologs in other yeasts) may be responsible for functions needed in the native environment of yeast but unnecessary under standard laboratory conditions. Still others may be required in the lab, but only after varying the nutrient conditions. The focus of recent chemical genomics work on subjecting yeast to an extremely broad range of chemical environments is helping to address these genes [57, 56], but auxotrophy in the deletion collection had precluded measurements of growth on simple but directly relevant metabolic conditions. Here we address the potential impact of these data on both uncharacterized genes and genes of little phenotypic consequence in standard conditions.

2.3.10 Novel effects for genes with high fitness in standard conditions

As described earlier, we observed that the number of significant effects in our data can be weakly predicted by single-mutant fitness in synthetic complete media. However, nearly 40% of the *S. cerevisiae* genome shows little to no such effect. Of the genes in this study with single-mutant fitness scores greater than 99% of wild-type under synthetic complete media, more than 50% of them (1548/2745; Fig. 2.8) show at least one significant slow-growth effect outside of glucose:ammonium. Multiple random assignments of the number of expected false positives (20% of effect counts listed in Table 1), demonstrate that only approximately 30% of genes should show an effect. Additionally, 5% (142/2745) show significant effects in five or more distinct non-glucose:ammonium conditions compared to a random expectation of 2.6×10^{-5} ($\ll 1/2745$). For example, *prs2Δ0* (the *PRS2* gene encodes one of the four phosphoribosyl-pyrophosphate (PRPP) synthetases encoded in the genome; these synthetases are required for nucleotide, histidine, and tryptophan biosynthesis); has a single-mutant fitness of 1.02 in

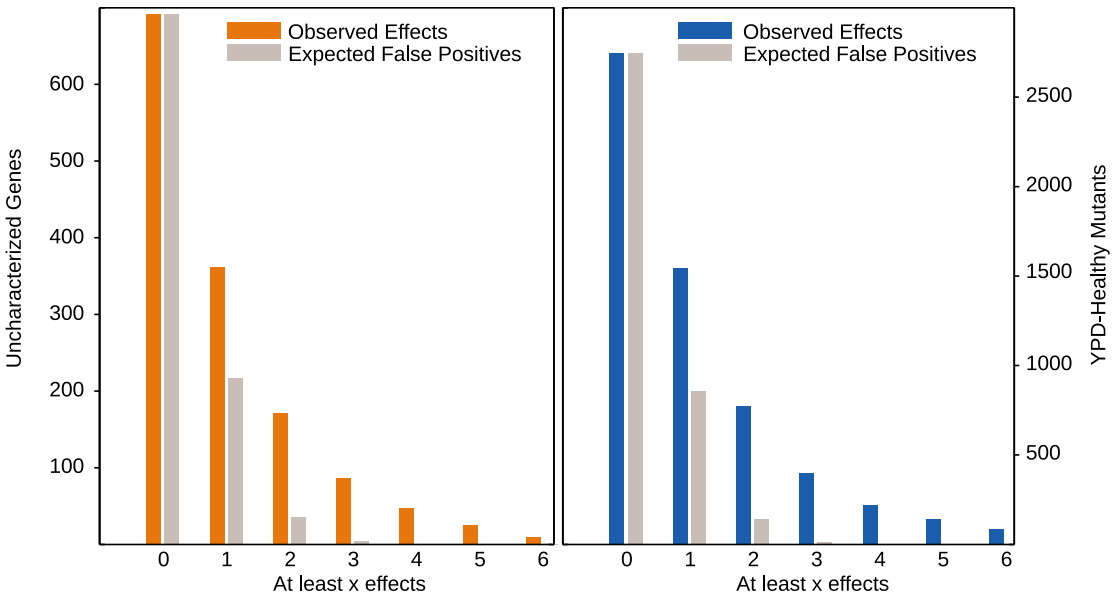

Figure 2.8: Counting effects for under-characterized genes. Histograms show the total number of mutants with at least x effects in our data from the set of uncharacterized genes (left, orange), and genes with little to no fitness defect on synthetic complete media (right, blue; single-mutant fitness $>$ 0.98% of wild-type). As a control, the expected number of false positives (20% of significant effects in each condition) were randomly distributed among all genes, and the number of effects for each gene was counted again. Gray bars show the mean of 1,000 such randomizations.

synthetic complete media [50] but shows significant growth defects in 14 different conditions. These conditions are highly coherent, including all seven galactose conditions, all ribose conditions (except ribose:arginine) and no conditions involving glucose except glucose:proline. *PRS2* is highly expressed under fermentative condititions [103]. Another example is *ICL1*, which facilitates a key reaction of the glyoxylate cycle, and shows slow growth effects in nine (non-glucose:ammonium) conditions despite a single-mutant fitness score slightly greater than that of wild-type under standard lab conditions (1.03) [50].

2.3.11 Novel phenotypes for uncharacterized ORFs

Approximately 13% of the *S. cerevisiae* deletion collection is composed of uncharacterized ORFs [59], 692 of which are included in this study. Nearly 25% of these uncharacterized genes show a significant effect in two or more non-glucose:ammonia conditions (172/692; Fig. 2.8) compared to the 4% expected given our FDR.

One such example with a very specific nitrogen sensitivity signature is *FMP32*. The *fmp32*Δ*0* strain displays dramatically decreased fitness under arginine and proline conditions. While the protein product of *FMP32* has been detected in highly purified mitochondria [104], the gene is otherwise uncharacterized. The *fmp32*Δ*0* strain was included in our liquid confirmation assay and these sensitivities were confirmed in this independent, small-scale assay (Fig. 2.9). This highly specific signature appears to be completely unique to the *fmp32*Δ*0* strain, as no other mutant in the collection shows a similar sensitivity profile.

The genes with the highest profile similarity to *FMP32* are *PUT1*, *PUT3*, and *RRF1* which have been previously implicated in proline utilization (*PUT1*, *PUT3*) [105] and mitochondrial ribosome recycling/mitochondrial protein synthesis during respiration (*RRF1*) [106, 107]. *PUT3* induces *PUT1* transcription when proline is present as the best available nitrogen source and the latter (along with *PUT2*) is responsible for the conversion of proline into glutamate for further use as a nitrogen source. Our analysis suggests that *FMP32* is similarly involved in the respiratory response under proline, though the reason for its additional sensitivity under arginine remains unclear. These examples show the utility of interactions between genes and simple environments in uncovering the function of both individual uncharacterized genes and genes without a previously observed fitness defect in more complete media.

2.3.12 Clustering of metabolic conditions reveals carbon source as primary factor driving mutant profiles

Just as gene-gene correlation predicts functional similarities, we expect a high correlation between condition pairs to reflect a substantial overlap in the cellular machinery required to utilize the provided carbon and nitrogen sources. When our matrix of z-scores is hierarchically clustered in both the gene and condition dimensions, a structure

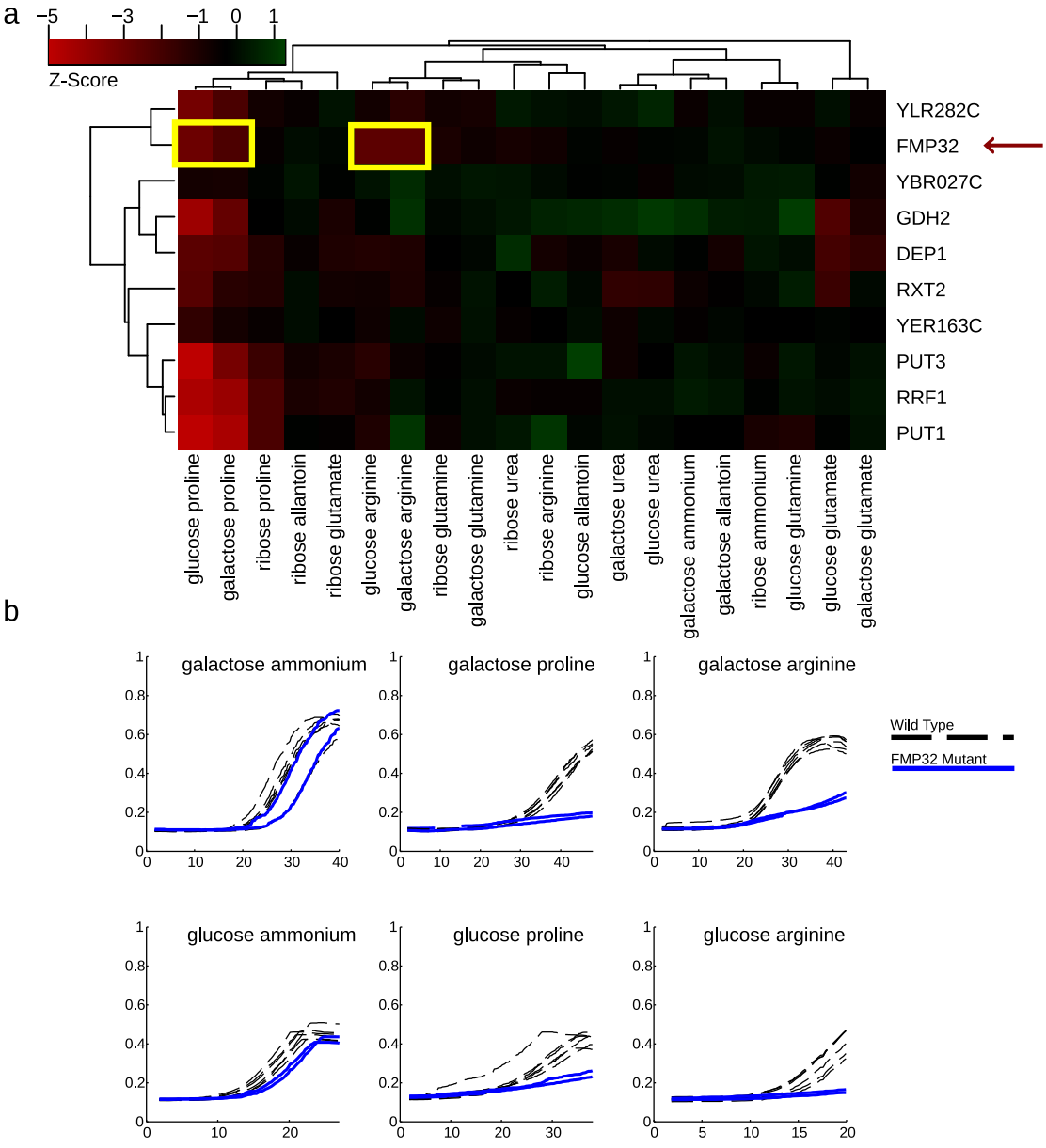

Figure 2.9: **(a)** z-score data show specific growth defects for the uncharacterized gene *FMP32* when grown on proline or arginine. **(b)** Liquid growth confirmations for effects highlighted in **(a)**. Two replicates of *FMP32* mutants are shown (blue line) along with six replicates of a wild-type strain (black dashed line) in two proline and two arginine conditions. The effects are pronounced when compared to observations in similar ammonium conditions.

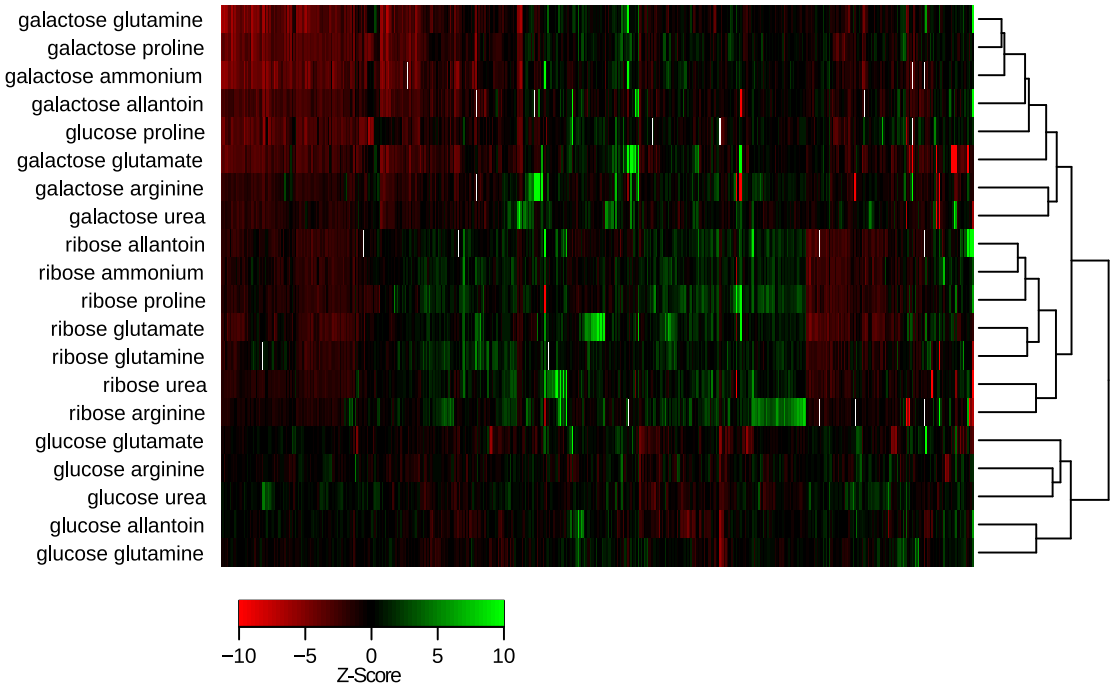

Figure 2.10: A clustergram of z-scores for the 500 mutants with the highest variance. The data have been hierarchically clustered in both dimensions. Conditions organize themselves primarily by carbon source, falling into three distinct clusters.

clearly driven by carbon sources emerges (Fig. 2.10). All of the glucose conditions cluster together, as do both the galactose and ribose conditions. The sole exception to this is glucose:proline, which falls in the galactose cluster. We attribute this observation to the fact that the utilization of proline as a nitrogen source requires some respiration. The glucose:proline signature reveals sensitivity in a number of respiratory deficient mutants, which is atypical for glucose conditions in general since fermentation is generally preferred over respiration when cells are grown on glucose. This respiration-dependent signature is strong enough to place the glucose:proline profile in the galactose cluster where one would expect a modest profile contribution from both respiration and fermentation related processes (Fig. 2.10), as is observed in growth on galactose [91].

2.3.13 Matrix factorization distinguishes carbon from nitrogen effects

Further examination of gene and environmental profiles after clustering revealed cases where a gene (for example, *FMP32*) exhibited an effect in multiple instances of a particular nitrogen source (for example, proline or arginine), but without a specific pattern with regard to carbon source (or vice versa). This is expected behavior for genes required for the utilization of a particular carbon/nitrogen source regardless of the context. In order to more formally extract a list of sensitivities for each source of carbon or nitrogen regardless of its partner, we employed a method known as Non-negative Matrix Factorization (NMF) [108, 109] to decompose our experimental data into a collection of characteristic source signatures. When a matrix of these source signatures is multiplied by a matrix describing the source composition in each of our conditions, the result should approximate our experimental observations. NMF allows us to run this multiplication in reverse and fit the source signatures as an unknown factor. Many of these source signatures demonstrate enrichment for related GO terms and KEGG pathways.

One example of a decomposed signature involves genes that are sensitive when glutamate is chosen as a nitrogen source. These genes are enriched for annotations relating to endocytosis, endosome and vacuole related transport, and retrograde transport ("GO:0007034 - vacuolar transport" $p < 3 \times 10^{-7}$, "GO:0016192 - vesicle-mediated transport" $p < 5 \times 10^{-8}$, "GO:0016197 - endosome transport" $p < 2 \times 10^{-7}$). Extracellular glutamate decreases cellular amino acid permease activity by redirecting intracellular trafficking of the permease Gap1 from the plasma membrane to the vacuolar membrane [110]. Many of the mutations in our glutamate signature increase Gap1 activity by misdirecting the protein to the plasma membrane [111]. Although *GAP1* is transcribed at equal levels in cells grown on urea and glutamate, permease activity in urea grown cells is 100 times higher than glutamate-grown cells [112]. Inappropriate Gap1 activity is toxic in the context of high concentrations of single amino acids [113], and we speculate that the inappropriate trafficking in these mutants causes high levels of permease activity that inhibit cell growth.

Many mutants (92) appear in both the galactose and ribose signatures, and overlapping GO enrichments in these conditions reveal many of these genes to have known involvement in various aspects of respiration. For example, enrichment for GO terms relating to mitochondrial organization and translation, as well as "aerobic respiration"

appear highly significant in both of these signatures (galactose $p < 4 \times 10^{-6}$, ribose $p < 7^{-11}$). Exceptions include GAL pathway mutants that fall uniquely into the galactose carbon signature ("galactose metabolic process" $p < 1.3 \times 10^{-4}$) and genes involved in acetyl-CoA biosynthesis that appear to be specifically sensitive under ribose ($p < 1.4 \times 10^{-6}$). As more complex environments are mapped, multivariate statistical techniques will become increasingly important in determining which environmental constituents are actually relevant to which experimental observations, and care should be taken when designing experiments to ensure their successful application (for example, complete combinatorial coverage of relevant environmental factors).

2.3.14 Environmental and genetic perturbations can provoke similar cellular states

Beginning to test the immense space of possible environmental and chemical conditions combined with experiments that have queried the space of genetic perturbations [66] allows us to investigate how these spaces interrelate. For example, if mappings can be found between them, we can apply knowledge from the already extensively mapped genetic perturbation networks to the intractable space of environmental variation. While the sensitivity profile for a given condition most certainly includes genes directly required for the processing of the provided raw materials (for example, the galactose metabolism pathway under galactose conditions), it also contains information about genes that, though not directly involved, are nonetheless indirectly required for optimal cell growth. These profiles then reveal much more than the functions of genes for which we measure a fitness defect, and in fact, give us a high dimensional fingerprint of the internal cellular state. We propose that genetic perturbations may put the cell into a very similar state as would an alteration of the environment. For example, the deletion of a gene that encodes a transporter may exhibit a profile that mimics the wild-type profile in an environment where the corresponding substrate is absent. Downstream consequences of the environment or genetic perturbation may cause subtle and seemingly unexpected sensitivities. Thus, genetic perturbation experiments and environmental perturbation experiments may both result in the same phenotypic profile. A similar principle has been demonstrated through the observation that deletion mutants with similar double-mutant sensitivity profiles tend to be functionally related

[66]. Parsons *et al.* [57] first applied this principle to predict drug targets, reasoning that a genetic sensitivity profile on a chemical that targets an individual gene would be similar to a sensitivity profile of a strain with the corresponding gene deleted. When we compared sensitivity profiles from our condition experiments to that of query-deletions crossed into the auxotrophic deletion collection via SGA [66], we found several interesting cases where genetic perturbation profiles significantly overlapped with sensitivity profiles from our environmental perturbations (see Appendix A.14). For example, the queries in the top 10% in terms of similarity to galactose:urea are enriched for members of the threonine and methionine biosynthesis pathway (*hom2*, *hom3*, *hom6*, *thr4*; Fig. 2.11; GO:0006566 - "threonine metabolic process" $p < 4.5 \times 10^{-2}$; KEGG: "glycine, serine and threonine metabolism" $p < 2.9 \times 10^{-2}$). The strength and specificity of this similarity is not driven by a handful of mutants in the collection, but instead by trends across a much larger set of genes. We speculate that the profile similarity in this case may be due to accumulation of aspartate, which is upstream of homoserine and threonine biosynthesis, and is excreted in part through urea production. Growth on urea in the setting of the respiratory growth of galactose may result in the accumulation of aspartate.

The idea of comparing environmental and genetic perturbations can be generalized to other genome-wide perturbation data as well. For example, we observe significant correlations between our glutamate signature and a rapamycin sensitivity profile as measured by two different chemical genomic screens (Hillenmyer *et al.* $p < 1 \times 10^{-18}$ [56] Parsons *et al.* $p < 1 \times 10^{-9}$ [57]). The enrichment for transport related terms observed in the glutamate signature (above), and its similarity to a rapamycin profile make sense given that rapamycin redirects trafficking of Gap1 from the plasma membrane to the vacuole [114]. Thus, the same set of mutations in vesicle trafficking that lead to inappropriate expression of Gap1 permease activity in cells grown on glutamate also cause inappropriate permease activity following rapamycin treatment.

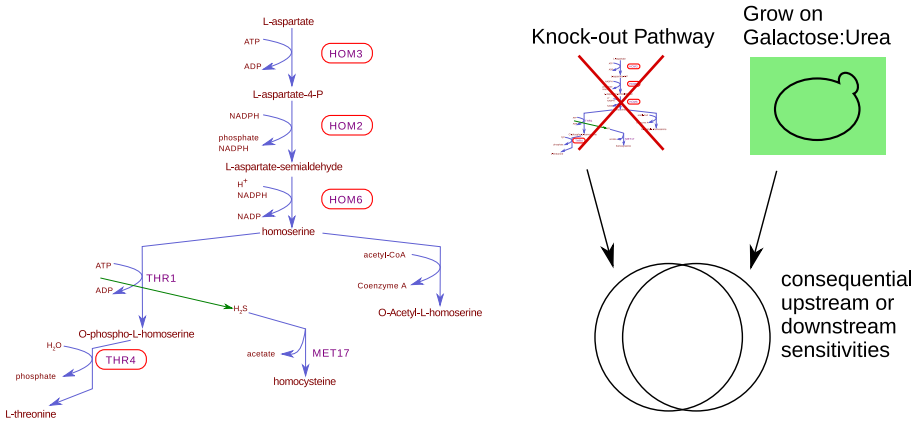

Figure 2.11: High dimensional sensitivity information for mutants in threonine biosynthetic pathway (circled in red) were obtained from SGA experiments [66]. These profiles correlate with the sensitivity profile obtained in this study when strains are grown on galactose:urea. This suggests a correspondence between the internal states of the cells when grown in a specific environment, and when subjected to a specific genetic perturbation. For example, *hom2*Δ, *hom3*Δ, *hom6*Δ, and *thr4*Δ mutants would all be expected to accumulate aspartate because these mutants shut down a major metabolic shunt for aspartate. The phenotypic similarity in genetic interaction space between these mutants and growth on galactose:urea suggests that growth on galactose:urea may cause the internal accumulation of aspartate or some other metabolic intermediate unique to the *hom2*Δ, *hom3*Δ, *hom6*Δ, and *thr4*Δ mutants.

2.4 Conclusions

The creation of the original yeast deletion collection has had a profound impact on the way in which reverse genetic experiments are performed. Yet despite a staggering number of successful studies, the inherent auxotrophies create a major blind-spot in a fundamental area of cellular function, and previous reviews of the topic have called for the creation and use of standardized prototrophic strains for metabolic experiments [86]. Recently, Mülleder and colleagues [115], have addressed the deletion collection auxotrophies by introducing a plasmid containing sequences for *HIS3*, *URA3*, *LEU2*, and *MET15*. The resource used in this study differs in that *URA3, LEU2*, and *MET15* are in their native genomic locations, with the exception of *HIS3* which is provided by *Schizosaccharomyces pombe HIS5* under the SGA reporter [88]. Without the necessity for plasmid selection, or possible effects on gene expression due to non-chromosomal location, we anticipate that our deletion collection will see frequent use by experimentalists.

The use of a genome-wide prototrophic strain collection enables truly informative sensitivity screening in metabolically controlled conditions. This represents a first step in probing how nutrients in the environment jointly affect cellular response with or without additional genetic perturbation. This study demonstrates that much work is yet to be done to understand growth in even simple environments. A solid grasp of the surprisingly complex responses to simple environments will add much needed context to studies done in more complex environments.

This study has demonstrated the potential of this collection, when screened against simple environments, to uncover phenotypes for hundreds of mutants that are phenotypically normal in standard lab conditions. We believe that the stock of simple experiments that might reveal a phenotype for these mutants has not yet been exhausted and expect that this whole-genome prototrophic collection will be an invaluable resource to the community. The rising number of metabolomics studies, fueled in part by the increasing accuracy of experimental mass-spectrometry, as well as the growing interest in metabolism as central to many common ailments in humans, make it more important than ever to properly design metabolically relevant experiments in the model eukaryote *S. cerevisiae*. Central to that goal is a version of the deletion collection that is

unhindered by historical auxotrophic requirements.

For example, while central metabolism is unrivaled among cellular processes with respect to our ability to make *in silico* predictions from constraint-based metabolic models, it is far from a fully understood system. Our results show a generally weak ability to predict condition-specific sensitivities, though performance is clearly above a random baseline. The prediction of condition-specific sensitivities is admittedly more difficult than the prediction of sensitivities in general, but it was our estimation that FBA and MoMA would be well suited to approximate our observations given our simple experimental setup. Their only moderate success in doing so demonstrates the current limitations of constraint-based modeling and the difficulty of relating models built from biomass predictions to quantitative growth rate data. There might be several possible reasons for the discrepancy between *in silico* and *in vivo* results. First, the success of predicting growth defects hinges on the proper formulation of biomass composition. While a single biomass composition is used for all our simulations, it likely changes across environmental conditions. Future studies could address this issue by measuring the composition of yeast cells under different nutrient settings. A second limitation of purely flux-based models is their inability to make predictions about components that have an indirect effect on metabolism. Consider for example the enrichment for transport related genes whose deletion confer glutamate-specific sensitivities. Their putative role in nutrient sensing and signaling reflect the fact that despite its constrained nature, the metabolic network operates as part of a much larger and more dynamic network. More generally, the basic constraint-based modeling approaches ignore regulatory mechanisms. Several attempts have been made to bridge this gap and they rely either on "omic" data to constrain the activity of specific reactions [116, 117, 118] or on integrating a mathematical representation of gene regulation with the metabolic model [119, 120, 121]. We feel that the availability of this whole-genome collection and accompanying growth data well suited to studies of metabolism will help the community to develop and test novel models and methods to better capture the operation of the greater cellular network.

Central to the understanding of the network as a whole, is the idea that a whole-genome screen reveals indirect as well as direct consequences of the perturbation tested. Positive gene-environment interactions under ribose conditions may well illustrate this

point. The median z-score for the 166 genes annotated to "chromosome segregation" in the Gene Ontology is negative for all seven galactose conditions, yet positive for all seven ribose conditions (binomial sign-test $p < 6.2 \times 10^{-5}$). We believe this shift may be explained by fundamental cellular rate limitations. Failure to segregate chromosomes in the midst of even moderate growth (for example, galactose) can have very severe consequences, ultimately limiting growth rate, whereas comparatively slow growth (for example, ribose) affords additional time for slowly segregating mutants to complete segregation. These mutants grow faster than we expect despite no apparent link between carbon metabolism and chromosome segregation. Thus growth rates under one condition disclose information about the interplay between a wide variety of cellular sub-systems, giving us a readout of the internal cellular state. Similarly, a mutant profile across many environments gives us information about how essential that gene may be in any of those various cellular states, in addition to elucidating any direct role that gene may have in direct utilization of the provided nutrients. Analysis of our growth data recapitulated the role of vesicle trafficking in the regulation of the amino acid permease Gap1, relating growth on glutamate to the drug rapamycin. This broader view of whole-genome screen information then allows for integration of profiles across different perturbation types (chemical, genetic, environmental), and should ultimately aid us in applying knowledge gained in one arena to observations made in another.

Chapter 3

Essential and non-essential genes in the complete genetic interaction network

3.1 Chapter Overview

Altering the environment reveals fitness phenotypes for many genes that are not required under standard conditions. However, extensive genetic redundancy may also contribute to the lack of observed phenotypes for single-mutants. To compensate for this redundancy, we must perturb multiple genes simultaneously. Multiple-perturbation studies date back several decades to experiments in the fruit fly [122]. In these studies, it was determined that mutations with no apparent effect could be lethal in combination, a relationship which came to be termed "synthetic lethality." In yeast, quantitative measurements of growth at the colony level enable higher resolution of effects. Not limited to just lethal and viable observations, we can characterize subtler faster-than-expected or slower-than-expected phenotypes. Synthetic Genetic Array analysis (SGA) is a robotic pinning procedure that automates the creation of yeast strains with two distinct genetic perturbations, paired with plate-imaging software that measures the resulting double-mutant fitness and identifies genetic interactions. This chapter presents analysis of the genetic interaction network in yeast, which is now nearly completely mapped.

When completed, the yeast genetic interaction map will be the only complete map of a eukaryote, and will therefore serve as the definitive model genetic interaction network. As such, we are interested in characterizing many of its general properties, which we expect to be conserved in other eukaryotes, including humans. These properties include basic measurements of how many genetic interactions there are in the yeast genome and how they are distributed. We are additionally concerned with the quality and informativeness of the data because of its definitive role representing an important and increasingly popular class of experimental data.

The work in this chapter represents only a small part of a very large collaborative effort. Because the included material represents only analyses for which I was primarily responsible, it forms an incomplete, and slightly less cohesive picture of our genetic interaction network mapping efforts. However, it will still serve to give a broad overview of genetic interaction networks and provide the foundation for the remaining chapters, which concern the genetic interactions of duplicated genes. My contributions to our efforts include the principle responsibilities for scoring raw genetic interaction data, and ensuring its technical and functional quality. This includes but is not limited to the identification and mitigation of non-biological systematic effects, and integration and normalization of data from different genetic interaction experiments or arrays. I also performed much of the fundamental characterizations of the complete genetic interaction network including drawing contrasts between its essential and non-essential components. These analyses include examining the total numbers of genetic interactions of various classes, their relative densities, and their ability to predict gene function both as individual interactions and as aggregated into profiles. I performed module analysis, examining the broad patterns within and between protein complexes, including specific complexes such as the proteasome. I also performed the clustering and hierarchy analysis of genetic interaction similarity data I generated in collaboration with Anastasia Baryshnikova.

My collaborators on the entire project are quite numerous. Principally among them are the rest of the genetic interaction team in my lab, Elizabeth Koch, Carles Pons, and Raamesh Deshpande. Anastasia Baryshnikova was also involved in many aspects of early analyses and participated directly in the creation of several figures that I have adapted for use in this chapter (Figs. 3.11, 3.13). This chapter has also been substantially

influenced by discussions with Charles Boone, his lab manager Michael Costanzo, and my advisor Chad Myers.

3.2 Introduction

3.2.1 Defining and interpreting genetic interactions

The term "genetic interaction" describes a relationship between two genes whereby the effect of their simultanious perturbation is surprising given the effects of their individual perturbations. This deviation from expectation indicates that the two genes have a joint impact on the phenotype of interest, and perhaps participate in a common cellular function. For example, one type of extreme genetic interaction, "synthetic-lethality" was characterized a half-century ago by Dobzhansky *et al.* [122]. In their fruit-fly experiments, Dobzhansky and others discovered that some mutations, while harmless on their own, were lethal in combination. Taken together, these interactions form a network that captures much of the of the cell's complex functional architecture.

There are several competing models for defining genetic interactions that differ in their theoretical and experimental properties [123]. The most widely used of these models agree that the effects of multiple independent perturbations will combine multiplicatively, but these models differ by whether they measure deviation from that expectation as a difference or as a ratio. The SGA scoring procedure uses a multiplicative null model, with the interaction term (ϵ) measured as a difference [50]. This definition can be seen in Eq. 3.1, where $\epsilon_{a,b}$ represents the genetic interaction score between genes a and b, f_a and f_b represent the single-mutant fitness scores of genes a and b respectively, and f_{ab} represents their double-mutant fitness. The resulting sign on the interaction score indicates the type of genetic interaction. Negative ϵ indicates a synthetic sick or synthetic lethal phenotype, in which the double mutant grew more slowly than expected, whereas a positive ϵ indicates a genetic interaction in which the double-mutant fitness exceeded expectation. Fig. 3.1 demonstrates an example of how the fitness of the double-mutant would determine a genetic interaction for a particular gene pair, according to Eq. 3.1.

$$\epsilon_{a,b} = f_{ab} - (f_a f_b) \tag{3.1}$$

The prevailing dogma of genetic interactions is that negative interactions (such as Dobzhansky's qualitative "synthetic-lethal" [122]) are thought to signify redundant relationships, such as those between genes in parallel pathways (Fig. 3.2). In this case, either of the single-mutants still has one functional pathway leading to the common function and the resulting single-mutants show (perhaps) no phenotype, but if both pathways are disabled simultaneously, fitness is adversely affected. Alternatively, positive scores reflect multiple perturbations to the same component, such as a protein complex or a linear pathway (Fig. 3.2). While each individual mutation disables the component, their coincidence doesn't confer any additional effect because the component is already non-functional.

There are many distinct mechanisms by which genes can be genetically related to each other. Two genes may even share mechanistic relationships that change depending on the phenotype being measured. These many classical genetic categories of interactions are reduced to two in the course of the SGA process, negative and positive. This reduction is one price of the scale of the experiment. However, the generality of the measurement (non-independent contributions to fitness) means many of these relationships can be captured in a single high-throughput assay that is relatively unbiased and captures all major cellular processes. Though an interaction may tell us that genes are related, we cannot know precisely how. This information is nonetheless informative. Genetic interactions show significant overlap with mechanistic information such as protein-protein interactions, similarity of expression patterns, and co-annotation to known functional processes. It has also been demonstrated that pairwise profile correlations are better at predicting functional relationships than individual interactions are [66]. This is in part because of the mechanistic uncertainty mentioned previously, but also because of the amount of noise generally found in biological experiments of this scale. Individual interactions are more sensitive to this noise than broad patterns are.

3.2.2 Generating genetic interactions in yeast

The robotic procedure by which SGA generates double mutants for genetic interactions in yeast is inherently asymmetric [89]. A single $MAT\alpha$ "query" strain is pinned out to every location on a plate. Meanwhile, $MAT\mathbf{a}$ "array" strains are pinned out to unique

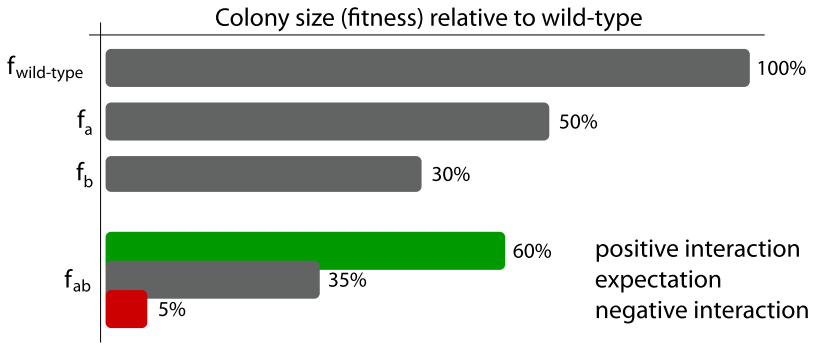

Figure 3.1:
Genetic interaction example. In this example aΔ and bΔ each have a single-mutant fitness score (f) which is less than wild-type. The expected double mutant fitness is derived from the multiplicative model shown in Eq. 3.1. Here, f_a is 0.7 and f_b is 0.5, thus the expected double-mutant fitness is 0.35. Deviations from that expectation result in either positive (green) or negative (red) genetic interactions.

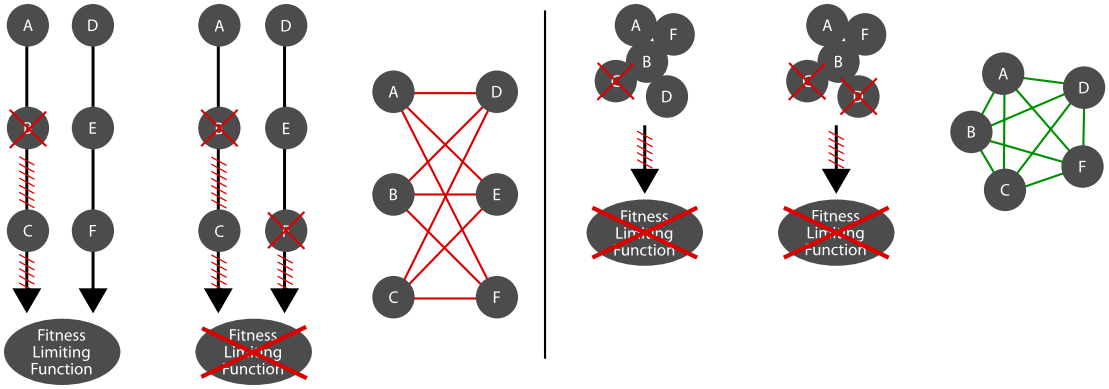

Figure 3.2:
Genetic interaction interpretations. Parallel pathways (left) feeding into the same function are expected to give rise to negative genetic interactions (red). A single perturbation (B) has no effect in the presence of an alternate route to the function. However, multiple complementary perturbations (B,F) cut all pathways feeding the function and we observe a corresponding phenotype. In this example, all complementary combinations should exhibit a negative interaction, forming a negative bi-clique.
Members of the same protein complex (right) give rise to positive genetic interactions (green). For example, a single perturbation of a complex member (C) is enough to compromise complex function. Additional perturbations (C,D) have no additional effect on the already non-functional complex. Their combined fitness effects would then be higher than a multiplicative expectation and the result is a positive genetic interaction. All complex member pairs exhibit this relationship and form a positive clique.

locations on a set of similar plates. An "array" is therefore an entire collection of single-mutant strains, but the term is sometimes used for a single gene on the array, to differentiate it from a single "query" gene. The query strains are then pinned on top of every array plate and a mating step ensues, followed by several rounds of selection. The result is a set of plates containing double-mutant colonies. The selection steps ensure that every colony is composed of haploids, and that all have been deleted for the same query gene. Additionally, each colony is deleted for a single array gene, depending on its plate-position.

Non-essential genes are perturbed in SGA by total deletions, so-called null alleles, where the entire open reading frame has been excised, and replaced by an appropriate selectable marker. It is often the case that a single copy of an essential gene can be deleted in a diploid, as long as one functional copy remains. However, the SGA procedure generally requires haploids, which precludes the possibility of including null alleles of essential genes. One approach to circumvent this problem is to alter the efficacy of the essential gene, but not so much as to be fatal. This can be done, for example, by perturbing its expression via mutations in regulatory DNA, by reducing translation through RNA interference (though *S.c.*, and many other organisms lack the requisite RNA interference machinery) or by a mutation to the gene itself which is subtler than an entire deletion.

In our SGA experiments, essential genes are represented by two alternative perturbation types. Temperature-sensitive alleles (TS), are typically coding-region mutants that show reduced growth at elevated temperatures. Recently, a collection of temperature sensitive (TS) alleles of essential genes has been released [124]. Strains in this collection have single mutations in essential genes are fatal at elevated temperatures. Presumably, these mutations slightly alter the structure of essential proteins, making them unstable when internal kinetic energy is high. Many of these mutants show reduced (though viable) growth at normal temperatures and so we can reasonably assume the function of the essential protein, or its stability has been at least slightly compromised. Alternatively, some essential genes are present in the form of "Decreased Abundance by mRNA Perturbation" (DAmP) alleles, in which untranslated 3' regions of essential have been disrupted, which results in functional proteins but at a theoretically reduced dosage [125]. These strategies allows us to pivot the SGA approach, which has proven

very successful in mapping the function of non-essential genes, toward the cells most important genes.

Data from a large set of queries is generally represented as a matrix, with each row corresponding to a query gene and each column corresponding to an array gene. Single rows are easy to add to the matrix, by screening a new query against an established array set, however to add columns, an entirely new array must be constructed, and all queries must be re-screened. There are two such arrays, referred to in this Chapter, the original *FG_array*, comprised of deletions for the bulk of non-essential yeast genes, and the *TS_array*, which contains mainly TS alleles of essential genes. Different (but highly overlapping) sets of queries have been screened against each of these arrays and so they must often be regarded separately. As data from each of the two arrays is initially slightly different, observations from the *TS_array* are first normalized to match the *FG_array* so as to make fair comparisons at matching thresholds (see Sec. B.1.1). A census of the number and type of query strains screened against each array, as well as the composition of the two arrays themselves is given in Table 3.1. SGA technology was originally developed to combine haploid mutants in the yeast deletion collection [88, 3]. This precluded the study of essential genes, which cannot be deleted in haploids. In yeast, nearly 20% of genes are essential [3], and these essential genes code for some of the cell's most important proteins. They are central to many fundamental processes, they tend to have more protein-protein interactions, and be more centrally located in cellular networks [65, 11]. Essential genes tend to have originated in more ancient common ancestors and so have higher rates of conservation between yeast and human. [60]. These properties make them tantalizing objects of study, however their essential nature presents a stumbling block to reverse genetics. By definition, the deletion of an essential gene results in inviability, so deletion-based perturbation assays can help us identify essential genes, but they often cannot give us insight into their specific function.

The inclusion of essential genes, opened up the possibility to map the entire yeast genetic interaction network, a first for any eukaryote. Specifically, it allows us to contrast the network properties of essential genes (and the interactions between them) to the non-essential genes, which was previously impossible. Because general properties of genetic interactions such as degree and modularity tend to be conserved to other

organisms [79], and because essential genes themselves are more often conserved, we expect the contrasting properties of essential and non-essential genes to be conserved also. As we seek to transfer knowledge from the model genetic interaction network to other organisms in an effort to understand how genes jointly affect complex phenotypes, information regarding this previously inaccessible segment of the interaction network will become increasingly valuable.

Mapping and understanding this network is important for a number of reasons. First, these networks give immediate insight into genes of unknown function. Genetic interactions measure the consequence to fitness of deleting or mutating a gene, and are therefore not dependent on any particular mechanism of gene function. This generalizability is especially helpful for uncharacterized genes as they may be involved any of the cell's diverse processes or structures. Genetic interactions can aid in the generation of specific functional hypotheses and appropriate assays can be used to follow up on mechanism.

Perhaps more importantly, we are mapping the underlying structure of complex phenotypes. Very few phenotypes are controlled by a single gene, and understanding how genotypes combine to produce a phenotype has broad implications. For example, the "missing heritability problem" refers to the fact that effects from individual genes fail to account for the heritability of complex phenotypes in humans. For example, a number of genome-wide association studies involving tens of thousands of people have identified at least 40 loci which are associated with human height. The heritability of this trait has been estimated to be about 80%, and yet these individual loci can explain only about 5% of the observed variance [126]. As we seek to understand the causes of ever more complex phenotypes, such as many heritable disorders in human, information regarding how genes jointly affect phenotypic outcomes can only become more important [71, 73, 72].

3.3 Results and Discussion

3.3.1 Overview of interactions discovered

We have screened approximately 16.6 million unique pairs of strains and discovered roughly 885,000 unique genetic interactions. This represents a 4.8-fold increase over

	Query Strains			Array Strains	
Mutant type	null	TS	DAmP	Deletion	TS
FG Array	3283	914	201	3827	0
TS Array	1194	865	762	176	788

Table 3.1: The section shows the allele type composition (perturbation) of each array set as well as the collection of queries screened against them. A null mutation refers to the complete deletion of a non-essential gene. TS strains carry alleles of essential genes with point mutations that render them non-functional at high temperature, but allow for partial functionality at nominal termperatures. DAmP strains carry mutations in untranslated regions adjacent to an essential gene which causes a reduction in expression (See Sec. 3.2.2).

previously available SGA data [66]. Out of the 5,959 genes in *S. cerevisiae* that are not annotated as dubious ORFs, we have a profile for 5,174 (86.8%). These include TS strains for 734 essential genes (69%), 585 of which are represented by multiple alleles with distinct perturbations, for a total of 1,129 distinct strain profiles of essential genes either screened as queries or on an array. Table 3.1 shows the number of strains in each dataset by perturbation type. Applying standard thresholds (described in [66, 50]) to the resulting genetic interaction scores yields hundreds of thousands of interactions between genes which are summarized by array set, and further by perturbation types in Table 3.2.

The large number of strains included provides excellent coverage of genes annotated to every major biological process in yeast, as well as a significant number of genes with unknown function (Fig. B.1). Despite the immense number of genetic interactions, the number of pairs tested means that interactions are still sparse. The inclusion of essential genes (both as queries and on an array) addresses a long standing blind spot in genetic interaction mapping, and the inclusion of multiple alleles of these essential genes may help us better understand the sequence to function relationships for these genes.

3.3.2 Assessment of experimental reproducibility

In a previous SGA publication, the recall characteristics of individual genetic interactions was estimated by measuring SGA's ability to recover interactions published in independent experiments [66]. However, as these data represent the vast majority of

		$\epsilon < -0.12$	$\epsilon < -0.08$	$\epsilon < 0$	$\epsilon > 0$	$\epsilon > 0.08$	$\epsilon > 0.16$
	total	343236	565999	1359321	1215443	334679	47401
By Array	FG array	210171	371120	1042900	945688	226029	30499
	TS array	133065	194879	316421	269755	108650	16902
By Type	Del-Del	139230	246682	747244	681247	144938	16023
	Del-TS	38291	57409	97614	78785	29307	4703
	TS-Del	82480	137458	291277	258302	87513	15992
	TS-TS	45294	62697	93038	83512	37281	5747
	DAMP-Del	14811	25407	65298	57922	15204	1970
	DAMP-TS	21614	34291	62078	52812	18698	2631
	Misc-All	1516	2055	2772	2863	1738	335

Table 3.2: The total number of negative and positive interactions is given using several standard thresholds for ϵ. All interactions counted have an additional p-value threshold of $p < 0.05$. The inner-most columns are then the most lenient and count supersets of the outer columns. In most analyses, the intermediate threshold is used ($|\epsilon| > 0.08$; $p < 0.05$)

available genetic interaction data we are without a complete "gold-standard" for comparison. The precision and recall of the SGA method is therefore difficult to estimate. In this section, I evaluate the present genetic interaction data both by standards previously used, as well as through analysis of biological replicate screens created for such a purpose.

