2016

Functional and evolutionary implications of in silico gene deletions

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

FUNCTIONAL AND EVOLUTIONARY IMPLICATIONS OF *IN SILICO* GENE DELETION STUDIES

by

CHRISTOPHER JACOBS

B.S., University at Buffalo, 2009
M.S., Boston University, 2014

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2016

Approved by

First Reader

Daniel Segrè, Ph.D.
Professor of Biology

Second Reader

Yu Xia, Ph.D.
Associate Professor of Bioengineering
McGill University, Faculty of Engineering

DEDICATION

In memory of my beloved nana, Margery Jacobs, who had always been disproportionately proud of all of my former accomplishments. If she were still with us, it would not be a stretch to imagine that the most people would hear about this work from her than anyone else.

ACKNOWLEDGMENTS

This thesis and the works described herein would not have been possible without substantial support from my family, friends, and colleagues. This is as much their accomplishment as my own.

I especially need to thank my parents, Timothy and Nancy Jacobs, who have remained as constant sources of encouragement, consolation, advice, and even patronage throughout the entirely of my life. I have always been, and will always be able to count on my brothers Timothy, Jonathon, and Colin, who are amongst my best of friends.

To my advisers, Daniel Segrè and Brandon Xia, I offer my profuse and sincere thanks. When I needed to take a sabbatical from the graduate program for health reasons, it was their words of kindness and encouragement were what made it the most difficult to leave and what eventually enticed me to return. The administrative members of bioinformatics graduate program also deserve much credit for offering their support during this time and for easing the transition from and back to my studies.

Finally, but not at all least, I need to thank my close friends, both those I work with in the lab and those I have met without. I wouldn't have managed it without all of you.

FUNCTIONAL AND EVOLUTIONARY IMPLICATIONS OF *IN SILICO* GENE DELETION STUDIES

CHRISTOPHER JACOBS

Boston University Graduate School of Arts and Sciences

and

College of Engineering, 2016

Major Professor: Daniel Segrè, Ph.D., Professor of Biology

ABSTRACT

Understanding how genetic modifications, individual or in combination, affect organismal fitness or other phenotypes is a challenge common to several areas of biology, including human health & genetics, metabolic engineering, and evolutionary biology. The importance of a gene can be quantified by measuring the phenotypic impact of its associated genetic perturbations "here and now", e.g. the growth rate of a mutant microbe. However, each gene also maintains a historical record of its cumulative importance maintained throughout millions of years of natural selection in the form of its degree of sequence conservation along phylogenetic branches. This thesis focuses on whether and how the phenotypic and evolutionary importance of genes are related to each other.

Towards this goal, I developed a new approach for characterizing the phenotypic consequences of genetic modifications in genome-scale biochemical networks using constraint-based computational models of metabolism. In particular, I investigated the impact of gene loss events on fitness in the model organism Saccharomyces cerevisiae, and found that my new metric for estimating the cost of gene deletion

correlates with gene evolutionary rate. I found that previous failures to uncover this correlation using similar techniques may have been the result of an incorrect assumption about how isoenzymes deletions affect the reaction they catalyze.

I next hypothesized that the improvement my metric showed in predicting the cost of isoenzyme loss could translate into an improved capacity to predict the impact of pairs of gene deletions involving isoenzymes. Studies of such pair-wise genetic perturbations are important, because the extent to which a genetic perturbation modifies any given phenotype is often dependent on the genetic background upon which it has been performed. This lack of independence within sets of perturbations is termed epistasis. My results showed that, indeed, the new metric displays an increased capacity to predict epistatic interactions between pairs of genes.

In addition to shedding light on the relationship between the functional and evolutionary importance of genes, further developments of our approach may lead to better prediction of gene knockout phenotypes, with applications ranging from metabolic engineering to the search for gene targets for therapeutic applications.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DNA ...	**d**eoxyribo**n**ucleic **a**cid
E-MAP .	**e**pistatic **m**icroarray **p**rofile
eQTL ..	**e**xpression **QTL**
FBA	**f**lux **b**alance **a**nalysis
GPR ...	**g**ene-**p**rotein-**r**eaction set/map
MCS ...	**m**inimal **c**ut **s**et
MOMA .	**m**inimization of **m**etabolic **a**djustment
PSI	**p**rotein **s**ynthesis **i**nhibitor
QTL ...	**q**uantitative **t**rait **l**ocus
RNA ...	**r**ibo**n**ucleic **a**cid
RNAi ...	**RNA i**nterference
SD	**s**ynthetic **d**efined (media)
SD−His	**SD** less/minus **h**istidine
SGA	**s**ynthetic **g**enetic **a**rray
SGD	***S**accharomyces* **G**enome **D**atabase
SNP	**s**ingle **n**ucleotide **p**olymorphism
SSL	**s**ynthetic **s**ick and/or **l**ethal
TAP-MS	**t**andem **a**ffinty **p**urification followed by **m**ass **s**pectrometry
TPR ...	**t**rue **p**ositive **r**ate
YKO ...	**y**east **k**nock**o**ut library
YPD ...	**y**east **p**eptone **d**extrose
YPLac .	**y**east **p**eptone **lac**tate

CHAPTER 1

INTRODUCTION AND BACKGROUND

Genetic modifications represent a central and pervasive concept in the study of biological systems. In genetics and molecular biology, deletions or perturbations of a gene are used as a fundamental tool for probing the gene's function [1]. In metabolic engineering and synthetic biology, genetic modifications are a key component of strategies for designing new cellular functions with specific practical tasks, such as the production of valuable molecular products [2, 3]. In evolutionary biology, genetic modifications are the basis for the generation of variation during reproduction, driving the process of natural selection.

Common to these different viewpoints in the study of genetic modifications in biology is the mapping that governs the relationship between genotype and phenotype. The precise nature and extent to which an organism's observable traits (the phenotype) are informed by its genetic constitution (the genotype) is still the subject of some debate [4, 5]. However, it is generally accepted that one can think of the connection between genotype and phenotype as a complex mapping [6, 7]. Through this mapping, whose mathematical formalization may be considered as one of the goals of systems biology, one could in principle predict whether any given genetic modifications is likely to result in phenotypic changes. The evaluation and improvement of a specific mathematical schema for performing this mapping constitutes one of the major outcomes of the work presented in this dissertation. In particular, I will describe a derivation of the functional importance of a gene from the quantitative impact its removal has on organism fitnes and establish that this formulation of gene importance correlates with evolutionary importance as derived by genetic

conservation.

To understand how conservation of a gene sequence conveys information about its importance, it is useful to remind ourselves that evolutionary adaptation proceeds through random mutation of genomes, which generates diversity of phenotypes in a population. The subsequent survival of the most fit individuals, through the process of natural selection, gives rise to a population which is increasingly adapted to its environment [8]. Importantly, the randomness of mutations gives rise to a large proportion of phenotypic changes that are deleterious or neutral. While deleterious mutations are usually quickly purged from the population [9, 10], nothing prevents the random accumulation of neutral alleles (genetic drift), which results in special signatures of genetic variation within and between species, useful as a basic reference for the study of the rate of evolution. In particular, by comparing genetic mutations that do and do not (neutral) affect the corresponding protein sequence, it is possible to estimate the degree of evolutionary pressure likely experienced by such gene. In other words, this evolutionary analysis of genetic sequences (the details of which are explained in more detail in Sub-Section 2.4.2) provides an estimate on the importance of a gene in terms of its degree of conservation across clades.

In this chapter of the dissertation, I will expand further on the ideas presented above. I will be providing an introduction to some of the specific areas relevant to this work by exploring the history of study into these topics — but only insofar as to provide a necessary background for the original work to be presented in the following chapters. It is important to note that I focus my attention on the model organism *Saccharomyces cerevisiae* (commonly known as baker's yeast). This choice of model organism is justified both by the long tradition of using this organism for basic genetic and evolutionary studies, as well as by the fact that yeast is central to

many biomedical and metabolic engineering studies, as we will see. The rest of this introduction will briefly review recent milestones in the study of gene deletions in yeast, and will provide an overview of genome-scale constraint-based mathematical modeling approaches that are used extensively later on in my own work.

1.1 Functional Discovery from Genetic Modifications

Genetic modifications are incredibly powerful tools for functional discovery of genes and other genetic sequences. By inducing genetic modifications in a controlled way at specific sites, such as when a whole gene is deleted from a genome, one may infer some aspects of the function of the missing gene from the altered phenotype that results. Even before it was made possible to specify a particular perturbation site with (relative) ease, functions could be annotated by genetic screening, a process whereby mutants with a particular phenotype were isolated from a population and the genetic differences which caused them subsequently identified [1]. Systematic application of these methods have been in use for decades [11], and represent *the* prototypical way by which molecular and systems biologists uncover the functional role of genes [12].

Owing to its role as a model organism of particular interest [13], the history of functional identification of genes in yeast dates back almost to the start of the practice itself and many of the techniques used for this purpose were first pioneered by their use in yeast [14, 15, 16]. Somewhat more recently, a concerted effort was undertaken to functionally characterize the entire yeast genome [17, 18], from which the first ever complete single-gene-deletion library (YKO or yeast knockout set) was constructed. The result of this effort, as well as those from many other similar studies (e.g. genetic screens using RNAi silencing [19]) have been collected online at

the *Saccharomyces* Genome Database (SGD) [20, 21].

Functional annotation of genes also has implications for the study of human health and disease. Obviously, it is no simple task to extrapolate from a phenotype caused by gene loss in yeast to what happens when a similar event occurs in humans. However, there are orthologs which are relatively well conserved between these two organisms — as much or more than 34% of yeast genes have homologs in the human genome [22], and as many as 17% of human disease genes are homologous to yeast genes [23]. Indeed, many of these genes are so well conserved that it is possible to replace several essential yeast genes with their human orthologs while maintaingin cell viability [24]. Furthermore, the mapping of orthologous genes between eukaryotes has been naturally extended to the concept of orthologous phenotypes (phenologs) [25, 26], whereby seemingly diverse phenotypes in different species are related by their shared association with overlapping sets of ortholgous genes. In this way, human diseases such as epilepsy could be related to alternative yeast phenotypes such as trichlormethine sensitivity. It is possible that by relating phenotypes across species in this way, one could develop novel treatments for human diseases based on induced "rescues" of wild-type yeast behaviors in mutants.

1.1.1 Unexpected Phenotypes Arising from Combinations of Perturbations

While individual genetic modifications are relevant in and of themselves, there is something fundamental about how such perturbations affect a system when performed in concert. These higher order effects are particularly important when the phenotype caused by multiple perturbations is different from the one expected, based on prior knowledge of the phenotypes caused by these same perturbations individually [27, 28]. Such deviation from expectation is generally referred to as epistasis.

Two genes, alleles, or genetic modifications displaying epistasis are also said to have a genetic interaction.

The reason epistasis is been increasingly studied in conjunction with systems biology is twofold. First, epistasis is fundamentally related to systems biology through its very definition. The behavior of the system may be drastically uncharacteristic of the behaviors of the individual components. To some extent, systems biology can be seen as the study of epistasis, i.e., of nonlinear, unexpected system-level behavior arising from combinations of components working together — the whole being more, or less, than the sum of its parts. Second, the analytical and experimental high-throughput methods of systems biology are very helpful for understanding epistasis at the cellular level. Some sophisticated high-throughput technologies have been specifically designed for the purpose of systematically measuring epistasis between many genes. Charles Boone's group pioneered the development of synthetic genetic arrays (SGAs) [29, 30], an automated approach for generating large-scale double-gene-deletion mutant libraries, which they then applied to three-quarters of the entire yeast genome [31]. The general SGA approach was subsequently modified to work well with smaller, designed subsets of yeast genes, forming E-MAPs (epistatic microarray profiles). These tools and their specific contributions to the study of genetic interactions are described in more detail later in (Chapter 3).

Epistasis also plays a crucial role in evolutionary biology. An abundant literature in population genetics has been dedicated to quantitatively understanding epistasis in natural populations [32]. Epistasis affects the topology and jaggedness of the fitness landscape [33, 34] and therefore the rate and properties of evolutionary adaptation. Sexual reproduction, still a perplexing phenomenon in evolutionary biology, may have evolved as a method to purge genomes of mutations through recombination

[35] in response to strong deleterious epistasis between loci [36], though this idea has been the subject of debate [37, 38].

Although different specific definitions and metrics for epistasis have been proposed in different contexts [39, 40], the intuitive idea of epistasis as a deviation from a null expected behavior is common to different fields, and constitutes an interesting bridge between systems biology and evolutionary biology. Broadly speaking, the study of genetic interactions represents a unique meeting point where biological organization principles and practical applications converge (Figure 3.1), impacting fields as diverse as functional genomics [30, 41, 31], drug development [42, 43, 44], and immunology [45, 46].

1.2 Metabolism As a Model System for Studying Genetic Perturbations

Metabolism is the network of all life-sustaining biochemical reactions that every living organism uses to transform available nutrients into energy and the molecular matter required for continued cellular maintenance and reproduction. Naturally, a system so central to the idea of life has been the subject of intense study. It is difficult to date with any precision the beginning of scientific investigation into metabolism, precisely because it was recognized as topic of investigation even when it was widely viewed to be a consequence of the so-called "vital force", a soul-like aspect that was thought to separate life from inanimate objects [47]. However, the modern study of metabolism is generally recognized to have begun around the time of Eduard Buchner's nobel-prize worthy work on "cell-free fermentation" (translated) [48] and the first formalization of the concept of enzymes, proteins that catalyze (accelerate) metabolic processes, in around the turn of the 20^{th} century. Today, there exist a

number of online and open-source resources designed specifically around enzymes (e.g. BRENDA [49, 50]), metabolic pathways (e.g. MetaCyc [51, 52] and Reactome [53, 54]), and other aspects of metabolism (e.g. KEGG [55, 56], which includes pathway maps, functional hierarchies, drug-interaction information and much more).

Metabolism is very well understood system, depsite all of its complexity, in part because of its long history of study. This is not to say that metabolic networks are are not diverse. Indeed, the individual metabolisms of organisms having been shaped by historical adaptation of separate organisms to distinct ecological niches for millenia. However, in general, metabolism as a whole is remarkably well conserved [57, 58] relative to the rest of the genome. That is, some metabolic genes and often entire pathways are identifiable as common to large swaths of the tree of life. Even metabolic genes which are found only within more specific phylogenetic clades tend to evolve slower than other genes in the genome. This is not to say that these genes cannot undergo rapid genetic modification. In fact, one of the hypothesized methods for how new metabolic functions can be gained is through so-called "neosubfunctionalization" [59], whereby a gene duplication event results in the accumulation of mutations in one or both paralogs (genes related by a duplication event) resulting in new functionality (neofunctionalization [60]) and/or increased specificy of the encoded enzymes (subfunctionalization [61]) to particular substrates.

1.2.1 Computational Models of Metabolism and Flux Balance Analysis

Alongside high-throughput experimental technologies, computational biology has developed several technologies which could be used to simulate large-scale biological systems and their response(s) to perturbations. In particular, the advent of whole-genome reconstructions of metabolic networks, such as the ones for *Escherichia coli*

[62, 63, 64] and *S. cerevisiae* [65, 66, 67], has made it possible to easily perform systematic and comprehensive computational screens of all possible single and double metabolic enzyme gene deletion phenotypes.

One approach that has now been amply used in this context is the framework of stoichiometric constraint-based models of metabolic networks, most notably flux balance analysis (FBA). FBA is used to predict growth rate and metabolic fluxes (steady state rates) within networks that encompass the whole set of metabolic reactions known to be possible in a given organism (hence "genome-scale") [68]. For a more comprehensive introduction to flux balance modeling, I refer the reader to available literature (e.g., [69, 70, 68, 71]). However, I do wish to stress here the fundamental assumptions behind FBA, as well as some of its limitations. FBA is based on two key simplifying assumptions. The first is that the metabolic network under study is at steady state, i.e., metabolite concentrations stay constant over time. While this is not true for individual cells, it is often a sensible assumption for populations of cells kept under stable conditions (e.g., bacteria or yeast in a continuous flow bioreactor). The second main assumption of FBA is that the system is operating close to a set of fluxes that makes it optimal for a given task (the objective function). FBA is therefore implemented as an optimization problem that identifies the optimal flux distribution, while obeying the mass-balance constraints of steady state and the constraints imposed by the available nutrients. This problem can be efficiently solved using linear programming. For microbial systems, the maximization of biomass production has been often used as an objective function. FBA has been used to adequately predict the growth rate and byproduct secretion rates in *E. coli* [72, 73] as well as the essentiality of metabolic genes under several growth conditions [74, 75]. Minimization of metabolic adjustment (MOMA), a variant of FBA, has

been introduced to provide an alternative to the unrealistic assumption that mutant strains should be able to maximize their growth rate upon the perturbation [71]. Instead, MOMA assumes that the internal control circuitry of the cell will tend to maintain the cell as close as possible to the flux state of the wild type while remaining compatible with the new constraints imposed by the deletion [65, 76].

Because of its high computational efficiency — a single FBA simulation may take less than 0.1 second — both of these methods are widely used in large-scale perturbation studies [77, 78, 79], including for predictions of phenotype in the presence of multiple simultaneous gene deletions [80] and/or across several environmental conditions [76, 81]. Briefly, one can use FBA as the computational analogue of a high-throughput growth-rate assay, by systematically computing the effects of single and double gene deletions in a given model organism. Then, one can use Equation 3.1, or variants thereof, to compute deviations from the multiplicative expectation. This type of analysis has been performed first in *S. cerevisiae*, for which highly curated and tested stoichiometric reconstructions have been published in recent years [82, 66, 83].

Finally, the present-day explosion in genomic [84] and metagenomic [85, 86] sequencing data has engendered even the high-throughput generation of such genome-scale models of metabolism. Online resources, such as the Model SEED are capable of automatically reconstructing a metabolic model from an assembled genome in only a couple of days [87]. KBase, the Department of Energy Systems Biology Knowledgebase, can do the same and even has applications which will perform the assembly of the genome from raw sequencing data. Both platforms also currently have some capacity to perform flux balance simulations. KBase also includes and online environment where its programs can be setup to chain together in a workflow that users can annotate. In short, there now exists the capacity to cultivate

an entire microbial community, sequence and assemble the genomes of the members, and automatically reconstruct metabolic models each in a matter of days. While the validated performance of these generated models is obviously less than their hand-curated counterparts, which have seen years of revisions and updates, these still represents powerful tools for use in the field of systems and computational biology.

1.3 Overview of the Topics Discussed Later in This Dissertation

The work presented in this disseration is focused on the study of genetic modifications, and more specifically on seeking a link between the metrics used across different areas of biology to quantify the effects of the loss of a gene in an organism. In particular, in Chapter 2 ("A New Metric for Gene Dispensability Anticorrelates with Evolutionary Rate"), I will use computational biology approaches to establish a previously sought, but elusive correlative relationship between the functional importance of a gene for the cell (e.g. the contribution to fitness) and the genomic signals that encode information about the gene's long-term evolutionary history. In Chapter 3 ("Organization Principles in Genetic Interaction
Networks"), I will explore what we refer to as the "three main principles of organization in genetic interaction networks", focusing on how it is that hierarchies of organizational units arise naturally from these principles and what we can learn about biological systems from their application and study.

What follows immediately in the sub-sections below are a pair of summaries of each of the chapters at large, including a restatement of the chapter objective in the context in which it will be presented and a very brief description of the ideas that will be discussed in more detail therein.

1.3.1 Synopsis of Chapter 2: Functional and Evolutionary Gene Importance

One way to estimate the contribution of any single gene to organism fitness is through calculating it's gene dispensability. In a similar vein, the evolutionary rate of a genetic sequence represents something of a historical record of a gene's dispensabilty, as evidenced by the frequency and type of its accumulated mutations (or relative lack thereof) over millions of years of evolution. Previous failure to identify a correlation between these two metrics has been ascribed to a real biological gap between a gene's fitness contribution "here and now" and that same gene's historical importance. In this chapter I will introduce a new method for calculating the cost of a gene deletion that I call "function-loss cost", which calculates the cost of a gene deletion event as the total potential functional impairment caused by the deletion. I demonstrate that this new metric displays significant anticorrelation with evolutionary rate. I then show that the improvement gained by using function-loss cost over gene-loss cost is the outcome of eliminating the assumption that isoenzymes provide unlimited capacity for backup. The relevance of function-loss cost of isoenzymes is also confirmed by the fact that this new metric improves the capacity of the flux balance model to predict epistatic interactions that include one or more isoenzymes. In addition to suggesting that the gene-to-reaction mapping in genome-scale flux balance models should be used with caution, our analysis provides new evidence that evolutionary gene importance captures much more than strict essentiality.