The estimate of recall in this section is more biological than technical in nature. That is, it gives a sense of the agreement of SGA interactions with other studies. By contrast, the precision estimate presented here is more technical. It demonstrates how often two SGA experiments agree with one another.

3.3.3 Correlation of reciprocal interactions

Each combination of one query and one array is performed in quadruplicate. However, these four colonies are adjacent to one another on the plate, and though they are useful in quality control procedures, they do not represent truly independent replicates. We therefore leveraged the substantial overlap between the query and array strain collections to assess the agreement when strain combinations are observed twice independently. On the *FG_array*, we have 2,651 strains screened as both a query and an array giving us ~3.5M pairs of reciprocal observations, while on the *TS_array* there were 713

replicate level to assess (n)	1	2	3	4	5	6
observations made at this level	155	250	310	235	115	25
unique queries contributing at this level	31	31	31	31	23	5

Table 3.3: To assess the impact of multiple replicates on SGA scoring accuracy, 31 queries were selected for an increased number of replicates. Queries were scored multiple times, each time grouped with different subsets of the available replicates. The table shows the number of observations within this set that are made after merging n replicates, as well has how many unique queries, of a possible 31, contribute to our observations at each level.

such strains and so ∼250K observation pairs. Correlations on these observations (after applying a lenient filter on interactions, $p < 0.05$) appear quite good ($r = 0.61, 0.77$, Fig. 3.3). This technical reproducibility of genetic interaction scores increases our confidence that significant effects called in our experiment are largely real, and form a legitimate basis for the investigation of the true genetic interaction network.

3.3.4 Replicate screening to estimate SGA precision and recall

To assess the effect of an increased number of replicates on technical estimates of precision and recall for SGA data, a subset of 31 queries was selected to be screened N times with N in $(5 \ldots 7)$. The normal procedure for merging replicate data within the SGA scoring pipeline is to use the mean over all replicates after all experimental effect corrections have been applied. The processing of these replicates was structured such that random groups of $n = (1 \ldots 4)$ replicates were treated as the same query (set A) while the remaining $N - n$ replicates were treated as a single separate query (set B). This procedure was repeated 5 times, and so a query with $N = 6$ replicates would be scored 10 times as random groups of 4 (with 5 from set A of $n = 4$, and 5 from set B of $n = 2$). The total number of observations we have for merging n replicates is given in Table 3.3.

Accurate measures of technical reproducibility are crucial, especially in experimental design stages (See Sec. 5.3.2). The estimates here are acceptable, given the high-throughput nature of the assay, and rarity of genetic interactions, but must be kept in mind when considering any individual genetic interaction.

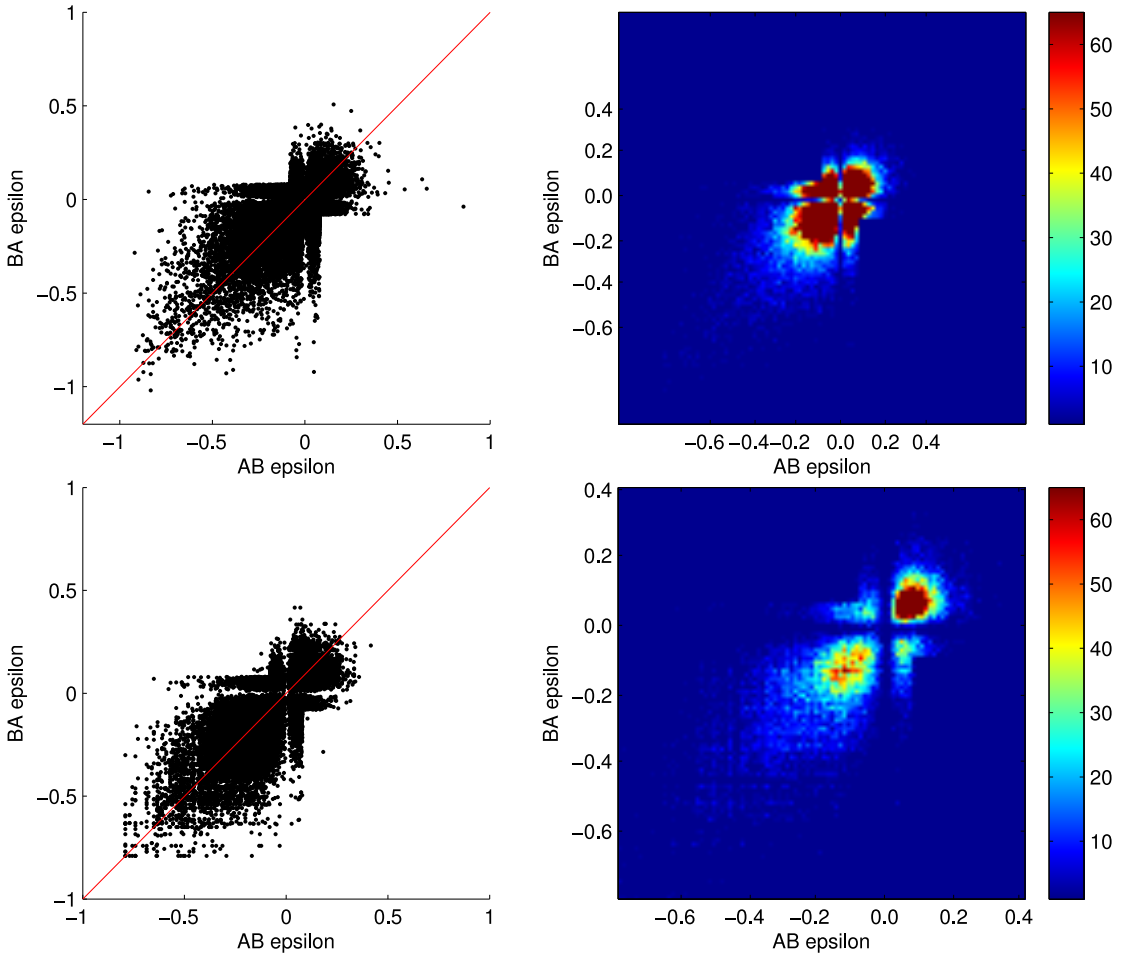

Figure 3.3: Correlation of AB - BA observations. ϵ scores gathered for pairs of interaction observations in which one gene (A) was screened as a "query" and crossed to the other on an "array." (B). These are plotted against the ϵ scores from the reciprocal observation (B×A) where they exist. (top) Data for the *FG_array*. $r = 0.61$ (Pearson) (bottom) Data for *TS_array*. $r = 0.77$ Left panels show a scatter plot of epsilon values, with a red line along $y = x$ for visual clarity. Right panels show the same data as a 2-dimensional histogram, with 50 bins along each axis.

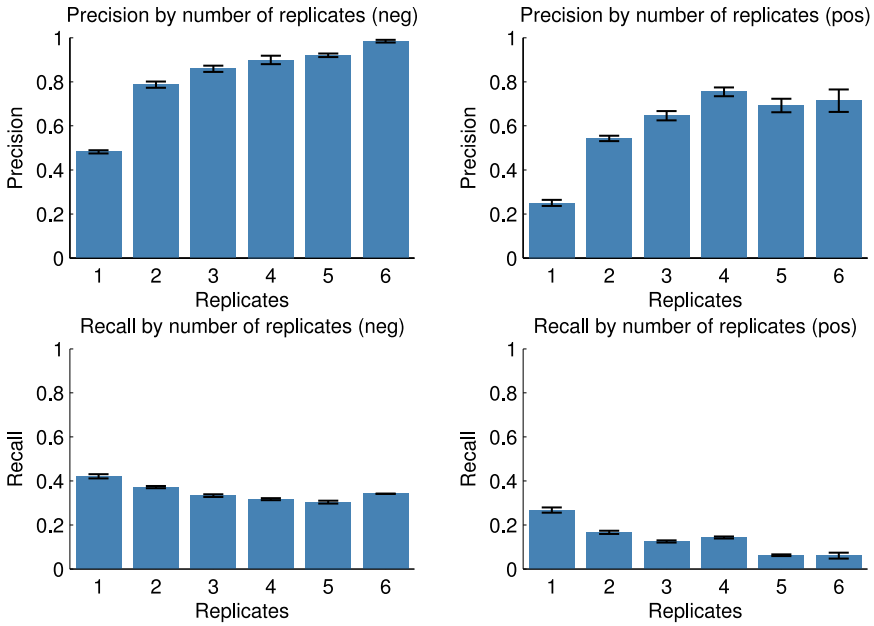

Figure 3.4: SGA precision and recall of biological replicates. 31 queries were selected for an increased number of replicates. Queries are grouped together multiple times and averaged as part of the normal SGA scoring pipeline. Results for queries are then compared against an SGA derived "gold-standard" (see Sec. B.1.2) to estimate precision and recall. Precision and recall estimates are aggregated for each query (median) and the bar shows the median and standard-error over all the queries.

3.3.5 Properties of the essential genetic interaction network

As the largest collection of genetic interactions to date, these data allow not only a reassessment of global genetic interaction structure, but a detailed comparison between essential and non-essential genes. Notably, this includes interactions between pairs of essential genes, which have never been measured on this scale.

3.3.6 Variations in genetic interaction density

The average density of genetic interactions measures between pairs of non-essential genes is slightly lower than previously reported (Fig. 3.5 NxN) [66]. This is likely due to the selection strategy of the previous experiment, which correctly predicted that genes with a single-mutant fitness defect would exhibit more genetic interactions and were prioritized for screening. Our estimate for non- essential interaction density is 2.0% for negative interactions and 1.2% for positive interactions. Strikingly, pairs of essential genes interact at a rate nearly five times higher (9.9% negative, 5.7% positive, Fig. 3.5 ExE). This suggests the essential genes tend to impinge upon a greater number of cellular functions, or that their disruption makes the cell vulnerable to a greater number of specific sensitivities. Essential genes tend to have more functional annotations, supporting a correspondence between genetic interaction degree and multi-functionality However, it is difficult to estimate the impact of investigation bias when comparing essential and non- essential functional annotations because non-essential genes are frequently screened in large genome-wide assays [3] whereas essential gene experiments must be targeted for study in a much different manner (e.g. suppressor screens), though subtle mutations of essential genes might be expected to show up for frequently in forward genetic screens. Additionally, essential genes, with their easily detectable phenotype (inviability) face an ascertainment bias regarding their inclusion in smaller scale studies, and large collections of functionally compromised essential alleles are only just becoming available [124].

3.3.7 Information content of essential genetic profiles

In addition to essential-gene genetic interactions being more numerous, they are more informative as well. Fig. 3.6 shows the value of individual interactions in the prediction

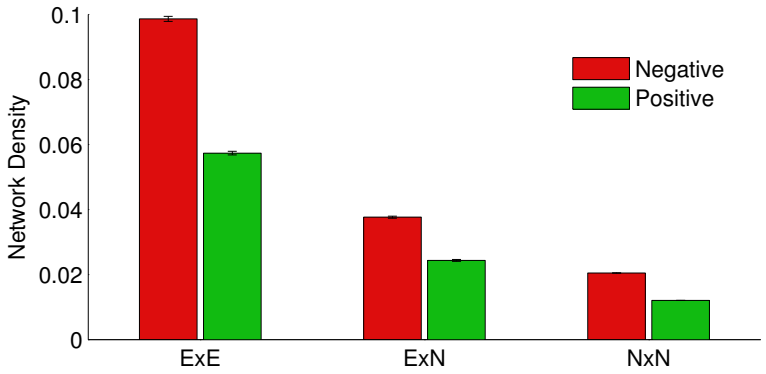

Figure 3.5: Genetic interaction network density. Each of three types of interactions are considered: interactions between essential genes (ExE), interactions linking essential and non-essential genes (ExN) and interactions between non-essential genes (NxN). Only TS alleles (not DAmP alleles) are considered for essential gene interaction densities. NxN and ExN data are taken from *FG_array*, ExE data are taken from *TS_array*, and "Network Density" is measured as the number of significant (intermediate) positive or negative interactions divided by the number of tested gene pairs. In cases where multiple alleles of the same gene were available, a single one was chosen at random. Error bar estimates were derived from multiple rounds of random allele selection.

of co-annotation to Gene Ontology terms, as well as in the prediction of protein-protein interactions. Pairs of non-essential genes (NN) with strong negative genetic interactions maintain a precision of 65% over the first 400 true positives, a 3.6-fold increase over background. Essential interactions have a precision of 80% over the first 1,000 true positives, though the background rate of functional relation among essential pairs is much higher (35%). In a surprising contrast to non-essential interactions, negative essential interactions do a much better job at predicting functional annotations and protein interactions, while positive interactions between essentials carry very little information at all according to this measure.

Overlap between negative genetic interactions and protein-protein interactions provides an even starker contrast between essential and non-essential genes. The strongest non-essential negative interactions (top 200) correctly predict a protein-protein interaction around 10% of the time, a high enrichment over the background rate of protein-protein interactions (0.18%), though their predictive power quickly falls off at higher

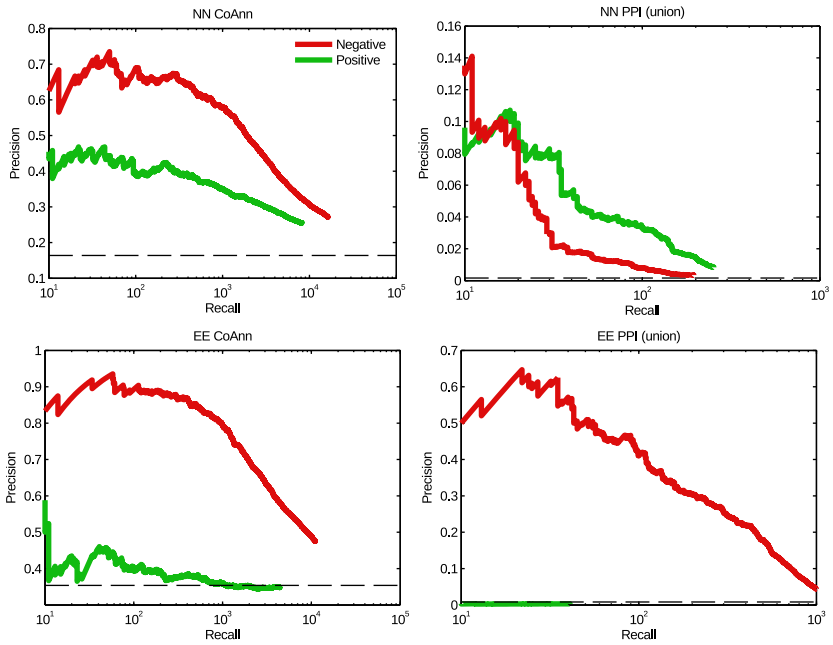

Figure 3.6: Precision-Recall plots for essential and non-essential genetic interactions. Negative interactions are shown in red, positive interactions in green. Interactions are ordered for prediction by the magnitude of epsilon after an intermediate threshold is applied ($|\epsilon| > 0.08,\ p < 0.5$). Prediction of co-annotation to an informative subset of the Gene Ontology is shown on the left, and prediction of protein-protein interactions (see Methods B.1.3) is shown on the right. The top two panels (NN) consider only interactions between pairs of non-essential genes, and the bottom two panels only consider essential pairs (EE). In each panel the background rate of true positives among all screened pairs is shown as a black dashed line.

values of recall (Fig. 3.6). Negative interactions for essential genes, by contrast, continue to be an excellent indicator for protein-protein interactions well into higher values of recall, and at lower values of recall, the precision of negative-essential interactions exceeds 40%, which is over 50 times the background rate of 0.73%.

3.3.8 Genetic interactions within and between functional modules

The inclusion of more essential genes as both queries and array genes allows us to examine whether long-standing dogmas about the structure of genetic interactions hold as true for essential interactions as they do for non-essential interactions. Previous work has established an enrichment for positive interactions "within" functional modules,

such as the fictional protein complex presented in the right panel of Fig. 3.2. Meanwhile, portions of the genetic interaction network falling "between" functional modules show a stronger enrichment for negative genetic interactions, resembling the parallel pathway example in the left panel of Fig. 3.2.

We examined the genetic interactions within and between a curated set of 420 protein complexes [50]. This protein complex standard represents an assignment of genes to specific functional modules of coherent function, which is independent of genetic interaction data, and 193 of these complexes include at least 4 gene pairs tested for genetic interactions. We find that genes pairs drawn from a single complex tend to show either negative or positive genetic interactions, and less commonly show both types (Fig. 3.7, top). We also find that the tendency to exhibit negative interactions is strongly related to the number of essential genes in the complex (Fig. 3.7, top). We further filtered the set of protein complexes to 96 that had 5 genes or more, and were comprised of mostly essential or mostly non-essential genes ($\geq 80\%$). After filtering, the set was roughly split between essential complexes (41) and non-essential complexes (55), and its segregated nature allows us to cleanly analyze essential, non-essential, and mixed-essential interactions within and between coherent functional modules.

We then examined each complex and determined how many of them are enriched for negative and positive genetic interactions (Fig. 3.7, center). The addition of such a huge number of genetic interactions to the network has impacted the average interaction density between non-essential genes (Fig. 3.5), however, the prevalence for positive interactions within non-essential modules remains, with 27.6% of non-essential complexes showing positive interaction enrichment, almost 2-fold more than the number showing enrichment for negative interactions (15.5%). These complexes are a subset of those shown in the top panel of Fig. 3.7, where we can see that non-essential complexes with an appreciable density of negative interactions have some positive interactions also (left of center). Genetic interactions within essential complexes tell a different story: 79% of essential complexes are enriched for negative interactions, while not one shows enrichment for positive interactions. Thus, the dogma for genetic interactions is reversed for essential genes, and is even more extreme. If the "negative-within" dogma for essential genes generalizes beyond protein complexes as we suspect, it would no doubt contribute

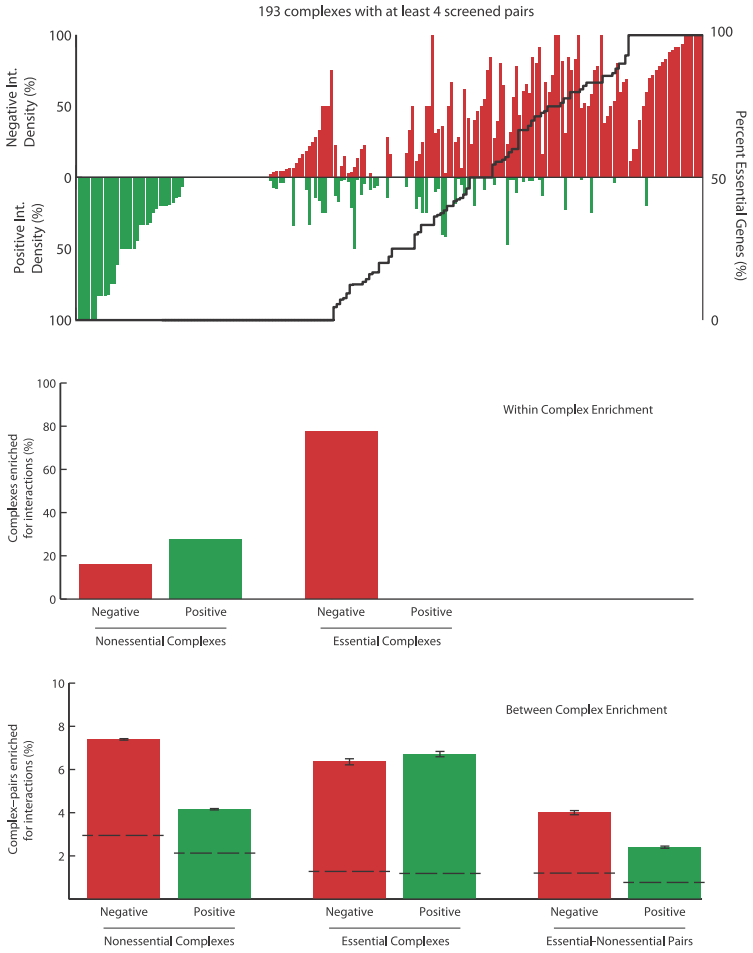

Figure 3.7: Genetic interactions and essential protein complexes. Top) The density of negative (red) and positive (green) significant genetic interactions for the 193 protein complexes with sufficient genetic interaction data. On the second Y-axis, the number of essential genes in each complex is shown (%, black line). Complexes are sorted by essential fraction, then negative density, then positive. Center) The fraction of complexes enriched for negative and positive interactions with themselves, based on a degree-controlling hyper-geometric test with a p-value threshold of $p < 0.05$. Results shown separately for non-essential and essential complexes, consisting of $< 20\%$ and $> 80\%$ essential genes respectively. Bottom) As in center, but considering interactions between pairs of complexes, including mixed pairs. Only one random TS allele of each essential genes was included, and error bars represent the standard error of each measure over 10 runs. Dotted lines show random expectation based on the same analysis performed on the genetic interaction network after a degree-preserving edge randomization. These lines were virtually indistinguishable from 0 in the center panel and removed for clarity.

to the increased predictive performance of negative ExE interactions in Fig. 3.6. In similar fashion, the near total absence of positive interactions within essential functional modules contributes to the decrease in prediction performance for ExE positives.

We then examined the nearly 1,000 pairs of complexes for interaction enrichment to test previous "between" module conceptions. We found that results for non-essential modules on the complete network reflect expectations derived from earlier incomplete networks (Fig. 3.7, bottom panel). Just under 7.5% of non-essential complex pairs are enriched for negative interactions, a 6-fold increase over random expectation (dotted line). While non-essential complex pairs also showed enrichment for positive interactions (4.2%) the effect is not quite as strong (4.9-fold over background). As in the "within" module analysis, essential modules show a striking contrast in characteristics. About 6% of essential complex pairs show enrichment for negative interactions, and the number of pairs showing positive enrichment is only slightly higher (6.5%).

Interaction enrichments between mixed pairs of essential and non-essential complex show a slight preference for negative enrichment, echoing the results for pairs of non-essential complexes, yet the rates are much lower for mixed pairs than they are for pairs of either other type.

These results have several important consequences and begin to paint a picture of how essential and non-essential components of the genetic interaction network connect to themselves and to each other. Most importantly, the previous rule of thumb: "positive-within, negative-between", already an over-simplification, is categorically reversed when applied to essential genes.

Perhaps the near uniform behavior of essential complexes in this regard suggests they have a more constrained or fragile structure. Let a "fragile" complex be a complex that is is non-functionalized if a single constituent gene is deleted. A fragile complex performing a non-essential function would exhibit positive interactions just as suggested in Fig. 3.2. A fragile complex performing an essential function would not survive any single deletion, and its members would all be labeled as essential genes. Our data substitutes temperature sensitive alleles for deletions in the case of essential genes, subject to the constraint that such alleles cannot be found for every essential gene. These alleles are selected in a procedure with viability as a prerequisite, and so a fragile complex of

even essential function can sustain a single perturbation of this type almost by definition. However, nearly every possible combination of these perturbations in a single complex is enough to result in cell death. While non-essential complexes have some negative interactions within themselves, they never reach this fever pitch, suggesting that essential complexes perform essential functions (obviously) but have a higher rate of "fragility" than non-essential complexes.

On the other-hand a "robust" complex, perhaps with a less intricate structure, could survive a single deletion. These complexes would not contain many essential genes (by definition) regardless of the importance of their function. These complexes would show some negative interactions when multiple perturbations are introduced. These interactions would be severe if the complex function were essential, and more subtle otherwise. Or they may show very few interactions at all, consistent with the majority of non-essential complexes showing neither negative nor positive enrichment. This reasoning unifies the contrasting patterns for essential and non-essential within complex interactions and suggests the existence of three protein complex classes in the cell, depending on the structural and functional characteristics: "fragile-essential", "fragile-non-essential", and "robust-non-essential", with a fourth class "robust-essential" being a sort of contradiction-by-definition.

The contrast in complex-complex interaction also begs some speculation. The relative equity of positive and negative interaction between essential complex pairs compared to non-essential complex pairs is especially surprising given the differences in predictive performance of the different interaction types seen in Fig. 3.6. It suggests that essential complexes often "communicate" by coherent sets of positive interactions, yet because positive interactions so seldom capture local functional information, these communications likely connect more distal cellular functions. Further, the decrease in "communications" of either type between essential and non-essential complexes suggests a sort of two-tiered network, where core functional processes communicate to carry out essential cellular function amongst themselves while non-essential processes do the same for less important functions with limited cross-talk between the two sectors.

There may be an evolutionary explanation at work. To draw an analogy to computer science, "kernel" functions handle processes that are most crucial to operation, while the implementation details differ, kernel responsibilities are largely the same from one

platform to another, regardless of the application. Direct access to system memory sectors is an example of a function which cannot be allowed to fail, and is entrusted only to the kernel. These are differentiable from "shell" functions, which are more adaptable, fault-tolerant, and specific to the application at hand. User applications such as an email program are examples of shell functions. In a computer system, the distinction is explicit and intentional, and this property allows great flexibility in the programs we use (and create) every day, allowing them to quickly adapt to meet changing requirements without running the risk of disrupting essential functions. Of course, operating systems do not evolve in the strict biological sense, but decisions made in the design of modern operating systems do reflect lessons learned from previous versions. Perhaps the slight schism we see between the essential and non-essential segments of the genetic interaction network reflects an evolutionary solution to a similar problem. It may be that the segment of the network responsible for core cellular processes is kept isolated from segments governing environmental responses so that the latter can evolve more quickly without fatal disruptions.

3.3.9 The predictive power of essential profiles

Previous work has shown that the potential for gene function prediction of genetic interaction profiles, even using very simple methods, far exceeds that of individual interactions [66]. This has to do with many factors. First and foremost, it is directly related to the precision of the genetic interactions we call. Our precision at determining real interactions is only about 50% (See Sec. 3.6), and of course, not all real interactions will belong to co-annotated genes. However, the coherent structures formed in the genetic interaction network can help us filter out the excess noise. Fig. 3.8 shows an example of how properties like those seen in Sec. 3.3.8, monochromaticity and consistent module-module interactions, can lead to high similarity scores when measured over a multi-dimensional profile. Profile similarity scores in the SGA network are based on concordance of thousands of observations. Even real genetic interactions may sometimes connect very distantly related processes by some unknown mechanism, yielding what this type of analysis would consider a false positive. However, it is far less likely that a pair will show the same pattern of mechanistic relationships (and therefore genetic interactions) over a large number secondary genes unless the pair is much more closely

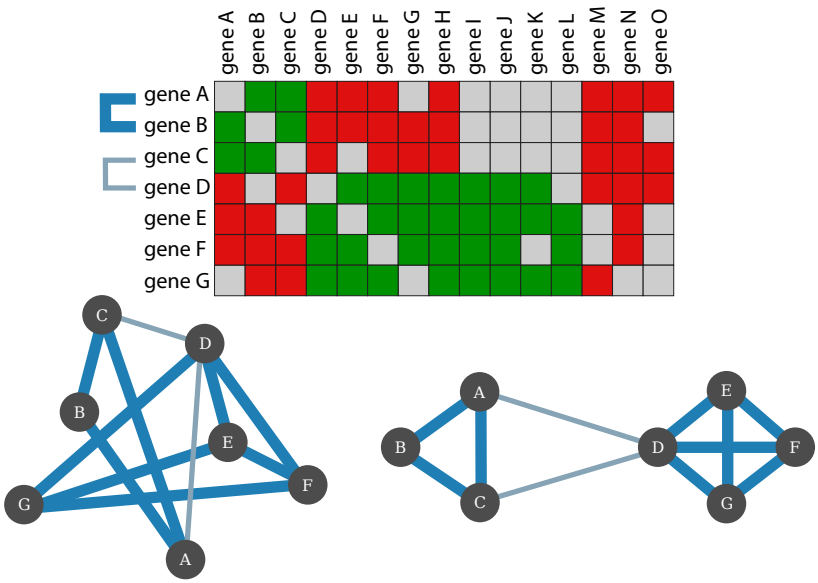

Figure 3.8: Profile similarity example. The tendency of genetic interactions to form modular structures (Fig. 3.2) makes genetic interaction profile similarity a powerful predictor of modular function. In this example, two positive interaction cliques (A–C, D–G) also form bi-cliques of negative interactions. As a result of this structure, genes A and B have a very high similarity and can be inferred to be co-modular, whereas genes C and D are in opposing modules and have a very low similarity score. A weighted network of all pairwise similarities (below left) is easily constructed, and a simple layout algorithm (such as a spring model) can be applied to reveal the modular structures in the data (below right).

related, thus these coherent relationships within and between modular structures make profile similarity highly accurate at function prediction via guilt-by-association. Additionally, by predicting relationships from thousands of observations instead of one or two, the predictions become much more robust to even very high levels of experimental noise [127].

We therefore set out to compare the predictive power of our essential vs non-essential queries. However, as noted in Sec. 3.3.6, essential genes are co-annotated to functional categories at a very high rate, so we classified each array gene and ran the experiment separately for the essential and non-essential arrays. Classification was done using a k-nearest neighbor (KNN) classifier to evaluate similarities between array-gene profiles. We could then compare the predictive power of sets of query genes (matrix rows) by

whether or not their inclusion had a positive impact on the similarity scores of array-gene pairs (matrix columns, See Methods B.1.5).

Fig. 3.9 shows the results of these predictions. The best predictions for both array gene function come from using our entire set of queries as features to correlate over, and this holds true for both non-essential array genes on the *FG_array* and essential array genes on the *TS_array* (black lines, left and right respectively). Functional predictions are performed independently for each Gene Ontology term the plot summarizes the performance by counting how many GO terms (X) achieve a precision of Y, at 25% recall (See Methods). When using all queries together as features to predict non-essential gene function, we can make predictions with 30% accuracy for more than 200 GO terms, and when predicting essential gene annotations nearly 300 GO terms reach that level of precision. However, using only 100 deletion queries those numbers drop to only about 50 and 125 GO terms respectively ($Y = 0.3$, blue lines). DAmP alleles of essential genes do not provide any more information than deletion mutants by this measure (green lines). 100 temperature sensitive queries, on the other hand provide much more useful features to correlate over, giving us 100 and 250 GO terms with 30% precision at 25% recall (red lines, non-essential and essential array prediction respectively).

The results show that, regardless of whether you are predicting the function of essential or non-essential genes, genetic interaction profiles of essential genes give you more predictive power than non-essential profiles per screen (Fig. 3.9) That is, data from our temperature sensitive queries allow us to make more correct predictions, across a larger set of functional categories than do an equivalent number of non-essential deletions. By this same measure of comparison, our DAmP perturbations of essential genes are no more informative than standard deletions, an observation that agrees with an assessment of individual interaction quality for DAmP alleles (data not shown).

3.3.10 The proteasome as an essential hub

The ubiquitin-proteasome system helps maintain control over cellular function by regulating the degradation of hundreds of different proteins in a highly specific and time dependent fashion. The proteasome is generally comprised of two subassemlies, the regulatory particle (19S), and the core particle (20S), which can be found in the nucleus and/or the cytoplasm of all eukaryotes, archaea, and some bacteria. The S. cerevisiae

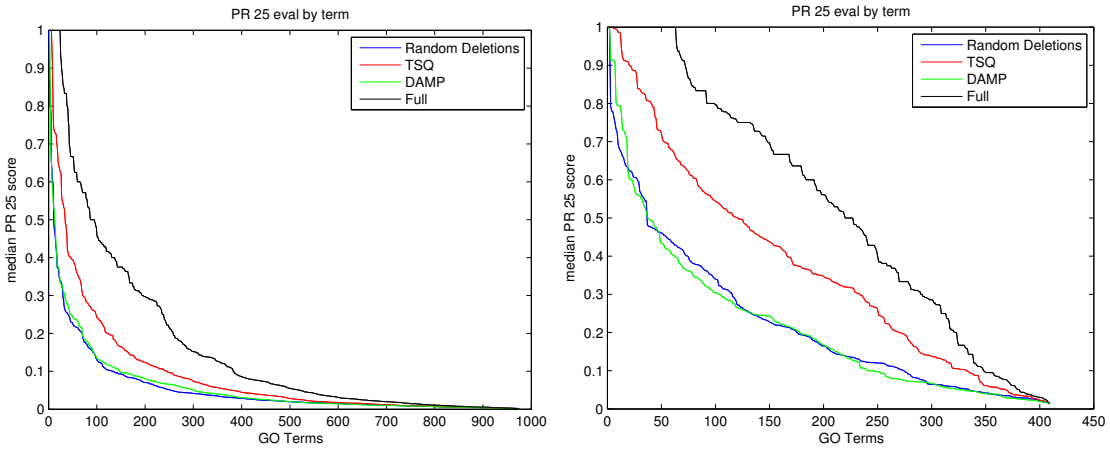

Figure 3.9: Left: Predicting non-essential co-function for genes on the *FG_array*. Right: Predicting essential array co-function on *TS_array*. The number of GO Terms (X) that have a median PR25 score (precision at 25% recall, Y) is shown when array-array profile similarity is calculated using different sets of queries. For each perturbation type, 100 random queries are selected and used to predict the function of all array genes (see Methods B.1.5). This procedure is repeated for 50 iterations and curves represent the mean results. The performance when using all queries combined is shown in black for comparison. The total number of queries available of each type, for each dataset can be found in Table 3.1.

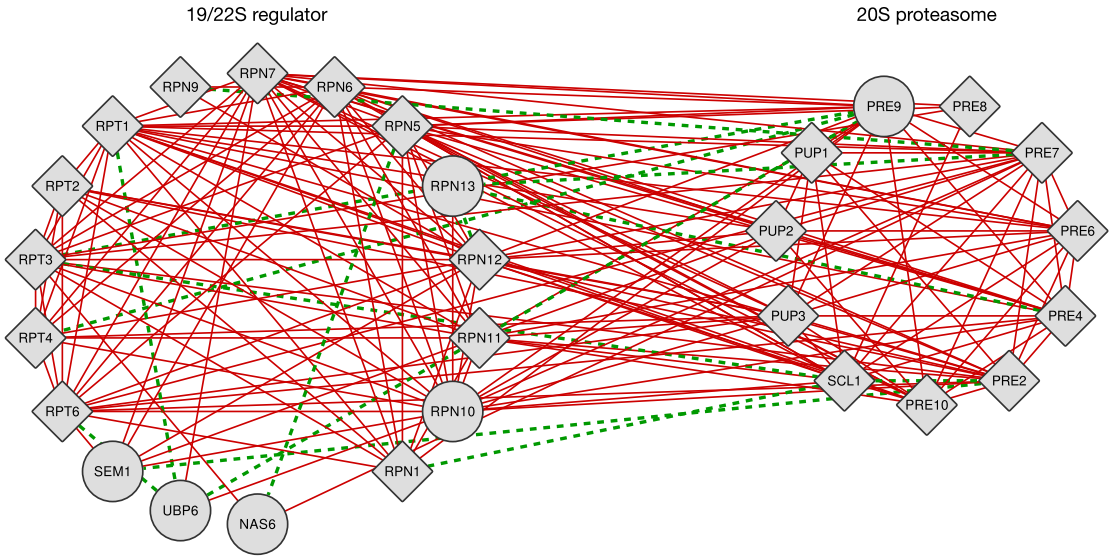

Figure 3.10: Proteasome genetic interactions. The two subassemblies of the proteasome show a significant enrichment for negative genetic interactions (red, solid), but also show a few positive interactions (green, dashed). Negative interaction density within the 19/22S, within the 20S, and between the two are all significantly above background by Fisher's exact test ($p < 10^{-121}$, 10^{-63}, 10^{-65} respectively). Diamonds represent essential genes, and circles represent non-essential genes.

proteasome is comprised of over 30 distinct proteins, and by this measure is the most complex protease known [128].

We have genetic interaction data for 28 of subunits annotated to the two main subassemblies, shown in Fig. 3.10. These two subassemblies are comprised almost entirely of essential genes, and each of them is highly enriched for negative interactions, consistent with observations from Sec. 3.3.8. Furthermore, they are highly enriched for negative interactions between them, consistent with their tightly cooperative role as pieces of the proteasome as a whole.

The proteasome also shows itself to be an important hub in the genetic interaction network. Consistent with its role as a major regulator of many diverse biological processes, the proteasome shows enrichment for both positive and negative genetic interactions with genes from almost every major category of cellular function. Furthermore, many of these interactions form characteristically coherent patterns. Fig. 3.11 shows the 25 complexes from our 430-complex standard with which the proteasome shows a

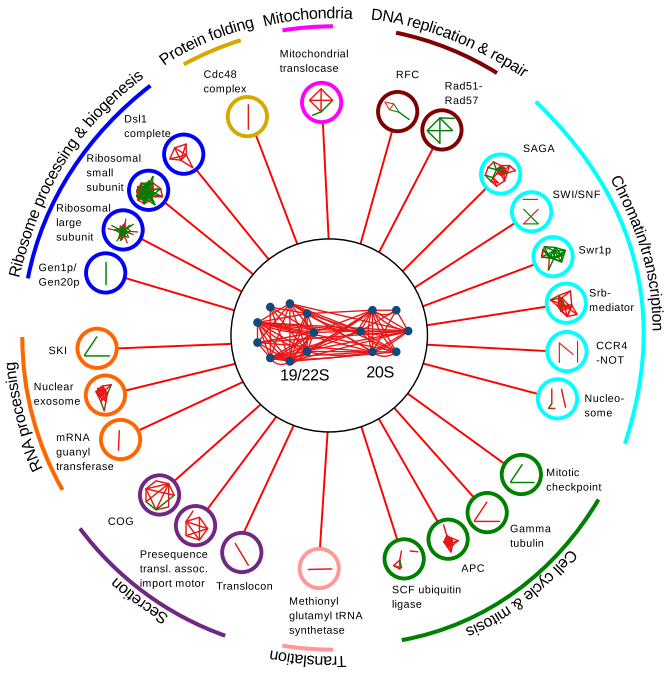

Figure 3.11: Proteasome module-module genetic interactions. The proteasome (center) is enriched for negative genetic interactions with many other annotated complexes. Each of these are shown, with their own "within" genetic interactions inset, and sorted by broad cellular process. Among them are representatives from every level of protein homeostasis control.

significant enrichment for negative interactions. They are sorted loosely according to high-level cellular processes and each is shown with its own genetic interactions inset. Perhaps most notably among them are representatives from every major process that (along with the proteasome itself) regulate protein homeostasis. These include complexes involved in chromatin and transcription, RNA processing, ribosome processing and biogenesis, translation, secretion, and protein folding.

Owing to its highly conserved nature, its important role in the process of protein homeostasis, and its high degree of structured genetic interactions, the proteasome represents an important addition to the complete genetic interaction network. Thus aside from interesting differences in the structural properties of essential genetic interactions and their increased predictive performance, the inclusion of essential genes in genetic interaction experiments yields a single huge advantage: the inclusion of processes with

a high proportion of essential genes. That is, the inclusion of essential genes not only allows us to contrast them with their non-essential counterparts, but provides a first look at genetic interactions for structures that are comprised almost entirely of essential genes, like the proteasome.

3.3.11 Hierarchical structure in the genetic interaction network

Protein complexes can give valuable insight into the modular structures of genetic interactions. However, they constitute only a small fraction of the genome, and so give only a limited view provided by direct physical interactions. Additionally, they capture functional modules at one fixed level of specificity, with no flexibility to examine functional modules of various scope. In order to examine genetic interaction patterns at various resolutions, we first calculated similarities between all pairs of array genes using data from both the *FG_array* and the *TS_array*. We then hierarchically clustered these similarity scores (i.e. calculating correlations of correlations) and thresholded the linkages at several levels of resolution (See Table B.1). The result was 5 "levels" of perfectly nested clusters.

Level 1 is the broadest cluster definition. It contains a single cluster to which all genes belong. Level 2 clusters represent broad cellular processes, and are enriched for genes belonging to similarly broad functional categories such as metabolism, RNA processing, chromatin/transcription and ER-Golgi trafficking.

Fig. 3.12 shows a matrix of the similarity data after hierarchical clustering. Overlaid along the diagonal are boxes (beginning at level 2 in red) showing the various levels of cluster resolution. Fig. 3.13 shows an alternative network representation of the data in which nodes (genes) have been colored according to their level 2 membership. This view confirms that many of the structures visible in the hierarchically clustered version also present themselves in a spring-layout network visualization, a common tool for exploring network structure.

Fig. 3.12 also shows an enlarged view of one of the level 2 clusters which highlights more specific levels and shows modular correlation structure (green blocks off the diagonal). There are hundreds of clusters at more specific levels. For example, level 5 clusters show enrichment for extremely specific GO terms, such as sister chromatid cohesion, and intracellular protein transmembrane transport, and frequently contain only

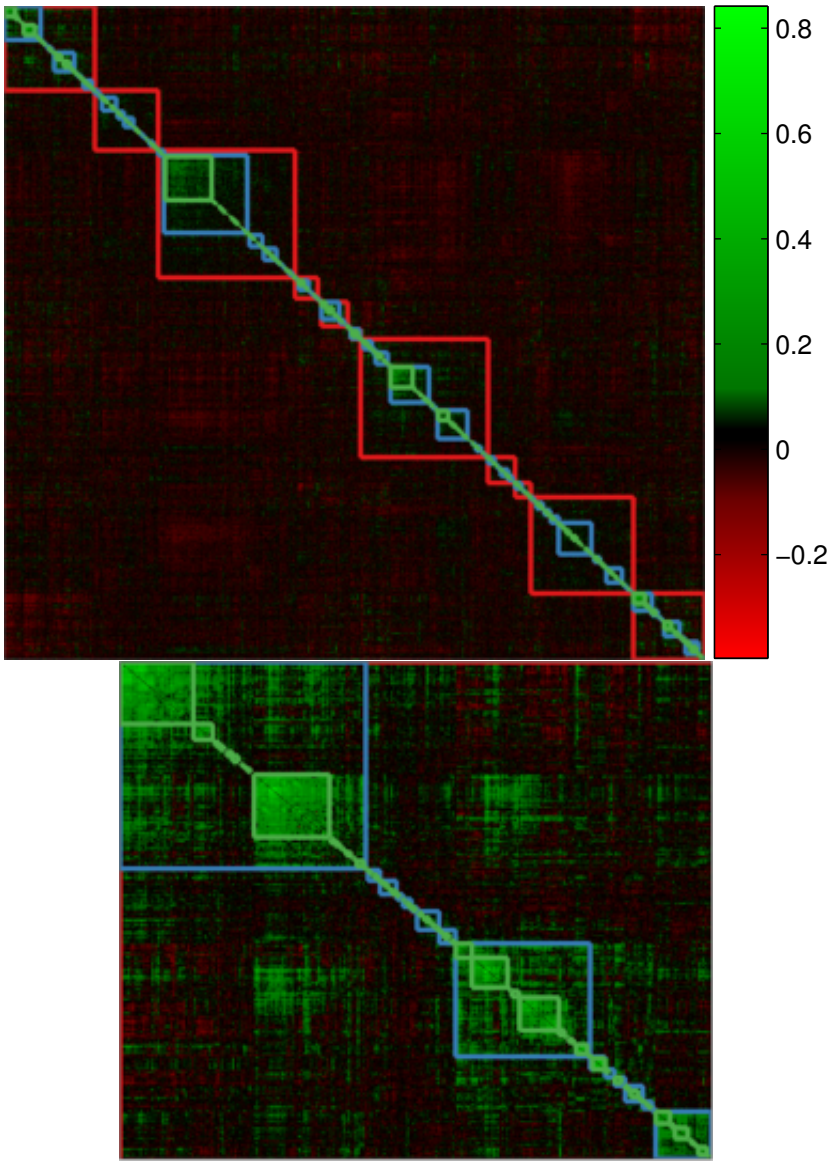

Figure 3.12: Defining hierarchical clusters. Above) Data from both arrays was integrated to produce a similarity matrix with all array genes on both axes. This matrix was clustered and four nested "levels" were defined by thresholding linkages. Level 2 is comprised of 11 large clusters (red boxes). Level 3 (blue) further partitions each into smaller segments and so on to 4 (green). Colors show Pearson correlations between pairs of array genes, though the data are organized by row-wise similarity (correlation of correlations), which is symmetric. Below) A magnification of the upper-left red-box.