1.3.2 Synopsis of Chapter 3: Organization Arising from Genetic Interactions

Understanding how genetic modifications, individual or in combinations, affect phenotypes is a challenge common to several areas of biology, including human genetics, metabolic engineering, and evolutionary biology. Much of the complexity of how

genetic modifications produce phenotypic outcomes has to do with the lack of independence, or epistasis, between different perturbations: the phenotypic effect of one perturbation depends, in general, on the genetic background of previously accumulated modifications, i.e., on the network of interactions with other perturbations. In recent years, an increasing number of high-throughput efforts, both experimental and computational, have focused on trying to unravel these genetic interaction networks. Here, I provide an overview of how systems biology approaches have contributed to, and benefited from, the study of genetic interaction networks. I focus, in particular, on results pertaining to the global multilevel properties of these networks, and the connection between their modular architecture and their functional and evolutionary significance.

CHAPTER 2

A NEW METRIC FOR GENE DISPENSABILITY ANTICORRELATES WITH EVOLUTIONARY RATE

System-level metabolic network models enable the computation of growth and metabolic phenotypes from an organism's genome. In particular, flux balance approaches have been used to estimate the contribution of individual metabolic genes to organismal fitness, offering the opportunity to test whether such contributions carry information about the evolutionary pressure on the corresponding genes. Previous failure to identify the expected negative correlation between such computed gene-loss cost and sequence-derived evolutionary rates in Saccharomyces cerevisiae has been ascribed to a real biological gap between a gene's fitness contribution to an organism "here and now" and the same gene's historical importance as evidenced by its accumulated mutations over millions of years of evolution. Here we show that this negative correlation does exist, and can be exposed by revisiting a broadly employed assumption of flux balance models. In particular, we introduce a new metric that we call "function-loss cost", which estimates the cost of a gene loss event as the total potential functional impairment caused by that loss. This new metric displays significant negative correlation with evolutionary rate, across several thousand minimal environments. We demonstrate that the improvement gained using function-loss cost over gene-loss cost is the outcome of eliminating the assumption that isoenzymes provide unlimited capacity for backup. The relevance of function-loss cost of isoenzymes is also confirmed by the fact that our new metric improves the capacity of the flux balance model to predict epistatic interactions that include one or more isoenzymes. In addition to suggesting that the gene-to-reaction mapping in genome-scale

flux balance models should be used with caution, our analysis provides new evidence that evolutionary gene importance captures much more than strict essentiality.

2.1 Introduction

Quantitatively assessing the contribution of each gene to the overall fitness of an organism is an ongoing challenge in evolutionary and systems biology [88]. A classical, bioinformatics estimate of this contribution has been the evolutionary rate of the gene in question, which is based on genetic sequence conservation patterns amongst phylogenetically related genes [89, 90, 91, 92]. This evolutionary rate metric serves as a historical record, providing a retrospective cumulative quantification of the importance of a gene. In contrast, systems biology methods are able to specifically quantify, for each gene, its current contribution to overall organism fitness by directly measuring [17, 18] or estimating [77, 78] the fitness defect caused by the removal of that gene. The natural question arises of whether the current contribution of a given gene to organism fitness, i.e. its dispensability, correlates with its historical importance. It is non-trivial whether such a relationship should exist, because the dispensability of any one gene at any set time point may be influenced by many complex factors, including the environmental condition(s) and its interactions with any other genes within the genome, whose effects cannot be discerned from evolutionary rate. This question has been previously addressed in the model organism *Saccharomyces cerevisiae* (baker's yeast) [93, 81], for which fitness defect scores upon gene deletion have been experimentally measured in a systematic and comprehensive way [17, 18, 94, 95]. Interestingly, a significant negative correlation between gene evolutionary rate and experimentally measured gene dispensability is detectable, although the signal is weak (Spearman's $\rho \approx -0.2$).

In addition to the high-throughput experimental techniques used to quantify gene dispensability at the genome scale, constraint-based modeling techniques — such as flux balance analysis (FBA) [68] — may be used to efficiently generate such data *in silico* [79]. Flux balance models have been shown to successfully recapitulate several experimental observations, including growth phenotypes under various environmental conditions and gene essentiality in select lab conditions [75, 83, 96]. However, one of the puzzling failures of FBA techniques has been precisely the lack of even moderate correlation between predicted gene dispensability and evolutionary rate [81]. This lack of correlation has been ascribed to a number of possible reasons, including lack of knowledge about the most relevant environmental conditions to be used in simulations, and the complex condition-dependence of gene essentiality.

Here we present an alternative metric for measuring gene dispensability using FBA, which we call "function-loss cost" (Figure 2.1, blue indicators). As opposed to the standard "gene-loss cost" (Figure 2.1, red indicators), our new metric estimates the total cost of a gene's deletion by integrating the fitness costs of removing each enzymatic function associated with that gene from the FBA model, even if alternative isoenzymes exist for a given reaction. Surprisingly, by using function-loss cost as our measure of gene dispensability, we are able to observe a negative correlation between the impact of gene deletion and gene evolutionary rate (Figure 2.2A, blue distribution).

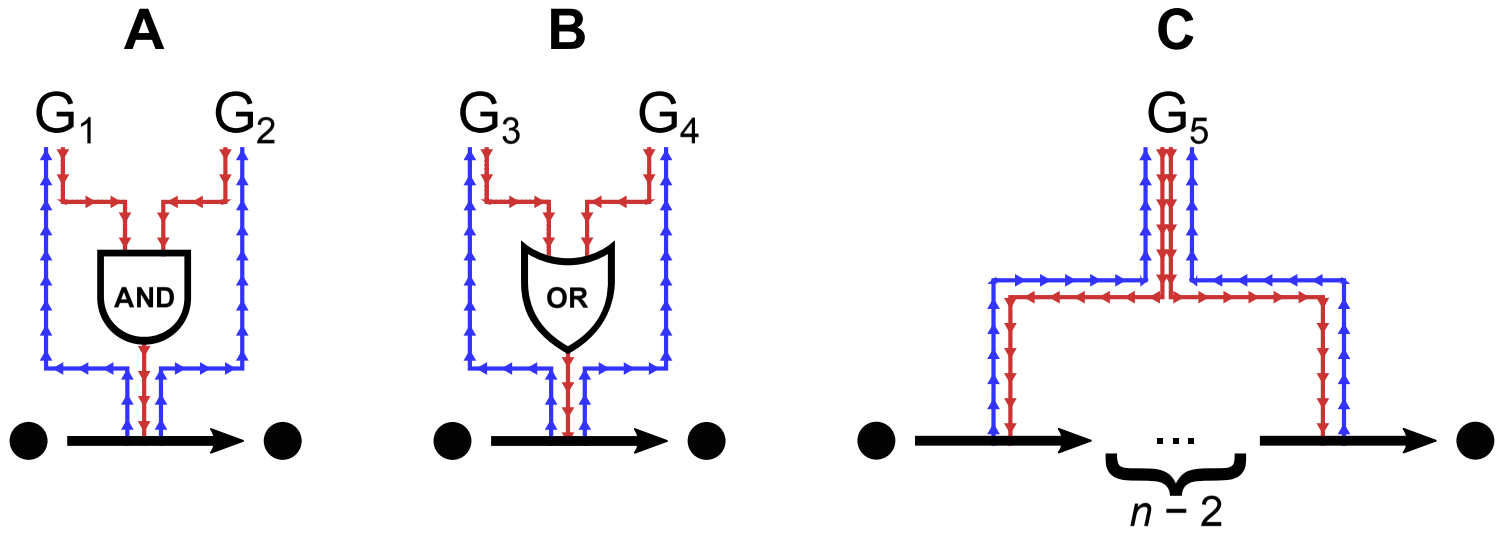

Figure 2.1: Comparison of the gene dispensability metrics function-loss cost and gene-loss cost. Each toy scenario (A-C) represents a possible gene-to-reaction mapping configuration in its simplest form. Gene-loss cost (red indicators) propagates gene deletions "downwards" through logic gates to determine which reaction(s) are removed from the network, which in turn determine model fitness predictions. Function-loss cost (blue indicators) conceptually reverses this process, first calculating the fitness cost of removing each reaction in the network and then propagating these costs "upwards" to all associated genes, whereby they are summed together. For enzyme complexes (A), gene-loss cost and function-loss cost are identical and are equal to the fitness cost of associated reaction's removal. For isoenzymes (B), the gene-loss cost is zero is all cases (because either gene will satisfy the logic gate's requirement that at least one enzyme is present), however the function-loss cost is as in scenario (A). For multi-function enzymes (C), gene-loss cost is determined by the maximum cost of removal of any of its associated reactions, while function-loss cost is equal to the total summed cost of all its associated reactions' removal cost.

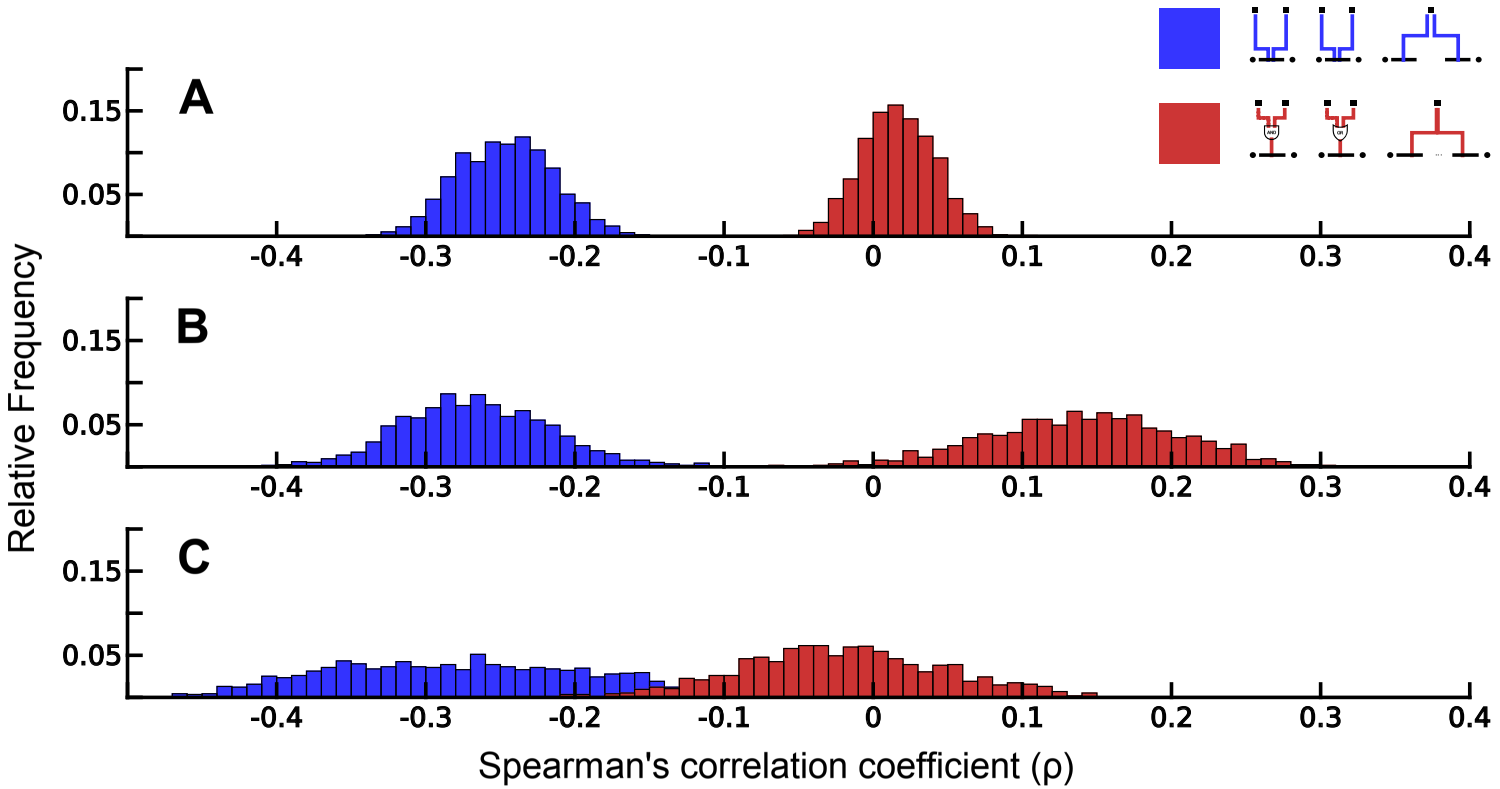

Figure 2.2: Frequency distributions of gene dispensability and evolutionary rate correlations. Gene evolutionary rate was calculated per gene as the average ranked $\frac{d_N}{d_S}$ ($\frac{K_a}{K_s}$) ratio between the *S. cerevisiae* gene and its ortholog in five related yeast species. A gene's deletion impact score was predicted with FBA using gene-loss cost (red distributions), function-loss cost (dark blue distributions), and a hybrid score derived that uses function-loss cost for isoenzymes and gene-loss cost for all other genes (light blue distribution). Rows show distributions for: (A) all genes, (B) the fastest-evolving genes in each gene set, and (C) slowest-evolving genes in each gene set.

2.2 Results

2.2.1 Gene-Loss Cost and Evolutionary Rate Do Not Correlate in Minimal Environments

In prior work, it was established that gene-loss cost, as estimated by flux balance genome scale models of metabolism, correlates poorly with gene evolutionary rate [81]. These prior calculations had been performed for a large number ($\sim 10^4$) of randomly generated combinations of environmentally available metabolites, and using different variants of the FBA objective function (including the standard maximization of biomass production flux [68] and the minimization of metabolic adjustment upon gene deletion [71]). We started by confirming these results, using a recently updated stoichiometric reconstruction [96], a different strategy for choosing a large number of environmental conditions, and independently computed evolutionary rates.

In particular, to impose environmental constraints in our FBA calculations, we generated 1,536 minimal media, each containing a nitrogen and a carbon source, in all possible combinations (see Methods and for details and [97] for similar uses of this strategy). Gene-loss costs were calculated across all metabolic enzyme genes and environments, using the standard FBA protocol for gene knockouts (see Methods and [98]). Evolutionary rates for *S. cerevisiae* metabolic genes were calculated using a modified version of $\frac{d_N}{d_S}$ from orthologs in five related species spanning a phylogenetic timetable of roughly 10–100 million years (see Methods and [99]). Our results (Figure 2.2A, red distribution) verify that gene-loss costs do not correlate with gene evolutionary rate (Spearman's ρ ranging between $-.05$ and 0.1).

Notably, in contrast to FBA calculations previously used for this type of analysis, by limiting each minimal environment to a single source of carbon and a single source of nitrogen, we are able to verify that no specific resources is significantly more likely

to have been historically important in yeast metabolic adaptation. For example, for one such fixed nitrogen source (ammonium), the total per-carbon distribution of correlations spans only about 4 standard deviations of the total distribution across all environments — from $\rho = -0.027$ (for the carbon source xanthosine) to $\rho = 0.06$ (for D-xylose), whereas the correlation for glucose is only $\rho = 0.002$. Thus no individual carbon source stands out as historically much more relevant than the average. Furthermore, at the model level, the use of minimal media strictly enforces a kind of resource scarcity. In absence of this scarcity, the FBA model can reroute metabolic fluxes to use alternate resources at no cost, masking the effect of blocking individual pathways with a gene deletion.

2.2.2 A newly defined function-loss cost anticorrelates with evolutionary rate

Given the lack of correlation observed between FBA-computed gene-loss cost and gene evolutionary rate, we asked ourselves whether any step in the FBA calculation could potentially distort the estimation of the cost of gene deletion. We ended up focusing our attention on the gene-to-reaction mapping, which, in the FBA knockout calculation, translates the deletion of a gene into the corresponding flux constraints that block (potentially multiple) reactions associated with that gene (Figure 2.1). This mapping, expressed using simple Boolean logic, plays a particularly important role for reactions are catalyzed by multiple enzymes (isoenzymes, Figure 2.1B) or by enzyme protein complexes (Figure 2.1C). For two isoenymes catalyzing the same reaction, for example, deletion of one the two enzymes has no effect on the corresponding flux in a traditional FBA knockout calculation, because the other enzyme is assumed to provide full backup functionality. However, abundant experimental evidence suggests that this backup effect is often limited, or condition-dependent

[100, 101, 102]. The cumulative effect of this discrepancy in genome scale calculations could be quite significant, given that almost one third of the metabolic enzymes in *S. cerevisiae* are members of isoenzyme sets (and thus would end up incurring no cost whatsoever under standard FBA knockout calculations). We thus hypothesize that fixing this oversimplification in the assessment of gene-loss cost could have a non-negligible effect on the above-mentioned correlation estimate.

In defining a new score for the functional cost incurred upon gene deletions, we also wanted to take into account the fact that multi-functional enzymes (i.e., enzymes that catalyze multiple distinct reactions, Figure 2.1C) may be under more evolutionary pressure to maintain their function(s) than genes performing only a single function, especially if all such functions are essential.

These considerations led us to define a new metric predicting the impact of gene deletions in genome-scale models. In particular, we define the function-loss cost of a gene as the sum of all costs incurred by removing each individual reaction catalyzed by the gene from the network (see also Methods and Figure 2.1), with the assumption of zero backup capacity by isoenzymes. The distribution of the newly introduced function-loss cost is substantially different from the distribution of gene-loss cost computed before (Figure 2.2A, blue distribution). Notably, for any gene that does not belong to the set of isoenzymes or to the set of multi-functional enzymes, the function-loss cost is identical to the gene-loss cost. This set includes genes that encode for proteins that are members of a complex (Figure 2.1A), or any gene which participates in only a one-to-one mapping with a reaction (not shown).

Interestingly, using our new function-loss cost metric as the measure of gene dispensability, we can rescue the expected negative correlation between this measure and gene evolutionary rate (Figure 2.2A, blue distribution). In fact, the mean

anticorrelation between these data ($\rho = -0.24$) is even stronger than the anticorrelation observed between gene evolutionary rate and experimentally-measured gene essentialities, even though strict gene essentiality prediction accuracy obtained using function-loss cost is not improved relative to the accuracy obtained using gene-loss cost (Table S2.2). Note that the distribution of correlations between function-loss costs and evolutionary rates across different environments is similar (and similarly narrow) relative to the distribution previously obtained for gene-loss cost, indicating the recovery of anticorrelation obtained with the function-loss cost is not strongly dependent on nutrient choice.

2.2.3 Isoenzymes play a special role in determining the anticorrelation with evolutionary rate

As a next step in our analysis, we set out to examine the contributions from isoenzymes and multi-functional enzymes to the improved negative correlations. Since very few genes belong both to the class of isoenzymes and to the class of multi-functional enzymes, it was straightforward to separate out the contributions of each of these two classes to the computation of function-loss cost. We thus recomputed the impact of gene deletion in two variant ways. First (hybrid-loss score 1), we computed the gene deletion impact using the gene-loss score (the old, traditional metric) for all genes except the multi-functional enzyme genes, for which we use the new function-loss score. Conversely, in a separate calculation (hybrid-loss score 2), we computed the gene deletion impact using the gene-loss score for all genes except the isoenzyme-associated genes, for which we use the new functional-loss score.

By recomputing the correlation between evolutionary rates and the hybrid scores we found that hybrid-loss score 2 displays a negative correlation very similar observed

with the full function-loss score (Figure S2.7, whereas hybrid-loss score 1 displays no correlation (inferable from Figure S2.7, data not otherwise shown). This indicates that incorrectly accounting for the effect of isoenzyme deletion has a prominent role in the capacity to discern the relationship between the impact of gene deletion and evolutionary rate. In turn, this suggests that deletion of an isoenzyme is costly, corroborating previous arguments that true redundant functional backup is should not be evolutionarily sustainable [103].

In order to gather further insight into the relationship between different enzymes in an isoenzyme set, we tested the correlation between function-loss cost and evolutionary rate for different specific choices of enzymes within each set. Specifically, for each isoenzyme set, we identified the enzyme which is most conserved (slowest evolutionary rate), and the one that is least conserved (fastest evolutionary rate). Thus, across all isoenzyme pairs, we could collect a subset of all fast evolving and slow evolving isoenzymes.