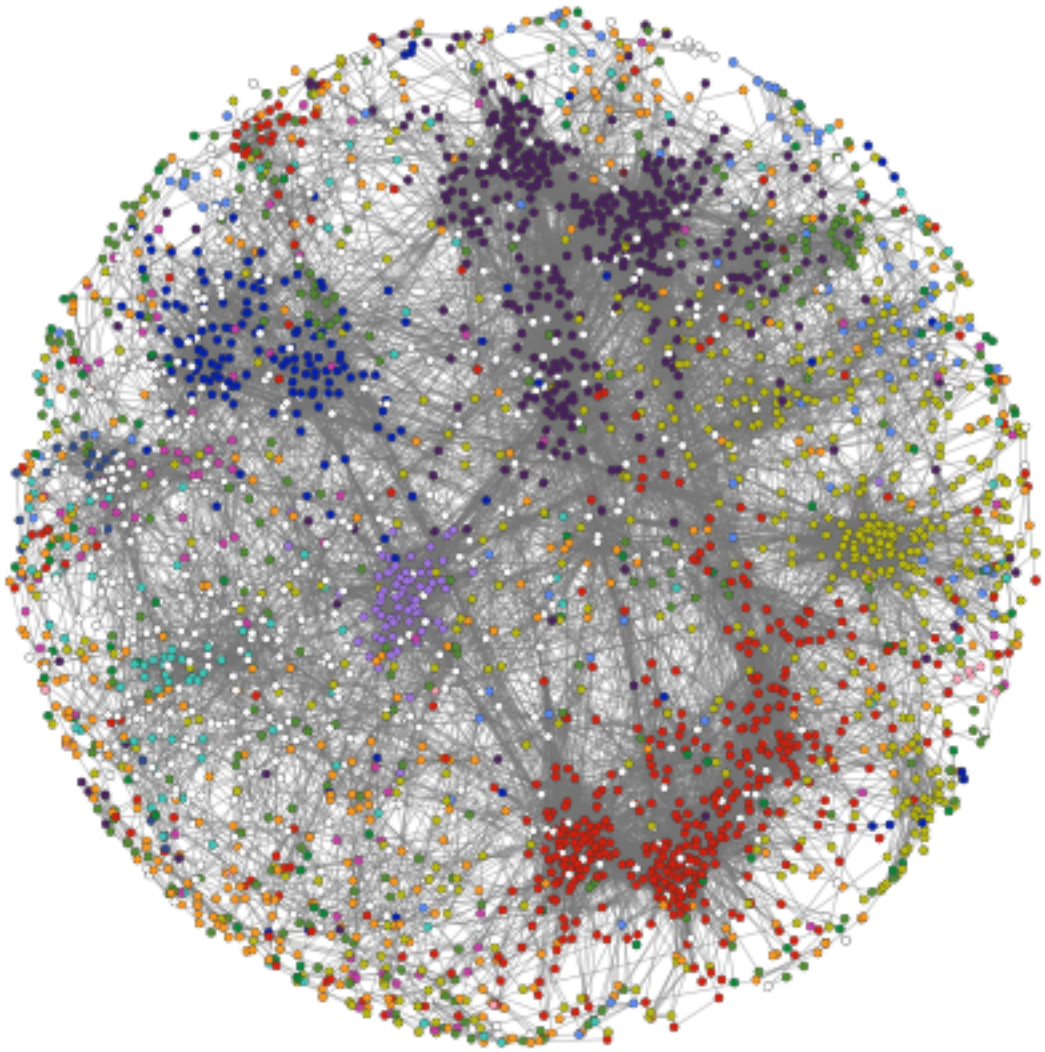

Figure 3.13: A map of cellular function. Each gene in the genetic interaction network is represented as a single node with edges connecting nodes with high genetic interaction profile similarity. Nodes are then laid out using a spring embedded model, then colored by their membership in one of 14 "Level 2" clusters (Red boxes in Fig. 3.12). These clusters form clearly coherent structures in the spring model layout.

two or three genes. To facilitate fair comparisons of interaction degree at different levels, we designed filtering criteria and only included genes that participated in functionally relevant clusters at every level (See Sec. B.1.6).

With functional modules characterized at regularly defined levels of functional specificity, we are equipped to answer questions about where individual interactions fall in relation to modules at each level. We first asked what the density of genetic interactions was within clusters at each level. Fig. 3.14-A shows a simplified diagram of the hierarchical clustering concept. At each level, the figure shows the density within functional modules. Densities are shown directly as a function of functional specificity in panel B. Negative interactions for both essential and non-essential genes become more frequent with increasing functional specificity, and do so at a relatively constant rate. Results for positive interactions among non-essential genes behave similarly, though the rate increases more quickly. Interestingly, positive interactions between essential genes (green solid line), actually get sparser with an increase in functional specificity. These results are consistent with the observations made in Sec. 3.3.8 where non-essential complexes showed enrichment for both negative and positive genetic interactions within themselves, but essential complexes only showed enrichment for negative interactions. These results suggest that these properties are not confined to protein complexes, but describe more general properties of functional modules which hold over both broad and narrow definitions of functional modules. Put another way, structural properties of the genetic interaction network are not solely the pervue of small, specific, sets of genes with well-defined function. They are apparent at every level, and describe connections within and between even broad functional neighborhoods.

Fig. 3.14 poses the question, if two genes fall into the same cluster at level X, what is the probability of them having a genetic interaction? However, an equally important question comes in the form of the converse question, if two genes share a genetic interaction, how likely are they to belong to the same cluster at level X? Or, put another way, what fraction of genetic interactions are actually functionally specific. Fig. 3.15 answers the latter question for varying definitions of functional specificity derived from our hierarchical clusters. It shows that even when using level 2, the loosest definition of functional "locality," less than half of all genetic interactions are local, regardless of sign or essentiality of participants. At the levels of locality generally

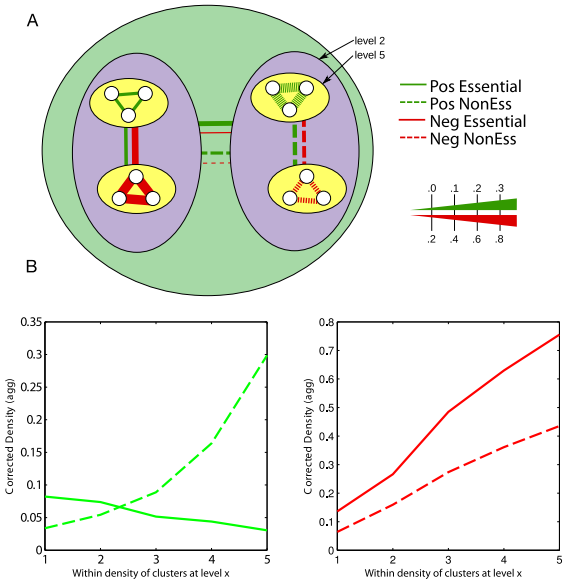

Figure 3.14: Interaction density at several levels of functional specificity. A) Conceptual rendering of nested hierarchical structures, showing real values for genetic interaction densities at levels 1, 2 and 5. Each interaction type is shown once at every level, and edge width denotes density. B) Interaction density measurements for all hierarchy levels.

considered in these studies (such as those described by protein complexes; levels 4–5) as many as 80–90% of interactions can go unaccounted for.

These observations have broad implications for the interpretations of genetic interactions, which are often based on intuitions like those seen in Fig. 3.2 and developed in Sec. 3.3.8. While those models are certainly useful, and in one sense correct, they can be misleading because they actually account for a small fraction of the genetic interaction network. Instead, the majority of genetic interactions connect genes "long-distance." Interestingly, some measure of this distance can be found in the strength of genetic interactions. Fig. 3.15-B shows the fraction of genetic interactions falling within functional clusters at levels 1–5, binned by the strength of the genetic interaction. The relative number of screened pairs in each level form a baseline for random expectation (bars labeled "rnd"). The figure shows that more than half of extremely strong negative interactions ($\epsilon < 0.64$) fall within a cluster at the highly specific level 5. This represents a striking enrichment over expectation and demonstrates power in the quantitative nature of genetic interactions obtained through SGA. The trend is quite smooth, with

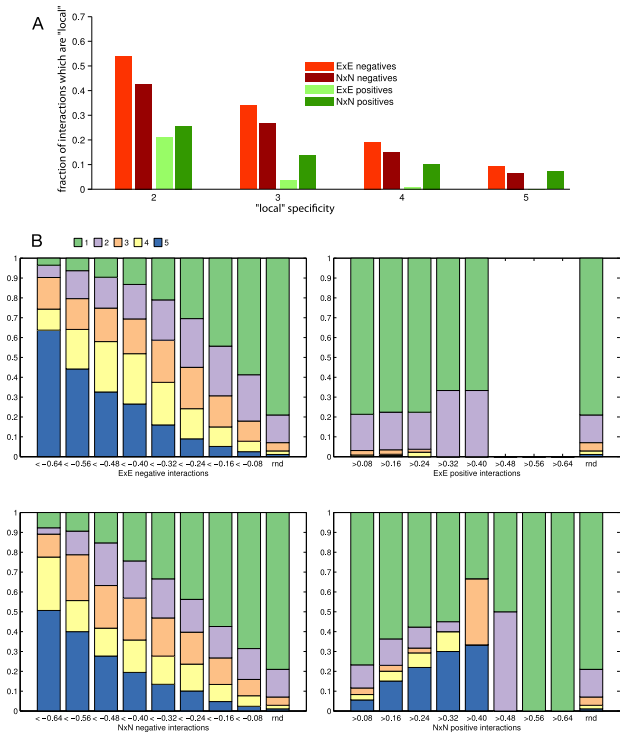

Figure 3.15:
Genetic interaction locality. A) Each level in the hierarchical scheme is in turn treated as a definition of "local" and the plot shows the fraction of all mapped interactions which are local by that definition. B) All significant genetic interactions are binned by magnitude, and within each bin the fraction of genetic interactions falling within cluster at each level is shown. The result demonstrates the extent to which the magnitude of a genetic interaction has power to predict its own functional specificity.

less and less information about the locality of both essential (ExE) and non-essential (NxN) negative interactions until the interaction magnitude falls to insignificant levels. Positive interactions between non-essential genes also show a trend between interaction magnitude and locality, though the effect is truncated as a result of SGA's reduced dynamic range for measuring positive interactions. Still, nearly a third of strong positive genetic interactions ($\epsilon > 0.4$) fall within a highly localized, level 5 cluster. Positive interactions for essential genes once again prove the exception. Across a wide range of values for ϵ, very little can be inferred about interaction locality. This is largely due to the fact that such functionally specific positive interactions between essential genes are so rare, exhibiting a decrease in density, even as that density is measured in smaller and smaller volumes.

3.4 Conclusions

The mapping of the first eukaryotic genetic interaction network is almost complete. This work represents not only a new scale in genetic interaction analysis, but also, for the first time, the systematic mapping of genetic interactions between essential genes. Analysis of these essential gene interactions has shown striking differences between the essential and non-essential segments of the network. The inclusion of essential genes now allows us to embed the study of some of the most important cellular components, such as the proteasome, in with more established methods of analysis for non-essential genetic interactions.

Essential genes were shown to have more interactions, and these interactions are more informative than those provided by non-essential genes. This increased informativeness extends beyond individual interactions, as we have shown that essential-gene profiles provide a boost to functional predictions of other essential and non-essential genes alike. The character of these interactions is also quite different from that of non-essential interactions, an observation which forces us to update well established conceptual models about the meaning behind negative and positive interactions respectively.

More broadly, the completion of the genetic interaction network allows us to study its structural characteristics at the broadest levels. Here we use clustering analysis to

demonstrate the hierarchical organization of cellular function, and describe the density and locality of genetic interactions at multiple levels of resolution, again uncovering a fundamental difference between essential and non-essential genetic interactions.

Chapter 4

Genetic interactions and the evolutionary trajectories of duplicate genes

4.1 Chapter Overview

This chapter continues the theme of the last chapter concerning genetic interactions derived from fitness observations of double-mutant yeast strains. Whereas the last chapter gave a broad overview of the entire genetic interaction network, this chapter focuses in the genetic interactions of duplicated genes. The central theme is that a single gene often has the capacity to compensate for the functions of a closely related duplicate sister, and that this capacity has consequences in the genetic interaction network. This buffering ability causes functions which are common to both duplicate copies not to manifest as genetic interactions, causing their overall degree to be lower than expected and reducing their genetic interaction profile similarity. Instead, genetic interactions are best suited as indicators of which of the pairs functions have diverged. I show this divergence to be asymmetric, that this asymmetry is correlated to asymmetries in other evolutionarily relevant properties, and offer a computational model encoding the minimal set of assumptions required to produce asymmetric divergence.

This chapter has been adapted from a previously published study entitled "Genetic

interactions reveal the evolutionary trajectories of duplicate genes" [78], on which I was the first author. The article version was published in 2010 in Molecular Systems Biology.

The statistical analysis in the paper was carried out by me, though discussions regarding the development of the models and their refinements generally involved my advisor, and my colleague Jeremy Bellay. Specifically, the buffering model as presented was initially developed by myself. The model describing self sustaining asymmetry and the discrete version of the accompanying proof were written by me, as was all of the necessary simulation code performed all of the statistical analysis. Data was collected in various forms from many previous publications, and most but not all, of that data was collected by me. Exceptions to this generally represent contributions from other lab members who are using the data for unrelated projects. First drafts of all biological examples were written by me, then sent to corresponding experts for additions and revisions.

Jeremy Bellay helped in the development of the models and with the writing in many intermediate drafts. Gabriel Musso and Balazs Papp are two biologists with expertise in the study of duplicate genes. They helped me form my early intuitions as I got up to speed in the field of duplicate gene evolution, and they also provided insight about potential directions of particular interest to scientists in relevant communities (e.g. the evolutionary biologists, duplicate gene researchers, and the yeast community at large). They also provided critical comments on the final draft, and a few of the intermediate drafts. Franco Vizeacoumar is an expert in the field of chromosome segregation and was brought on to help us develop one particular biological example (Cik1/Vik1), which was moved into the supplementary material for the final draft. Anastasia Baryshnikova and Michael Costanzo were the two principle authors on the SGA paper from which the data was taken [66]. Their principle contributions to this paper came mainly in the final stages before submission where Anastasia helped with aesthetic modifications to figures, and Michael helped in the development of biological examples, and editing the final draft. Charles Boone and Brenda Andrews are principle investigators at our collaborating labs at the University of Toronto. They provided critical feedback on the interpretation and presentation of the models we developed, and along with Chad Myers, oversee many projects seeking to explore and understand genetic interactions.

4.2 Introduction

Gene duplication is a primary mechanism for generating functional novelty, because it allows for the relaxation of selective constraints and thus provides an opportunity for functional innovation or specialization [74]. Genome sequencing studies in several species have revealed that a sizable fraction of many genomes are duplicated and that paralogous genes retain a relatively high degree of sequence similarity [42, 43]. In addition to the similarity of nucleotide/amino-acid sequence, functional genomic studies have identified significant overlap between duplicate genes in terms of their physical interactions [129, 130, 131, 132], fitness effects [133], metabolic activity [134, 90] and gene expression patterns [135], providing further evidence to suggest that functional similarity among duplicate gene families has been actively retained for over millions of years [42, 136].

Genetic interaction analysis offers another means to assess functional relationships between duplicated genes. A genetic interaction refers to an unexpected phenotype not easily explained by combining the effects of the individual genetic variants [137]. This phenomenon is also generally referred to as epistasis by the statistical genetics and evolution communities and can refer to phenotypes that are either aggravated (synergistic combinations) or alleviated (antagonistic combinations) in combination with other variants. Synthetic lethality represents an extreme form of negative genetic interaction in which mutation of a single gene, although having little or no effect on the organism, results in cell death when combined with mutation of a second gene [122, 138]. Negative genetic interactions are often taken as evidence of a functional relationship and, as a result, can be used to directly assess the extent of functional redundancy between genes. Indeed, a systematic survey identified negative interactions between 35% of gene pairs arising from the whole-genome duplication (WGD) event [100]. This rate represents an approximately 20-fold enrichment over random pairs and confirms that functional redundancy is pervasive among duplicate pairs [139, 140, 100]. Despite this wealth of data, we lack models that reconcile the long-term preservation of redundancy among duplicate genes with their patterns of functional divergence.

Synthetic genetic array (SGA) methodology enables large-scale analysis of genetic interactions in yeast [88, 83, 66], which can extend our view beyond individual duplicate

pair interactions to systematically examine the subsets of genetic interactions between duplicate genes and the rest of the genome. Analogous to studies based on protein-protein interactions (PPIs), the number of negative genetic interactions for a given duplicate pair and the extent to which their interactions overlap should provide insight into functional similarities and relationships between duplicate gene pairs. Furthermore, genes belonging to the same biological pathway or protein complex often share similar profiles or patterns of genetic interactions [83]. As a result, genes can be assigned into specific pathways or complexes by virtue of their genetic interaction profile similarity, as measured across a large fraction of the genome [83, 66]. This approach was adopted to examine the interaction profiles for 90 duplicate genes within a functionally biased subset of gene deletion mutants queried against itself [141]. This analysis showed that even though duplicate genes display negative genetic interactions with each other, they also appear to behave like singleton genes, in that they exhibit numerous unique genetic interactions; the authors suggest that duplicates are functionally redundant but have divergent roles because they often fail to provide a genuine backup when another gene is deleted [141].

In the current work, we explore evidence for duplicate gene redundancy in their genetic interaction profiles and further explain the previously observed lack of similarity among the interaction profiles of duplicate gene pairs [141]. Specifically, we propose that the established ability for many duplicate genes to buffer one another under certain conditions should cause genetic interactions related to common functions to be hidden from our experimental method. Furthermore, as duplicates evolve away from complete redundancy, non-overlapping genetic interactions should appear, reflecting their divergent roles. We find evidence to support these hypotheses in a genome-wide collection of quantitative genetic interactions in *Saccharomyces cerevisiae* [66]. We show that exceptions to the model provide insight into evolutionary mechanisms of duplicate gene retention by distinguishing partially redundant genes maintained because of their functional divergence [74, 142, 143, 144, 145] from those pairs retained because increased gene dosage is beneficial to the organism [146, 147, 141]. Finally, we provide evidence based on genetic interaction profiles supporting an asymmetric model of divergence, and show a connection between genetic interaction asymmetry and other physiological and phylogenetic properties.

4.3 Results and Discussion

4.3.1 A hypothesis about the buffering of genetic interactions after gene duplication

We hypothesize that immediately after a duplication event, duplicate genes are identical and presumably redundant, and thus, the only genetic interaction that either paralog exhibits should be with its sister gene (Fig. 4.1A and B). Such a scenario cannot persist without selection pressure to maintain the now redundant copies [148]. As the pair diverges, the selective pressures that maintained the ancestral gene will begin to act on each duplicate copy individually, creating unique genetic interactions (Fig. 4.1C). Implicit in this hypothesis is the fact that genetic interactions are buffered and undetectable immediately after a duplication event, and then are gradually revealed in one sister duplicate or the other as the pair diverges (Fig. 4.1C). The interactions that emerge after duplication may include the original ancestral genetic interactions that were buffered by the duplication or they may reflect a new function unique to one member of the pair, instances of sub- or neo-functionalization, respectively.On the basis of this hypothesis in which common functions are buffered, genetic interactions should reveal how paralogs have diverged, but seldom reveal their common functions. Requisite to this reduction in common interactions is the ability of a duplicate gene to partially compensate for the loss of its sister, which has been well established in previous studies (Fig. C.1A; [133, 141, 139, 140, 100].

4.3.2 Large-scale SGA data confirms an enrichment of negative genetic interactions among duplicates

To first affirm previous evidence for duplicate redundancy, we extracted genetic interactions for 576 duplicated *S. cerevisae* gene pairs (461 WGDs and 115 small-scale duplicates (SSD); see Materials and Methods C.6) from our recent quantitative and genome-scale SGA analysis [66]. This study captures both negative interactions, those in which the double mutant was less fit than expected (synergism of mutation effects), and positive interactions, those in which the double mutant was more fit than expected (antagonism of mutation effects). Because our SGA study focused on only genetic interactions involving two genes, we restricted our analysis to two-gene duplicate families.

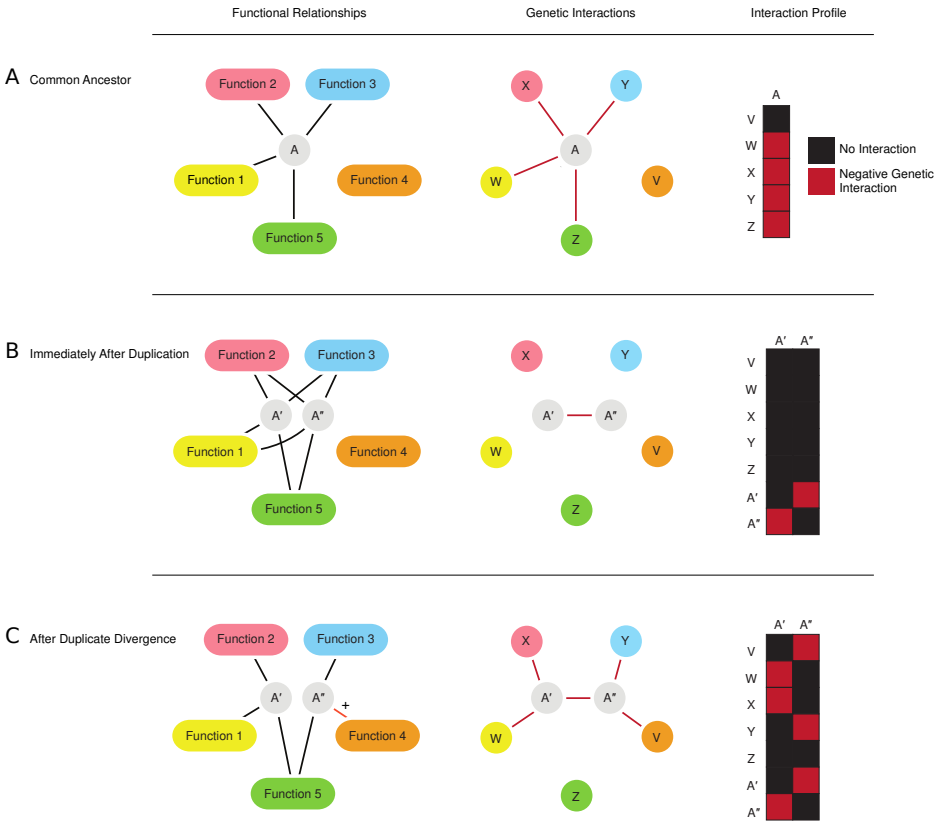

Figure 4.1: A model for the buffering of genetic interactions by partially redundant genes. The figure illustrates the relationship between a functional membership network, the observable genetic interaction network and the corresponding genetic interaction profiles, over the course of a duplication event and subsequent divergence. **(A)** Gene A has no redundant partner and its set of functional relationships is revealed through negative genetic interactions. The interaction profile for gene A is complete. **(B)** Immediately after duplication, genes A' and A" are fully redundant and their functional relationships are shared. Because each is capable of performing their common functions without the other, the deletion of A' and A" have negligible effects and do not exhibit negative interactions with any other genes. However, the simultaneous deletion of A' and A" reveals the original phenotype of their ancestor, and thus shows a negative genetic interaction. **(C)** A' and A" diverge, the redundancy becomes incomplete and unique deletion consequences emerge for each duplicate. Some of the negative genetic interactions observed for the ancestor gene A are not observed following duplication and divergence; for example, despite the functional relationship between A' and A" and Z, negative interactions are not observed with Z. A" has evolved a new relationship with function 4(+). A' lacks this ability and thus we see a genetic interaction between A" and V.

A primary requisite of the duplicate buffering hypothesis is that sister duplicates should show negative genetic interactions with each other, indicating at least partial redundancy among paralogs (Fig. 4.1C). We found a striking enrichment for negative genetic interactions between sister duplicates (67/205 pairs; 33%; Fig. 4.2A; Table 4.1), which was consistent with previous findings (35% [100]; 34% [140]; 55% [139]). This is substantially higher than the negative genetic interaction rate among randomly selected gene pairs (1.8%; [66], as well as the corresponding rate between physically interacting pairs (7%; $p < 5 \times 10^{-23}$; Fig. 4.2A; see Materials and Methods C.6) or pairs sharing specific functional annotations (4%; [94]). Although enrichment was observed for both WGD and SSD paralogs, the genetic interaction rate was significantly higher among WGD pairs ($p < 5 \times 10^{-2}$; Fig. 4.2B; see Materials and Methods C.6), supporting the greater retained functional overlap observed in general among WGD paralogs [130, 149]. However, when ribosomal duplicates are removed from consideration, the difference between WGD and SSD is no longer significant (See Appendix C.1 for more information on ribosomal duplicates).

4.3.3 Genetic redundancy between duplicates causes disparate interaction profiles

Our hypothesis about duplicate gene buffering suggests that duplicate genes will show fewer genetic interactions with other genes, because they functionally buffer one another (Fig. 4.1). Indeed, we found that duplicate genes, on average, exhibit 34 interactions compared with 55 interactions observed for singletons when assayed against a set of $\sim$ 1,700 functionally diverse query mutant strains ($p < 6 \times 10^{-16}$; Fig. 4.2C). Notably, the decrease in negative genetic interactions is more apparent on gene families consisting of more than two members. Only 5% (29/554; $p < 1 \times 10^{-27}$; see Materials and Methods C.6) of duplicates belonging to large gene families exhibit negative genetic interactions with each other, illustrating the impact of higher-order buffering and/or condition-specificity among repeatedly duplicated genes. To control for the tendency of certain classes of genes toward duplication [150, 151], we examined the number of genetic interactions (union) across a range of double-mutant fitness values, and confirmed that the deficit in genetic interactions is not due to a bias in duplicates toward gene pairs that are not important under the experimental conditions studied (Fig. C.1B).

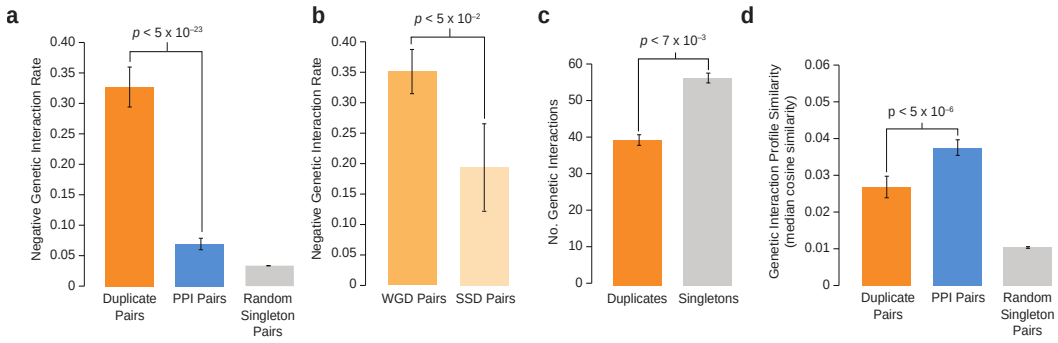

Figure 4.2: The distribution of genetic interactions supports the duplicate buffering hypothesis. **(A)** The proportion of negative interactions among screened pairs for duplicate pairs, singleton pairs with a protein-protein interaction (Materials and Methods C.6) and random singleton pairs. Error bars represent the error on a binomial proportion ($p < 5 \times 10^{-23}$; Binomial proportion test). **(B)** The proportion of negative interactions among duplicate pairs differs between modes of duplication. Whole- genome duplicates (WGD) exhibit a slightly higher rate of negative interaction than their small-scale duplication (SSD) counterparts ($p < 5 \times 10^{-2}$; Wilcoxon rank-sum). The rate of negative interactions within SSD pairs is still much higher than related singletons (Fig. 4.2A), indicating that the functional overlap observed within duplicate pairs is not solely driven by WGD pairs. **(C)** The number of genetic interactions (both positive and negative) is plotted for all non-essential duplicates and singletons. Genes shown represent those found on the SGA deletion array and thus the counts represent the number of query genes with which a given array gene shows an interaction (see Materials and Methods C.6). Means are shown and error bars represent one standard deviation of the mean over 1000 bootstrapped samples of the distribution. ($p < 6 \times 10^{-16}$; Wilcoxon rank-sum) **(D)** Although duplicate genes show far greater profile similarity than random pairs, they show significantly less similarity than physically interacting pairs ($p < 5 \times 10^{-6}$; Wilcoxon rank-sum). Median cosine similarity is shown (Materials and Methods C.6). Error bars represent the standard deviation of the median over 1000 bootstrapped samples.

In addition to fewer genetic interactions, our hypothesis suggests that sister duplicates should not share many interactions in common despite common function (Fig. 4.1C). Indeed, we found that sister duplicates share an average of 1.2 negative genetic interaction partners, whereas genes encoding physically interacting proteins (a proxy for functionally related genes) share an average of 7.2 negative interactions (see Materials and Methods C.6). This trend extends beyond the counting of discrete interactions to more continuous measures of genetic interaction profile similarity. Duplicate pairs exhibit lower interaction profile similarity than functionally related gene pairs or genes encoding physically interacting proteins ($p < 5 \times 10^{-6}$; Fig. 4.2D; C.6). The lack of genetic interaction profile similarity among a number of partially redundant duplicate pairs was previously observed in Ihmels *et al.* [141], in which the authors attribute the phenomenon to incomplete buffering, that is, divergence. Differing genetic interactions certainly convey differentiation of function; however, our updated model (Fig. 4.1) allows us to additionally explain how profile dissimilarity can also be a consequence of retained functional overlap. Thus, genetic interaction profiles for duplicate pairs are dissimilar, both for reasons of functional redundancy and divergence.

4.3.4 Dosage duplicates are exceptions to the buffering model

Assuming duplicate redundancy, our hypothesis about duplicate gene buffering suggests that only genetic interactions resulting from functional divergence will be observable. However, this reasoning should not apply to an important class of duplicate genes, namely, those selected for increased protein product [74, 141]. For example, Ihmels *et al.*, noted that duplicates expressed in high abundance have retained very similar expression profiles, indicating the cell's need for both copies simultaneously. In general, if the cell benefits from higher gene dosage immediately on duplication, then the overlapping function of the duplicate copies is not truly redundant and should induce interactions in both sisters' profiles. Indeed Ihmels *et al.* [141], noted several examples of high-abundance duplicates with significantly correlated genetic interaction profiles. Thus, dosage duplicates appear to behave differently in the genetic interaction network than duplicates retained because of functional divergence.

To determine whether genetic interaction profiles could generally distinguish duplicates under dosage selection, we first compiled a set of likely dosage-related duplicates

based on independent phylogenetic and genomic data (see Materials and Methods C.6). Using a combination of sequence and gene expression- related metrics, we defined a class of 80 putative "dosage" duplicate pairs. Importantly, this class was enriched for known dosage-mediated paralogs [146, 147, 141]. For example, 23 of the 80 pairs were ribosomal duplicates, which represents a significant enrichment ("Translation" GO term; $p < 3 \times 10^{-5}$; hypergeometric cdf). Furthermore, deletion of one of the dosage paralogs resulted in a more severe fitness defect than other paralogs, suggesting that the dosage duplicates tend to lack the redundancy exhibited by other duplicates (Figs. C.2B,C). The overall proportion of dosage pairs in our set is relatively low ($\sim 14\%$), but this is likely a conservative estimate for duplicates in general (Fig. C.2A). Indeed, we found that dosage duplicates exhibit strikingly different characteristics in the genetic interaction network. Specifically, dosage duplicates show significantly greater genetic interaction profile similarity than other duplicates (Fig. 4.3A). In fact, dosage duplicates are statistically indistinguishable from highly correlated singleton gene pairs that encode physically interacting proteins (Fig. 4.3A; $p > 0.4$; Wilcoxon rank-sum test; Materials and Methods C.6).

We speculated that the buffered interactions of non-dosage duplicates (for example, A'–Z and A"–Z in Fig. 4.1C) could be present in the genetic interaction profiles of functionally related genes that lack a duplicated partner. To identify these functionally related "proxy" genes, we focused on genes encoding proteins that exhibit physical interaction with both protein products of a duplicate gene pair (Fig. 4.3B; Materials and Methods C.6). We reasoned that these proxy proteins may have physically interacted with the ancestor of the duplicates and, thus, have a genetic interaction profile resembling that of the ancestor gene. Subsequent to duplication, either these interactions were distributed uniquely between the modern copies (sub-functionalization) or new functions arose (neo-functionalization) as the pair diverged. Comparing the genetic interaction profiles of the duplicate genes with their corresponding proxy, we found that the large majority of divergent duplicate gene profiles are more similar to the proxy gene profile than to their corresponding sister's profile (Fig. 4.3C). In contrast, dosage-mediated duplicates more often show higher profile similarity to each other than they do to the proxy gene (Fig. 4.3C), suggesting that these genes tend not to buffer one another. Thus, genetic interaction profile similarity appears to be an effective way to

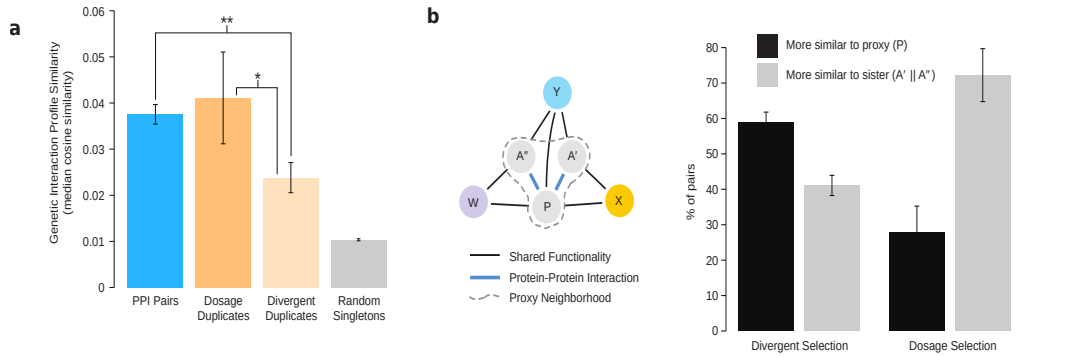

Figure 4.3: Global and local genetic interaction similarity comparisons support selection distinction. **(A)** Profile similarity shown as in Fig. 4.2D. Duplicate pairs have been separated into dosage and non-dosage (divergent) classes (Materials and Methods C.6). Divergent duplicates show significantly less profile similarity than either dosage duplicates or singletons showing a physical interaction ($p < 5 \times 10^{-2}$; $p < 1 \times 10^{-3}$; Wilcoxon rank-sum test). Dosage duplicates are not statistically distinguishable from physically interacting singletons. **(B)** A hypothetical functional network is shown that contains a duplicate pair (A'/A"). A proxy gene (P) is identified by finding a protein that shares protein-protein interactions with both duplicates (see Materials and Methods C.6), and P is used to approximate the genetic interaction profile of the common ancestor (that is, A). The number of times a duplicate's similarity with its sister exceeded its similarity with P is shown as a percentage, and error bars represent error on a binomial proportion. Dosage and divergent pairs are counted separately. In terms of genetic interaction profiles, divergent pairs more closely resemble their common neighbor than they do each other. In contrast, dosage pairs more closely resemble each other. The probability that these two classes come from the same binomial distribution is small ($p < 9 \times 10^{-5}$).

distinguish dosage duplicates from duplicates undergoing functional divergence.

4.3.5 Duplicates exhibit asymmetric genetic interaction patterns

On the basis of the buffering model, genetic interaction profiles should reflect the unique roles of duplicate genes undergoing functional divergence. Ohno [74] hypothesized that once a duplicate begins to accumulate mutations, the selection pressure will focus on the duplicate retaining the ancestral function and, therefore, most of the divergent changes should be confined to one copy. Although controversial [152, 153, 154, 155], evidence supporting such asymmetric divergence has been extracted from duplicate sequence data [156, 157, 42, 158], PPIs [152, 159], and expression patterns [135, 160].

The distribution of genetic interactions within each duplicate pair strongly supports a model of asymmetric evolution. We examined the ratio of unique negative genetic interactions for each pair of duplicates (max:min, see Materials and Methods C.6) and found that the ratio exceeds 4:1 for 430% of gene pairs surveyed (109/351), and more than 17% (60/351) of duplicate pairs exhibit a ratio greater than 7:1 (Fig. 4.4A). The observed interaction ratios are significantly greater than expected under a null model of symmetric interaction ($p < 1 \times 10^{-100}$; Wilcoxon rank-sum test; see Materials and Methods C.6), suggesting that genetic interactions tend to appear preferentially in one member of each duplicate pair.

We suspected that the asymmetric distribution of genetic interactions could be partially explained by asymmetric rates of sequence evolution, which provide an independent measure of selection pressure. Previous work showed a correlation between protein dispensability and evolutionary rate among duplicate genes [161]. A recent study of WGD pairs has also shown that both sisters undergo a period of accelerated change, but while one of them evolves much slower and is preferentially retained across different yeast species, the other evolves much faster and is preferentially lost [155, 158]. Interestingly, we found a related trend in which the rapidly evolving member had fewer genetic interactions than the more slowly evolving partner in 34/51 of previously defined asymmetric duplicate pairs ([42]; $p < 0.02$; binomial). The bias was more pronounced for pairs whose unique genetic interaction degree ratio exceeded 7:1. In this case, the rapidly evolving member was associated with a lower interaction degree for 27/38 pairs belonging to this group (Fig. 4.4B; $p < 7 \times 10^{-3}$). Furthermore, there was a significant correlation between the disparity in sequence evolution rates and the asymmetry of interaction degree ($r = 0.318$, $p < 0.03$), suggesting that the magnitude of asymmetry in genetic interaction degree was predictive of asymmetry in selection pressure acting on duplicate gene sequences. Interestingly, the set of duplicates with asymmetric evolution rates is significantly depleted for dosage-mediated pairs ($p < 2 \times 10^{-3}$; hypergeometric cdf; See Appendix C.2).

In searching for physiological evidence to corroborate the marked asymmetry in interaction degree, we examined PPIs involving gene pairs with the most extreme ratio of genetic interactions (7:1). Of these, 35 pairs exhibit at least one PPI for each member, and for 25/35 (71%) of these pairs, the partner with more genetic interactions also

tended to have retained or gained more physical interactions ($p < 9 \times 10^{-3}$; binomial; Fig. 4.4B). Genetic interaction degree asymmetry as a measure of selection pressure is also predictive of measurements of single-mutant fitness, wherein we observed that the partner with more genetic interactions has a larger impact on fitness when deleted ($p < 2 \times 10^{-8}$; binomial; Fig. 4.4B). We observed a similar trend with the number of chemical environments in which each duplicate sister displays a phenotype [56], wherein the duplicate sister with the higher genetic interaction degree generally had a higher chemical-genetic degree ($p < 3 \times 10^{-5}$; binomial; Fig. 4.4B; see Materials and Methods C.6). Interestingly, these trends between duplicate sisters mirror similar trends related to genetic interaction degree across the whole genome [66, 162].

We also found that WGD sisters with more genetic interactions tend to have higher sequence similarity to the remaining member of the pair in other WGD species (*S. castellii*, $p < 2 \times 10^{-3}$; *Candida glabrata*, $p < 1 \times 10^{-2}$; binomial; Fig. 4.4B; see Materials and Methods C.6). Specifically, in 11 of 13 instances in *S. castellii* and in 12 of 16 such cases in *C. glabrata*, the higher degree sister showed higher sequence identity to the single remaining WGD sister. Additionally, the duplicate sister with more genetic interactions tended to have a greater mRNAexpression level [163] for 32 out of the 51 pairs (63%; $p < 0.046$; binomial), although this difference was not significant in an independent expression level study [164]. Interestingly, we found that the rate of negative interactions between sisters in the asymmetric set was 46%, which is no less than the background rate for duplicates (Fig. C.3A), indicating retained functional overlap for even these highly skewed pairs.

The asymmetric distribution of genetic interactions among duplicate pairs motivated us to question whether the overall deficit of genetic interactions among duplicate genes is a result of buffered interactions distributed in both duplicate copies evenly or rather in only one paralog. Strikingly, we found that, on average, one of the two duplicates had a comparable or larger number of interactions than singletons while the sister has significantly fewer interactions (Fig. 4.4C). The slightly higher number of interactions for the high-degree duplicate gene appears to be a result of an important bias among the ancestors of the duplicates, as they became statistically indistinguishable from singleton genes after controlling for gene importance (Fig. C.3B). Thus, the overall deficiency of duplicate genes for genetic interactions (Fig. 4.2C) as well as the asymmetric distribution

of modern interactions (Fig. 4.4A) suggests that the majority of the interactions of the common ancestor are associated with a single member of the pair.

4.3.6 Dissecting the divergent functions of duplicates through genetic interaction profiles

Genes belonging to the same biological pathway or protein complex tend to share similar patterns of genetic interactions, and similarity between genetic interaction profiles has proven effective for predicting gene function and defining pathway and complex membership [66]. In this study, we exploited genome-wide genetic interaction profiles along with specific interactions to identify the functional differences that distinguish divergent gene pairs. For example, *SSO1* and *SSO2* encode SNARE proteins, core components critical for the specificity of membrane fusion and intracellular transport in eukaryotic cells [165, 166]. Although vesicle fusion with the plasma membrane is dependent on either *SSO1* or *SSO2* gene function, previous studies have shown an *SSO1*-specific requirement for prospore membrane formation during sporulation [167, 166]. We noticed that genes involved in chitin biosynthesis (*CHS3, CHS5* and *SKT5*) and polarized cell growth (*BUD6, BEM3*, and *AXL2*) shared genetic interactions in common with *SSO1* ($r > 0.14$; Table 4.1; see Materials and Methods C.6) but not with *SSO2* ($r < 0.04$), suggesting a specific role for *SSO1* in these processes during vegetative growth. These genetic interaction profile similarities support previous observations from high-content screening experiments, indicating that *SSO1* is important for normal actin localization, and deletion of *SSO1* results in more severe actin mislocalization (21%) compared with a *sso2*Δ mutant strain ([168] 4%; Fig. C.4).

We found that *SSO1* and *SSO2* also varied extensively in terms of their interaction degree. In fact, the ratio of *SSO1:SSO2* interactions was among the most asymmetric, with 149 negative interactions for *SSO2* compared with only 15 negative interactions involving *SSO1*. Consistent with evolution of a condition-specialized function, previous studies suggest that functional divergence has led to a more prominent sporulation-specific function for *SSO1* [167, 166]. The reduced number of interactions observed for *SSO1* may reflect its specialized function, in part, because genetic interactions were

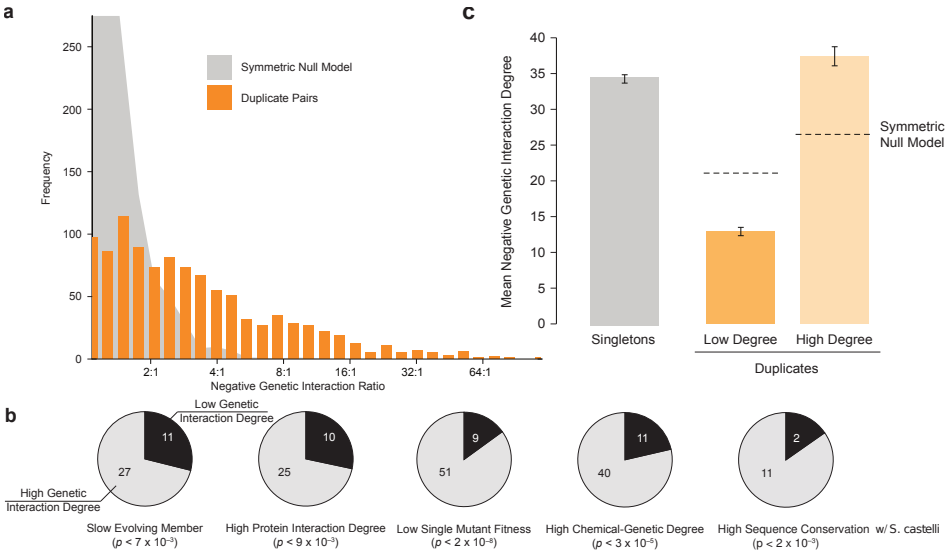

Figure 4.4: Genetic interactions provide evidence for asymmetric functional divergence. **(A)** A histogram of the duplicate interaction degree ratio. The ratio is defined for unique interactions with the higher degree in the numerator. Pairs included must have at least 10 total interactions between them, with each member having at least one interaction. Shown for comparison is another degree ratio histogram in which interactions for every duplicate pair are redistributed to either member with equal probability (symmetric null model). **(B)** Relating selection pressure measures on asymmetric duplicate pairs. Pairs with a unique interaction ratio exceeding 7:1 (60 pairs) are compared across several different sequence or functional genomic data sets. Each gene was put into the high or low interaction degree bin by comparison with its sister. Each pair was then examined for agreement in directionality with the indicated data set. For example, in 27 out of 38 pairs, the sister with higher genetic interaction degree also has a lower rate of sequence change. Comparisons with < 60 pairs reflect missing pairs in the secondary data set. Also shown are p-values resulting from a binomial test in which genetic interaction degree is assumed independent of the other data type. **(C)** The number of negative genetic interactions for singletons and duplicates. Each duplicate pair was sorted by genetic interaction degree and means are shown. Dotted lines represent the same process applied to the simulated distribution from Fig. 4.4A. The difference between high-degree duplicates and singletons is significant (34.9 versus 37.2; $p < 5 \times 10^{-8}$; Wilcoxon rank-sum); however, the mean number of singleton interactions is reduced by a large portion of singletons with no measurable deletion effect, and the significant difference presented here subsides when controlling for gene importance (Fig. C.3B).