Notably, when computing the correlation between function-loss cost and evolutionary rate with the inclusion of slow-evolving isoenzymes only, we found an average correlation of $\rho = 0.28$. This correlation is even more negative than what found for the whole set (Figure 2.2C). Conversely, we find that by similarly selecting for the fasting-evolving isoenzyme from each isoenzyme set, the correlation distribution does not significantly shift in any direction (Figure 2.2B). Prior work had long ago established that different isoenzymes catalyzing the same reaction evolve at different rates [104], and it has been observed more recently that asymetric evolutionary rates between gene duplicates this could be interpreted as a signal of neosubfunctionalization [105, 59]. Our analysis reveals for the first time that, in computing the anticorrelation between function-loss score and evolutionary importance, once we

exclude the fast-evolving isoenzymes, the correlation becomes even more negative. This suggests that historical (long-term evolutionary) importance of slow-evolving (i.e. highly conserved) genes carries more information about their experimentally measurable essentiality relative to fast-evolving counterparts.

2.2.4 Function-loss cost improves predictions of epistatic interactions involving isoenzymes

Given that our newly defined metric, function-loss cost, significantly affects the way one evaluates the effects of single gene deletion, it is interesting to ask how such new definition affects the ability to recapitulate the effects of multiple simultaneous gene deletions. Whether the combined effect of pairs of genetic perturbations is predictable from knowledge of each individual effect, constitutes a question with broad implications. In fact, deviations from simple expectations (i.e. epistasis) can significantly affect evolutionary processes [33], and can provide valuable functional information about the underlying system [106]. Previous work has investigated the capacity of FBA models to predict epistatsic interactions between metabolic enzyme genes [107], largely motivated by the availability of extensive experimental datasets of genetic interactions in *S. cerevisiae* [31]. Here, we test the efficacy of function-loss cost in predicting epistasis between gene knockout pairs, using a slightly modified to work with the multiplicative null model of epistasis used in studies such as these (see Methods).

Interestingly, modified function-loss cost more accurately predicts the existence of an epistatic interaction between gene pairs than gene-loss cost (Figure 2.3). The sensitivity of the model nearly doubled with respect to each interaction category, with the exception of non-interacting pairs, for which the model already presents

very high sensitivity.

This is especially true of interactions involving an isoenzyme — in fact, every correctly predicted interaction that was not correctly predicted by gene-loss cost simulations involved an isoenzyme. The fact that isoenzymes are overrepresented in this category is perhaps unsurprising, given than under standard gene-loss cost protocols, an isoenzyme may only be predicted to exhibit epistasis under very narrow circumstances, namely (1) it must be one of isoenzyme pair, (2) the secondary mutation must be the deletion of the partner isoenzyme, and (3) the reaction they catalyze must incur a cost penalty when blocked.

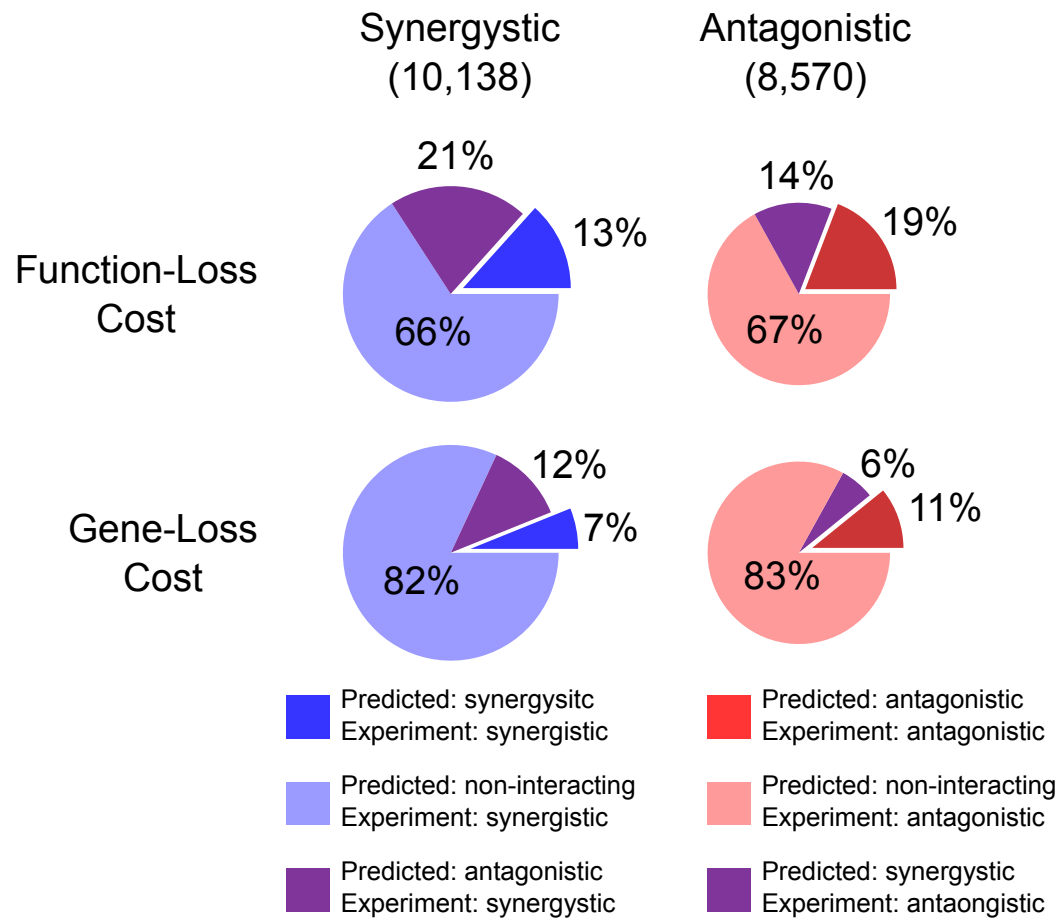

Figure 2.3: Comparison of correctly predicted epistatic interactions per gene. Each row represents one of the two gene dispensability measures in comparison: either function-loss cost (top) or standard gene-loss cost (bottom). Columns represent experimental classification of an epistatic interaction: synergistic (left) or antagonistic (right). The total number of these interactions are listed at the top just below the headers. The pie chart in each quadrant represents the capacity of FBA to predict these interactions. In each pie, the offset slices show the true positives, while the other two slices show the false negatives (and which category they were incorrectly assigned). The size of each chart is determined by the number of interactions in each whole pie relative to the sum total number of interactions in all both pies in the row (pie size is constant across rows). Not shown, non-interactions (see Figure S2.8).

2.3 Discussion

We have introduced function-loss cost, a new metric for quantifying the impact of the deletion of a gene based on genome-scale models of metabolism. This metric is similar to previously estimated gene-loss impacts, except for the modification of some of the basic assumptions on how the deletion of a gene translates into reaction flux constraints. The modification that ends up being responsible for recovering the expected correlation between gene deletion impact and evolutionary rate is the assumption on how isoenzyme deletion affects the corresponding reaction flux. While previous calculations assume that each enzyme in a set of isoenzymes can unconditionally perform the function in the absence of the other isoenzymes, the algorithm we use here assumes that deletion of each isoenzyme causes a complete loss of function for the cell. Based also on multiple types of analyses and observations [100, 108, 102], one would obviously expect the reality to be a complex combination of the above assumptions: different isoenzymes may respond differently to different environmental perturbations, and provide backup to each other to varying degrees. What our results indicate, however, is that *on average* the assumption that each isoenzyme fulfills an essential metabolic role is more consistent with the evolutionary record than the opposite assumption of isoenzymes being unconditionally, individually dispensable.

From the perspective of flux balance modeling, our analysis suggests that extra caution should be used when applying the classical gene-to-reaction mapping relationships to estimate the effect of gene loss, especially when using these models to understand evolutionary aspects of metabolism. As to whether our newly suggested way to deal with isoenzyme deletion will be helpful in comparisons with experimental gene deletion studies, this requires additional evaluations.

With respect to epistasis, we have shown that our approach makes it possible to

use flux balance analysis to predict interactions between isoenzyme genes. In prior calculations, using the standard gene-to-reaction mapping (Figure 2.1B), it would have been possible to detect such interactions only between two isoenzymes that are the only two catalysts for a given reaction. In any other case (e.g. interaction involving a single isoenzyme and another arbitrary enzyme), the complete backup assumption of isoenzyme sets would completely mask any possible interaction. One should keep in mind, conversely, that our method naturally tends to overpredict of such interactions. In the future, by integrating high throughput experimental data (such as epistasis measurements) and network structure information, it may be possible to rewrite reaction-specific gene-to-reaction relationship (using AND or OR) further improve model prediction capacity.

This could prove to be a very important development for the use of constraint-based models as tools in the future study of genetics, especially in the area of biomedicine. Double gene deletions that result in cell death (synthetic lethal deletions) are an important avenue of cancer research, where the ability to induce lethality only in a within subpopulation of cells that carry specific mutations already by inducing a perturbation to the entire population is of obvious benefit. Similarly, research into other metabolic diseases, such as fructose intolerance, could benefit from increased ability to predict unexpected losses of some non-growth phenotypes due to a double gene perturbation event.

2.4 Materials & Methods

2.4.1 Yeast metabolic model and genes used in this study

This study was conducted using the Yeast 6 metabolic model of *Saccharomyces cerevisiae* metabolism [96], which may be obtained from (specifically, version 6.06).

This model specifies a metabolic network consisting of 1888 reactions between 1488 metabolites, a set of 904 enzyme-encoding genes, and a set of Boolean expressions associating each reaction to all possible subsets of genes that are required for catalysis (the gene-to-reaction mapping, as known as the gene-protein-reaction expression map or GPR). We identified blocked reactions in the model (reactions incapable of carrying flux) using a previously established method [109] and subsequently purged all genes associated only with these reactions from our analyses. All subsequent analyses presented in this section and the results presented in this paper were made using this modified subset of 703 genes. For the specific cases of correlating gene-loss cost and function-loss cost with evolutionary rate, restricting our analyses to a subset of metabolic genes did not significantly impact the outcome (Figure S2.6).