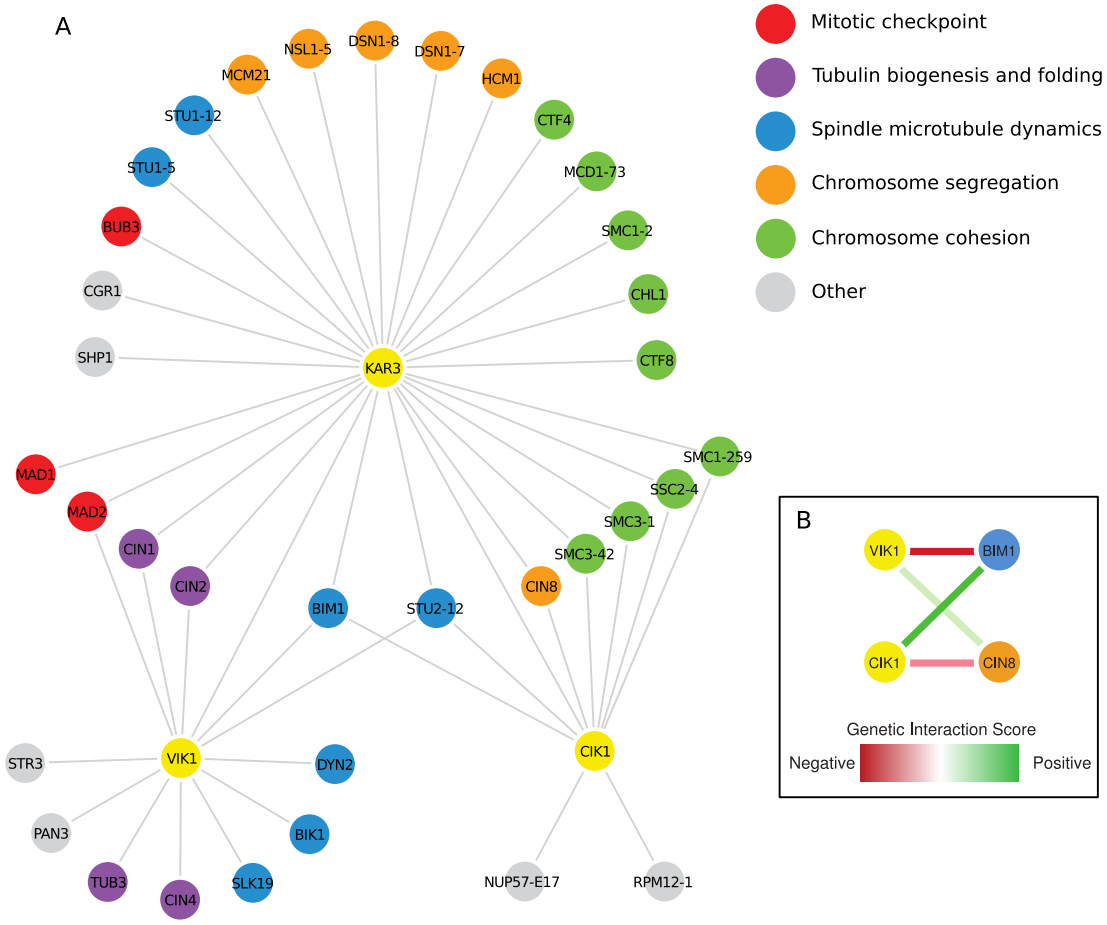

Figure 4.5: Functional analysis of duplicate pair *CIK1–VIK1* **(A)** Genetic interaction profile similarity. Similarity scores were taken from Costanzo *et al.* [66] and represent a combination of array side and query side correlations (Materials and Methods C.6). Nodes shown include all first neighbors of the three primary genes of interest (*CIK1, VIK1* and *KAR3*). A threshold of 0.2 was used as in Costanzo *et al.* [66] and edges between first neighbors of genes of interest have been removed for clarity. **(B)** Genetic interactions. SGA genetic interaction scores from Costanzo *et al.* [66] highlight differences between *CIK1* and *VIK1*. Green lines represent positive interactions, whereas red lines represent negative interactions. The opacity of the line is proportional to the strength of the interaction.

	CHS3	*CHS5*	*SKT5*	*BUD6*	*BEM3*	*AXL2*
SSO1	0.212	0.200	0.245	0.288	0.144	0.194
SSO2	-0.016	-0.060	-0.001	0.032	-0.082	-0.001

Table 4.1: Select profile correlations for duplicates *SSO1* and *SSO2* This table shows differences in composite profile correlations which highlight functional differences between *SSO1* and *SSO2*. Scores shown reflect both array side and query side interactions and are taken from Costanzo et al. 2010[66]. *SSO1* shows high profile similarity with genes involved in chitin biosynthesis (*CHS3, CHS5, SKT5*) and polarized cell growth (*BUD6, BEM3, AXL2*), suggesting a specific role for *SSO1* in these processes during vegetative growth. *SSO2* lacks genetic interactions in common with these genes and thus exhibits very poor similarity. Genetic profile similarities reported here support previous observations from high-content screening experiments indicating *SSO1* is important for normal actin localization and deletion of *SSO1* results in more severe actin mis-localization (21%) compared to a *sso2*Δ mutant strain (4%, Fig. C.4)

mapped under vegetative conditions when sporulation is not required. In a similar example, highly asymmetric genetic interaction degree may reflect sporulation or meiosis-specialized function for cell wall assembly duplicates *GAS1* and *GAS2*, suggesting that this may be a common basis for imbalances in genetic interaction degree (See Appendix C.3).

Genetic interaction profile examination yielded another interesting example in duplicate pair *CIK1/VIK1*. Comparison of profile similarity and interaction degree of *CIK1* and *VIK1* demonstrates the ability of genetic interaction analysis to distinguish subtle functional differences between paralogous genes. *CIK1* and *VIK1*, which arose from the WGD event, encode kinesin-associated proteins that form separate heterodimeric complexes with Kar3, a minus-end-directed microtubule motor protein, to mediate a diverse set of microtubule-dependent processes [169]. Despite strong sequence and structural similarities, *CIK1* and *VIK1* exhibit different genetic interaction profiles, suggesting that these proteins have specialized functional roles. Although both proteins depend on physical interaction with Kar3 for proper function, *CIK1* has more genetic interactions in common and is more closely correlated to the *KAR3* interaction profile (*CIK1–KAR3*; $r = 0.5$; see Materials and Methods C.6) compared with its duplicate *VIK1* (*VIK1–KAR3*; $r = 0.3$). Consistent with closely related interaction profiles (Fig. 4.5A), *kar3*Δ and *cik1*Δ deletion mutants share several phenotypes including abnormally short spindles, chromosome loss and delayed cell cycle progression [170, 169]. In contrast, a

*vik1*Δ mutant strain does not exhibit any overt phenotype [169].

In addition, *VIK1* and *CIK1* differ in their gene expression and protein localization [169]. Interestingly, we found that *CIK1* and *KAR3* interaction profiles more closely resemble the profiles of genes involved in chromosome cohesion and segregation (GO:0000070; $p < 8 \times 10^{-8}$; hyper-geometric cdf; Fig. 4.5A), whereas *VIK1* was more correlated to genes involved in microtubule assembly and stabilization (GO:0007017; $p < 2 \times 10^{-8}$; Fig. 4.5A). Our findings support a previous hypothesis [169] and suggest that the Cik1–Kar3 and Vik1–Kar3 heterodimers serve distinct, yet related, roles during cell division. In addition to profile similarity, examination of individual genetic interactions also highlight potential functional differences between these microtubule motor-associated proteins. We noticed strong asymmetry in the ratio of *CIK1:VIK1* interaction degree and, consistent with a more severe deletion phenotype, we found that *CIK1* has 4.5-fold more negative genetic interactions than *VIK1*. Interestingly, several genetic interactions connecting *VIK1* and *CIK1* to common partners differ in their type. In particular, the plus-end microtubule motor encoding gene, *CIN8*, shares a modest positive genetic interaction with *VIK1*, whereas a *cik1*Δ–*cin8*Δ double mutant displayed a synthetic sick/lethal phenotype (Fig. 4.5B). Findings derived from our large-scale survey of genetic interactions support previous observations that disruption of *VIK1*, but not *CIK1*, partially suppresses the temperature-sensitive growth defect of a *cin8–3 kip1*Δ double mutant [169]. One role for the Kar3 microtubule motor during vegetative growth is thought to involve opposing the action of the Cin8 and Kip1 motor proteins. The *VIK1*-specific positive genetic interactions reported here and elsewhere [169] suggest that a *CIN8* and *KIP1* antagonistic function may be unique to the Vik1–Kar3 heterodimer, thus distinguishing between Vik1–Kar3 and Cik1–Kar3-related functions. In another example, we found that *BIM1* shared a positive interaction with *CIK1* (*bim1*Δ suppressed the *cik1*Δ growth defect) and a negative interaction with *VIK1* (Fig. 4.5B). Bim1 is a microtubule-binding protein that localizes to the plus end of the microtubules where it is required for proper positioning of the nucleus during nuclear migration [171, 172]. Recent studies have shown that Bim1 also localizes to the spindle midzone to stabilize microtubules during anaphase [173]. Interestingly, Kar3 also exhibits different sub-cellular localization patterns that are dependent on physical interaction with Vik1 or Cik1. During vegetative growth, Kar3 associates with the

spindle midzone in a Cik1-dependent manner [174], whereas the Kar3–Vik1 heterodimer localizes to the spindle poles [169, 175]. Although the nature of the genetic interactions is unclear, the negative interaction between *BIM1* and *VIK1* might reflect the failure in nuclear positioning due to unstable microtubules while positive interaction observed between *BIM1* and *CIK1* might reflect opposing functions involved in stabilizing and destabilizing the microtubules [174, 173].

In both pairs of duplicates we investigated in detail (*SSO1–SSO2* and *CIK1–VIK1*), the duplicate genes exhibited a strong negative interaction between sisters. This suggests that despite evidence for functional specialization and dramatic asymmetry in their overall interaction degree, sister duplicates retain the ability to partially compensate for the loss of one another, and this trend appears to be relatively common across duplicates in yeast (Fig. C.3A). We also noted that, although genetic interactions can resolve functional differences between sisters, in these cases, the differences appear to be relatively subtle: context or conditional specialization in the case of *SSO1–SSO2* and localization specialization in the case of *CIK1–VIK1*.

4.4 Conclusions

We examined how partial redundancy and the functional divergence of duplicate gene pairs relates to their genetic interaction profiles.We found evidence for the hypothesis that immediately after duplication, duplicated gene pairs will mask each other's interactions with other genes, and that as the pair evolves apart, interactions reappear, highlighting functional differences between them.We have also shown that genome-wide genetic interaction profiles provide insight into the mechanisms of duplicate gene evolution by distinguishing duplicate pairs maintained for gene dosage effects from those retained because of functional divergence. These findings clarify previous observations about the surprising prevalence of genetic interactions for apparently redundant duplicate genes [141], and provide evidence that they do indeed reflect functional redundancy as well as functional divergence. Finally, we also showed that a disproportionate distribution of genetic interactions among gene pairs supports the asymmetric evolution of duplicate genes whereby one member of a duplicate pair is under stronger selective

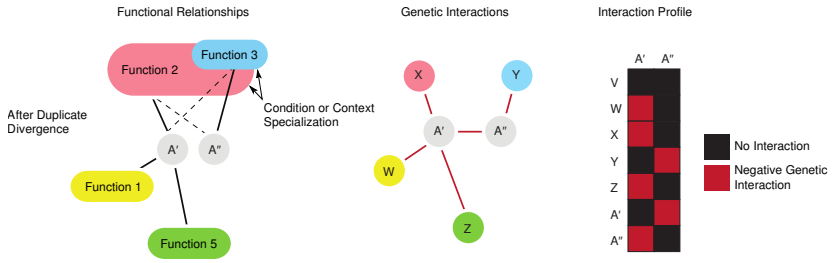

Figure 4.6: Updated model of asymmetric duplicate genetic interaction evolution. Asymmetry is rapidly established through the absence of purifying selection on a duplicate pair, but in rare cases, the quickly evolving duplicate confers a fitness advantage through functional or context specialization (Function 3). Subsequent selection on Function 3, however, also maintains a limited capacity of duplicate A" to carry out Function 2 (dotted lines). In this scenario, there is overlap in function, but the efficacy of the duplicate pair with respect to a particular function differs, and so the buffering is asymmetric. Fewer genetic interactions are observed for A" either because of its less constrained function or because of its role in other environmental or developmental contexts.

pressure. The skewed distribution is correlated with differences in rates of sequence evolution, PPI degree, single-mutant fitness defects and sensitivity to a variety of chemical environments, suggesting that one member of the gene pair assumes a predominant role under standard vegetative growth conditions.

Previous studies suggest that the asymmetric accumulation of loss-of-function mutations in many duplicate pairs is established quickly based on sequence evidence from the WGD event that indicates that the identity of the quickly evolving sister is consistent across several yeast species [154, 155, 158]. On the basis of these observations combined with results from this study, we propose a refined model of duplicate evolution (Fig. 4.6). Following a duplication event that does not provide a dosate-dependent fitness advantage, we argue that one member of a duplicate pair should accumulate loss-of-function mutations more quickly due to relaxed purifying selection alone (See Appendix C.4; Figs. C.5,C.6). In essence, a degenerate paralog is more accommodating of mutations and stands a higher chance of sustaining a mutation affecting any remaining redundant functions (See Appendix C.4). In many cases, the fast evolving duplicate meets the common fate of non-functionality and eventual gene loss. If early function loss is complementary, the pair is put on a path toward functional partition. Gene properties

that are necessary for multiple functions may be preserved in both copies if previous mutations caused these functions to fall to different sisters. Such an arrangement would render a complete functional divergence impossible. We note that this natural progression of asymmetry should occur for any duplication event, either whole-genome or small-scale, although the means of preservation of a duplicate pair might be distinct depending on the context. Presumably, in some cases, sister duplicates simply maintain complementary but essential roles despite their asymmetry, whereas in other cases, the asymmetric configuration provides some fitness advantage that ultimately enables a selective sweep.

We cannot rule out the possibility that neo-functionalization may have a role in the preservation of some duplicate pairs and their subsequent asymmetric evolution, but if that is the case, the quickly evolving duplicate appears to take on a more inconspicuous functional role in most pairs. Our data argues against dramatic neo-functionalization and instead suggests that the rapidly evolving duplicate retains a subset of the ancestral function for which it has become optimized (Fig. 4.6). Importantly, despite specialization, the high rate of negative genetic interactions observed between asymmetric duplicate pairs (Fig. C.3A) indicates that the lower degree sister often retains some ability to compensate for the loss of the more constrained sister. We do not interpret this as evidence for selection on their redundancy, rather that the function or context for which the quickly evolving duplicate has been specialized allows or requires it to at least partially maintain the ancestral role (Fig. C.5).

Our observations are consistent with previously proposed models of sub-functionalization, including the Duplication-Degeneration-Complementation and Escape from Adaptive Conflict models [142, 176, 75]. Both these schemes describe ancestral functions being split between duplicates, the latter allowing for optimizations previously constrained by other functions. Indeed, we identified several gene pairs in the yeast genetic interaction network that support specialization driven by adaptation to different environmental or developmental conditions, leading us to speculate at a special case of the Escape from Adaptive Conflict or Duplication-Degeneration-Complementation models may apply to a large fraction of duplicates in *S. cerevisiae*, in which this specialization is driven by adaptation to different environmental or developmental conditions.

For example, several of the most asymmetric pairs involve a gene specialized for sporulation or meiosis. Sporulation requires formation of a membrane structure known as the prospore membrane, which is dependent on the Sso1–Spo20 t-SNARE complex.

Although in vitro experiments indicate that both Sso1 and Sso2 can bind to Spo20 to form a functional t-SNARE, the Sso2–Spo20 complex exhibits much weaker membrane-fusion capacity and, thus, may explain why only Sso1 is able to support sporulation [177]. Furthermore, studies have shown that Sso1 can interact with phosphatidic acid, which is necessary for Spo20 localization and function [177]. Although the exact cause of functional divergence remains unclear, it is possible that the *SSO1* gene product acquired a specialized role after duplication, which is important for modulating *SPO21* function in non-dividing cells. This example supports our model illustrating that changes in protein function are often relatively subtle, and condition or developmental specialization may instead be the driving force behind duplicate gene retention. Although genome sequences provide a wealth of information about gene ancestry, they fail to address the functional efficacy of genes on which selection ultimately acts. Network analysis of PPIs [13] provide a complementary view, but common physical interactions shared by a duplicate pair still do not reveal whether interaction with a specific member of a duplicate pair has a functional consequence to the cell under a given experimental condition. Genetic interactions address both of these shortcomings by revealing exactly which relationships have an impact on fitness, and which do not, and thus provide a powerful perspective for understanding duplicate gene evolution.

Chapter 5

Complete functional profiles of paralogs revealed through trigenic genetic interactions

5.1 Chapter Overview

Like the two previous chapters, this chapter considers genetic interactions, which result from surprising combinations of lower order phenotypes. Here I follow the previous chapter's focus on duplicate genes, and their missing genetic interactions to its natural conclusion by exploring triple-mutant combinations involving many of the same duplicate pairs. The resulting three-gene (or trigenic) genetic interactions have not yet been studied at scale and shed additional light on the processes of duplicate evolution and divergence.

The experiments in this section were in large part conceived by me, and grew out of my previous work from the preceding chapter. The experimental design is the product of joint work between Elena Kuzmin and myself. I also developed the trigenic scoring model and was principally responsible for processing the trigenic interaction data and relevant control data. I performed all of the statistical analysis and functional predictions in this chapter, and developed the tools for computational simulation of subfunctionalization.

Elena Kuzmin served as the project's lead biologist in Toronto. She constructed all of the novel strains used in the project, and oversaw their progress through the SGA procedure. She was also heavily involved in the conception and design of the experiment, including the fundamental adaptations of SGA protocols to triple mutants. Additionally, she performed all of the validation experiments in this chapter. Raamesh Deshpande developed the tool which was used to select informative genes for inclusion in the mini-array. Justin Nelson developed a visual screening system for reviewing colony images during initial design stages. Michael Costanzo, Charlie Boone, and my advisor—Chad Myers—provided crucial input and direction throughout the project.

5.2 Introduction

Gene duplication has long been recognized as one of the primary sources of new genetic material in many genomes [74]. The processes by which genes come to have duplicate copies are reasonably well understood, but the longterm consequences of gene duplication, and the rules governing the retention and divergence of pairs that survive the process are not. Many duplicate pairs (or paralogs) show evidence of substantial levels of retained functional overlap despite millions of years of opportunity for divergence. For example, all genes in an ancestor of yeast were duplicated in what is called a whole-genome duplication event (WGD). Despite the approximately 100 million years since the WGD event in yeast, nearly 35% of WGD pairs show a synthetic lethal relationship (negative genetic interaction See Fig. 4.2). The level of retained sequence conservation, common protein-protein interactions, metabolic activity, gene expression patterns, and conserved sub-cellular localization patterns also suggest a surprising level of functional overlap between extant sister paralogs.

In contrast to these potential mechanisms of functional overlap (or divergence), genetic interactions give us an indication of the actual functional consequence of genetic perturbations, without the mechanistic underpinning. In the previous chapter, I laid out evidence for retained functional overlap between paralogs in the genetic interaction network. In addition to the high rate of synthetic lethality, these included an above-average single-mutant fitness for paralogs, a reduced genetic interaction degree, and a lower than expected profile similarity. These observations were consistent with the

expectation that functions that were retained in both members of a paralog pair, would fail to exhibit genetic interactions, and that genetic interactions instead would inform us only about divergent functions in each pair.

If functions carried out by both members of a duplicate pair fail to show interactions because of buffering, we can recover them by removing the buffer. By deleting both paralogs simultaneously, we can recover genetic interactions between the pair (as a single functional unit) and any third gene of interest. We have termed these three-gene genetic interactions as "trigenic" (as opposed to double-mutant or "digenic" interactions), and here present the first attempt to discover them in high-throughput. Only two studies have previously attempted to characterize the trigenic interactions of duplicate pairs, and between them they address only three duplicate pairs. This study attempts to map the trigenic interaction for over 200 pairs, including every pair of WGD duplicates for which the double mutant is not already inviable. To accomplish this feat, we required not only new experimental procedures, but a new model to remove the pairwise, or "digenic," components from each of our triple-mutant observations.

Not only are frameworks for the generation of trigenic data scarce, but systems for their interpretation are also only just emerging. Here we show that different classes of trigenic interactions exist, and they have different relative frequencies. Each of them has different possible interpretations, a property they have in common with recent differential digenic interaction studies but that has thus far gone under-appreciated.

Additionally, I showed in Chapter 5, that duplicate pairs that show a high degree of divergence tend to diverge asymmetrically. This asymmetry is much more pronounced than expected and the directionality of genetic interaction asymmetry between sisters agrees with other physiological and evolutionary measures, such as protein-protein interaction degree and sequence similarity with an non-duplicated ortholog in a related species of yeast (See Fig. 4.4). I go on to show that asymmetry is the result of multiple gene functions overlapping in the regions of gene sequence which carry them out.

Some computational models of duplicate divergence are sufficient to explain asymmetric divergence while others do not. Additionally, some of these models have difficulty in explaining the amount of retained common function in many duplicate pairs. The addition of trigenic interaction data to the genetic interaction profiles of duplicate genes fills in the missing information about ancestral function. Now when considering the

functional divergence of a paralog pair, we have all of the functions in hand, and can obtain a true estimate of functional divergence versus retained common function. In this chapter, I explore the relationship between divergence and common function in an attempt to unify these models and explain why some duplicate pairs diverge completely, while others retain such a pronounced level of common functionality (and therefore buffering ability).

5.3 Results and Discussion

5.3.1 Trigenic scoring model

Previous work

Low throughput triple mutant analysis has been used previously to reveal buffered functions with respect to various phenotypes in *S. cerevisiae* (e.g. [178, 179]), including cases involving duplicate pairs (e.g. [180, 181]). These examples have all targeted individual gene triads with respect to a qualitative phenotype of interest such as invasive growth or viability. More recently, attempts have been made to increase the throughput of triple mutant analysis by fixing two members of the triad and considering many possible candidates for the third [83, 182, 183]. These studies have examined a small number of pairs (3, 8, and 2 respectively), mostly focusing on duplicate partners, and the analysis of triple mutants has been either qualitative or semi-quantitative. Just as the power of double-mutant analysis has been greatly improved by a shift to quantitative measurements [50], so too will the power of triple-mutant experiments, but first we need a theoretically sound model for trigenic interactions.

Two earlier studies have extended quantitative systems designed for digenic interactions to gene triads. In his PhD dissertation [182], Musso used an SGA approach to construct genome-wide array screens for 8 pairs and scored the results by subtracting the two corresponding single-mutant control profiles. Thus, the trigenic interaction score (which we denote by τ) between query genes i, j and array gene k would be:

$$\tau_{i,j,k} = \epsilon_{ij,k} - \epsilon_{i,k} - \epsilon_{j,k} \tag{5.1}$$

where $\epsilon_{ij,k}, \epsilon_{i,k}$, and $\epsilon_{j,k}$ scores come from the established digenic SGA scoring procedure applied to a double-mutant query and two single-mutant control queries respectively. As we are concerned with finding interactions which are not easily explained by any of the constitutive double mutants, and the contribution from an interaction between the two query genes has been removed as part of normal scoring, it remains to remove significant effects appearing in the other two double mutants. A similar intuitive approach was applied by Haber *et al.* [183] to E-Map derived S-scores [184] (roughly analogous to SGA ϵ) by subtracting the stronger (more negative) of the two digenic interactions:

$$S_{i,j,k} = S_{ij,k} - \min(S_{i,k} - S_{j,k}) \tag{5.2}$$

Updated extension of digenic methodology

While each of these approaches does a reasonable job in removing the strongest digenic effects, neither of them use the same quantitative framework upon which the digenic models are built when extending consideration to the third gene. As a result, while the strongest effects are likely to be correctly identified, the trigenic score itself does not correctly capture differences in the same way as the digenic score does. As an example, consider the definition of ϵ under SGA methodology:

$$\epsilon_{i,j} = f_{ij} - (f_i f_j) \tag{5.3}$$

Here the expected double-mutant fitness of genes i and j is defined as the product of their single-mutant fitnesses $(f_i f_j)$ and ϵ is defined as the difference between the observed double-mutant fitness (f_{ij}) and that expectation. In other words, ϵ has a precise definition as a deviation from expectation with respect to fitness. For the same to be true of trigenic interaction scores, We have to redefine our trigenic equation to fit this same definition. Expectations are defined as multiplicative combinations in fitness (relative to wild-type), which are generally observed to be the case for pairs of unrelated genes. The expected fitness for a triple mutant deleted for three completely independent genes is therefore straightforward: $f_{ijk\ expected} = f_i f_j f_k$. However, we wish to remove influence from cases where two of the genes are not independent (say i and

j). In this case, the expected fitness of the triple mutant would not be the product of all three single mutants, but instead the product of the interacting double mutant and the unrelated single mutant: $f_{ijk\ expected} = f_{ij}f_k$. By solving Eq. 5.3 for the double-mutant fitness term, and subsequent substitution, we can rewrite this using only single-mutant fitnesses and pairwise epsilons:

$$f_{ijk\ expected} = (f_i f_j + \epsilon_{i,j})f_k \tag{5.4}$$

Thus we can see that as a result of our choice of a subtractive model in digenic space (Eq. 5.3), digenic interaction effects are scaled by the fitness of non-interacting genes when determining expectation. A trigenic interaction term for genes i, j and k, after removing a possible digenic influence would then be:

$$\tau_{i,j,k} = observed - expected = f_{ijk} - (f_i f_j + \epsilon_{i,j})f_k \tag{5.5}$$

By invoking a symmetric argument for the other two possible digenic contributions, and rearranging the terms for clarity, we arrive at the following equation for trigenic interactions:

$$\tau_{i,j,k} = f_{ijk} - f_i f_j f_k - \epsilon_{j,k}f_i - \epsilon_{i,k}f_j - \epsilon_{i,j}f_k \tag{5.6}$$

Two issues remain. First, it would be convenient for our equation to deal with epsilon scores and single-mutant fitnesses, as opposed to double- and triple-mutant fitnesses, easily accomplished as these quantities are trivially related to one another. And second, we must rearrange terms such that all of the quantities are either a single-mutant fitness, or an epsilon included in the trigenic experiment. So, assuming that there is a double-mutant query ij as well as two single-mutant control queries i, and j, along with gene k on the array, then all three single-mutant fitnesses are available as well as three epsilons ($\epsilon_{i,k}, \epsilon_{j,k}$, and $\epsilon_{ij,k}$). The last of these ($\epsilon_{ij,k}$) is defined as an interaction between mutant ij and mutant k as in Eq. 5.3:

$$\epsilon_{ij,k} = f_{ijk} - f_{ij}f_k \tag{5.7}$$

Then, by solving Eqs. 5.6 and 5.7 for f_{ijk}, and setting them equal to each other, we can solve for our trigenic interaction term ($\tau_{i,j,k}$) from known quantities.

$$\tau_{i,j,k} = \epsilon_{ij,k} - \epsilon_{i,k} f_j - \epsilon_{j,k} f_i \tag{5.8}$$

The final result, Eq. 5.8, is very similar to Eq. 5.1. Indeed, as single-mutant fitnesses approach 1, as in the case of fully redundant duplicates, the two estimates converge. The difference stems from the distribution of the fitness parameter introduced in Eq. 5.4, and is a result of our choice of subtraction for capturing the final difference from expectation. Other models for quantifying digenic interactions provide alternatives to Eq. 5.3, and each of these models has some benefit in terms of its biological, experimental, or theoretical properties [123]. Each of these models extends to trigenic interactions in different manners, with some more or less suited to the transition. Regardless, of the merits of each model in digenic space, a trigenic extension should be derived in a fashion consistent with its source. Eq. 5.8 represents a model for trigenic interactions which follows directly from our definition of digenic interactions, and therefore provides the best framework for our analysis.

5.3.2 Experimental Approach

Dual survey design

To accommodate the increased cost stemming from the need for additional replicates and controls, we adopted a two-pronged approach. First, a small number of duplicate pairs from diverse functional categories, were screened as double-mutant queries against the whole genome. The resulting "pilot-study" survey gave very detailed profiles for a small set of duplicate pairs, and is better suited for specific functional inquiries because nearly all possible interactions for those pairs are tested. Secondly, a much larger set of pairs was screened as double-mutant queries against a smaller array ($\sim$ 18% of the genome, See Sec. D.1). This "mini-array" is better suited to making generalizations about trigenic interactions and the buffering capacity of duplicate pairs because many

more duplicate pairs have been tested (though at a lower resolution).

The pilot-study consisted of 14 paralog pairs which were initially selected to try and cover the spectrum of trigenic interaction behavior. They consist of pairs with varying levels of sequence similarity, varying inter-paralog genetic interaction scores, and the majority of them were well characterized functionally. Notably, they contained the pair *CLN1/CLN2*, which had been previously tested for qualitative trigenic interactions and would serve as a benchmark of our scoring method. These 14 pairs were screened against both available SGA arrays (4,632 total array genes, See Sec. 3.2.2). Screening these pairs and their controls against all available SGA array genes means their profiles can be integrated with existing digenic interaction data and correlated with 1,346 single-mutant queries screened against the same array set.

The mini-array survey consists of 203 paralog pairs. All whole-genome duplicate (WGD) pairs or genetically interacting small-scale duplicate (SSD) pairs were considered for screening. Many duplicate pairs exhibit a strong synthetic-lethal genetic interaction, meaning that their double-mutant fitness is often quite low. Pairs with a measureed double-mutant fitness less than 0.7 were deemed too sick for the scoring procedure. These 203 double-mutant queries were screened against a custom mini-array, which contained a mix of essential and non-essential genes (temperature-sensitive and null mutations, respectively) from the two larger full-genome arrays. In total, we crossed these 203 double-mutant query pairs to 1,178 functionally informative array strains, testing for approximately 240,000 trigenic interactions. Additionally, each double-mutant query pair required two single-mutant control queries (406 in all), and each of these 609 query screens was performed in at least two replicates.

This dual approach allows us to demonstrate the power of trigenic interactions as a tool for functional investigation using complete profiles from the pilot study, while simultaneously conducing a broad survey of the trigenic interaction space for a large number of duplicate pairs using the mini-array survey.

Quality control and additional replicates

To discover trigenic interactions in the context of SGA, we have to screen three separate query strains. The first is the double-mutant query in which both paralogs have been deleted. Then we must screen each paralog again as its own single-mutant control.

To obtain final trigenic interaction scores (τ) we first scale each single-mutant control profile by the single-mutant fitness of the reciprocal paralog, then subtract both of them from the double-mutant query profile per Eq. 5.8. In effect, each observation in the final trigenic interaction profile is a composite of three experiments. This multiple observation paradigm presents potential problems with regard to the expected accuracy of each SGA profile. For example, to confidently call a specific interaction observed in the double-mutant query profile "trigenic", we have to be confident that the triple-mutant observation was not a false positive, but we also need to be confident that the double-mutant observations in the control profiles did not produce a false negative. As a consequence of this reliance on multiple observations we cannot say that reliability estimates for digenic interactions are accurate for trigenic interactions. Although the observations and the scoring model are quantitative, it is instructive to see how multiple observations impact coarser estimates of accuracy such as qualitative true positive and false positive rates.

If we assume that a double-mutant query profile, before adjustment, has similar characteristics as a double-mutant query profile, then we can calculate our expected trigenic interaction precision after adjusting for observations in the control queries. For example, let us assume that a double-mutant query profile has approximately the same sensitivity and specificity for real effects, both digenic and trigenic, as a single-mutant query. Let us assume also that they have approximately the same number of false positives. In this case, the double-mutant query profile, before removing digenics observed in the controls, will be composed of: *i*) real trigenic interactions, *ii*) real digenic interactions that were false negatives in the control screens, *iii*) real digenic interactions that were true positives in the control screens, and *iv*) false positives.

The size of class i, true trigenic interactions, depends on the real number of trigenic interactions (T), and the recall of a single screen (R). The size of class ii, digenic interactions that were recovered in the double-mutant query profile but missed in the single-mutant controls, is dependent on the number of digenic interactions, (D), and recall (R). Class iii represents the number of digenic interactions that were recovered in both the single- and double-mutant screens, and so is a function of R^2.

$$\begin{aligned} (Real\ trigenic)\ i &= R \cdot T \\ (Digenic,\ missed\ in\ controls)\ ii &= R \cdot D \cdot (1-R) \\ (Digenic,\ seen\ in\ controls)\ iii &= R^2 \cdot D \\ (False\ positives)\ iv &= (i+ii+iii) \cdot \tfrac{(1-P)}{P} \end{aligned} \tag{5.9}$$

Since classes i, ii, and iii, all represent real effects of various types, so the precision in a technical sense follows Eq. 5.10. However, the measure we are most interested in is the precision after removing any digenic interactions seen in the single-mutant query controls. After removing interactions recovered from the controls, the precision of trigenic interactions can be expressed as in Eq. 5.11

$$technical\ screen\ precision\ (P) = \frac{i+ii+iii}{i+ii+iii+iv} \tag{5.10}$$

$$trigenic\ interaction\ precision = \frac{i}{i+ii+iv} \tag{5.11}$$

If we assume that the number of trigenic interactions (T) is on the same order as the number of digenic interactions (D), then we can use Eqs. 5.9 to rewrite Eq. 5.11 as a function of the precision and recall of individual screens.

$$trigenic\ interaction\ precision = \frac{P}{P(2-R)+2(1-P)} \tag{5.12}$$

From our previous digenic experiments, we have empirical estimates for the technical precision and recall of an SGA screen as a function of the number of replicates (See Fig. 3.4). Using the values of technical precision and recall for a single replicate ($P_1 = 50\%$ and $R_1 = 40\%$ for negative interactions) we estimate our precision for identifying true trigenic interactions is only 28%. So our reliance on multiple observations for trigenic interactions has reduced our performance (in terms of precision) by almost half. Fig. 3.4 shows that screening additional replicates can greatly increase our technical precision, at little cost to recall. If we use estimates for screens with two replicates ($P_2 = 80\%$, $R_2 = 37\%$) our trigenic interaction precision increases to 47%, representing about the same level of quality as our previous digenic experiments which are conducted with one replicate of each query. A third replicate for each query would bring this overall measure

up to 50%, a marginal return in data quality for substantial increase in cost.

To summarize, as a consequence of defining trigenic interactions in terms of a combination of multiple SGA query screens (one double-mutant query, and two single-mutant query controls) the we see a reduction in data quality as the errors from each screen compound. To offset this reduction in quality of the screen combination, we must boost the quality of each individual screen by adding more replicates. Based on previous empirical estimates of precision and recall as function of replicate quantity, we conclude that two replicates of each query are required to make the quality of trigenic interactions in this study comparable to the quality of digenic interactions in previous studies. To that end, we have screened each query twice; bringing the total number of screens for each paralog pair to six.

Replicate data and reproducibility

The addition of a third perturbation represented a non-trivial change to the SGA experimental protocol, and because the bulk of the data was collected on a new array configuration, we used our replicate screens to assess the reproducibility of the method. Each double-mutant query, as well its two accompanying single-mutant controls was screened twice (See Sec. 5.3.2), and Fig. 5.1 shows the aggregate correlation between replicates, separately for single- and double-mutant queries. Single-mutant control screens have a strong Pearson correlation of 0.38, while double-mutant query screens are even more reproducible with a replicate correlation of 0.57. This likely reflects the fact that double-mutant queries show more interactions than single-mutant controls, both as a result of new trigenic interactions as well as a higher density of digenic interactions recovered from multiple perturbations, as epsilon scores for double-mutant queries in Fig. 5.1 have not yet been adjusted to remove digenic interactions seen in the controls. Similarly, we assessed the average replicate correlation on a per-query basis. Fig. 5.2 shows histograms for average pairwise replicate correlation scores, and here again we see a high degree of reproducibility. The higher average pairwise correlation for double-mutant queries in Fig. 5.2 agrees with the higher overall correlation shown in Fig. 5.1. However, the histograms show the contrast between single- and double-mutant query reproducibilities more directly. This difference is likely due to the increased amount of signal in double-mutant queries, as a result of more digenic and trigenic actions being present in

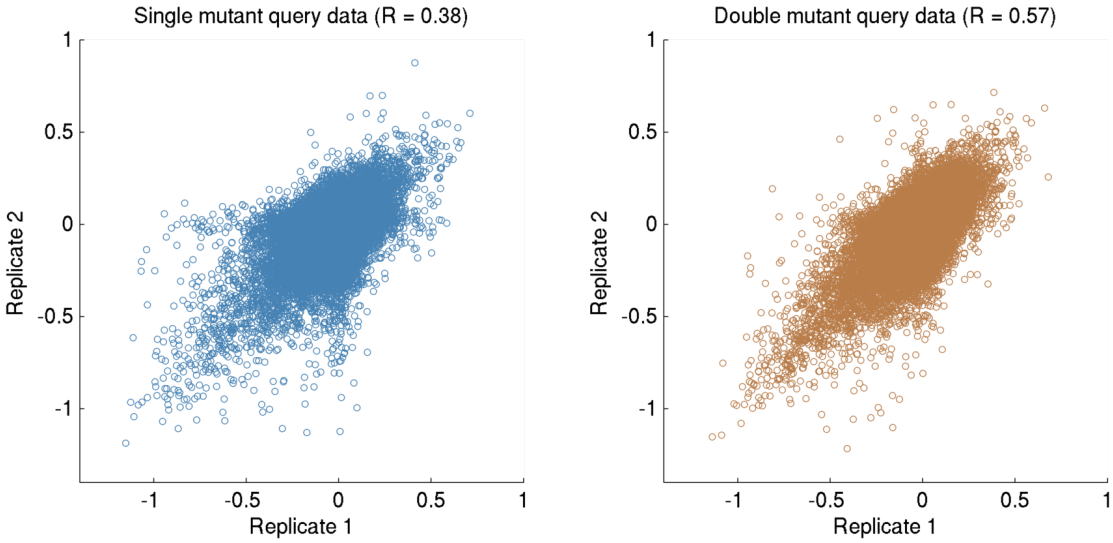

Figure 5.1: Replicate data from trigenic mini-array. For each query screened against the mini-array two replicates were selected at random. Epsilon scores tend to agree very well from one replicate to another.

the profile. The agreement between replicate screens demonstrates that the changes in the SGA protocol to accommodate a third genetic perturbation have been successful in that they are reliably producing genetic interaction data.

5.3.3 Double-mutant query profile includes digenic interactions

Central to our model of trigenic interactions is the assertion that a double-mutant query will show the same interactions as its two constituent single-mutant queries, with the addition of novel trigenic interactions. Given the significant error rates observed even in triplicate screens, we set out to demonstrate this property before we could justify applying the model laid out in Eq. 5.8. We measured how many digenic interactions in the two single-mutant controls were recovered in the profile of the double-mutant query. Fig. 5.3 shows the results of this analysis. In over a third of the double-mutant query screens, we recovered 43% of the digenic interactions observed in either of the two single-mutant query controls. Estimates for the precision of negative interactions in SGA screens with three replicates are just over 80% (See Fig. 3.4). If only 80% of the observed digenic interactions are real, that should be our upper-bound for recall in a

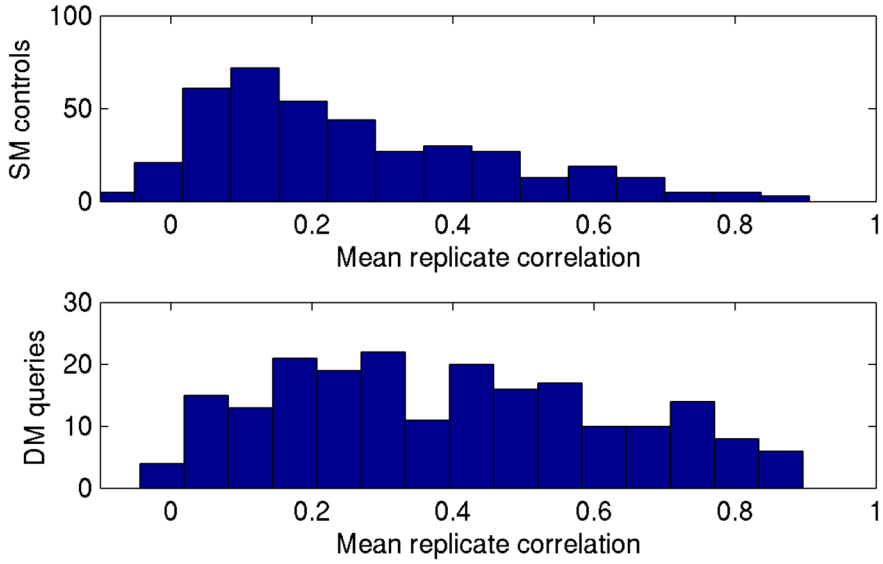

Figure 5.2: Per query replicate correlations. For each query, Pearson correlation coefficients were calculated for all available replicate pairs, and a histogram of the means is shown.

second (in this case, triple-mutant screen. This assumes that false positives are rare and their coincidental overlap is rarer still, which is justifiable given that all interactions, both genuine and spurious, comprise less than 5% of the data. The similarly derived recall estimates for a three-replicate screen is just under 40%. So if our hypothesis about digenic interaction re-occurrence in the triple-mutant screen is true, our expected recall would be $80\% \times 40\% = 32\%$. The mean recall for negative digenic interactions in a triple-mutant profile is 31%, very close to our expectation. Following another calculation using estimated parameters for positive interactions we see an expected recall of just under 10%, yet we observe a mean recall of 21%. These observations are consistent with the hypothesis that digenic interactions from both paralogs appear in the double-mutant query profile before adjustment, and that we can recover these interactions subject to the known reliability of individual SGA screens.

CLN1+CLN2 validation

After confirming that the raw digenic interaction scores (ϵ) were sufficiently reproducible, and of high enough quality to determine the presence of trigenic interactions

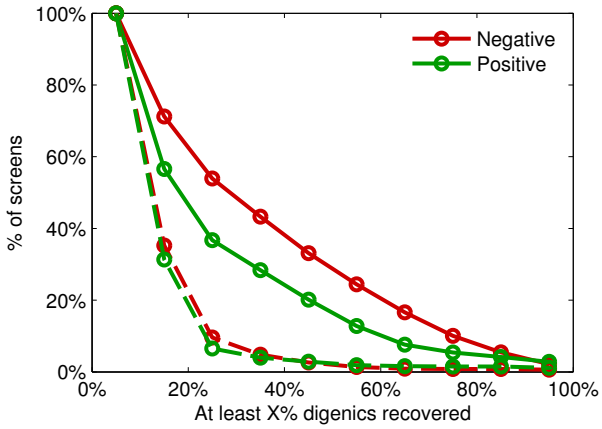

Figure 5.3: Double-mutant query recovers digenic interactions. For each of the 203 double-mutant queries in the mini-array survey, we compiled a union profile of digenic interactions from the two single-mutant control queries and assessed the ability of the double-mutant query to recover that set. The figure shows a normalized histogram for positive and negative interactions separately. For example, nearly 60% of trigenic screens recover over 20% of digenic interactions observed in the two separate single-mutant control screens.

(τ), we applied the model laid out in Eq. 5.8 to obtain trigenic interactions for the double-mutant query *CLN1+CLN2*. As stated above, *CLN1* and *CLN2* were specifically chosen for inclusion because they had previously been screened qualitatively for negative trigenic interactions by Zou *et al.* [185]. Using the pilot study set to maximize the overlap with previous work, we were able to re-test 29 of the 36 *CLN1+CLN2* trigenic interactions reported in Zou *et al.* That study was interested in genes required for optimal growth in the absence of *CLN1* and *CLN2*, but did not differentiate between digenic and trigenic effects. Indeed, our examination of colonies grown from spores isolated via tetrad dissection found six of these previously reported interactions were the result of single- or double-mutant effects and were not truly trigenic. Several others were untestable in SGA for technical reasons, for example their proximity to a locus containing a selection marker required for SGA. The remaining 18 negative genetic interactions formed our gold-standard for *CLN1+CLN2* and we were able to recover 12 of them with our automated scoring pipeline, which makes the best available estimate of trigenic interaction recall for our method 66%. This is higher than estimates for SGA

recall of digenic interactions in single replicate studies (See Fig. 3.3.4, bottom-left), however it should be noted that those results were based on an SGA-derived gold standard instead of a smaller qualitative set of the strongest interactions and so are only loosely comparable.

We also conducted a series of confirmation experiments to estimate the precision of our trigenic screen. Of the 57 strongest interactions measured for *CLN1+CLN2* on the FG_array ($\tau \lessapprox -0.2$), 82% of them were confirmed by random spore analysis. This estimate is only slightly below precision estimates reported for the strongest digenic interactions (89% for $\epsilon < -0.12$; [66]). Taken together, these results indicate that we can successfully recover known trigenic interactions using our novel experimental and computational pipeline, and that those interactions recovered are of a very high quality.

5.3.4 Summary of trigenic interactions discovered

In all, we screened for approximately 272,000 possible trigenic interactions between a diverse set of 203 double-mutant queries and the rest of the genome. In that space we discovered approximately 8,500 novel trigenic interactions, and 55% of them were negative. The total number of trigenic interactions discovered at a standard intermediate threshold of $|\tau| > 0.08;\ p < 0.05$ is shown in Table 5.1.

Estimates of trigenic interaction density vary slightly between the two collections. Negative density in the pilot-study is 1.65% whereas density in the mini-array is slightly higher (1.72%). However, the array genes chosen for the mini-array tended to show more interactions than the genome average, as did the 14 queries chosen for the pilot-study, and these biases are unlikely to offset one another exactly. For example the same density measure for the higher degree queries (from the pilot-study) when measured by themselves on the mini-array is 4.48% while those same queries from the pilot-study measured against only arrays in both collections is 2.97%, indicating there are moderate differences between the two experiments. Given these biases, the most conservative estimate of trigenic interaction density is under 2%, which is below estimates of digenic interaction density between all pairs of non-essential genes (See Chapter 3 Fig. 3.5). Also the true trigenic interaction density considers all possible triads of candidate genes, whereas here we are forcing two of them to be related (duplicates), so values for unrelated triads may be even lower yet.

Previous work has shown that number of interactions for each gene (its degree) can vary dramatically, and while genetic interaction networks are not exactly scale-free (with degrees following a power-law distribution) they none the less have a few nodes with high degree (many interactions) and a majority of nodes with few interactions [66]. While duplicate genes have fewer digenic interactions on average [66, 78], their degrees are nonetheless distributed in the same manner. The degree distribution of negative trigenic interactions is also very similar to that of negative digenic interactions with a small number of hubs and a large number of relatively unconnected genes. (See Fig. 5.4). Also, the two measures are related in that duplicate pairs that show more digenic interactions, tend to show more trigenic interactions ($p < 5 \times 10^{-8}$, Pearson). This may indicate that the number of digenic interactions we observe for a given pair may be a good indicator of the level of activity or importance of that pair under our experimental conditions. Keeping in mind that trigenic interactions (by definition of Eq. 5.8) have had the influence of digenic interactions removed, this means that the dominant factor in how many interactions we observe (trigenic or digenic) is the activity or importance of the pair, while their buffering potential has a second-order effect. However, there are many pairs which appear to exhibit predominately digenic or trigenic interactions. If paralog sisters show many digenic interactions but few trigenics, it would suggest that most of their function was already revealed in the digenic study, and that the two do not buffer each other. Conversely, if the pair show many trigenic interactions instead, it suggests that their functions significantly overlap and both must be perturbed to see an effect. Three diverse examples are shown in Fig. 5.4. The first pair, *OAF1+PIP2* (bottom left), show very few trigenic interactions in relation to their number of digenics. They also show a modest overlap in their digenic interaction profiles, which is rare for paralog pairs. This may suggest they are retained for reasons of dosage amplification. As explained in Chapter 4, pairs retained for dosage amplification may not be expected to buffer one another strongly despite their functional overlap. Oaf1p and Pip2p are transcription factors that regulate genes involved in peroxisome organization and biogenesis [186]. Normally they bind to form a heterodimer, but Oaf1p has been shown to form a functional homodimer in the absence of Pip2p [187]. On the other end of the spectrum is the familiar pair of cyclins *CLN1+CLN2* (bottom right), which show a very large number of trigenic interactions and almost no digenic

	Queries	Pos	Pos Ess	Pos NonEss	Neg	Neg Ess	Neg NonEss
mini-array survey	203	3043	738	2305	3718	982	2736
mini-array query average		15	3.6	11.4	18.3	4.8	13.5
pilot study	14	796	547	249	919	444	475
pilot study query average		56.9	39.1	17.8	65.6	31.7	33.9

Table 5.1: The total number of discovered negative and positive trigenic interactions, given seperately for the mini-array survey and the pilot study for the standard thresholds of $|\tau| > 0.08;\ p < 0.05$. See Sec. 5.3.2 for details including the number of array genes tested in each set. Pos and Neg refer to the total number of positive and negative interactions. Ess and NonEss displays how that total breaks down when only essential and non-essential array genes are considered. Query average rows give the mean number of interactions for each double mutant query in each set.

interactions, which is consistent with their near complete ability to compensate for one another. The pair displayed in the middle show a relatively modest number of trigenic interactions and a few shared digenic interactions, indicating some retained functional overlap, but a great deal of divergence as well.

These examples illustrate a broad variation in trigenic interaction behavior relative to digenic interactions. If we consider all observed interactions as a complete representation of pair function, then the *trigenic proportion* of those interactions reflects the ability of paralogs to functionally buffer each other. We will return to this measure to assess evolutionary divergence below(See Sec. 5.3.8).

5.3.5 Functional validation of novel trigenic interactions

To ascertain if trigenic interactions are as functionally relevant as digenic interactions at the same threshold, we used them to predict functional relationships as captured in the Gene Ontology (See Fig. 5.5). The differences for negative interactions in the pilot study are not very pronounced, both are hugely informative and in this case, it can be said that their performance is about equal. The excellent performance of digenic interactions in the pilot study is in part due to our selection bias toward well-annotated queries which were known to behave well in SGA. The functional information in the resulting pilot study trigenic interactions therefore meets a very high standard. The blue line shows the typical performance for paralog digenic interactions, functionally informative but at a rate less than two fold over background. In this more general case,

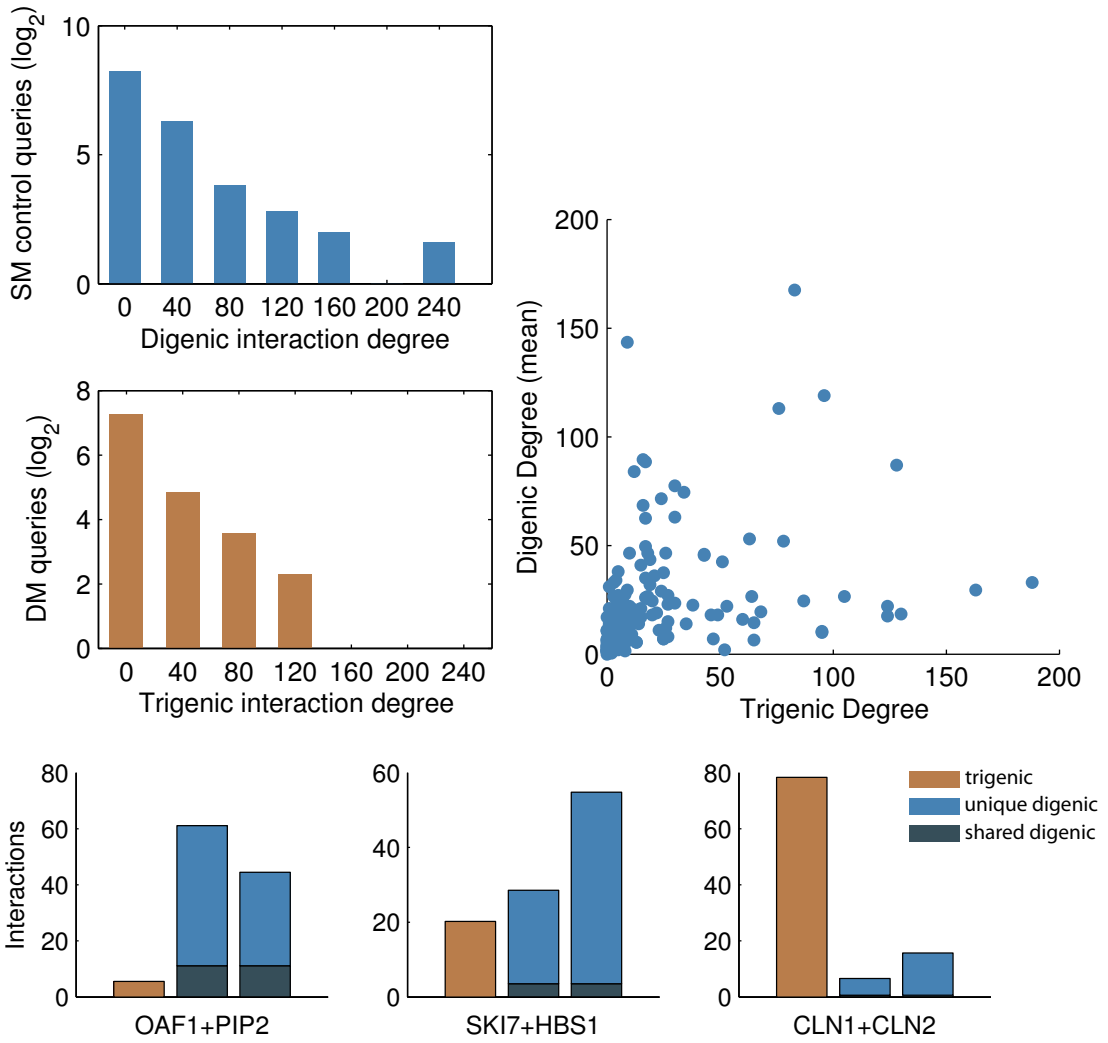

Figure 5.4: Trigenic degree distribution from the mini-array survey. (Left top) A histogram of the (negative) digenic degree distribution of all 406 single- mutant (SM) control queries. The Y-axis (number of queries in each bin) has been $\log_2$ transformed. (Left middle) A similar histogram showing the (negative) trigenic degree distribution of 203 double-mutant (DM) queries. (Right top) A scatter plot showing the trigenic degree of each double-mutant query and the corresponding digenic degree (mean) of its two single-mutant control queries. The Pearson correlation between the two is 0.37 ($p < 5 \times 10^{-8}$). A similar plot for all 14 pairs in the pilot study can be found in Fig. D.1 (Bottom) Trigenic/Digenic degrees for three duplicate pairs. For each pair the number of trigenic interactions for the double mutant is shown in orange, the number of digenic interaction unique to each single-mutant control is shown in light blue, and the number of digenic interactions which appear in both controls in dark blue.

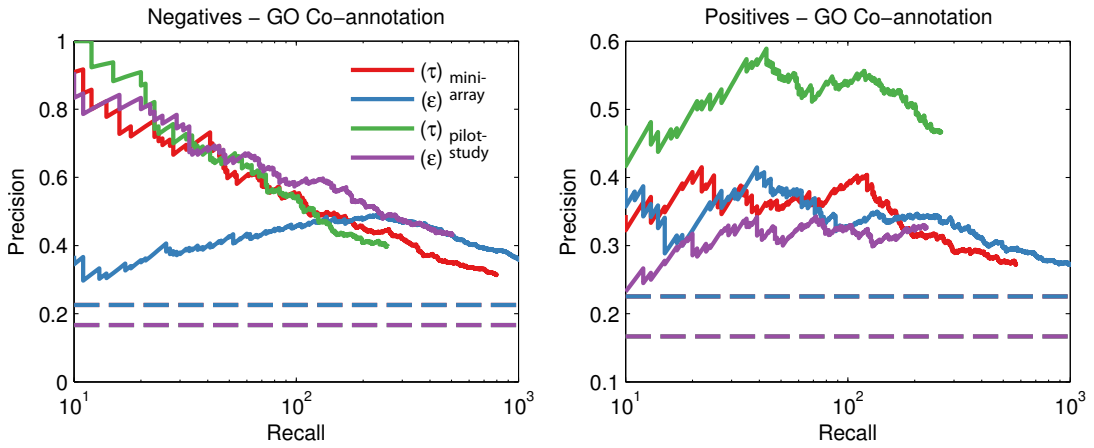

Figure 5.5: Novel trigenic interactions predict gene function. Precision and recall are shown for trigenic interactions (τ) and digenic interactions (ϵ) in each survey. The prediction standard is based on co-annotation a functionally informative subset of "process" terms in the Gene Ontology. A trigenic interaction is considered a true positive if the array gene is co-annotated with both query paralogs, though in practice this is not much more stringent that requiring only one. The background expectations for co-annotation differ between the two surveys and are shown as dotted horizontal lines.

negative trigenic interactions (red line) again do exceedingly well at predicting known associations.

Positive trigenic interactions are not as informative as negative interactions, but still perform above random expectation (Fig. 5.5, right). This observation is consistent with digenic interactions in this and previous studies (See Fig. 3.6) [66]. For positive interactions in the general case (mini-array) trigenic interactions are about as informative as digenic interactions. Both are about 1.5-fold over background expectation, and the relatively flat slope of the precision-recall plots indicates that the magnitude of the interactions does not carry much additional information, in stark contrast to negative interactions, but again consistent with previous digenic studies. Interestingly, positive trigenic interactions for the pilot study have a precision of nearly 50% (right, green), or more than two-fold over expectation. Some of these positive trigenic interactions align with negative digenic interactions (left, purple) and represent cases where array associations could be observed in digenic interactions but these interactions did not combine as expected when an additional paralog was deleted (See Sec. D.2). In other words,

because the sign and magnitude of a trigenic interaction are dependent on digenic interactions between pairwise triad members, the interpretation of the trigenic interaction must be as well.

5.3.6 Different sub-types of trigenic interactions

When measuring digenic interactions, many different types of biological relationships get summarized by a single number and then reduced to three classes: negative interactions, positive interactions, and non-interactions. For example, consider positive digenic interactions, where the double mutant grows faster than expected given the single mutants. It may be that both single-mutants show a phenotype, but their combined phenotype is not any stronger; this is the relationship suggested by Fig. 3.2. In classical terminology, one phenotype is said to "mask" the other. Another positive interaction may result from two slow-growth phenotypes that combine with a result that grows at the wild-type rate. These phenotypes are said to "suppress" one another. Moreover, these two cases may result in equivalent ϵ scores, thus their interpretation may require additional comparisons of wild-type, single- and double-mutant fitness scores.

Similar questions arise in the case of a triple mutant. To examine the interplay between our observations in the digenic and trigenic spaces, we plotted the components of the trigenic scoring equation (Eq. 5.8) against one another, and defined 14 regions in the resulting two dimensional space which may have different biological properties. Fig. 5.6 shows a map of these regions and the distribution of data among them from mini-array survey, as well as their overall ability to predict co-annotation to Gene Ontology terms. The total number of negative trigenic interactions listed in Table 5.1 for the mini-array survey is equal to the total number of points above and to the left of the red line ($\tau = X - Y < -0.08$). This single class (negative trigenics) can thus be partitioned into six sub-types of interactions (regions A, B, C, F, G, & J). The X-axis gives the strength of the triple-mutant score before we apply control data to adjust for digenic components, and the Y-axis gives the magnitude of that digenic interaction adjustment. For example, if a duplicate pair shows no digenic interactions with a particular array gene, the score will fall near $Y = 0$; if then their double mutants shows a strong negative interaction with that array gene, the point will be fall to the left and be classified in region F. Since the strength of the trigenic interaction depends on both of these factors

($\tau = X - Y$), a similar trigenic score can be obtained, for example, when a weaker triple-mutant score is paired with a positive digenic interaction between one paralog and the array gene (which would fall into region A), though these scenarios have different biological implications. When viewed in this way, several classes of region stand out as potentially interesting.

For example, region F represents the most intuitive negative trigenic interactions, such as *CLN1-CLN2-CLN3*, where the total digenic adjustment is small because neither paralog shows a digenic interaction with the array gene yet the triple-mutant score is substantially negative. Thus region F, and its mirror region for positives (H) represent *qualitative* trigenic interactions of the type that have been discovered by eye in Zou *et al.* [185]. This is the most populous class of trigenics in our study and cases of complete three-way redundancy would fall here. Notably, as most of these relationships actually require a triple-perturbation to see, it would be surprising if they did not have a lower rate of annotation.

In contrast to the qualitative class, region J indicates combinations where the digenic contribution was negative, but not strong enough to account for the triple-mutant score. Region J (and it's reflection region E) represent a class of *quantitative-agreement* interactions which all of the digenic and trigenic signs agree, but the digenic magnitudes alone can not account for the triple-mutant phenotype. This type of interaction does very well at predicting co-annotation relationships (Fig. 5.6), as well as PPI relationships. Interestingly, the negative cases of quantitative-agreement interactions (J) are very enriched for cases where the array gene has a physical interaction with one, but not both paralogs, while the positive version (E) shows the opposite trend (Fig. D.3). This class of interaction suggests cases where paralogs have the ability to partially buffer a particular genetic interaction. Perhaps, these interactions represent functions where dosage plays a role, thus while both paralogs can perform the common function in question, neither of them can fully compensate if the other is totally deleted. Quantitative-agreement is the second most populous class of trigenic interactions. A related class, and the next most numerous, is *quantitative-disagreement*. This class is comprised of regions C and L, and made up of cases where the digenic interaction scores were in agreement, but the resulting trigenic score had an opposing sign. For example,

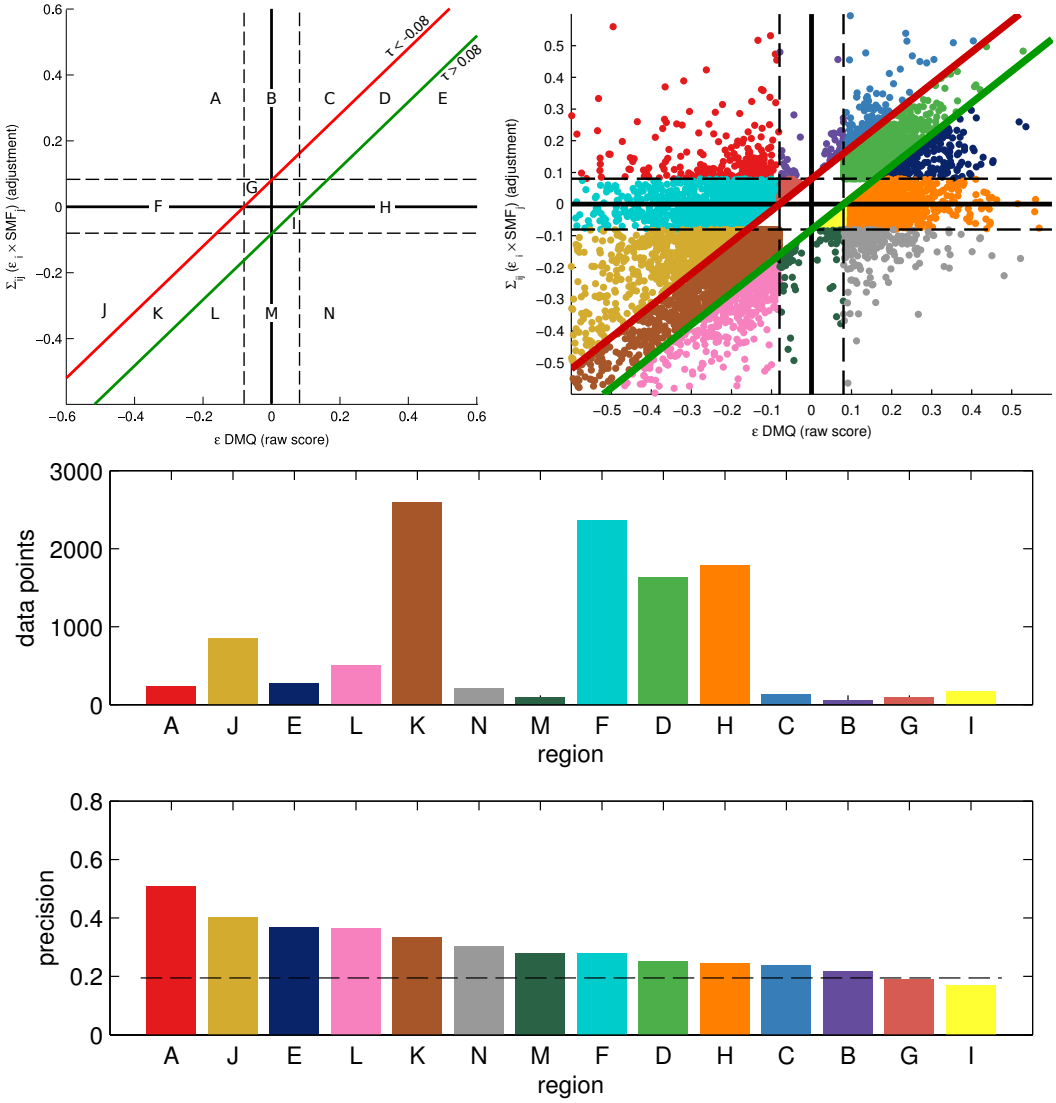

Figure 5.6: Map of trigenic interaction sub-types. (Top left) A map labeling various interesting regions in trigenic interaction space. The X-axis shows raw triple-mutant ϵ scores, before their trigenic adjustment according to Eq. 5.8. The Y-axis shows the adjustment value, which is the the sum of digenic influences. The red and green lines show the thresholds for trigenic interactions $\tau = X - Y$. (Top right) All trigenic interaction data with a significant p-value ($p < 0.05$) for all regions labeled. (Middle) A summary of the number of data points shown in each labeled region. (Bottom) Sorted precision of each region in prediction co-annotation between the array gene and both query genes. The expected co-annotation rate (derived from unlabeled regions) is shown as a dotted line. Similar results for PPI data are shown in Fig. D.3.

in region L, the digenic contribution is extremely negative, and the unadjusted triple-mutant ϵ score was also negative, but not as negative as expected and so a positive trigenic interaction results. Three of these four quantitative regions (J, E, L, but not C) do extremely well at predicting co-annotation in the gene ontology.

The final, and rarest, class of trigenic interactions are *contradictory*, and fall into either region A or N. In these regions, both the single- and double-mutant query screens give strong interactions but they oppose one another in direction. The curious cases which were alluded to in Sec. 5.3.5 and confirmed in Sec. D.2 fall into region A, where the digenic screens indicated a positive component, but a negative triple-mutant ϵ was measured, with an extremely negative trigenic interaction (τ) as a result. Interestingly, the cases in region A who reverse their sign, although not terribly numerous, appear to carry the highest functional signal. The sign reversal may be an indicator of exceptionally complex relationships between the three genes. A hypothetical example is shown in Fig. 5.7. In this example, we must reconcile the fact that gene **b** interacts with each individual paralog positively, but shows a negative (trigenic) interaction with them as a unit. If all three genes participate in some common essential function but only one is required, we would expect a negative trigenic interaction. Furthermore, if the paralogs regulate each other with positive feedback, and b suppressed them both in balance, we might expect positive interactions between b and each paralog as the deletion of the suppressor offsets the effects of breaking the feedback loop.

Importantly, many of the classes which outperform the classical interactions found in regions F and H represent relationships which could already be inferred from digenic studies (e.g. A, J, E, L), and while they represent some of the most intriguing examples, they may reveal more mechanistic information than novel functional relationships. Also, there are other ways these interactions might be partitioned, and even this scheme could be fine tuned, for example by incorporating single mutant scores or treating double-mutant scores individually instead of summing them.

The problem of differential interaction interpretation has previously been largely ignored even in differential digenic studies. However, the consideration of multiple observational combinations which may result in the same overall score will become increasingly important as the number of simultaneous experimental perturbations (and therefore phenotype combinations) increases. Frameworks for the exploration of these

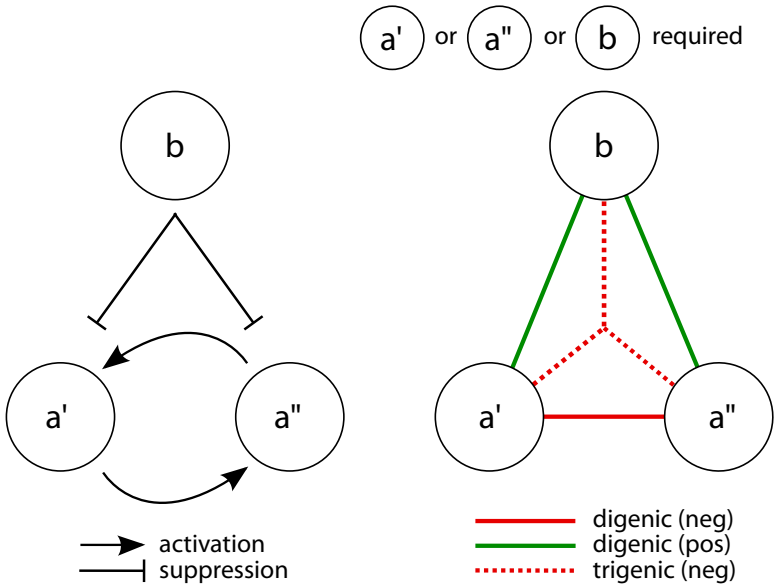

Figure 5.7:
A hypothetical arrangement for trigenic interactions in region A. Suppose **a'** and **a"** are paralogs and perform a redundant function with gene **b** such that one of the three is required for some essential function, giving rise to a trigenic interaction. In this case a regulatory relationship such as the one shown on the left my give rise to the apparently contradictory genetic interaction shown on the right.

possible combinations, and their potential biological significance, such as that presented in this section may become crucial if we are to learn as much from trigenic interactions (or indeed from differential interactions) as we have from digenic interactions.

5.3.7 Trigenic proportion as indicative of buffering capacity

The existence of trigenic interactions and their functional relevance is established, so aside from elucidating previously hidden relationships, what can they tell us about duplicate pairs? We reasoned in the previous chapter that the duplicate sisters that were most adept at compensating for one another, would be missing the most digenic interactions, and would therefore have the most trigenic interactions. To characterize this property for the duplicate pairs in this study we devised a simple measure, *trigenic proportion* (Eq. 5.13), which captures the fraction of all of the pair's interactions that are digenic. If the union of the two digenic profiles along with their shared digenic profile captures the ancestral profile, and common retained functions are buffered in the digenic profiles but are revealed in the trigenic profile, then this measure also reflects the fraction of retained functional overlap, capturing the variation observed in Fig. 5.4 in a single continuous measure.

$$trigenic\ proportion = \frac{|\tau_{ij}|}{|\tau_{ij} \cup \epsilon_i \cup \epsilon_j|} \tag{5.13}$$

A histogram of this measure for all 203 pairs in the mini-array survey is shown in Fig. 5.8. The measure does indeed show a great amount of variation, with just under half of pairs showing a small trigenic proportion ($< 30\%$), but a substantial tail with the other half of pairs displaying profiles that are least 30% trigenic or more. In the next several sections we will focus on this property and its potential physiological determinants.

5.3.8 Properties which correlate with trigenic proportion

We first filtered our pairs to include only those for which we can be confident in the measured proportion. To accomplish this, we removed any pair which showed less than

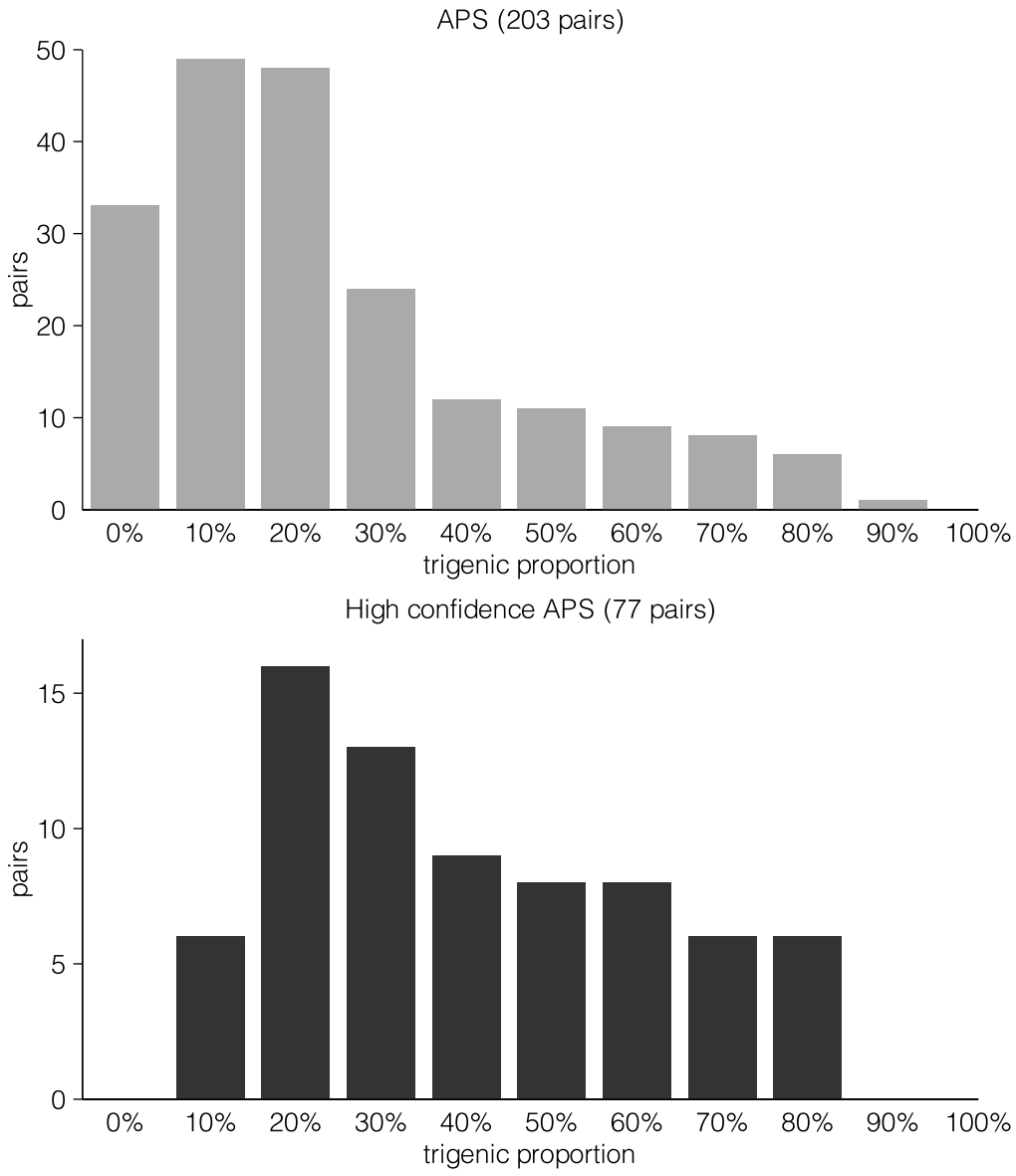

Figure 5.8: Distribution of trigenic proportion for 203 pairs in the mini-array survey (top), and for a subset of 77 pairs that show > 10 trigenic interactions (bottom). Trigenic proportion (Eq. 5.13) measures the fraction of total interactions displayed by a pair that are trigenic. Members of pairs with a high trigenic proportion are expected to have increased ability to buffer the loss of their partner.

Trigenic interactions > 10 (77 pairs)	Spearman ρ	p-value
Digenic negative path length	-0.51	4.6×10^{-3}
Total unique GO "function" annotations	-0.41	2.9×10^{-4}
Divergence asymmetry (Kellis 2004)	-0.40	1.9×10^{-3}
Paralog digenic ϵ	-0.38	2.8×10^{-3}
Divergence asymmetry (ANNG, AMM*)	-0.36	2.5×10^{-3}
Double-mutant fitness	-0.33	1.1×10^{-2}
Total unique INTERPRO domains	-0.26	2.4×10^{-2}
Digenic degree asymmetry $> 7:1$ (VanderSluis 2010)	-0.23	4.7×10^{-2}
Similar localization (Marques 2008)	0.32	5.8×10^{-3}
SGA profile similarity (array)	0.36	7.7×10^{-3}

Table 5.2: A selection of paralog-pair features that show significant correlations with trigenic proportion as defined in Eq. 5.13. Results shown are for a high-confidence subset of 77 pairs which show at least 10 trigenic interactions. Similar correlations for all pairs are shown in Table D.1. *: Personal communication.

10 trigenic interactions, ensuring that both the numerator and denominator in the proportion would have sufficient signal. The 77 resulting pairs show a slightly more uniform distribution of trigenic proportions (see Fig. 5.8), which can be advantageous in correlation analysis. We then assembled a database of features curated from high throughput studies which might be relevant to evolutionary divergence. Table 5.2 shows those features that have a significant correlation with this measure. Our measure is derived from genetic interactions, and we see several other genetic interaction measures in agreement. Each of these genetic interaction measures are derived from an independent experiment (data from Chapter 3), performed by the same lab.

A digenic interaction score between the paralogs themselves is the best single indicator of trigenic proportion, one of the strongest correlations over the filtered mini-array survey pairs ($\rho = -0.38$, $p = 2.8 \times 10^{-3}$; Spearman), supporting the hypothesis that a strong negative interaction score is an excellent predictor of retained functional overlap, and that this functional overlap does indeed translate to an increase in the number of trigenic interactions (relative to digenic). Indeed, the high rate of synthetic sickness/lethality among duplicate pairs observed in the previous chapters was one of the motivations for this study and this result supports the central hypothesis of Chapters 4 & 5. A highly related measure, the double-mutant fitness of each pair, yields another significant correlation ($\rho = -0.33$). Pairs with a low double-mutant fitness also

commonly have a genetic interaction (that is, when the single mutants are healthy), so again, this supports the idea that a digenic interaction provides direct evidence of functional overlap which then manifests more specifically in the trigenic interaction network. Furthermore, correlation between double-mutant fitness and trigenic interaction degree ($r = -0.47$, $p = 1.3 \times 10^{-4}$; Pearson) echoes previous results relating single-mutant fitness to digenic interaction degree [66]. Another correlate derived from separate SGA data is profile similarity as measured between paralog array profiles ($\rho = 0.36$; Table 5.2). This result makes sense given the known relationship between genetic interaction profile similarity and functional overlap which applies broadly to all genes pairs [66]. However, we previously predicted this correlation to have the opposite sign for paralog sisters, reasoning that very closely related paralog sisters would buffer common interactions completely, and their profile similarities would consequently be greatly reduced. This was an attempt to explain why profile similarity for duplicate pairs seemed lower than expected, and very few pairs having many significant interactions in common [78]. Surprisingly, we instead see more trigenic interactions for paralogs with a higher profile similarity (relative to other paralogs). This indicates that the interactions which are buffered as the result of functional overlap are not masked completely, but instead are quantitatively reduced, often to the point of insignificance. This unexpected result underscores the importance of not only measuring interactions quantitatively, but that reasoning about them only qualitatively can be misleading.

The final, and strongest, SGA-derived measure to appear is shortest path-length on the negative genetic interaction network ($\rho = -0.51$). This measure describes the number of digenic interactions in previously observed data are needed to connect the sister paralogs to one another, and is a convenient short-hand for a combination of other features. By definition, duplicate pairs with a direct genetic interaction (i.e. "digenic ϵ") have a path-length of 1; pairs without a direct interaction, but who share common interactions with one or more third-parties have a path-length of two (having many common interaction partners also means profile similarity will be high). Path-lengths are seldom longer than 3 due to the small-world nature of the genetic interaction network.

Other properties which appear in Table 5.2 also support the hypothesis that functionally divergent pairs show a low proportion of trigenic interactions whereas functionally similar pairs show a higher proportion of trigenics. For example, a measure of sub-cellular localization pattern conservation developed by Marques *et al.* [145] shows a positive correlation with trigenic proportion ($\rho = 0.32$; Table 5.2). Localization can be a key factor in paralog specialization, and because a paralogous protein cannot buffer the functions of a deleted sister unless it is in the right compartment, it makes sense that conserved localization profiles would be a requisite for conserved function or buffering ability. Another relevant functional measure of divergence can be taken from protein sequences directly. Two similar measures of sequence divergence asymmetry give significant correlations with trigenic proportion. The first was published by Kellis *et al.* in 2004 [42], and captures the rates of evolution of each paralog using information from the non-WGD species *Kluyveromyces waltii*. Kellis *et al.* then divided the rate of one sister by the other to detect cases where one paralog was evolving much more quickly than the other. The correlation between this measure and trigenic proportion is negative ($\rho = -0.4$, $p = 1.9 \times 10^{-3}$) indicating that pairs which have diverged asymmetrically have a low trigenic proportion, and hence a limited capacity to buffer one another. A similar measure, devised by Alex Nguyen and Alan Moses, uses sequence from a number of closely related yeasts to calibrate the expectation of evolutionary rates for protein binding domains within each paralog, then measures deviations in these rates for each sister against one another (see D.1). This measure is similarly designed to capture asymmetric instances of sequence evolution and also shows a negative correlation with trigenic proportion ($\rho = -0.36$, $p = 2.5^{-3}$).

We also observed a relationship with one Gene Ontology-based measure: the total number of annotations to the GO "Molecular function" ontology from both sisters (union) shows a strong negative correlation ($\rho = -0.41$) with trigenic proportion. This may indicate that pairs that can no longer buffer one another have diverged by gaining new functions. However, an alternative, and in our opinion more likely hypothesis, is that pairs with a larger number of ancestral functions have had more opportunity to sub-functionalize via common degenerative mutations. This latter hypothesis seems better in line with the observation that the paralog that is evolving more quickly tends to

have fewer genetic interactions, fewer protein-protein interactions, and fewer chemical-genetic interactions [78]. Notably absent from the list of correlations is any significant relationship with the number of GO "Biological process" annotations. If GO "Molecular function" annotations indeed capture physical mechanism, whereas "Biological process" annotations captures physiological consequences, these results would suggest that the number of cellular processes that a pair impinges upon is less important for its evolutionary trajectory than the number of physical mechanisms by which it participates in those processes. Although selective pressure is applied based on how well a gene fulfills a role in one or more processes, genetic mutations ultimately occur in a more tangible, mechanistic way. For example, a newly duplicated pair that performs two duties via two discrete binding domains, may have a higher probability of sub-functionalization (and therefore, long-term pair retention) than a pair that performs two duties via a single, highly constrained, binding domain. In the former case, the two processes are mechanistically separable, but in the latter, they are entangled.

In other words, typical measures of multi-functionality and pleiotropy are insufficient as a measure of a pair's ability to sub-functionalize. In order to partition two roles to separate paralog copies, the roles themselves must be separable, and a measure that summarizes potentially separable mechanisms, (e.g. GO "Molecular function") will be more successful at capturing the opportunity for divergence of a pair than a measure that counts the downstream consequences of those mechanisms (e.g. GO "Biological process" annotations). This mechanistic interpretation is in agreement with the negative correlation between trigenic proportion and the total number of protein domains (union) a paralog pair has. We counted the number of unique domains annotated to either member of a pair using predictions from the INTERPRO database, and this count has a significant negative correlation with trigenic proportion ($\rho = -0.26$), which is consistent with the expectation that paralog pairs with ample opportunity to partition mechanistic functions via sub-functionalization will do so, while those with fewer avenues for divergence will either retain more functional overlap or lose one sister to eventual degeneration.

Taken together, these results begin to form a picture of which pairs of paralogs will tend to diverge via sub-functionalization, and therefore be retained, if they have the potential. The amount of this potential depends on whether the functions they perform

can be partitioned as a result of sequence mutations. A pair with many responsibilities, all carried out by the same crucial sequence segment, has no opportunity to survive as a sub-functionalized pair, unless those responsibilities can be partitioned temporally or spatially instead, for example through divergence in localization patterns.

In Chapter 4, I laid out a mechanistic model predicting that duplicates will diverge asymmetrically given a set of assumptions that very much resemble the GO properties of pairs with the lowest trigenic proportion (See Secs. C.5, C.6). This model explains why paralog sequence divergence is asymmetric, and how, once established, this asymmetry perpetuates itself in support of this model. Digenic interaction degree asymmetry, which is a boolean measure described in Chapter 4, also shows a significantly negative correlation ($\rho = -023$) with trigenic proportion. In Chapter 4, I also showed a connection between this interaction degree asymmetry (as measured for paralogs on the array in a separate experiment) and several other physiological measures that could be viewed as constraints on evolutionary divergence. Two of those measures were sequence-based, and suggest that the member of an asymmetric pair with more genetic interactions, tends to evolve (or degrade) more slowly than the other member (See Fig. 4.4). The other measures (protein-protein interaction degree, single-mutant fitness, and chemical-genetic degree), all give some measure of evolutionary constraint or consequence, and show that the direction of these asymmetries in evolutionary constraint agree with the direction of asymmetries in genetic interaction degree.

5.3.9 Modeling evolutionary divergence

Several observations from the previous section suggest that physiological properties can predict the evolutionary trajectory of a duplicate pair. These include the number of functions a gene carries out, which provide opportunities for sub-functionalization, and whether those functions are entangled with one another, which presents a potential obstacle to divergence. In this section, we present a computational framework for simulating the evolutionary divergence of paralog pairs, and explore the requirements for duplicate pair fates. The framework is based on the Duplication-Mutation-Complementation model (DMC) [143]. In the DMC model, sisters are functionally identical immediately after duplications, and begin to acquire degenerative mutations, which disable one or

more of the now-buffered ancestral functions. The sister only survive as a pair if mutations accumulate in both sisters in a complementary fashion, such that each performs a subset of ancestral functions, and both are required for complete functionality. However, the generality of the framework presented here, specifically the pliable definition used for individual gene functions, allows us to incorporate elements from another model, Escape from Adaptive Conflict (EAC) [176]. The EAC model describes a potential scenario where a gene that performs multiple functions is prevented from acquiring mutations that would be beneficial to one function, because they would negatively affect the other. The two functions are said to be in adaptive conflict with one another, and a duplication event can resolve this conflict by enabling the two copies to specialize to one of the related functions. The source of this conflict may be, for example, an overlap in sequence regions crucial to the performance of each function. Our abstract representation of the sequence-to-function relationship described below can capture aspects of either of these models and is therefore able to explain a wide range of post-duplication outcomes.

Duplicate representation and evolution

The framework generates and evolves genes pairs in the following way. First, it creates a gene with a fixed length. Length here refers to the number of positions that can suffer a potentially debilitating mutation and is analogous to the sequence length of a real gene. Then the framework assigns hypothetical functions to contiguous regions of the gene. Each region is necessary for the gene to carry out that function. Since each position in a hypothetical gene is independent of every other, these functional regions need not be contiguous but here are represented as contiguous to speed computation and simplify visualization. More importantly, positional independence allows a contiguous region to represent functional regions of actual genes or proteins which are often not contiguous. The number of functions assigned to a new gene is a parameter, and the length of each functional region, as well as its position within the gene, are chosen at random. At this point the gene is duplicated resulting in a "left" and a "right" paralog. Fig. 5.9 shows two examples of duplicate pairs, each with seven functions. They are shown immediately after duplication, with all functions initially in-tact for both paralogs, as well as after evolutionary divergence.

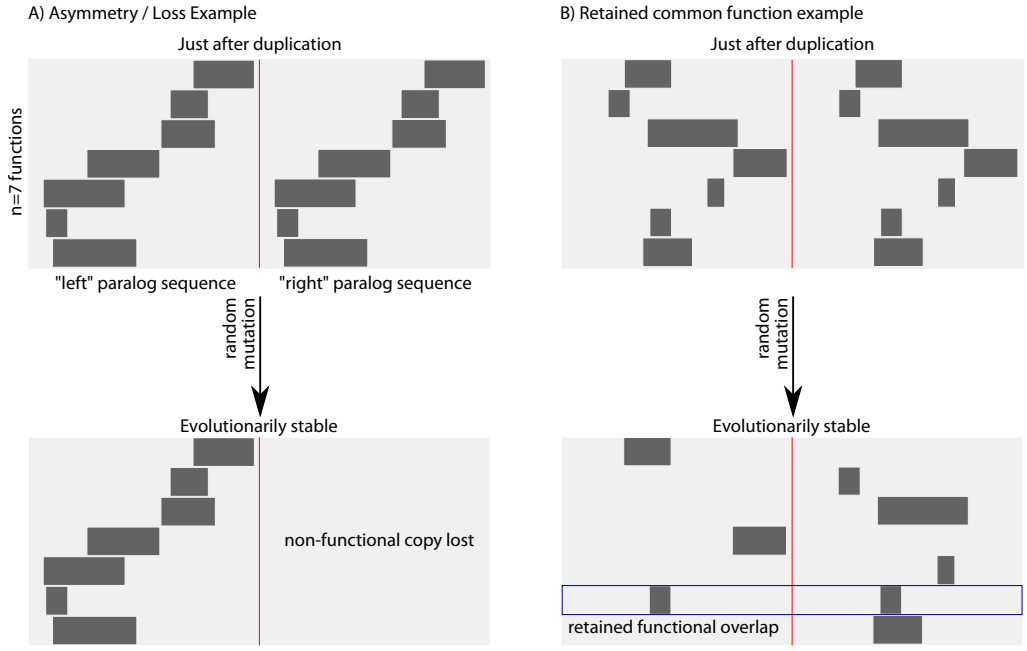

Figure 5.9: Divergence modeling examples. A) An example of a paralog pair with 7 functions (rows), all carried out by variously sized, potentially overlapping, regions of sequence (dark bands). Initially after duplication, both the "left" and "right" paralogs are intact, and can carry out all functions. The pair is left to evolve through random mutations, until it reaches an evolutionarily stable-steady state, which can sustain no further mutations without loss of function. In this case, one paralog has completely degenerated, and the other has reverted to singleton status.
B) Another paralog pair example generated with the same parameters as in A). In this case the pair achieves a more equitable division of labor with each paralog carrying out several unique functions after reaching steady-state. Additionally, the members of this pair have retained a common function (blue box, sixth row). Though displayed here as a distinct function, this may represent a type core functionality common to the functions in the first, third, and seventh rows. For example, the sequence region covered by the common function may be a the catalytic site, while the other "functions" are responsible for different targeting that site to catalyze specific reactions.