2.4.2 Calculation of gene evolutionary rates

The evolutionary rates of all metabolic genes included in this study were derived following the procedure described in [99]. Briefly, an adjusted $\frac{d_N}{d_S}$ [91] ratio (hereafter referred to simply as k) was calculated for each *Saccharomyces cerevisiae* model gene from its corresponding ortholog in five related yeast species: *Saccharomyces bayanus*, *Saccharomyces castellii*, *Lachancea kluyveri* (formerly *S. kluyveri* (28)), *Saccharomyces mikatae*, and *Saccharomyces paradoxus*. This provides, for each gene g, five separate strain-dependent measures of evolutionary rate ($k_g^{S.bayanus}$, $k_g^{S.castellii}$, etc.). To obtain a single representative rate for each gene $\hat{k}_g$, we first grouped all values of k by strain, converted these sets to rank order, and then took the average rank of each gene across these sets; that is, $\hat{k}_g =< rank(\{k_g^y \mid y = x\}) >$ where x is a yeast strain. Some strains did not contain an appropriate ortholog for every gene in our set, however, no gene was without at least one ortholog from which to

derive a k. Note that throughout the paper we refer to the averaged evolutionary rate rank scores as the evolutionary rates. Importantly, since all our correlations involving evolutionary rates are rank-based measures, this does not affect the outcome of these calculations.

2.4.3 Prediction of gene-loss costs for S. cerevisiae metabolic genes

The gene-loss cost of each gene is calculated as the relative loss in predicted fitness of the gene-knockout mutant as compared to the predicted fitness of the wild-type yeast. Fitness predictions for the wild type and all mutants were obtained using standard flux balance analysis (FBA), which has been previously described in [68]. Briefly, FBA calculates the rate of flow (i.e. flux) of metabolites through each reaction (v_i) in the metabolic network in such a way as to maximize the flux through a pseudo-reaction describing organism growth ($w = v_{\text{biomass}}$). Constraints may be imposed on each reaction, such that the minimum and/or maximum flux allowed through it is bounded ($\alpha_i \leq v_i \leq \beta_i$). Gene deletions are translated, through the gene-to-reaction mapping, to constraints on some number of reactions (possibly zero) which limit the flux through these reactions to zero ($\alpha_i = 0 \leq vi \leq \beta_i = 0$). With fitness taken to be the flux through the biomass reaction, the normalized gene-loss cost of any gene g can be expressed as:

$$c_g^{\text{GLC}} = \frac{w_{\text{w.t.}} - w_{\Delta g}}{w_{\text{w.t.}}} \tag{2.1}$$

2.4.4 Prediction of function-loss costs for S. cerevisiae metabolic genes

The function-loss cost for each gene is calculated as the sum total of the individual costs of removing each function (reaction) the gene is responsible for from the model one-by-one, where an individual cost is represented by the fitness loss of the single-

reaction knockout mutant relative to the wild type as predicted by FBA. For this purpose, a gene g is said to be responsible for a reaction r if the gene appears anywhere in that reaction's associated GPR expression. This translates to a fairly simple adaptation of the gene-loss cost metric, which can be expressed as:

$$c_g^{\mathrm{FLC}} = \sum_{\{r|\langle r,g\rangle \in \mathrm{GPR}\}} \frac{W_{\mathrm{w.t.}} - W_{\Delta r}}{W_{\mathrm{w.t.}}} \tag{2.2}$$

2.4.5 Generation of environmental conditions for gene-deletion impact simulations

Environmental conditions for flux balance simulations were generated by an adaptation of a previously defined heuristic for determining minimal media that support growth [97]. First, an initial minimal medium was manually defined for the model, such that each primary nutrient (e.g. carbon and nitrogen) was provided by only a single metabolite. Our initial medium consisted of glucose, ammonium (NH4+), inorganic phosphate and sulfate, oxygen, and minerals (Table S2.3). We then identified alternative carbon-providing metabolites by removing glucose from this initial medium and exhaustively testing all other metabolites for growth. Similarly, nitrogen-providing metabolites were identified by the removal of ammonium and subsequent testing of metabolites. Our final set of minimal media was constructed by taking all pair-wise combinations of carbon-providing and nitrogen-providing metabolites, together with the secondary metabolites listed previously, for which the wild-type model predicted positive growth (Table S2.3).

Simulations were also conducted on several non-minimal environments representing common lab-growth media. Such, so-called "rich media" were defined manually for YPD, YPLactate (both D- and L-Lactate), an SD complete and SD-His complete media (Table S2.3). Maximum import rates were restricted based on the measured

uptake rate of glucose by *S. cerevisiae* grown in YPD where this rate is limiting.

2.4.6 Calculation of epistasis

Epistatic interaction scores were calculated for each possible pair-wise interaction between genes using the standard method [110]. In these studies, epistasis (ε) between any pair of genes i and j, is defined as

$$\varepsilon = W_{ij} - W_i \cdot W_j \tag{2.3}$$

where W_i and W_j represent the relative fitness of each single-gene deletion mutant and W_{ij} is the relative fitness of the double-gene deletion mutant. Our function-loss cost metric had to be slightly modified to fit this definition of an epistatic interaction. By default, the relative fitness of any mutant W_i may be derived from the cost of gene deletion (c_i) as $W_i = W_0 - c_i$, where W_0 represents the wild type fitness. This works well with standard gene-loss cost. However, in the case of multi-functional enzymes, the function-loss cost of a knockout (c_i^{FLC}) may exceed one, which causes problems with the expected fitness of the double knockout mutant ($W_i \cdot W_j$) under the multiplicative model. For this reason, all function-loss costs were bounded to a maximum value of one, such that the relative fitness of any mutant is bounded between one and zero, inclusive ($0 \leq W_i \leq 1$).

2.4.7 Comparison of predicted epistasis with experimental data

In order to assess the validity of our predicted epistatic interaction scores, we compared our predictions against a data set for which these scores have been computed from experimentally observed fitness [31]. We limited our comparisons to genes for which the experimentally observed fitness of deletion mutant was no greater than the

fitness of the wild type, because FBA is incapable of predicting increases in fitness due to gene deletions (in the absence of other types of perturbations). We started by classifying all calculated epistasis scores, from both data sets, using the standard method [110] based on the sign of ε. Each score implied either a synergistic interaction (negative ε), an antagonistic interaction (positive ε), or a non-interaction (ε close to zero). Performance was measured by testing whether or not the epistatic classification predicted using FBA techniques matched the experimental classification.

2.5 Supplemental Figures and Tables

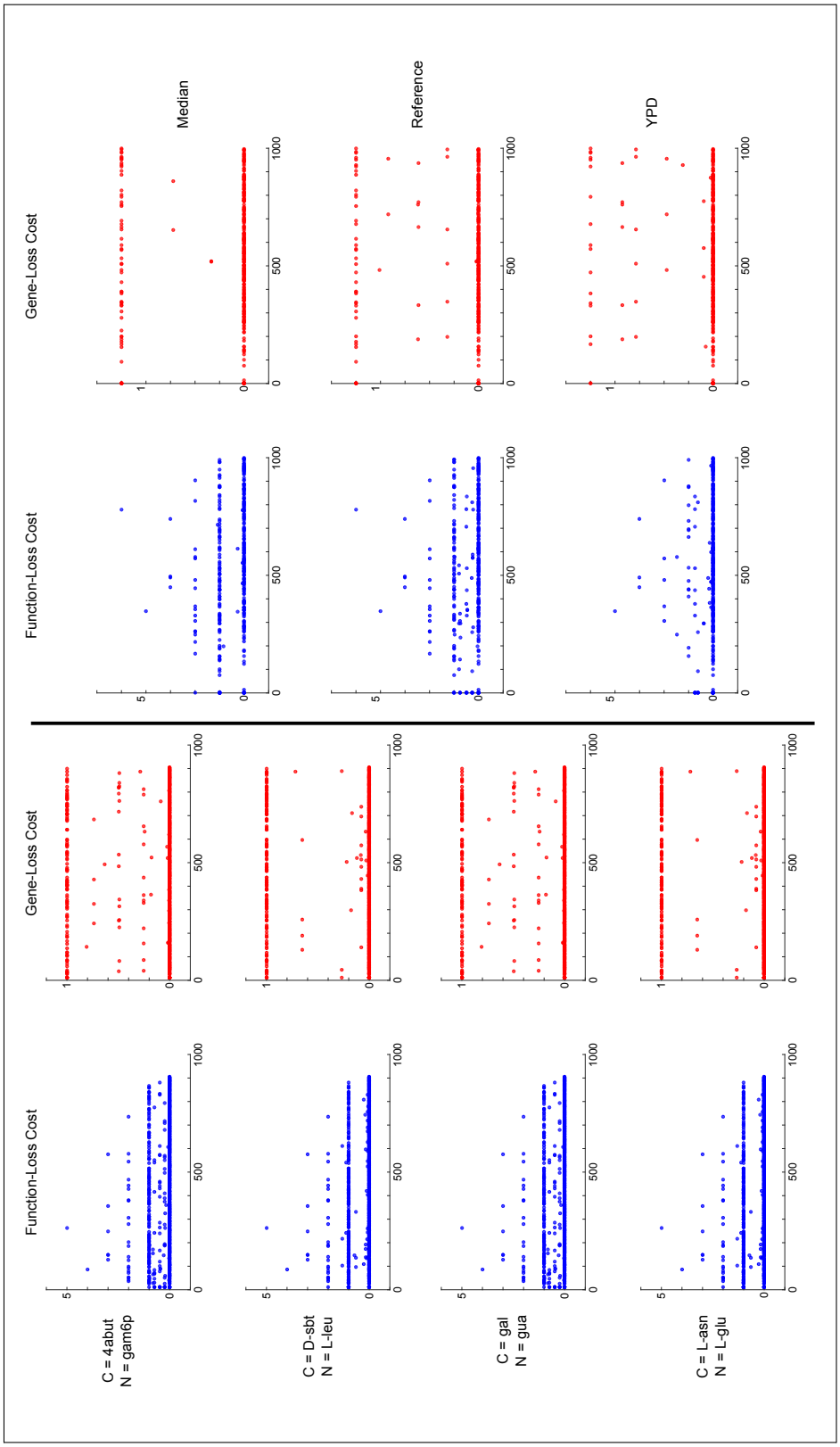
Function-Loss Cost
Gene-Loss Cost
C = 4abut
N = gam6p
C = D-sbt
N = L-leu
C = gal
N = gua
C = L-asn
N = L-glu
Median
Reference
YPD

Figure S2.1: Sampling of gene dispensability vs. evolutionary rate plots. These plots show the specific function-loss cost vs. evolutionary rate (blue dots) and gene-loss cost vs. evolutionary rate (red dots) plots. The four plots on the left side of the figure compares the dot plot generated by function-loss cost and gene-loss cost by testing in the same media. The plots were selected as the most extreme ρ-generating media from both distributions (function-loss cost, top and bottom; gene-loss cost top middle and bottom middle). The three plots on the right show the dot plots for YPD rich media, the reference media (the default carbon/nitrogen pair from which all other media were generated, see Table S2.3) and the median ρ-producing media set.

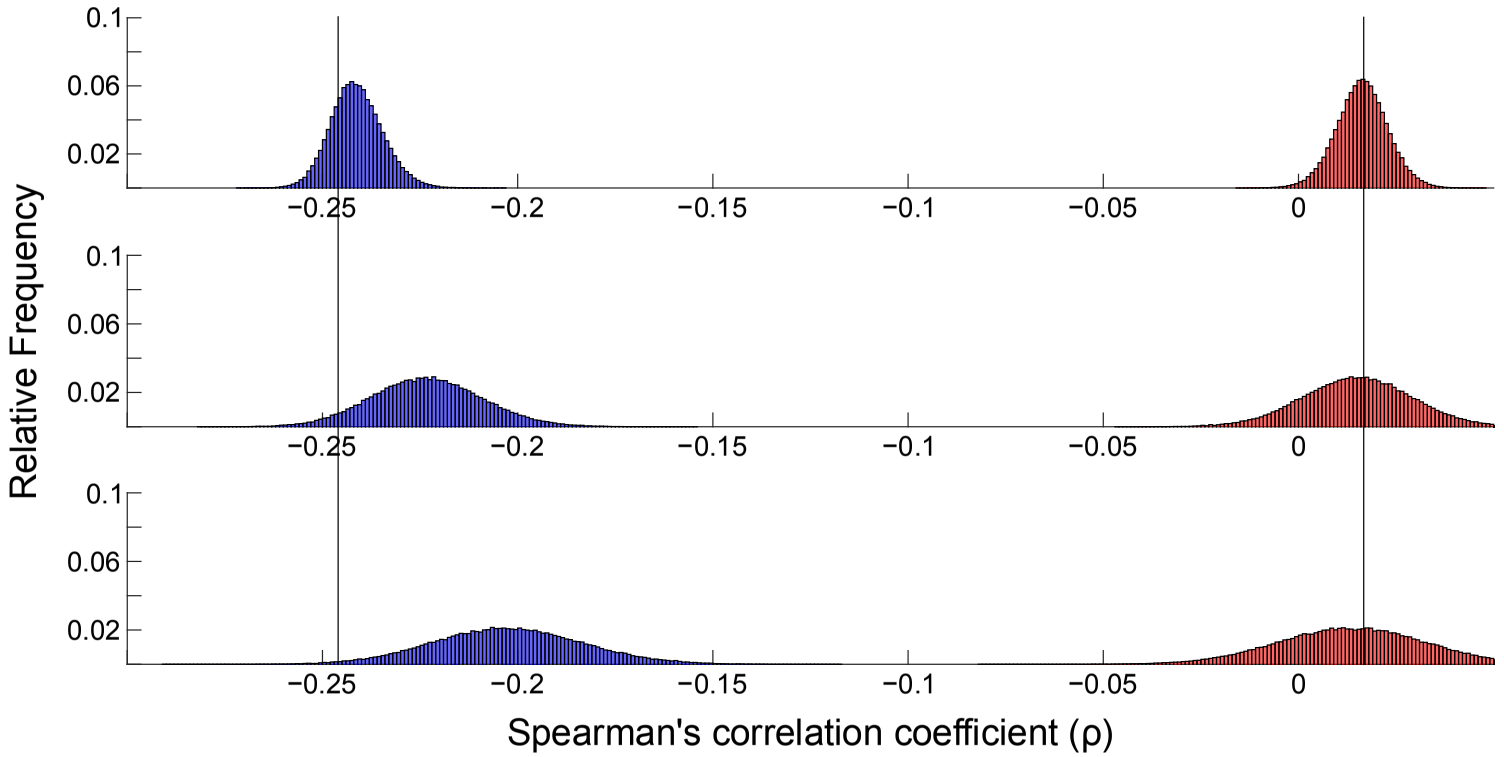

Figure S2.2: Robustness of Spearman's ρ correlation for the median minimal media. Each pair of distributions represents 100K recalculations of the Spearman's rank correlation coefficient (ρ) between evolutionary rate and two measures of gene dispensability — function-loss cost (blue) and gene-loss cost (red) — after swapping 1% (A), 5% (B), or 10% (C) of evolutionary rate ranks. The median minimal media was taken as the minimal media set whose Spearman's ρ was closest to the median value in the original distributions (Figure 1A): acetate (carbon source) and L-asparagine (nitrogen source) in the function-loss cost distribution and L-glutamate (carbon) and adenosine 3',5'-bisphosphate (nitrogen) in the gene-loss cost distribution.

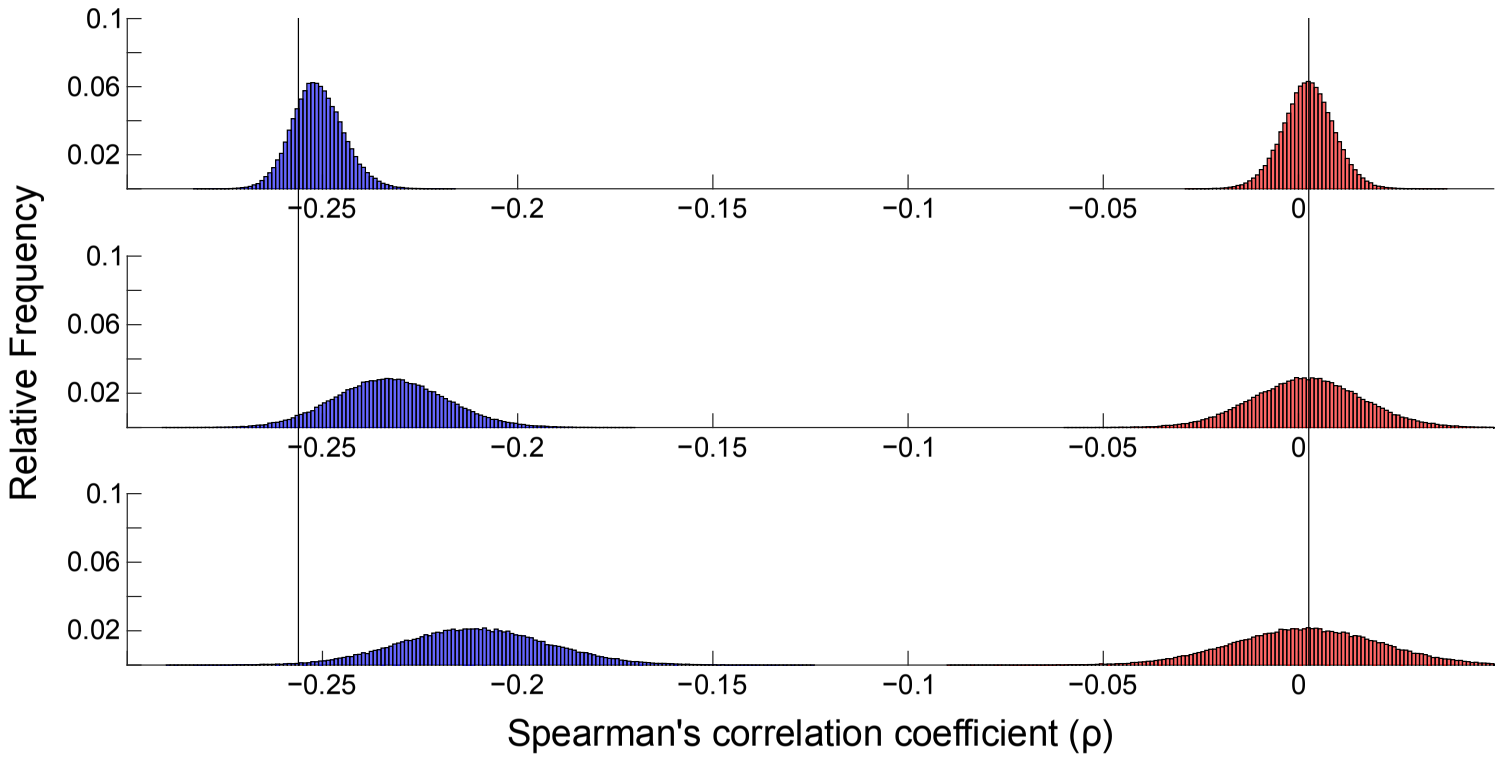

Figure S2.3: Robustness of Spearman's ρ correlation for the reference minimal media. Each pair of distributions represents 100K recalculations of the Spearman's rank correlation coefficient (ρ) between evolutionary rate and two measures of gene dispensability — function-loss cost (blue) and gene-loss cost (red) — after swapping 1% (A), 5% (B), or 10% (C) of evolutionary rate ranks. The refernce minimal media contains glucose and ammonium.