Simulated paralogs are computationally evolved by degenerative mutations at a constant rate. First, either the right or left paralog is selected with equal probability. Then a uniformly random position along that paralogs sequence is randomly chosen for mutation. If the mutation falls within one or more functional regions, those functions are considered to be disabled for that paralog, and the regions are removed. Each mutation then has three possible outcomes. A mutation is "silent" if it falls in a position which is not involved in any functions. These mutations are evolutionarily neutral and result in no changes. A mutation is "divergent" if it disables a function that can still be performed by the other paralog. In these cases, the affected function is removed from the mutated paralog, and the role is assumed by its sister. The result of a "divergent" mutation is therefore an increased level of functional divergence between the two sisters. A mutation is "deleterious" if it would disable a function that is not covered by the other sister. The framework assumes that a duplicate pair is expected to retain all ancestral functions, so any loss of a function that cannot be performed by the other sister is deleterious. Lineages harboring these mutations would be out-competed in a population and therefore have no effect on long-term pair evolution. These mutations can then safely be discarded. A pair has reached "steady-state" when there exists no remaining possible "divergent" mutations, and therefore no further possibility of change. Fig. 5.9 shows the evolutionarily stable, steady-states for two paralog pairs generated by the model with identical parameters.

Asymmetry confirmed

We applied this model to see if it would generate the type of asymmetries we observed in genetic interactions in the previous chapter, and whether initial functional overlap would again be the requisite factor in determining the level of asymmetry. Fig. 5.10 shows the results for 50,000 simulated duplicate pairs, each of which began with 20 functions. A separate 50,000 pair control group was generated in which functional regions were not allowed to overlap. Each group was evolved to steady-state, and we counted the number of functions that each paralog could still perform. Both sets distributed functions equitably between left and right paralogs as expected (left, median bias = 0.5), however the group with overlapping functional regions had a broader bias distribution, indicating that extreme asymmetries were more common. If functions do

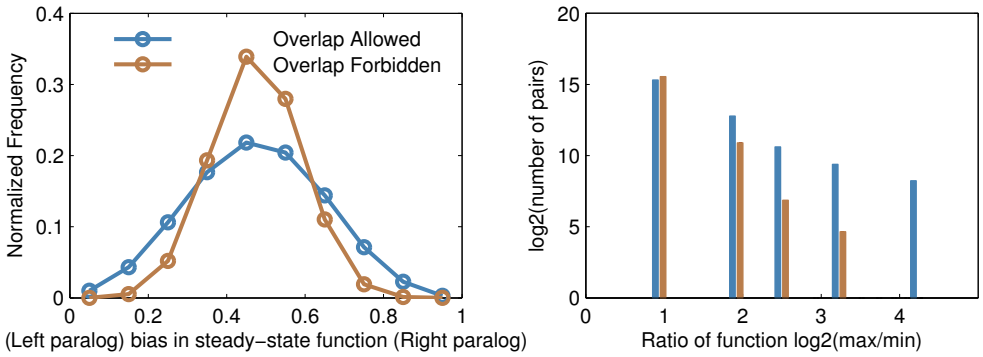

Figure 5.10:
Divergence model confirms asymmetry. Left) the distribution of functions between "left" and "right" paralogs after reaching steady-state. The peak at 0.5 indicates that each function is as likely to sub-functionalize to one paralog as to the other, while the breadth of the distribution describes the degree to which asymmetries occur. The blue line indicates results from a model which allowed functional regions to overlap one another, while the yellow line describes a model where such overlaps are explicitly forbidden. Moderate asymmetry is expected according to a binomial distribution; extreme cases become more common when overlap in functional regions are allowed. Right) A log-log histogram of function ratios for the same set as in (left). The ratio of functions remaining after steady-state (max/min) is log_2-scaled and binned, and the number of pairs in each bin is also log_2-scaled. The yellow bars form a straight line in agreement with the corresponding distribution in the left panel being a simple binomial. In contrast, the blue bars show a much heavier skew toward higher asymmetric ratios.

not overlap with one another in sequence space, they get partitioned to each paralog with equal probability, and the number of functions each paralog can perform relative to its sister follows a perfect binomial distribution. Conversely, if functions are entangled with one another, they get partitioned in groups. The right panel of Fig. 5.10 shows the distribution of function ratios on a log-log plot, where the no-overlap cases follow a straight line, and the cases with functional entanglement have a tail in the more extreme ratios. This model therefore agrees with our previous work, demonstrating initial functional entanglement as required for asymmetric divergence.

Requirements for sustained functional overlap

We have established the ability of paralog sisters to share functional ability despite their millions of years of opportunity to diverge. We reasoned that the initial complexity of functional region overlap for newly duplicated pairs would constrain steady-state solutions, and govern the possible degrees of functional overlap after divergence had run its course. In the trivial case, with functions forbidden from overlapping in sequence space, complete sub-functionalization is inevitable. This fact can be deduced directly from the rules of the model. With no overlap, each mutation can disable at most one function. If both paralogs can still perform a common function, then any mutation affecting that function is by definition "divergent," and the presence of a "divergent" mutation indicates that further evolution toward the steady-state is possible. The amount of initial functional entanglement therefore governs the potential number of retained common functions.

This gives rise to an apparent contradiction. Multi-functionality is positively correlated with asymmetry because asymmetry is self-perpetuated and upper-bounded only by the total number of functions (Sec. C.4). Retained overlap is positively correlated with structural entanglement for the reasons discussed in the previous paragraph. Further, multi-functionality and entanglement are trivially related to each other, as adding more functions to a fixed length gene can only increase the entanglements. However, our functional data suggest that retained common function (trigenic proportion) and asymmetric divergence (as measured by sequence or genetic interaction degree) are, in fact, negatively correlated. To reconcile these observations, we propose that the evolutionary fate of a duplicate pair is not governed by multi-functionality or structural entanglement in isolation, but instead by these two factors in relation to one another, and we illustrate this idea in Fig. 5.11-A.

Because of the inherent relationship between the number of functions, and their overlap in sequence space, gene pairs will tend to fall along a diagonal (of unknown positive slope) in panel A of Fig. 5.11-A. It is each pair's unique functional properties which cause deviations from that diagonal and describe their post-duplication divergence. If a pair has many functions, and those functions are all easily partitioned by degenerative mutations, then the pair will sub-functionalize completely, and symmetrically, retaining no functions in common. A slight increase in functional entanglement is

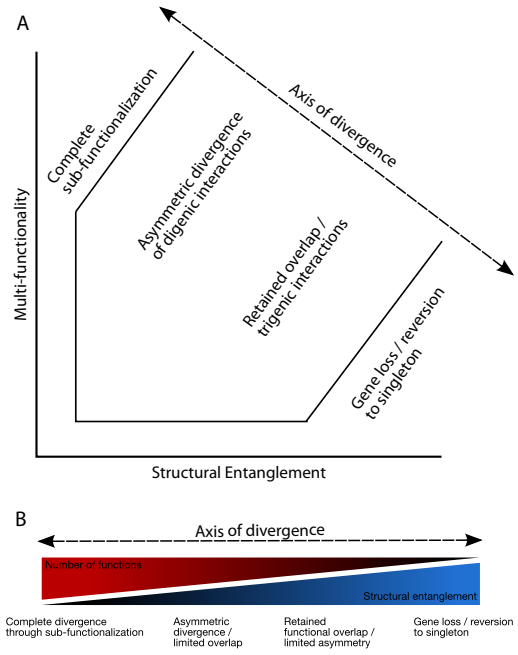

Figure 5.11: Hypothetical map of duplicate divergence space. A) A two-dimensional map of the relationship between paralog multi-functionality, and structural entanglement. Here, structural entanglement refers to inseparability of functions in sequence space, captured in our model by the degree of functional region overlap in the pre-duplication singleton. The map also shows various familiar post-duplication outcomes which form a single axis of divergence.
B) The axis of divergence reduces multi-functionality and structural entanglement to a single feature according to their relative quantities, as they jointly affect the evolutionary outcome.

sufficient for asymmetry to take hold and simultaneously introduces a chance the pair will retain common functions. At the other extreme, if a pair has only a few functions that are very entangled, they will quickly become the responsibility of one sister while the other becomes completely non-functional, reverting the pair back to a singleton. However if the amount of entanglement is slightly lower, or the number of functions sufficiently high, the pair will find some functions it can partition some functions and retain many others in both copies. Measuring these quantities relative to one another thus provides a single informative axis describing the divergence potential of a paralog pair (Fig. 5.11-B).

The model is able to simulate paralog pairs and place them on this map. Using the number of simulated functions as the measure of multi-functionality, and the percentage of sequence positions participating in more than one functional region as a measure of structural entanglement, we can test our hypotheses about asymmetry and retained common function. For a range of multi-functionality parameters (3–30) we simulated the evolution of 5,000 paralog pairs, and binned them by their initial structural entanglement, thus generating a two-dimensional grid of binned samples over the space. We could then calculate the fraction of pairs that revert to singletons, measure the average asymmetry, or count the average number of retained functions in each bin. Fig. 5.12 shows these measures for each bin, plotted against the bin's location on the divergence axis. Fig. 5.11-A predicts that pairs with a high score on the divergence axis (that is down and to the right) will more commonly end up as singletons because they have few functions that are difficult to separate. Indeed, we find that the average rate of pairs converting to singletons is much higher for those regions of the map. The map also predicts that retained common functions should likewise be more frequent for higher divergence axis scores, and again simulations bare that out. Further, the map shows lower divergence axis scores being associated with asymmetry, and the simulations show this property to be true also.

Modeling conclusions

If we attempt to apply our real trigenic interaction data to the map shown in Fig. 5.11, the results are encouraging. Using GO function annotations to asses multi-functionality, and GO component annotations to represent structural entanglement as a constraint

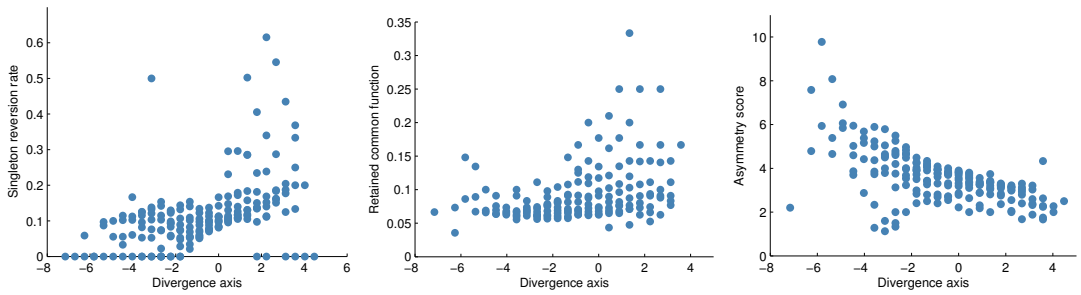

Figure 5.12: Modeling the axis of divergence. Paralogs were generated and evolved to cover as much of the space in the divergence map as possible. The space was then binned in two dimensions, and summary statistics were gathered for each bin, and plotted against the bins score on the axis of divergence. Left) Axis score versus gene loss / reversion to singleton status. The plot confirms that the rate at which paralogs are converted back into singletons, is much higher for pairs with high divergence-axis scores (that is pairs with fewer, highly entangled functions).
Center) Axis score versus retained common functions. The plot confirms that pairs with a higher axis score result in a broader range of retained common function. Notably, the increase in the potential for retained common function begins at about -2, which is left of the position where gene loss becomes increasingly common.
Right) Axis score versus asymmetry. The plot confirms that the average functional asymmetry is higher for simulated pairs with a lower axis scores. Pairs with functions that are numerous, or easily sub-functionalized to one paralog are prone to asymmetric divergence.

to sub-functionalization, we can achieve correlations between divergence axis scores and trigenic proportion ($\rho > 0.27,\ p < 0.021$, Spearman), and divergence asymmetry ($\rho > 0.38,\ p < 10^{-3}, (AN)$). We also find a significant relationship between divergence axis score, and the "dosage" pair classification used in Chapter 4 ($\rho > 0.31,\ p < 0.007$, Spearman), suggesting that perhaps the divergence axis may be more predictive of pairs retained for dosage sensitivities than the absence of trigenic interactions, as we predicted earlier.

Gene duplication events have several possible outcomes. For any particular paralog pair, which of these comes to pass is a result of their functional properties and, to a large extent, chance. Understanding the relative impact of different functional properties and the role of chance in duplicate evolution and divergence will bring new insight to an old evolutionary problem. In this section we tried to understand the interplay between opportunities for, and obstacles to, divergence via sub-functionalization. Importantly, we demonstrated that a key component in retained common functions of the sort that might give rise to trigenic interactions was the amount of structural entanglement relative to multi-functionality. More generally we developed a framework to study how the sequence-function relationships of a paralog pair affect their evolutionary trajectory.

5.4 Conclusions

In this study, we have conducted systematic analysis of triple-mutant perturbations in yeast at an unprecedented scale. We developed the experimental and theoretical systems with which the trigenic interaction network can be mapped, and discovered nearly 8,500 novel trigenic interactions in the process. These interactions were shown to have a high overlap with what little trigenic interaction data exists, and were otherwise shown to be of a quality similar to previous digenic interaction studies.

We explored novel types of trigenic interactions, which arise due to the combinatoric nature of higher order perturbations, and speculate as to their biological interpretations. Additionally, we demonstrated broad variation in trigenic proportion, which captures the extent to which a duplicate pair has retained common functions, and gave several examples of physiological properties that correlate with trigenic proportion in support of that hypothesis.

Finally, we developed an updated framework to explore paralog evolution as it relates to the evolutionary stability of retained common functions, and asymmetric divergence, and concluded that paralogs will follow an evolutionary path which depends on both their opportunities to diverge, as well as their freedom to do so. Our model and simulation results suggest that sub-functionalization will tend to partition ancestral function asymmetrically, unless the sequence-function relationship is sufficiently complex, in which case common functions can be retained in both paralogs giving rise to trigenic interactions.

Chapter 6

Conclusion and Discussion

In this dissertation, I have outlined several different projects with one single goal: to build a complete map of gene functions in a model organism through perturbation analysis. These efforts have combined several perturbation approaches, varying both the number of simultaneous perturbations as well as the environment under which they are tested. In Chapter 2, I described the first whole-genome survey of single-mutants in a truly minimal environment, where yeast are forced to exercise their full range of metabolic potential. In Chapter 3, I covered our efforts to construct a complete map of genetic interactions from double-mutant perturbations. In Chapter 4, I examined a specific segment of that genetic interaction network for insights into the mechanisms of duplicate gene evolution and divergence, while in Chapter 5, I extended the work on duplicate genetic interactions with the first genome-scale maps of triple-mutant interactions.

There is still much experimental and computational work to be done in pursuit of our goal: a complete understanding of a single model organism. The single-mutant study in Chapter 2 represents an important contribution to the study of perturbations in truly minimal environments. While simple environments are not likely to garner as much attention as more complex environments (i.e. drug treatments), it is essential to examine them for a more complete understanding of basic cellular operation. Our effort to disentangle the effects of multiple environmental factors, observed simultaneously, represents an important contribution, but additional computational work will be needed to scale these methods to more complex environments.

Of course, network models of metabolism are not yet complete in yeast. I observed that despite their supposed dominance as *in silico* models of cellular processes, current metabolic networks have difficulty predicting the real world consequences of even simple perturbations in basic environments. Despite their shortcomings, these metabolic models do produce impressive results, and have made tremendous progress since their inception. What is required for these models to fulfill their purpose—aside from more experimental data in absolutely minimal environments—is a better understanding of how environmental elements interact in the metabolic network. Such interactions include those mediated by genes not currently represented in flux-based models such a transporters, transcription factors, and genes responsible for nutrient sensing and signaling. Work in this area has begun, however, methods for the incorporation of the necessary experimental observations into computational models requires additional attention.

The completion of digenic interaction mapping efforts in yeast has opened many new doors for computational discovery. These directions chiefly concern the structure of network interactions beyond simple local associations. My work shows that long-distance interactions comprise the majority of the edges of the genetic interaction network. Fundamental to the problem of unraveling complex phenotypes in humans and other organisms is an understanding of how influence aggregates over the entire network. Still, our work shows that even these long-distance functional interactions contain some structure and that they connect across broad functional processes in meaningful ways. Notably, the work in Chapter 3 describes meaningful network properties of essential genes. While their prominence in terms of network degree is unsurprising, given their essential nature, the character of their interactions is very different from that of non-essential genes. Whether or not these differences are conserved to other organisms requires further experimentation and computational analysis. If the properties are found to be universal, it will validate the necessity of including essential genes in experimental interaction maps of model organisms.

Similarly the analysis of gene duplication and its effects in the genetic interaction networks provides insights that extent beyond one organism. Much is still not understood about how these networks become so robust, and how they maintain robustness in the face of evolutionary pressure to simplify. The genetic interaction network does an excellent job at characterizing this robustness at the level of functional modules and

pathways, but is poorly suited to capture the most specific example of robustness: nodes of identical redundant function. For that task, we must counter gene-level redundancy with higher-order perturbations, such as the triple-mutant perturbations we use in the definition of trigenic interactions.

While mapping of the digenic interaction network is nearly complete, mapping of the trigenic interaction network is just beginning. The combinatorial explosion of high-order perturbations makes it unlikely that the trigenic network will be systematically mapped in any organism. However, an efficient sampling of trigenic interaction space may prove informative about how many trigenic interactions to expect, where to expect them, and how their structure differs from digenic interactions. Paralog pairs, with their known capacity for buffering, provide a natural subset of trigenic interaction space for systematic mapping, and while many of the pairs in this study showed numerous trigenic interactions (77/203), many showed few or none. This study will help researchers who wish to map trigenic interactions in other model organisms do so more efficiently by targeting pairs with characteristics outlined here. It will be interesting to see which of those properties are conserved, and to what extent.

The interpretation of different types of trigenic interaction, and also of differential digenic interactions, is another interesting direction worthy of further study. Knowing whether these sub-types correspond to distinct biological mechanisms, and do so reliably, would be preferable to simply applying intuitions and models built for digenic interactions.

Models for evolutionary divergence and retention provide many opportunities for both the simulation of data, and its comparison with real-world observations. In my estimation, the principle obstacle here is our inability to directly relate sequence to functional ability. Individual studies exist, with specific functional readouts such as enzyme catalytic activity, but a more generalized approach could prove very useful. For example, genetic interactions derived from point mutants may provide relevant data for such an approach, but current genetic interaction data is limited to mutants that have not yet been sequenced to determine exact genotype. Further, while we progress in our understanding of how sequence mediates function in many ways (for example, protein folding, binding, RNA splicing, post-translational modifications, chromatin organization, and so on), integrating genome-scale information concerning all of these processes

together in a meaningful way is only just becoming a possibility. A complete model of paralog evolution would need to account for each of these potential mechanisms of divergence and understanding their interactions will be important to making such a model successful.

The data in these study will prove useful both to researchers studying particular duplicate pairs, as well as those studying the general effects of duplication. Duplicated genes have been, and will continue to be, an important area of study, both for the insights they yield into the characteristic robustness of biological networks, as well as the light they shed on the evolutionary histories of those networks. A complete functional profile for each of these pairs represents an important step in the understanding of both of these areas, and should enable many future directions in the study of historical and extant networks.

The wealth of data available for model organisms such as *S. cerevisiae* brings exciting computational opportunities, but it also brings significant challenges. Complete functional characterization from individual genes, to functional modules, to entire cellular sub-systems continuously demands new computational methods such as those included here. Ultimately, these tools may help us to understand, not only how all the components come together to form a whole organism, but the processes by which evolution has shaped these components in relation to one another.

References

[1] R. Fleischmann, M. Adams, O White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J. Tomb, B. Dougherty, J. Merrick, and e. Al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512, July 1995.

[2] Granger G. Sutton, Owen White, Mark D. Adams, and Anthony R. Kerlavage. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. *Genome Science and Technology*, 1(1):9–19, January 1995.

[3] Guri Giaever, Angela M Chu, Li Ni, Carla Connelly, Linda Riles, Steeve Véronneau, Sally Dow, Ankuta Lucau-Danila, Keith Anderson, Bruno André, Adam P Arkin, Anna Astromoff, Mohamed El-Bakkoury, Rhonda Bangham, Rocio Benito, Sophie Brachat, Stefano Campanaro, Matt Curtiss, Karen Davis, Adam Deutschbauer, Karl-Dieter Entian, Patrick Flaherty, Francoise Foury, David J Garfinkel, Mark Gerstein, Deanna Gotte, Ulrich Güldener, Johannes H Hegemann, Svenja Hempel, Zelek Herman, Daniel F Jaramillo, Diane E Kelly, Steven L Kelly, Peter Kötter, Darlene LaBonte, David C Lamb, Ning Lan, Hong Liang, Hong Liao, Lucy Liu, Chuanyun Luo, Marc Lussier, Rong Mao, Patrice Menard, Siew Loon Ooi, Jose L Revuelta, Christopher J Roberts, Matthias Rose, Petra Ross-Macdonald, Bart Scherens, Greg Schimmack, Brenda Shafer, Daniel D Shoemaker, Sharon Sookhai-Mahadeo, Reginald K Storms, Jeffrey N Strathern, Giorgio Valle, Marleen Voet, Guido Volckaert, Ching-yun Wang, Teresa R Ward, Julie Wilhelmy, Elizabeth a Winzeler, Yonghong Yang, Grace Yen, Elaine Youngman, Kexin Yu, Howard Bussey, Jef D Boeke, Michael Snyder, Peter Philippsen,

Ronald W Davis, and Mark Johnston. Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, 418(6896):387–91, July 2002.

[4] B Dujon. European Functional Analysis Network (EUROFAN) and the functional analysis of the Saccharomyces cerevisiae genome. *Electrophoresis*, 19(4):617–24, April 1998.

[5] Bart Scherens and Andre Goffeau. The uses of genome-wide yeast mutant collections. *Genome biology*, 5(7):229, January 2004.

[6] David Botstein and Gerald R. Fink. Yeast: an experimental organism for 21st Century biology. *Genetics*, 189(3):695–704, November 2011.

[7] Albert-László Barabási and Zoltán N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, 5(2):101–13, February 2004.

[8] Chad L Myers, Drew Robson, Adam Wible, Matthew a Hibbs, Camelia Chiriac, Chandra L Theesfeld, Kara Dolinski, and Olga G Troyanskaya. Discovery of biological networks from diverse functional genomic data. *Genome biology*, 6(13):R114, January 2005.

[9] David Warde-Farley, Sylva L Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, Chris Grouios, Farzana Kazi, Christian Tannus Lopes, Anson Maitland, Sara Mostafavi, Jason Montojo, Quentin Shao, George Wright, Gary D Bader, and Quaid Morris. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(Web Server issue):W214–20, July 2010.

[10] Magali Michaut, Anastasia Baryshnikova, Michael Costanzo, Chad L Myers, Brenda J Andrews, Charles Boone, and Gary D Bader. Protein complexes are central in the yeast genetic landscape. *PLoS computational biology*, 7(2):e1001092, February 2011.

[11] H Jeong, S P Mason, A L Barabási, and Z N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–2, May 2001.

[12] Alexei Vázquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Modeling of Protein Interaction Networks. *Complexus*, 1(1):38–44, 2003.

[13] Aviva Presser, Michael B Elowitz, Manolis Kellis, and Roy Kishony. The evolutionary dynamics of the Saccharomyces cerevisiae protein interaction network after duplication. *Proceedings of the National Academy of Sciences*, 105(3):950–954, January 2008.

[14] Supporting Online Material, Science Web, Highwire Press, New York, and Avenue Nw. Evidence for network evolution in an Arabidopsis interactome map. *Science (New York, N.Y.)*, 333(6042):601–7, July 2011.

[15] Sara Sharifpoor, Dewald van Dyk, Michael Costanzo, Anastasia Baryshnikova, Helena Friesen, Alison C Douglas, Ji-Young Youn, Benjamin VanderSluis, Chad L Myers, Balázs Papp, Charles Boone, and Brenda J Andrews. Functional wiring of the yeast kinome revealed by global analysis of genetic network motifs. *Genome research*, 22(4):791–801, April 2012.

[16] Sebastian Wernicke and Florian Rasche. FANMOD: a tool for fast network motif detection. *Bioinformatics (Oxford, England)*, 22(9):1152–3, May 2006.

[17] Zahra Razaghi Moghadam Kashani, Hayedeh Ahrabian, Elahe Elahi, Abbas Nowzari-Dalini, Elnaz Saberi Ansari, Sahar Asadi, Shahin Mohammadi, Falk Schreiber, and Ali Masoudi-Nejad. Kavosh: a new algorithm for finding network motifs. *BMC bioinformatics*, 10:318, January 2009.

[18] A Goffeau, B G Barrell, H Bussey, R W Davis, B Dujon, H Feldmann, F Galibert, J D Hoheisel, C Jacq, M Johnston, E J Louis, H W Mewes, Y Murakami, P Philippsen, H Tettelin, and S G Oliver. Life with 6000 genes. *Science (New York, N.Y.)*, 274:546, 563–567, 1996.

[19] Stephen J Giovannoni, H James Tripp, Scott Givan, Mircea Podar, Kevin L Vergin, Damon Baptista, Lisa Bibbs, Jonathan Eads, Toby H Richardson, Michiel Noordewier, Michael S Rappé, Jay M Short, James C Carrington, and Eric J Mathur. Genome streamlining in a cosmopolitan oceanic bacterium. *Science (New York, N.Y.)*, 309(5738):1242–5, August 2005.

[20] Christina A. Cuomo, Christopher A. Desjardins, Malina A. Bakowski, Jonathan Goldberg, Amy T. Ma, James J. Becnel, Elizabeth S. Didier, Lin Fan, David I. Heiman, Joshua Z. Levin, Sarah Young, Qiandong Zeng, and Emily R. Troemel. Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth .

[21] F. R. Blattner. The Complete Genome Sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462, September 1997.

[22] Jane M Carlton, Robert P Hirt, Joana C Silva, Arthur L Delcher, Michael Schatz, Qi Zhao, Jennifer R Wortman, Shelby L Bidwell, U Cecilia M Alsmark, Sébastien Besteiro, Thomas Sicheritz-Ponten, Christophe J Noel, Joel B Dacks, Peter G Foster, Cedric Simillion, Yves Van de Peer, Diego Miranda-Saavedra, Geoffrey J Barton, Gareth D Westrop, Sylke Müller, Daniele Dessi, Pier Luigi Fiori, Qinghu Ren, Ian Paulsen, Hanbang Zhang, Felix D Bastida-Corcuera, Augusto Simoes-Barbosa, Mark T Brown, Richard D Hayes, Mandira Mukherjee, Cheryl Y Okumura, Rachel Schneider, Alias J Smith, Stepanka Vanacova, Maria Villalvazo, Brian J Haas, Mihaela Pertea, Tamara V Feldblyum, Terry R Utterback, Chung-Li Shu, Kazutoyo Osoegawa, Pieter J de Jong, Ivan Hrdy, Lenka Horvathova, Zuzana Zubacova, Pavel Dolezal, Shehre-Banoo Malik, John M Logsdon, Katrin Henze, Arti Gupta, Ching C Wang, Rebecca L Dunne, Jacqueline A Upcroft, Peter Upcroft, Owen White, Steven L Salzberg, Petrus Tang, Cheng-Hsun Chiu, Ying-Shiung Lee, T Martin Embley, Graham H Coombs, Jeremy C Mottram, Jan Tachezy, Claire M Fraser-Liggett, and Patricia J Johnson. Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. *Science (New York, N.Y.)*, 315(5809):207–12, January 2007.

[23] Kenneth H Wolfe. Comparative genomics and genome evolution in yeasts. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 361(1467):403–12, March 2006.

[24] Berit Olsen Krogh and Lorraine S Symington. Recombination proteins in yeast. *Annual review of genetics*, 38:233–71, January 2004.

[25] A Baudin, O Ozier-Kalogeropoulos, A Denouel, F Lacroute, and C Cullin. A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. *Nucleic acids research*, 21(14):3329–30, July 1993.

[26] Tomoya Baba, Takeshi Ara, Miki Hasegawa, Yuki Takai, Yoshiko Okumura, Miki Baba, Kirill A Datsenko, Masaru Tomita, Barry L Wanner, and Hirotada Mori. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology*, 2:2006.0008, January 2006.

[27] Feng Zhang, Morgan L Maeder, Erica Unger-Wallace, Justin P Hoshaw, Deepak Reyon, Michelle Christian, Xiaohong Li, Christopher J Pierick, Drena Dobbs, Thomas Peterson, J Keith Joung, and Daniel F Voytas. High frequency targeted mutagenesis in Arabidopsis thaliana using zinc finger nucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 107(26):12028–33, June 2010.

[28] Daniel J Dickinson, Jordan D Ward, David J Reiner, and Bob Goldstein. Engineering the Caenorhabditis elegans genome using Cas9-triggered homologous recombination. *Nature methods*, 10(10):1028–34, October 2013.

[29] Yuexin Zhou, Shiyou Zhu, Changzu Cai, Pengfei Yuan, Chunmei Li, Yanyi Huang, and Wensheng Wei. High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature*, 509(7501):487–91, May 2014.

[30] Ian Spencer Hornesey. *A History of Beer and Brewing*. Royal Society of Chemistry, Cambridge, 2003.

[31] Diego Libkind, Chris Todd Hittinger, Elisabete Valério, Carla Gonçalves, Jim Dover, Mark Johnston, Paula Gonçalves, and José Paulo Sampaio. Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 108(35):14539–44, August 2011.

[32] Stacia R Engel, Fred S Dietrich, Dianna G Fisk, Gail Binkley, Rama Balakrishnan, Maria C Costanzo, Selina S Dwight, Benjamin C Hitz, Kalpana Karra, Robert S Nash, Shuai Weng, Edith D Wong, Paul Lloyd, Marek S Skrzypek, Stuart R

Miyasato, Matt Simison, and J Michael Cherry. The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. *G3*, 4(1):389–398, December 2014.

[33] Carl. C. Lindegren. *The yeast cell, its genetics and cytology.* Educational Publishers, St Louis, 1949.

[34] F Sanger, G M Air, B G Barrell, N L Brown, A R Coulson, C A Fiddes, C A Hutchison, P M Slocombe, and M Smith. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265:687–695, 1977.

[35] C. J. Bult, O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J.-F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. M. Geoghagen, J. F. Weidman, J. L. Fuhrmann, D. Nguyen, T. R. Utterback, J. M. Kelley, J. D. Peterson, P. W. Sadow, M. C. Hanna, M. D. Cotton, K. M. Roberts, M. A. Hurst, B. P. Kaine, M. Borodovsky, H.-P. Klenk, C. M. Fraser, H. O. Smith, C. R. Woese, and J. C. Venter. Complete Genome Sequence of the Methanogenic Archaeon, Methanococcus jannaschii. *Science*, 273(5278):1058–1073, August 1996.

[36] The C. elegans Sequencing Consortium. Genome Sequence of the Nematode C.elegans: A Platform for Investigating Biology. *Science*, 282(5396):2012–2018, December 1998.

[37] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, N Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla,

K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowski, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, and J Szustakowki. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[38] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski,

G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen,

N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–51, February 2001.

[39] Kara Dolinski and David Botstein. *Orthology and functional conservation in eukaryotes.*, volume 41. January 2007.

[40] Mark Johnson, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L Madden. NCBI BLAST: a better web interface. *Nucleic Acids Research*, 36:W5–W9, 2008.

[41] Martha L Bulyk. Computational prediction of transcription-factor binding site locations. *Genome biology*, 5(1):201, January 2003.

[42] Manolis Kellis, Bruce W Birren, and Eric S Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature*, 428(6983):617–624, April 2004.

[43] Kevin P Byrne and Kenneth H Wolfe. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research*, 15:1456–1461, 2005.

[44] D. A. Lashkari, J. L. DeRisi, J. H. McCusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, and R. W. Davis. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences*, 94(24):13057–13062, November 1997.

[45] Patrick O. Brown and David Botstein. Exploring the new world of the genome with DNA microarrays.

[46] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, December 1998.

[47] Y Cheng and G M Church. Biclustering of expression data. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB.*

International Conference on Intelligent Systems for Molecular Biology, 8:93–103, 2000.

[48] S Raychaudhuri, J M Stuart, and R B Altman. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 455–66, January 2000.

[49] Jeremy Bellay, Gowtham Atluri, Tina L Sing, Kiana Toufighi, Michael Costanzo, Philippe Souza Moraes Ribeiro, Gaurav Pandey, Joshua Baller, Benjamin Vandersluis, Magali Michaut, Sangjo Han, Philip Kim, Grant W Brown, Brenda J Andrews, Charles Boone, Vipin Kumar, and Chad L Myers. Putting genetic interactions in context through a global modular decomposition. *Genome research*, June 2011.

[50] Anastasia Baryshnikova, Michael Costanzo, Yungil Kim, Huiming Ding, Judice Koh, Kiana Toufighi, Ji-Young Youn, Jiongwen Ou, Bryan-Joseph San Luis, Sunayan Bandyopadhyay, Matthew Hibbs, David Hess, Anne-Claude Gingras, Gary D Bader, Olga G Troyanskaya, Grant W Brown, Brenda Andrews, Charles Boone, and Chad L Myers. Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nature methods*, 7(12):1017–24, December 2010.

[51] Lourdes Peña Castillo and Timothy R Hughes. Why are there still over 1000 uncharacterized yeast genes? *Genetics*, 176(1):7–14, May 2007.

[52] E. a. Winzeler. Functional Characterization of the S.cerevisiae Genome by Gene Deletion and Parallel Analysis. *Science*, 285(5429):901–906, August 1999.

[53] Dong-Uk Kim, Jacqueline Hayles, Dongsup Kim, Valerie Wood, Han-Oh Park, Misun Won, Hyang-Sook Yoo, Trevor Duhig, Miyoung Nam, Georgia Palmer, Sangjo Han, Linda Jeffery, Seung-Tae Baek, Hyemi Lee, Young Sam Shim, Minho Lee, Lila Kim, Kyung-Sun Heo, Eun Joo Noh, Ah-Reum Lee, Young-Joo Jang, Kyung-Sook Chung, Shin-Jung Choi, Jo-Young Park, Youngwoo Park, Hwan Mook Kim, Song-Kyu Park, Hae-Joon Park, Eun-Jung Kang, Hyong Bai Kim, Hyun-Sam Kang, Hee-Moon Park, Kyunghoon Kim, Kiwon Song,

Kyung Bin Song, Paul Nurse, and Kwang-Lae Hoe. Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. *Nature biotechnology*, 28(6):617–23, June 2010.

[54] Christopher P Austin, James F Battey, Allan Bradley, Maja Bucan, Mario Capecchi, Francis S Collins, William F Dove, Geoffrey Duyk, Susan Dymecki, Janan T Eppig, Franziska B Grieder, Nathaniel Heintz, Geoff Hicks, Thomas R Insel, Alexandra Joyner, Beverly H Koller, K C Kent Lloyd, Terry Magnuson, Mark W Moore, Andras Nagy, Jonathan D Pollock, Allen D Roses, Arthur T Sands, Brian Seed, William C Skarnes, Jay Snoddy, Philippe Soriano, David J Stewart, Francis Stewart, Bruce Stillman, Harold Varmus, Lyuba Varticovski, Inder M Verma, Thomas F Vogt, Harald von Melchner, Jan Witkowski, Richard P Woychik, Wolfgang Wurst, George D Yancopoulos, Stephen G Young, and Brian Zambrowicz. The knockout mouse project. *Nature genetics*, 36(9):921–4, September 2004.

[55] Y. S. Rong and Kent G. Golic. Gene Targeting by Homologous Recombination in Drosophila. *Science*, 288(5473):2013–2018, June 2000.

[56] Maureen E Hillenmeyer, Eula Fung, Jan Wildenhain, Sarah E Pierce, Shawn Hoon, William Lee, Michael Proctor, Robert P St Onge, Mike Tyers, Daphne Koller, Russ B Altman, Ronald W Davis, Corey Nislow, and Guri Giaever. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science*, 320(5874):362–5, April 2008.

[57] Ainslie B Parsons, Renée L Brost, Huiming Ding, Zhijian Li, Chaoying Zhang, Bilal Sheikh, Grant W Brown, Patricia M Kane, Timothy R Hughes, and Charles Boone. Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nature biotechnology*, 22(1):62–9, January 2004.

[58] Aimée Marie Dudley, Daniel Maarten Janse, Amos Tanay, Ron Shamir, and George McDonald Church. A global view of pleiotropy and phenotypically derived gene function in yeast. *Molecular systems biology*, 1:2005.0001, January 2005.

[59] J Michael Cherry, Eurie L Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T Chan, Karen R Christie, Maria C Costanzo, Selina S Dwight, Stacia R Engel, Dianna G Fisk, Jodi E Hirschman, Benjamin C Hitz, Kalpana Karra, Cynthia J Krieger, Stuart R Miyasato, Rob S Nash, Julie Park, Marek S Skrzypek, Matt Simison, Shuai Weng, and Edith D Wong. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic acids research*, 40(Database issue):D700–5, January 2012.

[60] Timothy R Hughes. Yeast and drug discovery. *Functional & integrative genomics*, 2(4-5):199–211, September 2002.

[61] Raamesh Deshpande. *Computational methods to explore chemical and genetic interaction networks for novel human therapies.* Ph.d., University of Minnesota, 2013.

[62] Marc Vidal, Michael E Cusick, and Albert-László Barabási. Interactome networks and human disease. *Cell*, 144(6):986–98, March 2011.

[63] Yuanfang Guan, Chad L Myers, Rong Lu, Ihor R Lemischka, Carol J Bult, and Olga G Troyanskaya. A genomewide functional network for the laboratory mouse. *PLoS computational biology*, 4(9):e1000165, January 2008.

[64] Curtis Huttenhower, Erin M Haley, Matthew A Hibbs, Vanessa Dumeaux, Daniel R Barrett, Hilary A Coller, and Olga G Troyanskaya. Exploring the human genome with functional maps. *Genome research*, 19(6):1093–106, June 2009.

[65] Nevan J Krogan, Gerard Cagney, Haiyuan Yu, Gouqing Zhong, Xinghua Guo, Alexandr Ignatchenko, Joyce Li, Shuye Pu, Nira Datta, Aaron P Tikuisis, Thanuja Punna, José M Peregrín-Alvarez, Michael Shales, Xin Zhang, Michael Davey, Mark D Robinson, Alberto Paccanaro, James E Bray, Anthony Sheung, Bryan Beattie, Dawn P Richards, Veronica Canadien, Atanas Lalev, Frank Mena, Peter Wong, Andrei Starostine, Myra M Canete, James Vlasblom, Samuel Wu, Chris Orsi, Sean R Collins, Shamanta Chandran, Robin Haw, Jennifer J Rilstone, Kiran Gandi, Natalie J Thompson, Gabe Musso, Peter St Onge, Shaun Ghanny, Mandy H Y Lam, Gareth Butland, Amin M Altaf-Ul, Shigehiko Kanaya, Ali Shilatifard, Erin O'Shea, Jonathan S Weissman, C James Ingles, Timothy R Hughes,

John Parkinson, Mark Gerstein, Shoshana J Wodak, Andrew Emili, and Jack F Greenblatt. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440(7084):637–643, March 2006.

[66] Michael Costanzo, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D Spear, Carolyn S Sevier, Huiming Ding, Judice L Y Koh, Kiana Toufighi, Sara Mostafavi, Jeany Prinz, Robert P St Onge, Benjamin VanderSluis, Taras Makhnevych, Franco J Vizeacoumar, Solmaz Alizadeh, Sondra Bahr, Renee L Brost, Yiqun Chen, Murat Cokol, Raamesh Deshpande, Zhijian Li, Zhen-Yuan Lin, Wendy Liang, Michaela Marback, Jadine Paw, Bryan-Joseph San Luis, Ermira Shuteriqi, Amy Hin Yan Tong, Nydia van Dyk, Iain M Wallace, Joseph a Whitney, Matthew T Weirauch, Guoqing Zhong, Hongwei Zhu, Walid a Houry, Michael Brudno, Sasan Ragibizadeh, Balázs Papp, Csaba Pál, Frederick P Roth, Guri Giaever, Corey Nislow, Olga G Troyanskaya, Howard Bussey, Gary D Bader, Anne-Claude Gingras, Quaid D Morris, Philip M Kim, Chris a Kaiser, Chad L Myers, Brenda J Andrews, and Charles Boone. The genetic landscape of a cell. *Science*, 327(5964):425–31, January 2010.

[67] Matthew W Hahn and Andrew D Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular biology and evolution*, 22(4):803–6, April 2005.

[68] Xionglei He and Jianzhi Zhang. Why do hubs tend to be essential in protein networks? *PLoS genetics*, 2(6):e88, June 2006.

[69] Elena Zotenko, Julian Mestre, Dianne P O'Leary, and Teresa M Przytycka. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS computational biology*, 4(8):e1000140, January 2008.

[70] Andrew Chatr-Aryamontri, Bobby-Joe Breitkreutz, Sven Heinicke, Lorrie Boucher, Andrew Winter, Chris Stark, Julie Nixon, Lindsay Ramage, Nadine Kolas, Lara O'Donnell, Teresa Reguly, Ashton Breitkreutz, Adnane Sellam, Daici Chen, Christie Chang, Jennifer Rust, Michael Livstone, Rose Oughtred, Kara

Dolinski, and Mike Tyers. The BioGRID interaction database: 2013 update. *Nucleic acids research*, 41(Database issue):D816–23, January 2013.

[71] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53, October 2009.

[72] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating missing heritability for disease from genome-wide association studies. *American journal of human genetics*, 88(3):294–305, March 2011.

[73] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1193–8, January 2012.

[74] Susumu Ohno. *Evolution by Gene Duplication*, volume 9. Springer-Verlag, New York, 1970.

[75] Hideki Innan and Fyodor Kondrashov. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2):97–108, February 2010.

[76] Karin Voordeckers, Chris A. Brown, Kevin Vanneste, Elisa van der Zande, Arnout Voet, Steven Maere, and Kevin J. Verstrepen. Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication. *PLoS Biology*, 10(12):e1001446, December 2012.

[77] K H Wolfe and D C Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634):708–13, June 1997.

[78] Benjamin VanderSluis, Jeremy Bellay, Gabriel Musso, Michael Costanzo, Balázs Papp, Franco J Vizeacoumar, Anastasia Baryshnikova, Brenda Andrews, Charles Boone, and Chad L Myers. Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Molecular systems biology*, 6(1):429, November 2010.

[79] Elizabeth N Koch, Michael Costanzo, Jeremy Bellay, Raamesh Deshpande, Kate Chatfield-Reed, Gordon Chua, Gennaro D'Urso, Brenda Andrews, Charles Boone, and Chad L Myers. Conserved rules govern genetic interaction degree across species. *Genome biology*, 13(7):R57, July 2012.

[80] Haiyuan Yu and Mark Gerstein. Genomic analysis of the hierarchical structure of regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(40):14724–31, October 2006.

[81] Nicholas M Luscombe, M Madan Babu, Haiyuan Yu, Michael Snyder, Sarah a Teichmann, and Mark Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–12, September 2004.

[82] Benjamin VanderSluis, David C Hess, Colin Pesyna, Elias W Krumholz, Tahin Syed, Balázs Szappanos, Corey Nislow, Balázs Papp, Olga G Troyanskaya, Chad L Myers, and Amy A Caudy. Broad metabolic sensitivity profiling of a prototrophic yeast deletion collection. *Genome Biology*, 15(4):R64, 2014.

[83] Amy Hin Yan Tong, Guillaume Lesage, Gary D Bader, Huiming Ding, Hong Xu, Xiaofeng Xin, James Young, Gabriel F Berriz, Renee L Brost, Michael Chang, YiQun Chen, Xin Cheng, Gordon Chua, Helena Friesen, Debra S Goldberg, Jennifer Haynes, Christine Humphries, Grace He, Shamiza Hussein, Lizhu Ke, Nevan Krogan, Zhijian Li, Joshua N Levinson, Hong Lu, Patrice Ménard, Christella Munyana, Ainslie B Parsons, Owen Ryan, Raffi Tonikian, Tania Roberts, Anne-Marie Sdicu, Jesse Shapiro, Bilal Sheikh, Bernhard Suter, Sharyl L Wong, Lan V Zhang, Hongwei Zhu, Christopher G Burd, Sean Munro, Chris Sander, Jasper Rine, Jack Greenblatt, Matthias Peter, Anthony Bretscher, Graham Bell, Frederick P Roth, Grant W Brown, Brenda Andrews, Howard Bussey, and Charles Boone. Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808–13, February 2004.

[84] Sourav Bandyopadhyay, Monika Mehta, Dwight Kuo, Min-Kyung Sung, Ryan Chuang, Eric J Jaehnig, Bernd Bodenmiller, Katherine Licon, Wilbert Copeland, Michael Shales, Dorothea Fiedler, Janusz Dutkowski, Aude Guénolé, Haico van Attikum, Kevan M Shokat, Richard D Kolodner, Won-Ki Huh, Ruedi Aebersold, Michael-Christopher Keogh, Nevan J Krogan, and Trey Ideker. Rewiring of genetic networks in response to DNA damage. *Science*, 330(6009):1385–9, December 2010.

[85] Nathan D Price, Jennifer L Reed, and Bernhard ØPalsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature reviews. Microbiology*, 2(11):886–97, November 2004.

[86] Jack T Pronk. Auxotrophic yeast strains in fundamental and applied research. *Applied and environmental microbiology*, 68(5):2095–100, May 2002.

[87] Rachel B Brem, Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science (New York, N.Y.)*, 296(5568):752–5, April 2002.

[88] Amy H Y Tong, M Evangelista, A B Parsons, H Xu, G D Bader, N Pagé, M Robinson, S Raghibizadeh, C W Hogue, H Bussey, B Andrews, M Tyers, and Ch Boone. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550):2364–8, December 2001.

[89] Amy Hin Yan Tong and Charles Boone. Synthetic genetic array analysis in Saccharomyces cerevisiae. *Methods in molecular biology (Clifton, N.J.)*, 313:171–92, January 2006.

[90] Lars Kuepfer, Uwe Sauer, and Lars M Blank. Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome research*, 15(10):1421–30, October 2005.

[91] J. P. Barford and R. J. Hall. An Examination of the Crabtree Effect in Saccharomyces cerevisiae: the Role of Respiratory Adaptation. *Journal of General Microbiology*, 114(2):267–75, October 1979.

[92] Rupert Millard. Venn's four ellipse construction.

[93] Michael Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9, May 2000.

[94] Chad L Myers, Daniel R Barrett, Matthew A Hibbs, Curtis Huttenhower, and Olga G Troyanskaya. Finding function: evaluation methods for functional genomic data. *BMC genomics*, 7:187, January 2006.

[95] Evan S Snitkin, Aimée M Dudley, Daniel M Janse, Kaisheen Wong, George M Church, and Daniel Segrè. Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions. *Genome biology*, 9(9):R140, January 2008.

[96] Benjamin D Heavner, Kieran Smallbone, Brandon Barker, Pedro Mendes, and Larry P Walker. Yeast 5 - an expanded reconstruction of the Saccharomyces cerevisiae metabolic network. *BMC systems biology*, 6(1):55, January 2012.

[97] AR Ali R Zomorrodi and CD Costas D Maranas. Improving the iMM904 S. cerevisiae metabolic model using essentiality and synthetic lethality data. *BMC systems biology*, 4:178, January 2010.

[98] Jeffrey D Orth, Ines Thiele, and Bernhard ØPalsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–8, March 2010.

[99] Daniel Segrè, Dennis Vitkup, and George M Church. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–7, November 2002.

[100] Gabriel Musso, Michael Costanzo, Manqin Huangfu, Andrew M Smith, Jadine Paw, Bryan-Joseph San Luis, Charles Boone, Guri Giaever, Corey Nislow, Andrew Emili, and Zhaolei Zhang. The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome research*, 18(7):1092–9, July 2008.

[101] Maya Schuldiner, Sean R Collins, Natalie J Thompson, Vladimir Denic, Arunashree Bhamidipati, Thanuja Punna, Jan Ihmels, Brenda Andrews, Charles Boone, Jack F Greenblatt, Jonathan S Weissman, and Nevan J Krogan. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell*, 123(3):507–19, November 2005.

[102] Colm J. Ryan, Assen Roguev, Kristin Patrick, Jiewei Xu, Harlizawati Jahari, Zongtian Tong, Pedro Beltrao, Michael Shales, Hong Qu, Sean R. Collins, Joseph I. Kliegman, Lingli Jiang, Dwight Kuo, Elena Tosti, Hyun-Soo Kim, Winfried Edelmann, Michael-Christopher Keogh, Derek Greene, Chao Tang, Pádraig Cunningham, Kevan M. Shokat, Gerard Cagney, J. Peter Svensson, Christine Guthrie, Peter J. Espenshade, Trey Ideker, and Nevan J. Krogan. Hierarchical modularity and the evolution of genetic interactomes across species. *Molecular cell*, 46(5):691–704, June 2012.

[103] Christian R Landry, Julia Oh, Daniel L Hartl, and Duccio Cavalieri. Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene*, 366(2):343–51, February 2006.

[104] Joerg Reinders, René P Zahedi, Nikolaus Pfanner, Chris Meisinger, and Albert Sickmann. Toward the complete yeast mitochondrial proteome: multidimensional separation techniques for mitochondrial proteomics. *Journal of proteome research*, 5(7):1543–54, July 2006.

[105] H L Huang and M C Brandriss. The regulator of the yeast proline utilization pathway is differentially phosphorylated in response to the quality of the nitrogen source. *Molecular and cellular biology*, 20(3):892–9, February 2000.

[106] Emeline Teyssier, Go Hirokawa, Anna Tretiakova, Bradford Jameson, Akira Kaji, and Hideko Kaji. Temperature-sensitive mutation in yeast mitochondrial ribosome recycling factor (RRF). *Nucleic acids research*, 31(14):4218–26, July 2003.

[107] T Kanai, S Takeshita, H Atomi, K Umemura, M Ueda, and A Tanaka. A regulatory factor, Fil1p, involved in derepression of the isocitrate lyase gene in Saccharomyces cerevisiae–a possible mitochondrial protein necessary for protein synthesis in mitochondria. *European journal of biochemistry / FEBS*, 256(1):212–20,

August 1998.

[108] D D Lee and H S Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91, October 1999.

[109] Danial D. Lee and H. Sebastian Seung. Algorithms for Non-Negative Matrix Factorization. *Advances in Neural Information Processing Systems*, 13:556–62, 2001.

[110] Kevin J Roberg. Physiological Regulation of Membrane Protein Sorting Late in the Secretory Pathway of Saccharomyces cerevisiae. *The Journal of cell biology*, 137(7):1469–1482, June 1997.

[111] Esther J Chen and Chris A Kaiser. Amino acids regulate the intracellular trafficking of the general amino acid permease of Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 99(23):14837–42, November 2002.

[112] Michael Stanbrough and Boris Magasanik. Transcriptional and posttranslational regulation of the general amino acid permease of Saccharomyces cerevisiae. *Journal of bacteriology*, 177(1):94–102, January 1995.

[113] April L Risinger, Natalie E Cain, Esther J Chen, and Chris A Kaiser. Activity-dependent reversible inactivation of the general amino acid permease. *Molecular biology of the cell*, 17(10):4411–9, October 2006.

[114] Esther J Chen and Chris a Kaiser. LST8 negatively regulates amino acid biosynthesis as a component of the TOR pathway. *The Journal of cell biology*, 161(2):333–47, April 2003.

[115] Michael Mülleder, Floriana Capuano, Pnar Pir, Stefan Christen, Uwe Sauer, Stephen G Oliver, and Markus Ralser. A prototrophic deletion mutant collection for yeast metabolomics and systems biology. *Nature biotechnology*, 30(12):1176–8, December 2012.

[116] Scott a Becker and Bernhard O Palsson. Context-specific metabolic networks are consistent with experiments. *PLoS computational biology*, 4(5):e1000082, May 2008.

[117] Tomer Shlomi, Moran N Cabili, Markus J Herrgå rd, Bernhard ØPalsson, and Eytan Ruppin. Network-based prediction of human tissue-specific metabolism. *Nature biotechnology*, 26(9):1003–10, September 2008.

[118] Paul a Jensen and Jason a Papin. Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics (Oxford, England)*, 27(4):541–7, February 2011.

[119] M W Covert, C H Schilling, and B Palsson. Regulation of gene expression in flux balance models of metabolism. *Journal of theoretical biology*, 213(1):73–88, November 2001.

[120] Markus W Covert, Nan Xiao, Tiffany J Chen, and Jonathan R Karr. Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. *Bioinformatics (Oxford, England)*, 24(18):2044–50, September 2008.

[121] Sriram Chandrasekaran and Nathan D Price. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, 107(41):17845–50, October 2010.

[122] T Dobzhansky. Genetics of Natural Populations. XIII. Recombination and Variability in Populations of Drosophila Pseudoobscura. *Genetics*, 31(3):269–90, May 1946.

[123] Ramamurthy Mani, Robert P St Onge, John L Hartman, Guri Giaever, and Frederick P Roth. Defining genetic interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9):3461–6, March 2008.

[124] Zhijian Li, Franco J Vizeacoumar, Sondra Bahr, Jingjing Li, Jonas Warringer, Frederick S Vizeacoumar, Renqiang Min, Benjamin Vandersluis, Jeremy Bellay, Michael Devit, James a Fleming, Andrew Stephens, Julian Haase, Zhen-Yuan Lin, Anastasia Baryshnikova, Hong Lu, Zhun Yan, Ke Jin, Sarah Barker, Alessandro Datti, Guri Giaever, Corey Nislow, Chris Bulawa, Chad L Myers, Michael Costanzo, Anne-Claude Gingras, Zhaolei Zhang, Anders Blomberg, Kerry Bloom,

Brenda Andrews, and Charles Boone. Systematic exploration of essential yeast gene function with temperature-sensitive mutants. *Nature biotechnology*, 29(4), March 2011.

[125] Zhun Yan, Michael Costanzo, Lawrence E Heisler, Jadine Paw, Fiona Kaper, Brenda J Andrews, Charles Boone, Guri Giaever, and Corey Nislow. Yeast Barcoders: a chemogenomic application of a universal donor-strain collection carrying bar-code identifiers. *Nature methods*, 5(8):719–25, August 2008.

[126] Peter M Visscher. Sizing up human height variation. *Nature genetics*, 40(5):489–90, May 2008.

[127] Raamesh Deshpande, Benjamin VanderSluis, and Chad L Myers. Comparison of profile similarity measures for genetic interaction networks. *PloS one*, 8(7):e68664, January 2013.

[128] Daniel Finley, Helle D Ulrich, Thomas Sommer, and Peter Kaiser. The ubiquitin-proteasome system of Saccharomyces cerevisiae. *Genetics*, 192(2):319–60, October 2012.

[129] Anaïs Baudot, Bernard Jacq, and Christine Brun. A scale of functional divergence for yeast duplicated genes revealed from analysis of the protein-protein interaction network. *Genome Biology*, 5:R76, 2004.

[130] Yuanfang Guan, Maitreya J Dunham, and Olga G Troyanskaya. Functional analysis of gene duplications in Saccharomyces cerevisiae. *Genetics*, 175(2):933–943, February 2007.

[131] Gabriel Musso, Zhaolei Zhang, and Andrew Emili. Retention of protein complex membership by ancient duplicated gene products in budding yeast. *Trends in Genetics*, 23:266–269, 2007.

[132] Ilan Wapinski, Avi Pfeffer, Nir Friedman, and Aviv Regev. Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449(7158):54–61, September 2007.

[133] Z Gu, L M Steinmetz, Xun Gu, Curt Scharfe, R W Davis, and W H Li. Role of duplicate genes in genetic robustness against null mutations. *Nature*, 421(2):63–66, 2003.

[134] Balázs Papp, Csaba Pál, and Laurence D Hurst. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, 429(6992):661–664, June 2004.

[135] Zhenglong Gu, Dan Nicolae, Henry H-S Lu, and Wen Hsiung Li. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in Genetics*, 18(12):609–13, December 2002.

[136] Ran Kafri, Melissa Levy, and Yitzhak Pilpel. The regulatory utilization of genetic redundancy through responsive backup circuits. *Proceedings of the National Academy of Sciences of the United States of America*, 103(31):11653–8, August 2006.

[137] Scott J Dixon, Michael Costanzo, Anastasia Baryshnikova, Brenda Andrews, and Charles Boone. Systematic mapping of genetic interaction networks. *Annual Review of Genetics*, 43:601–625, January 2009.

[138] P Novick, B C Osmond, and D Botstein. Suppressors of Yeast Actin Mutations. *Genetics*, 121:659–674, 1989.

[139] Alexander DeLuna, Kalin Vetsigian, Noam Shoresh, Matthew Hegreness, Maritrini Colón-González, Sharon Chao, and Roy Kishony. Exposing the fitness contribution of duplicated genes. *Nature Genetics*, 40(5):676–681, May 2008.

[140] E Jedediah Dean, Jerel C Davis, Ronald W Davis, and Dmitri A Petrov. Pervasive and Persistent Redundancy among Duplicated Genes in Yeast. *PLoS Genetics*, 4(7):e1000113, July 2008.

[141] Jan Ihmels, Sean R Collins, Maya Schuldiner, Nevan J Krogan, and Jonathan S Weissman. Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. *Molecular Systems Biology*, 3(86):86, January 2007.

[142] A L Hughes. The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society B: Biological Sciences*, 256:119–124, 1994.

[143] A Force, M Lynch, F B Pickett, A Amores, Y L Yan, and J Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151:1531–1545, 1999.

[144] Gavin C Conant and Kenneth H Wolfe. Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics*, 9:938–50, 2008.

[145] Ana C Marques, Nicolas Vinckenbosch, David Brawand, and Henrik Kaessmann. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. *Genome Biology*, 9:R54, 2008.

[146] Fyodor A Kondrashov and Alexey S Kondrashov. Role of selection in fixation of gene duplications. *Journal of Theoretical Biology*, 239:141–151, 2006.

[147] Gavin C Conant and Kenneth H Wolfe. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Molecular Systems Biology*, 3:204, 2007.

[148] John Brookfield. Can genes be truly redundant? *Current Biology*, 2:553–554, 1992.

[149] Luke Hakes, John W Pinney, Simon C Lovell, Stephen G Oliver, and David L Robertson. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biology*, 8(10):R209, January 2007.

[150] Elizabeth Marland, Anuphap Prachumwat, Natalia Maltsev, Zhenglong Gu, and Wen-Hsiung Li. Higher gene duplicabilities for metabolic proteins than for non-metabolic proteins in yeast and E. coli. *Journal of Molecular Evolution*, 59:806–814, 2004.

[151] Xionglei He and Jianzhi Zhang. Higher duplicability of less important genes in yeast genomes. *Molecular Biology and Evolution*, 23:144–151, 2006.

[152] Andreas Wagner. Asymmetric functional divergence of duplicate genes in yeast. *Molecular Biology and Evolution*, 19(10):1760–1768, October 2002.

[153] Michael Lynch and Vaishali Katju. The altered evolutionary trajectories of gene duplicates. *Trends in Genetics*, 20(11):544–549, November 2004.

[154] Mario A Fares, Kevin P Byrne, and Kenneth H Wolfe. Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of Saccharomyces species. *Molecular Biology and Evolution*, 23:245–253, 2006.

[155] Kevin P Byrne and Kenneth H Wolfe. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics*, 175:1341–1350, 2007.

[156] Gavin C Conant and Andreas Wagner. Asymmetric sequence divergence of duplicate genes. *Genome Research*, 13(9):2052–2058, September 2003.

[157] Peng Zhang, Zhenglong Gu, and Wen-Hsiung Li. Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biology*, 4(9):R56, January 2003.

[158] Devin R Scannell and Kenneth H Wolfe. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Research*, 18(1):137–147, January 2008.

[159] Xionglei He and Jianzhi Zhang. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, 169(2):1157–64, March 2005.

[160] Itay Tirosh and Naama Barkai. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biology*, 8(4):R50, January 2007.

[161] Jing Yang, Zhenglong Gu, and Wen-Hsiung Li. Rate of protein evolution versus fitness effect of gene deletion. *Molecular Biology and Evolution*, 20:772–774, 2003.

[162] Ben Lehner. Genes Confer Similar Robustness to Environmental, Stochastic, and Genetic Perturbations in Yeast. *PLoS ONE*, 5:5, 2010.

[163] F C Holstege, E G Jennings, J J Wyrick, T I Lee, C J Hengartner, M R Green, T R Golub, E S Lander, and R A Young. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95:717–728, 1998.

[164] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320:1344–1349, 2008.

[165] Reinhard Jahn and Richard H Scheller. SNAREs–engines for membrane fusion. *Nature Reviews Molecular Cell Biology*, 7:631–643, 2006.

[166] Hui-Ju Yang, Hideki Nakanishi, Song Liu, James A McNew, and Aaron M Neiman. Binding interactions control SNARE specificity in vivo. *The Journal of Cell Biology*, 183:1089–1100, 2008.

[167] Jussi Jäntti, Markku K Aalto, Mattias Oyen, Lena Sundqvist, Sirkka Keränen, and Hans Ronne. Characterization of temperature-sensitive mutations in the yeast syntaxin 1 homologues Sso1p and Sso2p, and evidence of a distinct function for Sso1p in sporulation. *Journal of Cell Science*, 115:409–420, 2002.

[168] Yoshikazu Ohya, Jun Sese, Masashi Yukawa, Fumi Sano, Yoichiro Nakatani, Taro L Saito, Ayaka Saka, Tomoyuki Fukuda, Satoru Ishihara, Satomi Oka, Genjiro Suzuki, Machika Watanabe, Aiko Hirata, Miwaka Ohtani, Hiroshi Sawai, Nicolas Fraysse, Jean-Paul Latgé, Jean M François, Markus Aebi, Seiji Tanaka, Sachiko Muramatsu, Hiroyuki Araki, Kintake Sonoike, Satoru Nogami, and Shinichi Morishita. High-dimensional and large-scale phenotyping of yeast mutants. *Proceedings of the National Academy of Sciences of the United States of America*, 102:19015–19020, 2005.

[169] Brendan D Manning, Jennifer G Barrett, Julie A Wallace, Howard Granok, and Michael Snyder. Differential regulation of the Kar3p kinesin-related protein by two associated proteins, Cik1p and Vik1p. *The Journal of Cell Biology*, 144(6):1219–1233, March 1999.

[170] B D Page, L L Satterwhite, M D Rose, and M Snyder. Localization of the Kar3 kinesin heavy chain-related protein requires the Cik1 interacting protein. *The Journal of Cell Biology*, 124(4):507–519, February 1994.

[171] Jennifer S Tirnauer, Eileen O'Toole, Lisbeth Berrueta, Barbara E Bierer, and David Pellman. Yeast Bim1p promotes the G1-specific dynamics of microtubules. *The Journal of Cell Biology*, 145:993–1007, 1999.

[172] L Lee, J S Tirnauer, J J Li, S C Schuyler, J Y Liu, and D Pellman. Positioning of the mitotic spindle by a cortical-microtubule capture mechanism. *Science*, 287:2260–2262, 2000.

[173] Melissa K Gardner, Julian Haase, Karthikeyan Mythreye, Jeffrey N Molk, Marybeth Anderson, Ajit P Joglekar, Eileen T O'Toole, Mark Winey, E D Salmon, David J Odde, and Kerry Bloom. The microtubule-based motor Kar3 and plus endbinding protein Bim1 provide structural support for the anaphase spindle. *The Journal of Cell Biology*, 180:91–100, 2008.

[174] Lisa R Sproul, Daniel J Anderson, Andrew T Mackey, William S Saunders, and Susan P Gilbert. Cik1 targets the minus-end kinesin depolymerase kar3 to microtubule plus ends. *Current Biology*, 15:1420–1427, 2005.

[175] John S Allingham, Lisa R Sproul, Ivan Rayment, and Susan P Gilbert. Vik1 modulates microtubule-Kar3 interactions through a motor domain that lacks an active site. *Cell*, 128:1161–1172, 2007.

[176] David L Des Marais and Mark D Rausher. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, 454(7205):762–5, August 2008.

[177] Song Liu, Kirilee A Wilson, Travis Rice-Stitt, Aaron M Neiman, and James A McNew. In vitro fusion catalyzed by the sporulation-specific t-SNARE light-chain Spo20p is stimulated by phosphatidic acid. *Traffic Copenhagen Denmark*, 8:1630–1643, 2007.

[178] Beth A Montelone, Mer F Hoekstra, and Robert E Malone. Spontaneous Mitotic Recombination in Yeast: The Hyper-Recombinational. *Genetics*, 119(2):289–301, 1988.

[179] D J Keszenman, V A Salvo, and E Nunes. Effects of bleomycin on growth kinetics and survival of Saccharomyces cerevisiae: a model of repair pathways. *Journal of bacteriology*, 174(10):3125–32, May 1992.

[180] J G Cook, L Bardwell, S J Kron, and J Thorner. Two novel targets of the MAP kinase Kss1 are negative regulators of invasive growth in the yeast Saccharomyces cerevisiae. *Genes & development*, 10(22):2831–48, November 1996.

[181] S Izawa, K Maeda, K Sugiyama, J Mano, Y Inoue, and A Kimura. Thioredoxin deficiency causes the constitutive activation of Yap1, an AP-1-like transcription factor in Saccharomyces cerevisiae. *The Journal of biological chemistry*, 274(40):28459–65, October 1999.

[182] Gabriel Musso. *The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast*. Phd, University of Toronto, 2010.

[183] James E Haber, Hannes Braberg, Qiuqin Wu, Richard Alexander, Julian Haase, Colm Ryan, Zach Lipkin-Moore, Kathleen E Franks-Skiba, Tasha Johnson, Michael Shales, Tineke L Lenstra, Frank C P Holstege, Jeffrey R Johnson, Kerry Bloom, and Nevan J Krogan. Systematic triple-mutant analysis uncovers functional connectivity between pathways involved in chromosome regulation. *Cell Reports*, 3(6):2168–78, June 2013.

[184] Sean R Collins, Maya Schuldiner, Nevan J Krogan, and Jonathan S Weissman. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome biology*, 7(7):R63, January 2006.

[185] Jian Zou, Helena Friesen, Jennifer Larson, Dongqing Huang, Mike Cox, Kelly Tatchell, and Brenda Andrews. Regulation of Cell Polarity through Phosphorylation of Bni4 by Pho85 G1 Cyclin-dependent Kinases in Saccharomyces cerevisiae. *Molecular Biology of the Cell*, 20:3239 –3250, 2009.

[186] I V Karpichev and G M Small. Global regulatory functions of Oaf1p and Pip2p (Oaf2p), transcription factors that regulate genes encoding peroxisomal proteins in Saccharomyces cerevisiae. *Molecular and cellular biology*, 18(11):6560–70, November 1998.

[187] Joanna Trzcinska-Danielewicz, Takao Ishikawa, Arkadiusz Miciakiewicz, and Jan Fronk. Yeast transcription factor Oaf1 forms homodimer and induces some oleate-responsive genes in absence of Pip2. *Biochemical and biophysical research communications*, 374(4):763–6, October 2008.

[188] Patrick a Gibney, Charles Lu, Amy a Caudy, David C Hess, and David Botstein. Yeast metabolic and signaling genes are required for heat shock survival and have little overlap with the heat-induced genes. *Proceedings of the National Academy of Sciences*, 110(46):E4393–402, November 2013.

[189] Thouis R Jones, In Han Kang, Douglas B Wheeler, Robert a Lindquist, Adam Papallo, David M Sabatini, Polina Golland, and Anne E Carpenter. CellProfiler Analyst: data exploration and analysis software for complex image-based screens. *BMC bioinformatics*, 9:482, January 2008.

[190] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The KEGG resource for deciphering the genome. *Nucleic acids research*, 32(Database issue):D277–80, January 2004.

[191] Jan Schellenberger, Richard Que, Ronan M T Fleming, Ines Thiele, Jeffrey D Orth, Adam M Feist, Daniel C Zielinski, Aarash Bordbar, Nathan E Lewis, Sorena Rahmanian, Joseph Kang, Daniel R Hyduke, and Bernhard ØPalsson. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature protocols*, 6(9):1290–307, September 2011.

[192] Hermann-Georg Holzhütter. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *European journal of biochemistry / FEBS*, 271(14):2905–22, July 2004.

[193] Anne-Claude Gavin, Patrick Aloy, Paola Grandi, Roland Krause, Markus Boesche, Martina Marzioch, Christina Rau, Lars Juhl Jensen, Sonja Bastuck, Birgit

Dümpelfeld, Angela Edelmann, Marie-Anne Heurtier, Verena Hoffman, Christian Hoefert, Karin Klein, Manuela Hudak, Anne-Marie Michon, Malgorzata Schelder, Markus Schirle, Marita Remor, Tatjana Rudi, Sean Hooper, Andreas Bauer, Tewis Bouwmeester, Georg Casari, Gerard Drewes, Gitte Neubauer, Jens M Rick, Bernhard Kuster, Peer Bork, Robert B Russell, and Giulio Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, March 2006.

[194] Mohan Babu, James Vlasblom, Shuye Pu, Xinghua Guo, Chris Graham, Björn D. M. Bean, Helen E. Burston, Franco J. Vizeacoumar, Jamie Snider, Sadhna Phanse, Vincent Fong, Yuen Yi C. Tam, Michael Davey, Olha Hnatshak, Navgeet Bajaj, Shamanta Chandran, Thanuja Punna, Constantine Christopolous, Victoria Wong, Analyn Yu, Gouqing Zhong, Joyce Li, Igor Stagljar, Elizabeth Conibear, Shoshana J. Wodak, Andrew Emili, and Jack F. Greenblatt. Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. *Nature*, 489(7417):585–9, September 2012.

[195] Kirill Tarassov, Vincent Messier, Christian R Landry, Stevo Radinovic, Mercedes M Serna Molina, Igor Shames, Yelena Malitskaya, Jackie Vogel, Howard Bussey, and Stephen W Michnick. An in vivo map of the yeast protein interactome. *Science (New York, N.Y.)*, 320:1465–1470, 2008.

[196] Haiyuan Yu, Pascal Braun, Muhammed a Yildirim, Irma Lemmens, Kavitha Venkatesan, Julie Sahalie, Tomoko Hirozane-Kishikawa, Fana Gebreab, Na Li, Nicolas Simonis, Tong Hao, Jean-François Rual, Amélie Dricot, Alexei Vazquez, Ryan R Murray, Christophe Simon, Leah Tardivo, Stanley Tam, Nenad Svrzikapa, Changyu Fan, Anne-Sophie de Smet, Adriana Motyl, Michael E Hudson, Juyong Park, Xiaofeng Xin, Michael E Cusick, Troy Moore, Charlie Boone, Michael Snyder, Frederick P Roth, Albert-László Barabási, Jan Tavernier, David E Hill, and Marc Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science (New York, N.Y.)*, 322(5898):104–10, October 2008.

[197] Enrico Ragni, Thierry Fontaine, Carmela Gissi, Jean Paul Latgè, and Laura

Popolo. The Gas family of proteins of Saccharomyces cerevisiae: characterization and evolutionary analysis. *Yeast Chichester England*, 24:297–308, 2007.

[198] E Ragni, A Coluccio, E Rolli, J M Rodriguez-Pena, G Colasante, J Arroyo, A M Neiman, and L Popolo. GAS2 and GAS4, a pair of developmentally regulated genes required for spore wall assembly in Saccharomyces cerevisiae. *EukaryotCell*, 6:302–316, 2007.

[199] Evelyn Sattlegger, Mark J Swanson, Emily A Ashcraft, Jennifer L Jennings, Richard A Fekete, Andrew J Link, and Alan G Hinnebusch. YIH1 is an actin-binding protein that inhibits protein kinase GCN2 and impairs general amino acid control when overexpressed. *The Journal of Biological Chemistry*, 279:29952–29962, 2004.

[200] Adam M Deutschbauer, Daniel F Jaramillo, Michael Proctor, Jochen Kumm, Maureen E Hillenmeyer, Ronald W Davis, Corey Nislow, and Guri Giaever. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics*, 169(4):1915–25, April 2005.

[201] Zhenglong Gu, Andre Cavalcanti, Feng-Chi Chen, Peter Bouman, and Wen-Hsiung Li. Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. *Molecular Biology and Evolution*, 19:256–262, 2002.

[202] B Rost. Twilight zone of protein sequence alignments. *Protein Engineering*, 12:85–94, 1999.

[203] Matthew A Hibbs, David C Hess, Chad L Myers, Curtis Huttenhower, Kai Li, and Olga G Troyanskaya. Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, 23:2692–2699, 2007.

Appendix A

Appendix for Chapter 2

A.1 Construction of a prototrophic delection collection

As recently described [188], the strains in the standard *MAT***a** deletion collection (*MAT***a** *yfg*Δ*0::KanMX his3*Δ*1 leu2*Δ*0 met15*Δ*0 ura3*Δ*0*) [3] were mated to a *MAT*α *can1*Δ *::STE2pr–SpHIS5 his3*Δ*1 lyp1*Δ*0* strain, creating diploids (selection on minimal media + his + G418). These were sporulated and successive pinnings on selective media were used to select prototrophic *MAT***a** strains carrying each deletion allele. These prototrophic strains were organized into an array of 16 plates including one entire plate of the wild-type strain (*ho*Δ*::KanMX*), with additional wild-type replicates in each row and column of every plate (701 in all). The entire prototrophic collection is available upon request, as is the individual SGA-ready prototroph strain for crossing into other collections.

A.2 Media preparation

Minimal growth media were prepared using yeast nitrogen base (BD Difco, Sparks, Maryland, USA) with the specified carbon and nitrogen sources. Carbon sources included glucose, galactose, ribose, and glycerol. Nitrogen sources included ammonium, allantoin, arginine, glutamate, glutamine, proline, and urea. Carbon sources were provided at a concentration of 2%; nitrogen sources were 3.8 mM with respect to nitrogen.

A.3 Calculation of growth rate data

Sixteen 16×24 well plates were grown in 28 chemical conditions for 24–48 hours. Pictures were taken at 0, 5, 10, 24 and, in the case of glycerol, 48 hours. Each condition is composed of one carbon source and one nitrogen source. In total, 4,772 mutants were grown, and colony areas were extracted from tiff images by CellProfiler [189] and precise time points were take from EXIF data in the digital images. These values were used to compute an estimate of the growth rate of each colony equal to the slope of the least-squares linear fit of area (pixels) to time (seconds). Colonies with insufficient data were given a growth rate of NaN, colonies with a negative calculated growth rate were defined to have a growth rate of 0.

A.4 Definition and construction of a reference condition

Six replicates of the glucose:ammonium combination were merged to form a reference condition, establishing a baseline score for each deletion. The six replicates were first normalized to each other to control for differences in the overall scale of growth rates, then averaged together according to the following procedure. For each array plate (p) the glucose:ammonium replicate with the fewest missing data points was held out (PlateA) and the remaining five replicates were LOWESS smoothed (window size = 50% of available data) and normalized by:

$$GA : plate'_p = GA : plate_p \times \frac{Plate_A}{lowess(GA{:}plate_p)}$$

The result of this approach is quite robust to the choice of $Plate_A$, and so we used whichever replicate had the fewest number of missing values and would therefore provide the most complete LOWESS fit. After normalizing five replicates to the sixth, all six were averaged together to create one reference plate, and this procedure is repeated 16 times to create a glucose:ammonium reference for each array plate.

A.5 Normalization of experimental rates against reference

In every experimental condition(y), each plate was LOWESS smoothed (window size = 50% of available data) against the constructed glucose:ammonium reference plate, then

normalized:

$$Cond_y : plate'_p = Cond_y : plate_p \times \frac{GAref{:}plate_p}{lowess(Cond_y{:}plate_p)}$$

A.6 Recovery of missing data

In certain cases a growth rate of NaN was assigned to a colony due to insufficient data being collected by CellProfiler[189]. In an effort to recover any good data, these cases were visually inspected by five researchers operating independently and a vote was taken to determine whether to leave it as missing data (NaN) or assign it a growth rate of 0, indicating that the colony appeared to be correctly plated but non-viable. In total 1,362 of 2,601 colonies were recovered this way.

A.7 Transformation from normalized rates to z-scores

For each array plate, at each position, a strain-wise standard deviation is calculated across the residuals of the six glucose:ammonium (GA) replicates. Similarly, a plate-wise standard deviation is calculated that accounts for the general growth variation on the plate, separately for each condition. These are then combined, and a z-score measure is calculated for each strain on each experimental plate:

$$z = \frac{Cond_y{:}plate'_p - GAref{:}plate_p}{\sqrt{stddev(strain)^2 + stddev(plate)^2}}$$

These z-scores are an expression of the difference in magnitude and direction between the growth observed at each position of a plate under a given condition from the same position (and hence deletion) under the reference GA model.

A.8 Spatial smoothing procedure

The plate level spatial smoothing filter is similar to that found in [50]. First, temporarily replace any extreme values (top and bottom 5%) along with NaNs with the plate mean. Second, replace previous NaN positions with values from a two-dimensional symmetric gaussian filter. Third, compute and subtract the residual between the two-dimensional smoothed plate and its mean.

A.9 Choosing effect thresholds

Each condition had 701 wild-type replicates. The mean and standard deviation of the set of wild-type z-scores were used to define a normal distribution against which p-values for the experimental z-scores could be calculated. This information allowed the use of Benjamini-Hochberg procedure to establish condition-specific effect thresholds as a function of a desired FDR. See Table A.1

Condition	slow	fast
glucose ammonium01	-1.390896	1.247624
glucose ammonium02	-1.533445	1.541672
glucose ammonium03	-1.657256	1.651313
glucose ammonium04	-2.096102	NaN
glucose ammonium05	-1.104211	1.095158
glucose ammonium06	NaN	2.094176
glucose proline	-1.199736	1.235451
glucose glutamate	-1.069737	1.242443
glucose glutamine	-1.126722	1.231199
glucose arginine	-0.688915	1.167267
glucose urea	-1.389899	1.572141
glucose allantoin	-1.232272	1.33048
galactose ammonium	-1.078656	1.180252
galactose proline	-0.814656	0.923763
galactose glutamate	-0.616518	0.783427
galactose glutamine	-0.626531	0.862998
galactose arginine	-1.024493	1.047379
galactose urea	-1.452546	1.678125
galactose allantoin	-1.066719	1.290977
ribose ammonium	-1.18146	1.12634
ribose proline	-1.051107	0.671917
ribose glutamate	-1.113968	1.113366
ribose glutamine	-1.384957	1.268024
ribose arginine	-1.582742	1.444913
ribose urea	-1.227792	1.164294
ribose allantoin	-1.067669	1.025517

Table A.1: FDR 20% thresholds for z-score data. For each condition the z-score values at which a growth deviation was deemed significant are shown here.

A.10 Liquid growth confimation assay

The growth rate of 40 mutants in a liquid growth assay was measured across 20 of the experimental conditions excluding ribose:arginine and all glycerol pairings. Liquid culture assays were not performed for the ribose:arginine conditions because the combination of these carbon and nitrogen sources did not allow arginine to maintain adequate solubility over the duration of the experiment. The precipitation of arginine prevented accurate optical density readings from being obtained and thus these data were excluded from our subsequent analyses. Six replicate wells contained the wild-type strain and each mutant strain was represented twice. Cells were pre-grown on glucose:ammonia medium and diluted at a low density into the growth medium of interest. Growth rates were determined as the maximum optical density (saturation) divided by the time to saturation. A simple model was favored in order to robustly accommodate drastic differences in curve characteristics between fast growth and slow growth conditions (for example, galactose versus ribose).

We adjusted the liquid growth scores by dividing the mean of mutant growth slopes by the mean of wild-type growth slopes in the relevant condition. We further normalized these scores by dividing them by the corresponding adjusted mutant score in glucose:ammonium so they would reflect condition-specific effects, similar to our modified z-score derived from the agar experiment.

A.11 Gene Ontology and KEGG annotations

GO [93] and KEGG [190] annotations were downloaded in January 2011.

A.12 Constraint-based metabolic modeling (FBA/MoMA)

Two *S. cerevisiae* metabolic models were used for mutant biomass prediction. The Yeast Consensus Reconstruction version 5.35 (Yeast5) [96] and iMM904 [97]. Yeast5 consisted of 898 ORFs, 2,031 reactions and 1,594 metabolites and the iMM904 model contained 901 ORFs, 1,597 reactions and 1,234 metabolites. Default biomass descriptions were used for both models.

Wild-type biomass production flux for each condition was obtained using FBA [98] in MATLAB with the COBRA Toolbox [191], which assumes optimal biomass production (that is, maximum biomass yield). Mutant biomass flux was predicted using both FBA [98] in MATLAB with the COBRA Toolbox [191] and MoMA [99] in MATLAB with the ILOG CPLEX optimization suite. MoMA was formulated as a quadratic programming problem, whereby mutant fluxes were selected that minimized the Euclidean distance from an optimal wild-type flux distribution. The yeast wild-type flux distribution was calculated as a network flux solution producing maximal biomass flux, determined by FBA, with minimal total fluxes [192].

FBA and MoMA biomass fluxes were correlated with both raw and normalized (z-score) experimental growth rates using the Spearman rank correlation. Predicted biomass fluxes were also normalized for comparison to experimental growth rate z-scores (separately in each condition Y):

$$MutantFlux_{normY} = \frac{MutantFlux_{rawY}}{MutantFlux_{rawGlu:Amm} \times WildTypeFlux_{raw\ \ Glu:Amm}}$$

Prediction of positive z-scores was also carried out, though performance was generally below random expectation (Tables 2.5–2.6). This is likely due to the fact that many positive z-scores corresponded to raw growth rates for mutants that were faster than wild-type under the same condition, a consequence that FBA- and MoMA-based methods would find difficult or impossible to predict.

To calculate the effect of gene deletions on the metabolic network (Fig. 2.7), sets of producible metabolites were calculated for the complete model, and for a mutant with all four auxotrophic marker genes deleted. Producible metabolites were calculated for both iMM904 and Yeast5 model in the glucose:ammonium media condition by adding a special exchange reaction for each metabolite and iteratively optimizing flux exported through that reaction. If the export flux for a given metabolite exceeded 0.001 (with an upper and lower bound on internal reactions set to ± 1,000), it was classified as "producible." A non-zero threshold is required to limit false positives as a result of numerical errors. The threshold was determined to be robust by scaling the upper and lower bounds, as well as the threshold by a large constant and counting the number of producible metabolites. Obtaining consistent results in these experiments led us to conclude that numerical errors are an order of magnitude smaller than contributions

from stoichiometry.

A.13 Source signature decomposition via modified non-negative matrix factorization

Growth data was decomposed using a variant of NMF [108]. Following transformation to z-scores, the data were made binary using condition-specific FDR estimates as thresholds (20% FDR; Tables 2.1,A.1). The resulting boolean *Data* matrix was treated as numeric and served as the target for decomposition. Genes without any significant z-scores in any condition (empty rows) were removed, as were the columns involving growth on glycerol. We then defined a *Coefficient* matrix that related *Condition* rows in the data to their component *Sources*. This matrix then had C columns and S rows. For example, the glucose-urea column has a 1 in the glucose row and a 1 in the urea row. Our task is then to find a *Signatures* matrix ($Genes \times Sources$) such that the difference between the *Data* matrix and the *Sigantures-Coefficients* product is minimized:

$$Data(G, C) \approx Signatures(G, S) \times Coefficients(S, C)$$

To ensure linear independence among the columns of the *Coefficient* matrix, we removed all but one glucose:ammonium column (glucose:ammonium01), removing the same columns in the *Data* matrix. Traditional NMF would use a multiplicative update algorithm applied to both the *Signature* and the *Coefficient* matrix to find the best fit to the *data*; however, we chose to fix the *Coefficient* matrix at the initial defined values (0 or 1). This gives each *Signature* column equal weight and prevents over-fitting caused by the sparsity of the *Data* matrix and the dramatically different number of non-zero elements from one column to the next. The multiplicative update was applied for 20 iterations, though in practice the results converged in fewer than 10, and repeated trials from different random initializations of the *Signature* matrix showed the results to be quite stable. Genes were considered part of a signature if their value exceeded 0.4.

A.14 Comparison to SGA data

For the comparison to auxotrophic SGA data represented in Fig. 2.11, the SGA data were taken from [66]. The SGA data and the z-score data were independently normalized

so that row and column vectors had a euclidean length approximately equal to 1, and missing values were set to 0. Inner product was then used to measure the similarity between SGA "queries" and environmental profiles. The top 10% of queries in each condition were checked for enrichment for GO terms and KEGG pathway annotations, and the resulting p-values were Bonferonni corrected to account for the number of terms/pathways tested against.

A.15 Abbreviations

FBA, flux balance analysis; FDR, false discovery rate; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; MoMA, minimization of metabolic adjustment; NMF, non-negative matrix factorization; ORF, open reading frame; SGA, synthetic genetic array.

Appendix B

Appendix for Chapter 3

B.1 Supplementary materials and methods

B.1.1 SGA array normalization

In order to combine SGA data from each of our two array experiments in a meaningful way, we used the following procedure to adjust ϵ scores from the *TS_array* to resemble scores from the *FG_array*. The two datasets were intersected on their common set of 1,931 queries and 175 array genes. The result is approximately 316,000 pairs of matching ϵ observations. Direct scaling of one set of data to match the other via least-squares fit produced unsatisfactory results due to the high variance of the data relative to its correlation. We therefore decided to scale the data to minimize the differences in the distribution of common ϵ scores. We used quantile normalization on the common set of scores, and in the process built a table of quantile normalization values, which then could be applied to the remaining *TS_array* interactions. This procedure ensures an identical distribution of common scores, without constraining the entire distribution of *TS_array* scores. These two goals must be achieved in order to use the same threshold on both datasets, and to compare the resulting total number of interactions at a given threshold.

B.1.2 SGA gold standard definition

For queries that were screened at least 5 times, we created a gold-standard for both positive and negative interactions. Any interaction seen twice, at the intermediate threshold, is included in the standard. For queries with more than five replicates, the five replicates with the fewest missing values were chosen.

B.1.3 Protein-protein interaction data

For the purposes of this chapter, protein protein interaction data is taken from BioGrid [70] and represents the union of five high-throuput studies: Gavin *et al.* (PMID: 16429126 [193]), Babu *et al.* (PMID: 22940862 [194]), Krogan *et al.* (PMID: 16554755 [65]), Tarassov *et al.* (PMID: 18467557 [195]), Yu *et al.* (PMID: 18719252 [196]).

B.1.4 Gene Ontology terms for functional prediction

In order to make predictions across many diverse processes in a non-redundant way, we used a subset of the Gene Ontology referred to as the "fringe" set. GO terms in the fringe set (119) span the entire GO "process" tree and were selected to be specific enough to be functionally informative, yet large enough to be useful in prediction scenarios. They were also selected so as to be as non-overlapping as possible [94].

B.1.5 Array gene function prediction via KNN

Deletion array genes were classified into functional categories using a variation of k-nearest neighbors with leave one out cross-validation. Functional categories for classification were taken from a subset of the Gene Ontology (See Sec. B.1.4), and classification on each GO term was performed separately to accommodate multiple annotations. Terms with fewer than 10 participating genes were removed leaving 132 GO terms for prediction. Array genes with no annotations were deemed useless for prediction as these would provide no information for classification of other genes and would be impossible to correctly classify themselves. These array genes were therefore removed.

Each gene received a score for every GO term in following way. The K (=5) largest similarity scores between the gene in question and members of the term in question are summed. Similarity scores are calculated as inner products between array gene profiles,

using a subset of queries. Genes are then ranked within each term according to this summed similarity with known term participants. This process was then repeated using different subsets of SGA queries to calculate similarities between array genes. No more than one allele of each gene was used in either the DAmP or TS analysis, and one allele of each was selected at random.

To control for the number of available features of each respective query type, we selected 100 queries of each type at random. There are approximately 100 DAmP queries in the collection, so this represents the largest number of features at which performance can be measured fairly. Results shown in Fig. 3.9 represent means, bootstrapped over 50 iterations of random query-feature selection.

B.1.6 Hierarchical cluster filtering

The bounds on correlation coefficients which were used to define clusters at each level can be seen in Table B.1. Hierarchical clustering inevitably puts every gene in a cluster, however this can lead to clusters which are driven by noise in screens with little genetic signal and therefore carry no functional information. To mitigate the impact of these clusters on the overall analysis we filtered out genes based on two criteria: *i* genes must participate in at least one functionally informative and specific cluster. That is, they must belong to a cluster at level 3 or deeper with significant enrichment for annotations to a GO term. *ii* genes must be included in a cluster at level 5. A number of low-signal individual genes joined the only at relatively high linkage bounds, because they did not significantly correlate to any other genes. Mandating cluster membership at level 5 (and therefore at levels 1–4) ensures that we are always using the same set of genes, and therefore degree distributions, when making comparisons from one level to another.

Level	Linkage bound	Pearson equivalent	Clusters	Average size
1	2.00–0.95	-1.00–0.05	1	925
2	0.95–0.80	0.05–0.20	10	93
3	0.80–0.60	0.20–0.40	50	19
4	0.60–0.40	0.40–0.60	112	8
5	0.40–0.20	0.60–0.80	231	4

Table B.1: Hierarchical clusters after filtereing. The table shows the linkage thresholds used to determine each "level," as well as the number of clusters that result at each level, and their average size. Because genes which did not fall into any cluster at level 5 were filtered out, and each cluster contains all gene-members of its children clusters, the total number of genes included is the same at every level (401 essential genes, and 524 non-essential genes).

B.2 Supplementary figures

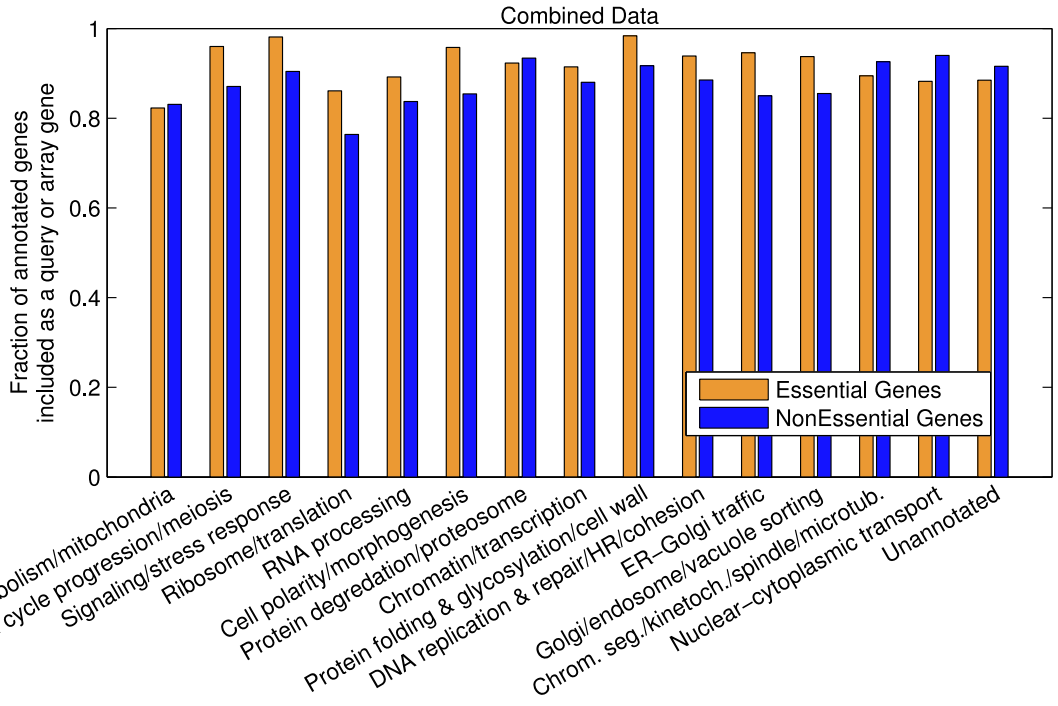

Figure B.1:
The proportion of non-dubious genes included as either a query or array (on either array) is shown for each of 14 broad functional categories. Also included is the proportion of genes with no functional annotation according to this scheme. The fraction is shown separately for essential and non-essential genes.The vast majority of genes from every major biological process are included in the experiment as are ∼90% of unannotated genes.

Appendix C

Appendix for Chapter 4

C.1 Ribosomal Duplicates

Ribosomal duplicates constitute a sizable fraction of whole-genome duplicates, and their direct impact on growth rate means they present a very strong signature within the genetic interaction network. They also (reassuringly) represent a significant portion of our defined dosage class. To ensure that our results are not overly influenced by this characteristic signature, we here present a summary of statistics which are affected by the removal of ribosomal duplicates from consideration. Statistical tests were repeated as in the main text with any duplicate pair with an annotation in the "Translation" Gene Ontology term being removed. First, the direction and significance of most of the core statistical results was maintained after removing the ribosomal duplicates: Fig. 4.2A: genetic interaction rate among duplicates; Fig. 4.2C: evidence for fewer genetic interactions among duplicates; Fig. 4.2D: evidence for lower profile similarity among duplicate pairs than protein- protein interaction pairs; Fig. 4.3B: the shared protein-protein interaction partner dosage/divergent selection analysis; Fig. 4.4A: asymmetry of duplicate pairs interaction degree; Fig. 4.4B: relationship between genetic interaction asymmetry and other functional data; and 4.4C: the relation between the high/low degree sister and the singleton average.

For Fig. 4.3A, which showed the difference in profile similarity between dosage and divergent pairs, the medians trend in the same direction, but the difference between dosage and divergent pairs is not significant after removal of the ribosomal pairs due

to a loss of many pairs in the dosage class. All other distinctions on that figure (e.g. the PPI/divergent difference) remain significant, and notably, Fig. 4.3B demonstrates a similar conclusion using a different approach and is statistically significant after removing the ribosome. Interestingly, the result from Fig. 4.2B, the difference in synthetic sick/lethal interaction rates between WGD and SSD pairs, appears to be explained by the ribosomal duplicates as this difference is no longer significant after removing the ribosome: 17 out of 23 (74%) screened ribosomal WGD pairs are synthetic sick/lethal, which is much higher than the rate for non-ribosomal WGD pairs (28%). Finally, the difference in the synthetic lethality rate for symmetric vs. asymmetric duplicates presented in Fig. C.3A becomes significant ($p < 1 \times 10^{-2}$) as many of the (often synthetic lethal) ribosomal duplicates fall into the symmetric class. Our conclusions based on this result (i.e. that asymmetric duplicates show negative interactions at least as frequently as symmetrically diverged duplicates) remain unchanged.

C.2 Sequence evolution rates support selection class distinction

Cross-referencing our dosage-mediated and divergent duplicate sets with slowly and quickly evolving pairs from Kellis *et al.* [42] revealed another connection in principle. Selecting pairs whose sequences appear to be diverging very slowly, we found an enrichment for paralogs in the dosage set. Specifically, out of the 372 pairs that existed in the referenced study and had appropriate classification data, 41 were classified as slowly diverging, and 45 were annotated as dosage. The overlap between these two sets (14 pairs) proved to be significant ($p < 7 \times 10^{-5}$; hypergeometric cdf). A similar comparison showed that all but three pairs from the quickly evolving set (totaling 89 pairs) belonged to the functionally divergent set ($p < 2 \times 10^{-3}$; hypergeometric cdf), again supporting the distinction between the two sets. Duplicate pairs that are performing the same functions, and therefore must be retained to maintain dosage levels, would be under equal and symmetric selection against change, and therefore exhibit a very slow rate of divergence. Meanwhile, pairs which are maintained because they are upholding even slightly different responsibilities would be under far less sequence preservation pressure and are therefore far less constrained.

C.3 Genetic interactions highlight the divergence of *GAS1* and *GAS2*

GAS1 and *GAS2*, are extremely asymmetric in the number of interactions they exhibit. Both of these genes are involved in the maintenance of the cell wall, but appear to be utilized under very different contexts [197]. *GAS1* has 139 negative genetic interactions, and the genes with which it interacts are enriched for annotations to GO processes relating to cellular structure and morphogenesis (GO:0032989; $p < 6 \times 10^{-4}$). It is required for cell wall assembly, and expressed during normal vegetative growth. *GAS2*, by contrast, has only 7 negative genetic interactions and is expressed exclusively during sporulation, where it is required for spore wall assembly [198]. These two genes may be performing very similar tasks in the construction of similar cellular structures, and yet they share only one negative genetic interaction with *YIH1*, which affects gene expression in response to starvation [199]. Presumably, starvation triggers the cellular switch from a context where *GAS1* is used in the construction and maintenance of normal cell and bud wall material, to a context where *GAS2* is instead used in the construction of spore wall. Before the small scale duplication event from which this pair arose, it is conceivable that these roles were upheld by a single ancestral *GAS* gene, and one modern copy now carries the burden of the responsibilities while the other operates on a very specific subset.

C.4 Self-reinforcing model of duplicate divergence.

We propose a model for self-reinforcing asymmetric divergence of duplicate genes which relies only on the relaxation of negative selective pressure resulting from genetic redundancy and loss-of-function mutations. The key observation of the model in comparison to previous attempts at explaining asymmetric divergence is that while mutations occur in sequence space, selection ultimately acts on function space, and thus, a single change at the amino acid level may affect multiple functions of a given protein. Similarly, a given function may have been lost due to any one of a number of mutations (Fig. C.5). Thus, we developed a simple discrete formulation of this model, assuming multiple functions (> 1) per protein, and that random mutations in sequence space may

have K (> 1) effects in function space. This represents an update of a previous model that attempted to show asymmetry as a result of only negative selection pressure and loss-of-function mutations [152].

However, this previous model assumed that either duplicate gene had an equal probability of acquiring an additional loss-of-function mutation, which is not realistic in the case that one gene has already lost much of its function due to degenerating mutations. Our assumption allowing single sequence mutations to affect more than one function predicts asymmetry without requiring the assumption of equally probable loss-of-function mutations. Consider two divergent duplicate genes with N functions. These N functions belong 3 different categories. Those lost in duplicate 1 (l_1), those lost in duplicate 2 (l_2), and those redundant functions, which are lost in neither (R)(Fig. C.6A). We assume that a mutation resulting in the loss of a function which has already been compromised in the sister duplicate will be deleterious and unsustainable. Given this formulation, the probability that the region l_1 increases via a sustainable mutation (with k out of K effects in region R) increases with l_1 itself (Fig. C.6B).

$$P(l_1 \ increase) = \sum_{k=1}^{K} \frac{\binom{R}{k} * \binom{l_1}{K-k}}{\binom{N}{K}}$$

We can further generalize this model beyond this discrete formulation by formalizing the relationship between mutations at the sequence level and their consequences at the functional level. If we have a duplicate pair G_1 and G_2, either may accumulate mutations freely immediately after duplication due redundancy provided by the sister gene. However, if one gene has a mutation that seriously impinges on one of its major functions (F_1), any mutation in the sequence regions that support F_1 that would lead to a similar loss of function in G_2 is selected against because the cell presumably would incur a fitness penalty if the function F_1 is lost in both sisters. At the same time, G_1 can have mutations in sequence regions that only effect F_1 or any remaining redundant functions shared by G_1 and G_1 (call this F_R). In this manner, G_1 continues to accumulate mutations that reduce it's functionality until it is completely non- functional, or until G_1 has a mutation that impinges on a different function F_2, causing the corresponding sequence that supports F_2 to be conserved in G_1.

If we let S_1, S_2, and S_R be the parts of the sequence that correspond to functions F_1, F_2 and F_R respectively, and assume that mutations happen with equal probability

at any point of the sequence, we can find the probability of having a mutation in G_1 that incurs an additional loss of function for G_1 without impinging on any functions already lost to G_2:

$$\mathrm{P}(S_R \cap S_2^c) = \mathrm{P}(S_R \cap (S_1 \cup S_2)^c) \quad [\mathbf{A}] \quad + \mathrm{P}((S_R \cap S_1) \cap S_2^c) \quad [\mathbf{B}]$$

The key point is that while **A** (the sequence regions that support only F_R but not F_1 or F_2) is the same for G_1 and G_2, **B** (the sequence regions that support F_R or F_1 but not F_2) is likely larger for G_1 than the equivalent term for G_2 ($\mathrm{P}((S_R \cap S_1) \cap S_2^c)$) if $S_1 > S_2$. Therefore a gene that has already lost more functionality is likely to have more sustainable mutations that result in an even greater loss of function. Yet the sequence-function relationship structure may forbid particular redundant functions from falling to one duplicate or the other (Fig. C.5). This framework then not only explains the asymmetry observed between duplicate pairs in function-based and sequence based studies, but also accounts for the high degree of retained functional overlap among even the most asymmetric duplicate pairs.

C.5 Supplementary figures

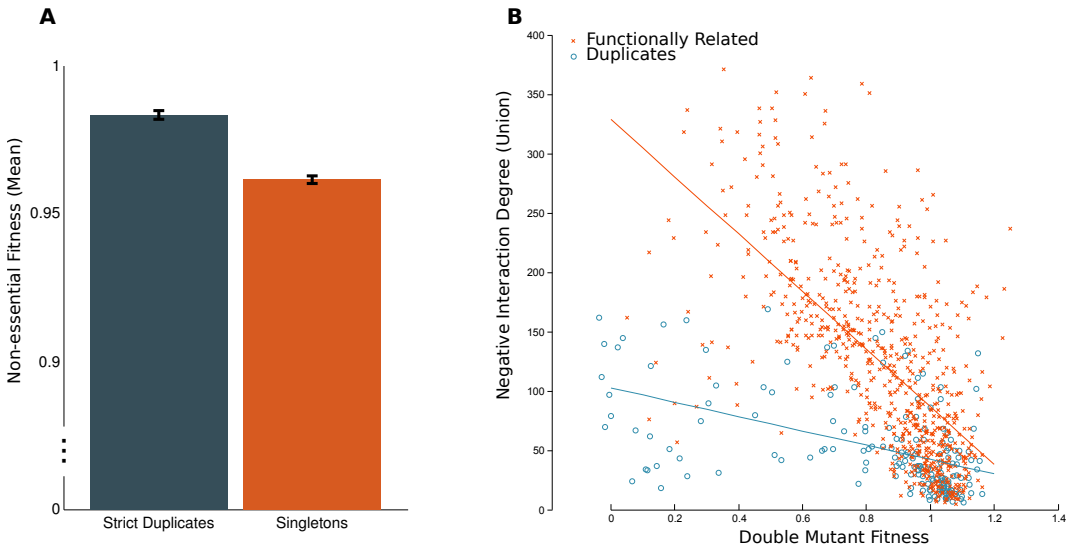

Figure C.1: **(A)** Duplicate deletion has less impact on cell fitness. Around 17% of all genes in the yeast genome are annotated as essential. The essential rate of duplicates in the small family ($n = 2$) set is much lower (5%). When deleted, small-family non-essential duplicates have less of an impact on cell fitness than do their singleton counterparts (ranksum $p = 6 \times 10^{-5}$). **(B)** Duplicate genetic interactions cannot account for their double-mutant fitness. In previous work we found a strong correlation ($r = 0.7$, [66] between the single-mutant fitness defect of a gene and its genetic interaction degree. We used this idea to control for the *importance* of each duplicate pair by fitting a linear model of the pair's double-mutant fitness (DMF) to the union of their genetic interaction degree. We then did the same for functionally related pairs and found that after controlling for the DMF, duplicate pairs had fewer interactions than expected (shown here). For example, the transcription factors *STP1* and *STP2* both activate the transcription of amino acid permease genes in response to extra-cellular stimuli, and exhibit an SSL interaction with each other. Their double-mutant fitness (0.48) would predict they show some 220 interactions. However, their combined profile contains only 103 SSL interactions (and they share only 6). Taking the DMF of a duplicate pair to approximate the SMF of the pair's ancestor [139], this result suggests that the union of their interactions (an approximation of the ancestor's interactions) is missing interactions supposedly possessed by the ancestor. We submit that these missing interactions are buffered due to retained redundancy, but represent real functional consequence as evidenced by their effect on double-mutant fitness.

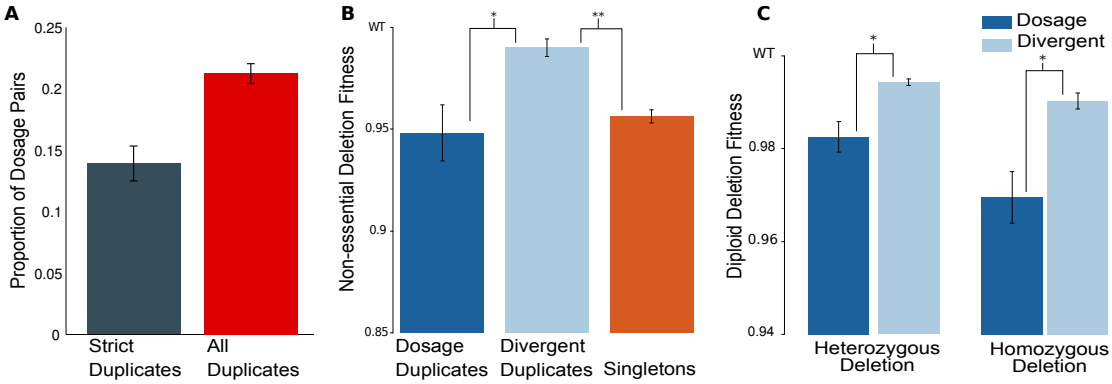

Figure C.2: **(A)** Proportion of Duplicates in "Dosage" class. Approximately 14% of duplicate pairs in this study fall into the dosage category. However, if we apply the threshold criteria derived from the small-family set (Methods) to duplicates in general, we see that about 21% of duplicates would belong to the dosage class. The difference between these two proportions is significant ($p = 3 * 10^{-5}$) as the inclusion of larger gene families naturally picks up many genes which have higher phylogenetic volatility scores. These are canonical dosage-mediated pairs, and may have been duplicated for the sole purpose of increasing product quantity. However, the thresholds determined on the small family set may not be appropriate for duplicates in general. Interestingly, SSD paralogs appear to have a higher proportion of dosage- mediated pairs than WGD pairs (22% vs 12%, $p < 5 \times 10^{-3}$, see Methods C.6). We speculate that this difference may stem in part from the unique balance opportunities that a whole-genome duplication event might provide, possibly allowing greater tendency towards functional specialization. **(B)** Dosage and Divergent genes show fitness differences in haploid deletion assays. The buffering model predicts dosage genes will have more of an impact than divergent genes when deleted individually. Means are shown and error bars represent the standard deviation on the mean over 1000 bootstrapped samples. Duplicates classified as "Dosage"(Methods C.6) have a significantly higher fitness impact than do other (Divergent) duplicates when deleted. (* $p < 6 \times 10^{-4}$; Wilcoxon rank-sum test) Duplicates retained for partial divergence show much less of an impact on fitness than do non-duplicates. (** $p < 2 \times 10^{-9}$) The difference in single-mutant fitness between Dosage duplicates and singletons is not significant. ($p > 0.3$) **(C)** Independent data confirms difference in deletion fitness for dosage and divergent genes The buffering model here presented predicts dosage genes will have more of an impact than divergent genes when deleted individually. We turned to an independent study [200] and found that this hypothesis is upheld in diploid yeast when deleting either one copy or both copies of a particular duplicate gene. Means are shown, and error bars represent the standard error on the mean over 1000 bootstrapped samples of the distribution. (* $p$values for significance shown are $p < 4 \times 10^{-2}, p < 1 \times 10^{-3}$ respectively; Wilcoxon rank-sum test)

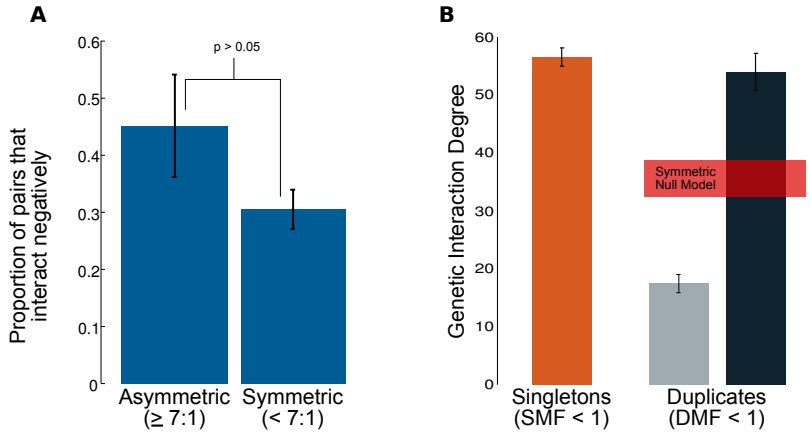

Figure C.3: **(A)** Asymmetric pairs show retained shared functionality. Proportion of synthetic sick duplicates by degree ratio is shown. Highly asymmetric duplicate pairs are no less functionally related than less asymmetric pairs. In fact, the rate of negative genetic interaction between pair members is slightly higher than for duplicates with a more balanced distribution of interactions, though the difference is not significant (See Sec C.1). Error bars represent the error on a binomial proportion and details for the binomial proportion significance test can be found in section C.6. **(B)** High degree sisters are statistically indistinguishable from singletons after controlling for importance. This plot differs from Fig. 4.4C only in that duplicate pairs were first restricted to those with a double-mutant fitness defect (DMF <1), and singletons were restricted to those with a single-mutant fitness defect (SMF <1). Each duplicate pair was then sorted by genetic interaction degree and aggregates are shown. Dotted lines represent the same process applied to a simulated distribution as in Fig. 4.4A. The difference between high degree duplicates and singletons is not significant (56.4 vs 53.9; $p > 0.2$; Wilcoxon rank-sum). The difference between singleton and duplicate interaction degree (Fig. 4.2C) is then generally attributable to one member of each pair.

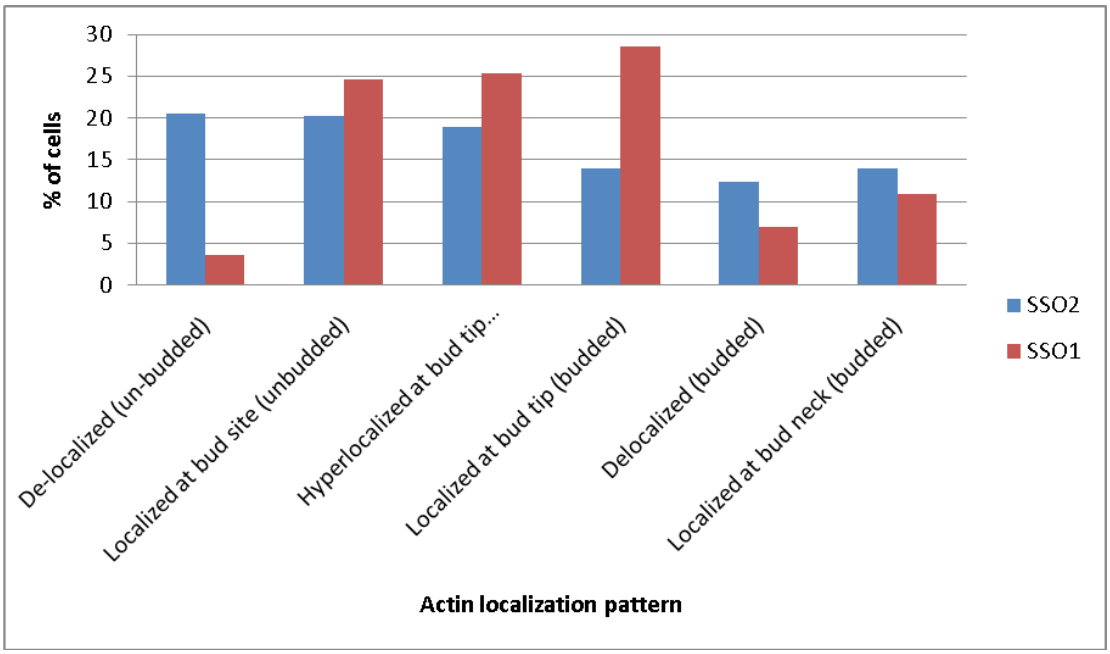

Figure C.4: Differences in *SSO1/SSO2* interaction profiles agree with localization patterns. *SSO1* shows high profile similarity to genes involved in chitin biosynthesis and polarized cell growth, which *SSO2* does not (Table 4.1). Actin localization patterns [168] support a unique roll for *SSO1* during polarized cell growth.

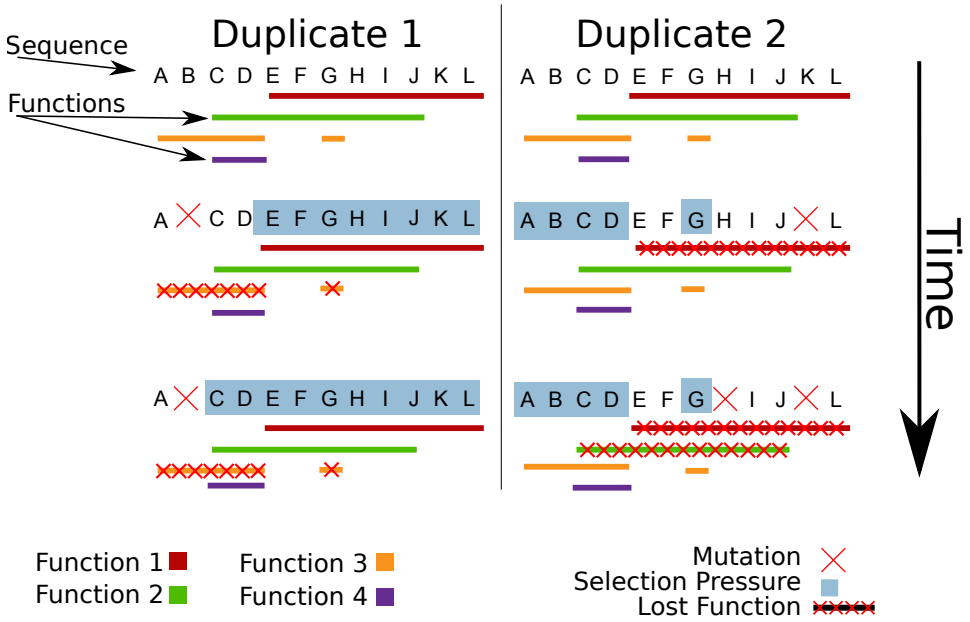

Figure C.5: A simplified model of how the functions of two duplicate genes evolve through time. The figure shows which regions of a sequence are essential for the performance of certain functions. At time 1 (the top-most panel) there is no selection pressure on the sequence of either of the duplicates. At time 2 however, duplicate 1 has lost function 3 by a mutation at B while duplicate 2 has lost function 1 by a mutation at K. Duplicate 1 now has selection pressure on E–L because these sequence regions support function 1 and the cell would lose function 1 if duplicate 1 has a mutation at any of these positions. Because duplicate 1 has less selection pressure, the probability of it having another sustainable mutation is higher than for duplicate 2, and it receives one at H at time 3. Now duplicate 1 supports functions 1,2, and 4, while duplicate 2 supports functions 3, 4 and no other loss of function mutations can occur for either duplicate. Note that function 4 is supported by both duplicates causing a negative genetic interaction if both are deleted.

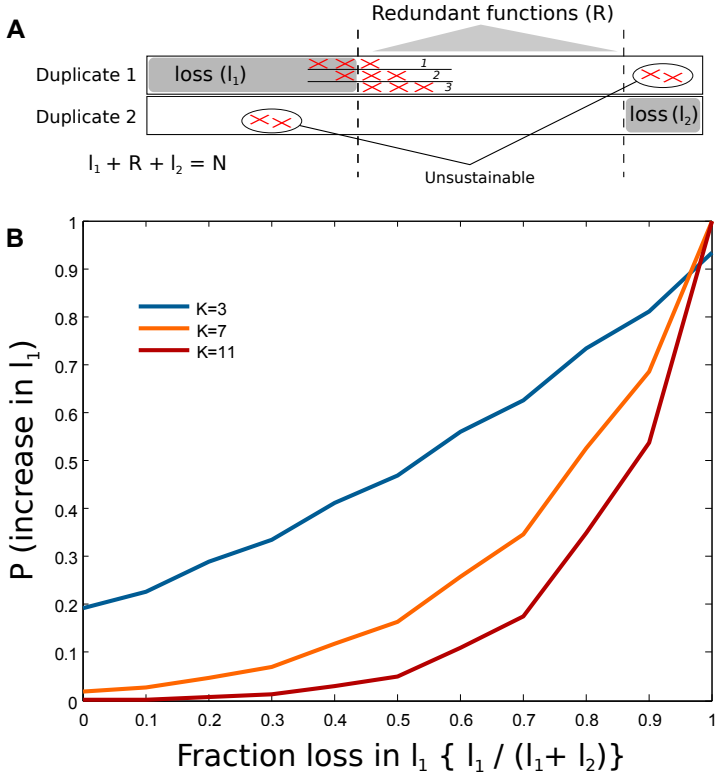

Figure C.6: **(A)** Asymmetric divergence model. The figure shows two duplicate genes, each of which has lost some of their once redundant functions. Mutations to sequence (not shown) cause the loss of multiple functions (red X). Functions must be maintained in one duplicate or the other to be sustainable. Shown are 3 possible arrangements for the loss of function affects of a single sequence mutation($K = 3$). The duplicate with greater loss has more possible sustainable arrangements in which loss is increased. In essence, the less functional copy is more accommodating to loss-of-function mutations in general, and stands a greater chance of losing further redundant function. **(B)** Probabilistic simulation of discrete asymmetric duplicate divergence. Probability of further loss in duplicate 1 (given a mutation in duplicate 1) as a function of duplicate 1's proportion of total loss, for various values of K. The line increases monotonically, indicating that the duplicate with greater proportional loss, has a higher probability of sustaining a mutation which increases loss. For this example $N = 60$ and $R = 35$, though the always increases property holds for any $N, R, K > 1$. The special case $K = 1$ illustrates the probabilities of an earlier model [152] in which a mutation only affects functions within R, in which case the probability depends only on R (equal for both duplicates), and thus the total proportion of loss has no effect.

C.6 Supplementary materials and methods

C.6.1 Definition of duplicates and singletons

The full list of duplicate pairs consists of those identified as the result of the WGD event, as reconciled from several sources [43]. Additionally, any pair of genes fulfilling established similarity requirements [201] was reasoned to be a duplicate pair resulting from a SSD event. Specifically, the gene pair must have a sufficient sequence similarity score (FASTA Blast, $E = 10$) and sufficient protein alignment length(>80% of the longer protein). The pair must also have an amino-acid level identity of at least 30%for proteins with aligned regions longer than 150 amino acid, and for shorter proteins, the identity must exceed $0.01n + 4.8L^{-0.32(1+exp(-L/1000))}$, where L is the aligned length and $n = 6$ [202, 201]. After combining pairs from theWGD event, with pairs determined through sequence alone (SSD), families with more than two members as a result of multiple pairings were completely removed from analysis to control for potential buffering from a third member affecting the interactions of the first two, and any gene not involved in any pairings was deemed an unambiguous singleton.

C.6.2 Functionally related pairs

As a proxy for non-duplicated yet functionally related gene pairs, we have used pairs that exhibited a PPI in at least one of two high-throughput TAP-MS studies [193, 65]. To increase the number of duplicate pairs considered in the analysis relating sistersister profile similarity to sisterproxy similarity,we did not limit PPI interactions to TAP-MS (see section C.6.3). Interactions for this analysis were included from BioGrid if they fell into one of the following categories: affinity capture-RNA, affinity capture-Western, two-hybrid, PCA, affinity capture-MS, co-fractionation, biochemical activity, co-crystal structure, co-purification, far western, FRET, proteinpeptide, proteinRNA or reconstituted complex.

C.6.3 Significance of binomial proportions

Synthetic sick/lethal proportion rates were tested under using the following normally distributed random variable:

$$Z_0 = \frac{P_1 - P_2}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where P_1 and P_2 are the binomial proportions in the respective classes and $\hat{P}$ is the binomial proportion of the combined set.

C.6.4 Genetic interaction data and profile similarity calculations

Genetic interaction data were taken from a recent global genetic interaction study [66]. For the presence or absence of individual interactions, such as calculating the proportion of synthetic lethal duplicates, or counting interaction degree for a given gene the following magnitude, and p-value thresholds were used ($|\epsilon| > 0.08$ and $p < 0.05$). When counting discrete interactions, column degree was used. Thus, only genes in the deletion array (3,885 genes) have valid degrees. This dimension was chosen to maximize the number of covered genes, as fewergenes (1,712) have been screened as queries. For assessing profile similarity, we first normalized the (unthresholded) data along both rows and columns and then used inner product between any pair of array genes as their profile similarity [202, 201].

C.6.5 Definition of dosage class

A duplicate pair was labeled as a dosage pair if it met two of the following three conditions: (1) The pair's representative ortho-group had a volatility score [132] in the top quartile. (2) The pair had a scaled difference in transcript quantity in the bottom quartile. Absolute expression data is taken from Holstege *et al.* [163] and scaled expression difference is defined as in Ihmels *et al.* [141]:

$$\textit{Scaled difference } (a, b) = \frac{|a-b|}{a+b}$$

(3) The pair had a scaled difference in expression stability in the bottom quartile, wherein stability for each gene is defined as the number of data sets out of a possible 127 from Hibbs *et al.* [203]) in which the expression of the given gene is in the bottom 2% for variance.

C.6.6 Ancestral proxies on the PPI network

To find suitable proxy genes for a given duplicate pair, we isolated the common interaction partners on the expanded physical PPI network for each pair with the assumption that interactions common to both paralogs are not likely to have evolved independently, and are therefore tied to one or more of the pairs ancestral functions. We then measured genetic interaction profile similarity between each paralog and the neighbor for comparison with profile similarity between the duplicates themselves. Results were averaged across all common partners for a given duplicate pair.

C.6.7 Genetic interaction degree asymmetry

To compare genetic interaction degree and rates of evolution, we used the original rates provided in the supplement to Kellis *et al.* [42]. This ratio was defined as the rate of the quickly evolving or derived function member divided by that of the slowly evolving or ancestral function member. To test for bias in which member of the pair had more interactions, we assumed a null model in which either gene was equally probable to have the most interactions.We obtained a P-value for this hypothesis using MATLAB's binomial cumulative distribution function *binocdf()*. The proposed ancestral gene generally has a higher degree; hence, the genetic interaction ratio for the pair was calculated with the ancestral function members property in the numerator.

C.6.8 Chemical-genetic degree

To ascertain the number of chemical environments under which a gene displayed a significant phenotype, we used the original data from Hillenmeyer *et al.* [56]. We counted the number of conditions in which the homozygous deletion displayed a significant p-value ($p < 0.05$) out of a possible 1,144. As above, we then used a binomial cumulative distribution to test whether the correspondence between the two data sets (the number of times the gene with more genetic interactions also had more chemical interactions) could be attributed to chance.

C.6.9 Phylogenetic comparison for asymmetric pairs

We compared the sequence similarity of the WGD pairs in *S. cerevisiae* with orthologs in other post-WGD species (*S. castellii, C. glabrata and S. bayanus*) in which one WGD copy had been lost as annotated in the Yeast Genome Order Browser [43]. For each such case, we produced an amino-acid sequence alignment between each *S. cerevisiae* gene and the out-group ortholog using the BLAST algorithm [40]. We then compared the percent identity score for each duplicate with the out-group ortholog. For every pair identified as asymmetric, we used a binomial test to ascertain whether the gene with more interactions was more similar to the orthologous gene, the null hypothesis being that the lower degree and higher degree genes have equal chance of a higher percent identity score with the orthologous gene. In *S. bayanus*, we found only three single orthologs to asymmetric WGD pairs in *S. cerevisiae*, and as such that data is not included.

C.6.10 Biological example profile similarity

Profile correlations for specific biological examples, *SSO1:SSO2*, *GAS1:GAS2*, and *CIK1:VIK1* were taken from the supplement to Costanzo *et al.* [66]. It represents a composite score using information from both array and query profiles in an attempt to give a uniform similarity score across all pairs of genes. Fig. 4.5A shows edges from this composite network involving *CIK1*, *VIK1* and *KAR3* using a correlation threshold of $r > 0.2$.

Appendix D

Appendix for Chapter 5

D.1 Supplementary methods

Smaller array justification and design

In order to generate representative profiles for the largest possible set of duplicate pairs for the APS, we designed a smaller "mini-array" which reduced the number of plates per query (and thus the cost) by a factor of ∼4.5. Strains included on the array were selected from two existing arrays. A total of 986 non-essential and 192 essential genes were selected based on their usefulness in predicting known functional annotations as assessed from existing genetic interaction data by a greedy algorithm [61].

To mitigate experimental effects related to the proximity of sick strains on the plate [50] I designed the array to keep apart strains with fitness defects, as well as those deleted for genes with overlapping linkage regions, which would be simultaneously sick for certain queries. For each of these I developed a score, and then examined the scores over 10,000 random configurations of the 1,178 genes.

$$\text{linkage penalty} = \sum_{chr=1-16,L\&R} \left[\sum_{i,j \subset \text{strains on arm } chr} \begin{pmatrix} \frac{1}{d_{i,j}} & for\ d < 4 \\ 0 & for\ d \geq 4 \end{pmatrix} \right]$$

$$\text{fitness penalty} = \sum \frac{|(1-f_i)(1-f_j)|}{d_{i,j}}$$

where $d_{i,j}$ is the Manhattan distance on the plate between strains i and j, while f_i denotes the single-mutant fitness of strain i. The linkage penalty discourages genes on the same arm of the same chromosome from being placed near each other. The fitness penalty discourages strains with fitnesses much lower than wild-type (1.0) from being placed near one another. Many random layouts were able to achieve a near zero linkage penalty score and from them we chose one with the lowest fitness penalty.

Domain divergence rates

Two domain divergence asymmetry measures are referenced in Table 5.2. The first is from Supplemental Online Materials from Kellis *et al.* 2004 [42].

The second, similar measure was devised by Alex Nguyen, in the lab of Alan Moses, and obtained via personal communication. Calculation of these relative rates was performed according to the following procedure:

First, obtain the rate of evolution for the domains (number of substitutions per site), subtracting the value of each sister from the other in order to gauge them independently. Each sister may have a slightly different set of species in the yeast clade with available sequence, so these values are not yet directly comparable, because the more species the clade has, the more substitutions will arise per site. So we normalize the number of substitutions per site by the expected number, which is basically the phylogenetic tree of the yeast species. This expected value is the relative expected rate of substitutions per site. These values require one further normalization step to ensure that the number of substitutions on the species before the duplication is equal for each sister. Normalization by this factor gives absolute expected rates for each paralog (without the influence of the other). Each observed rate is divided by the corresponding expected rate, and the final measure is the absolute difference in the two ratios.

A high value means that the two sisters are evolving differently, and our hypothesis is that they must have different function and therefore low trigenic proportion. A low value means either both sisters evolve equivalently, either faster or more slowly than expectation.

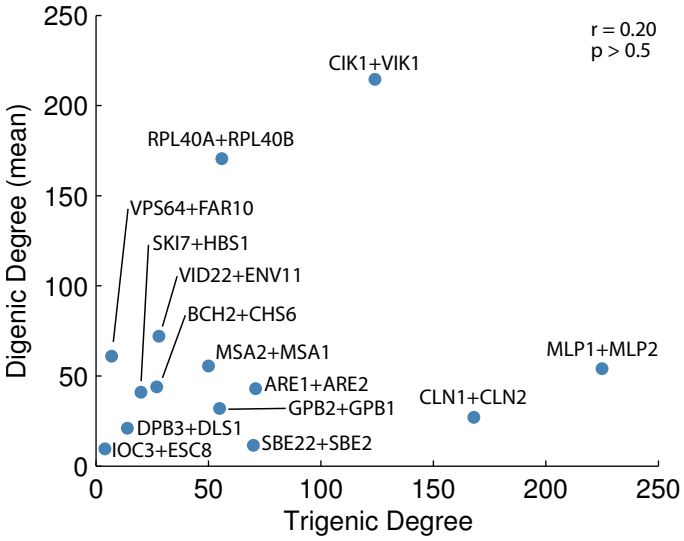

Figure D.1: Pilot study degree scatterplot. Trigenic degree for each double-mutant query in the pilot study is plotted against the mean of the two single-mutant controls.

D.2 Supplemental figures and tables for Chapter 5

Precision recall and digenic/trigenic overlap

Fig. 5.5 shows a relatively high precision for positive trigenic interactions on the WGS (right, green). This is in part due to an overlap between positive trigenic interactions and negative digenic interactions. The left panel of Fig. 5.5 shows an excellent precision for these negative digenics (left purple). This overlap may also exist in the APS trigenics without boosting positive precision (right, red) because the negative digenics are much poorer predictors (left, blue). Fig. D.2 shows the corresponding results if digenic/trigenic overlaps are explicitly forbidden.

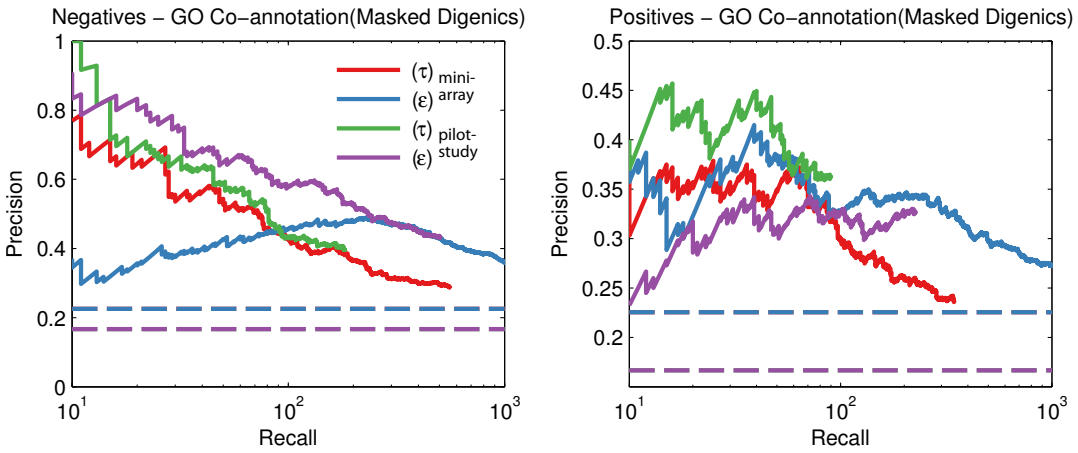

Figure D.2: Precision and digenic/trigenic overlap. The same as Fig. 5.5 with the exception that any significant digenic interaction causes the corresponding trigenic interaction, which by definition will oppose in sign, to be invalidated. Differences between the two figures are negligible, with the exception of positive trigenic precision in the pilot study (right, green).

Fig. 5.5 shows a relatively high precision for positive trigenic interactions on the pilot study (right, green). This is in part due to an overlap between positive trigenic interactions and negative digenic interactions. The left panel of Fig. 5.5 shows an excellent precision for these negative digenics (left purple). This overlap may also exist in the mini-array survey trigenics without boosting positive precision (right, red) because the negative digenics are much poorer predictors (left, blue). This figure shows the corresponding results if digenic/trigenic overlaps are explicitly forbidden.

No Trigenic Filter (203 pairs)	Spearman ρ	p-value
Paralog digenic ϵ	-0.37	1.2×10^{-6}
Double-mutant fitness	-0.33	1.5×10^{-5}
Digenic negative path length	-0.25	2.2×10^{-2}
Single-mutant fitness (mean)	-0.19	7.2×10^{-3}
Expression level (difference)	-0.18	2.1×10^{-2}
Divergent localization (Marques 2008)	-0.15	2.8×10^{-2}
Expression stability (mean)	0.18	1.1×10^{-2}
SGA profile similarity (array)	0.21	8.0×10^{-3}
Similar localization (Marques 2008)	0.25	3.6×10^{-4}

Table D.1: A selection of paralog-pair features that show significant correlations with trigenic proportion as defined in Eq. 5.13. Results shown are for all 203 pairs in the mini-array survey. Similar correlations for high-confidence pairs are shown in Fig. 5.2.

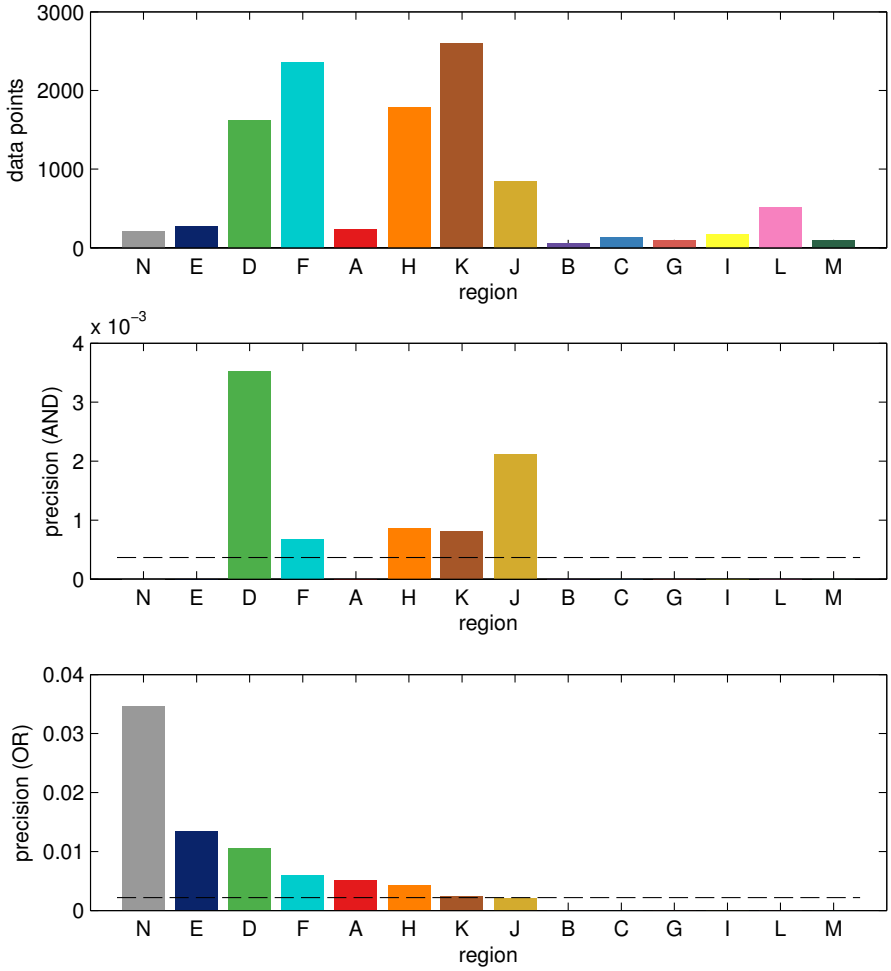

Figure D.3:
Trigenic interaction predict protein-protein interactions. The top plot shows the number of observations in each category (as in Fig. 5.6). The bottom two plots show enrichment for protein- proein interactions in each region. In the middle plot a data point was considered a true positive if the array gene shared a ppi interacion with both paralog sisters (AND model). This resulted in very sparse data, with a background rate for paralog-array interactions on the order of 10^{-3}. In the bottom plot, a data point was considered a true positive if the array gene shared a ppi interaction with either paralog sister (OR model), resulting in an order of magnitude increase in the expected rate, and several interesting differences in regional enrichments. Bars in each plot are sorted by precision using the (OR) model.