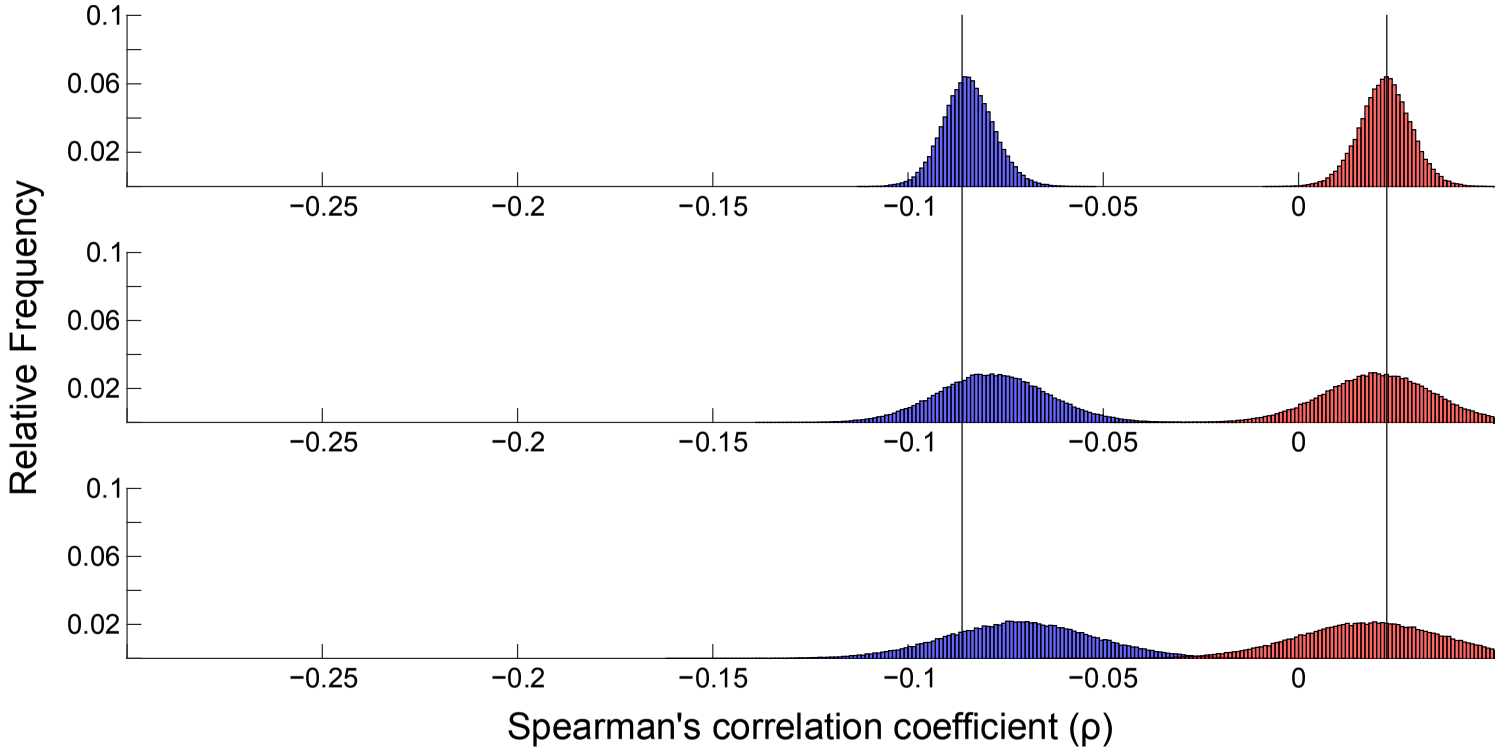

Figure S2.4: Robustness of Spearman's ρ correlation for YPD media. Each pair of distributions represents 100K recalculations of the Spearman's rank correlation coefficient (ρ) between evolutionary rate and two measures of gene dispensability — function-loss cost (blue) and gene-loss cost (red) — after swapping 1% (A), 5% (B), or 10% (C) of evolutionary rate ranks. The metabolic capacity of YPD is in Table S1

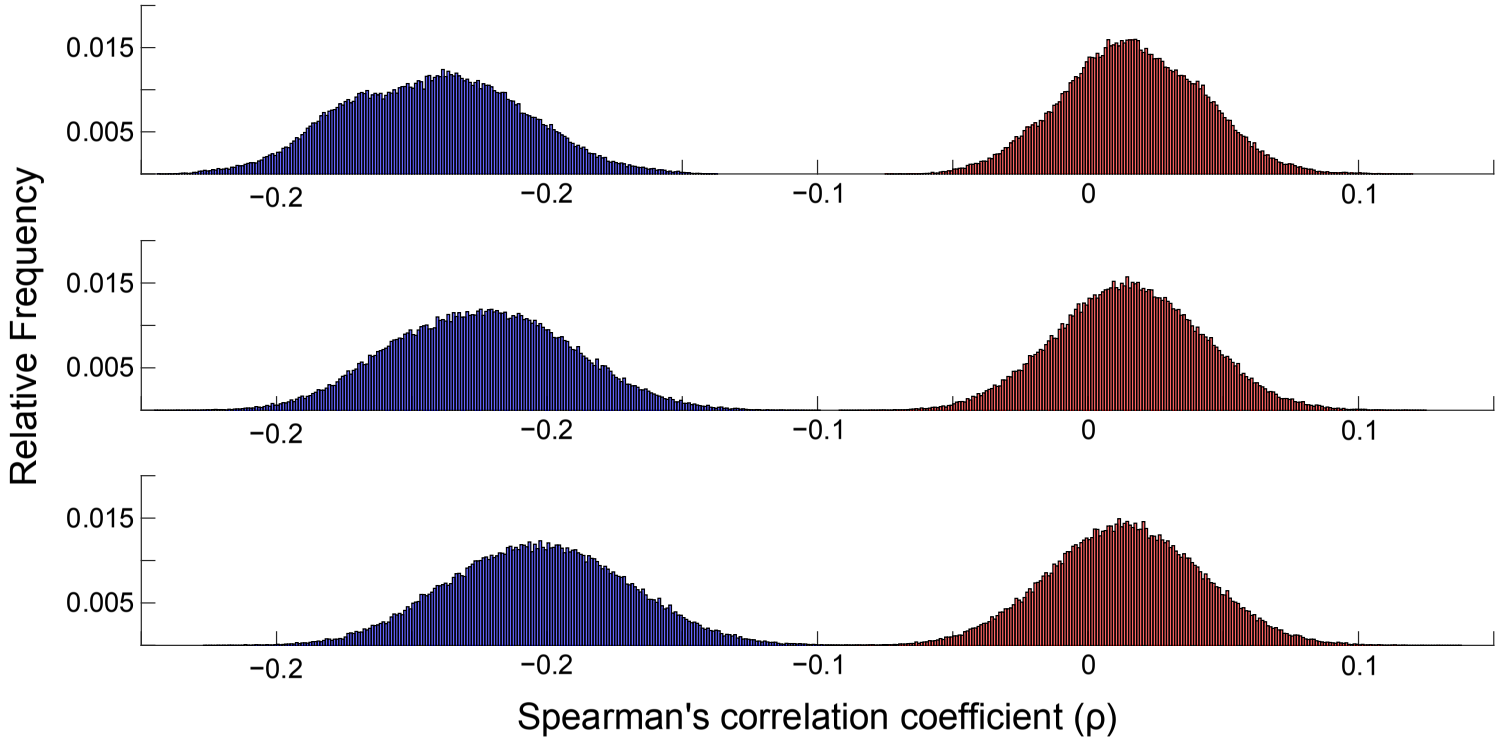

Figure S2.5: Robustness of Spearman's ρ correlation for random minimal media. Each pair of distributions represents 10M recalculations of the Spearman's rank correlation coefficient (ρ) between evolutionary rate and two measures of gene dispensability — function-loss cost (blue) and gene-loss cost (red) — after swapping 1% (A), 5% (B), or 10% (C) of evolutionary rate ranks. A new media was selected on each iteration.

Table S2.1: List of blocked genes in the Yeast 6 consensus model. Blocked reactions in flux balance analysis are those reactions in the netword which are incapable of carrying flux. This is the list of all the gene which catalyze these blocked reactions. Their calculated gene-loss cost and function-loss score must also be zero. Total of 201 genes.

Yeast 6 Blocked Genes					
YAR073W	YAR075W	YBL013W	YBL064C	YBL076C	YBR003W
YBR023C	YBR026C	YBR034C	YBR038W	YBR058C-A	YBR121C
YBR127C	YBR132C	YBR149W	YBR161W	YBR199W	YBR205W
YBR213W	YBR265W	YCL004W	YCL069W	YCR075C	YDL015C
YDL040C	YDL045C	YDL078C	YDL100C	YDL103C	YDL131W
YDL141W	YDL142C	YDL182W	YDL185W	YDL205C	YDR023W
YDR037W	YDR044W	YDR047W	YDR062W	YDR072C	YDR135C
YDR173C	YDR204W	YDR236C	YDR242W	YDR256C	YDR268W
YDR294C	YDR315C	YDR341C	YDR376W	YDR483W	YDR531W
YEL017C-A	YEL027W	YEL042W	YEL051W	YEL058W	YER014W
YER015W	YER061C	YER087W	YER119C	YER141W	YER175C
YFL001W	YFL017C	YFL022C	YGL040C	YGL063W	YGL067W
YGL119W	YGL184C	YGL225W	YGL245W	YGR020C	YGR065C
YGR094W	YGR096W	YGR143W	YGR147C	YGR171C	YGR185C
YGR247W	YGR255C	YGR264C	YGR267C	YGR286C	YHR011W
YHR019C	YHR020W	YHR026W	YHR039C-A	YHR043C	YHR044C
YHR067W	YHR068W	YHR091C	YIL078W	YIL134W	YIL164C
YIL168W	YJL068C	YJL071W	YJL101C	YJL126W	YJL134W
YJL139C	YJR066W	YKL055C	YKL080W	YKL132C	YKL140W

Table S2.1 continued on next page...

...Table S2.1 continuation from previous page.

YKL194C	YKL203C	YKR053C	YKR061W	YKR069W	YKR093W
YLL012W	YLL018C	YLL048C	YLL057C	YLL061W	YLR060W
YLR109W	YLR138W	YLR151C	YLR172C	YLR189C	YLR195C
YLR201C	YLR240W	YLR260W	YLR285W	YLR299W	YLR305C
YLR307W	YLR308W	YLR351C	YLR372W	YLR382C	YLR447C
YML086C	YML110C	YMR013C	YMR054W	YMR088C	YMR113W
YMR207C	YMR226C	YMR272C	YMR293C	YMR296C	YMR319C
YNL003C	YNL009W	YNL029C	YNL073W	YNL192W	YNL247W
YNL267W	YNL292W	YNR041C	YNR056C	YNR057C	YNR058W
YOL033W	YOL049W	YOL096C	YOL097C	YOR040W	YOR099W
YOR125C	YOR168W	YOR171C	YOR176W	YOR221C	YOR222W
YOR241W	YOR270C	YOR278W	YOR332W	YOR335C	YPL040C
YPL053C	YPL057C	YPL069C	YPL097W	YPL104W	YPL134C
YPL160W	YPL172C	YPL212C	YPL234C	YPL244C	YPL252C
YPL268W	YPR033C	YPR036W	YPR047W	YPR081C	YPR128C
YPR159W					

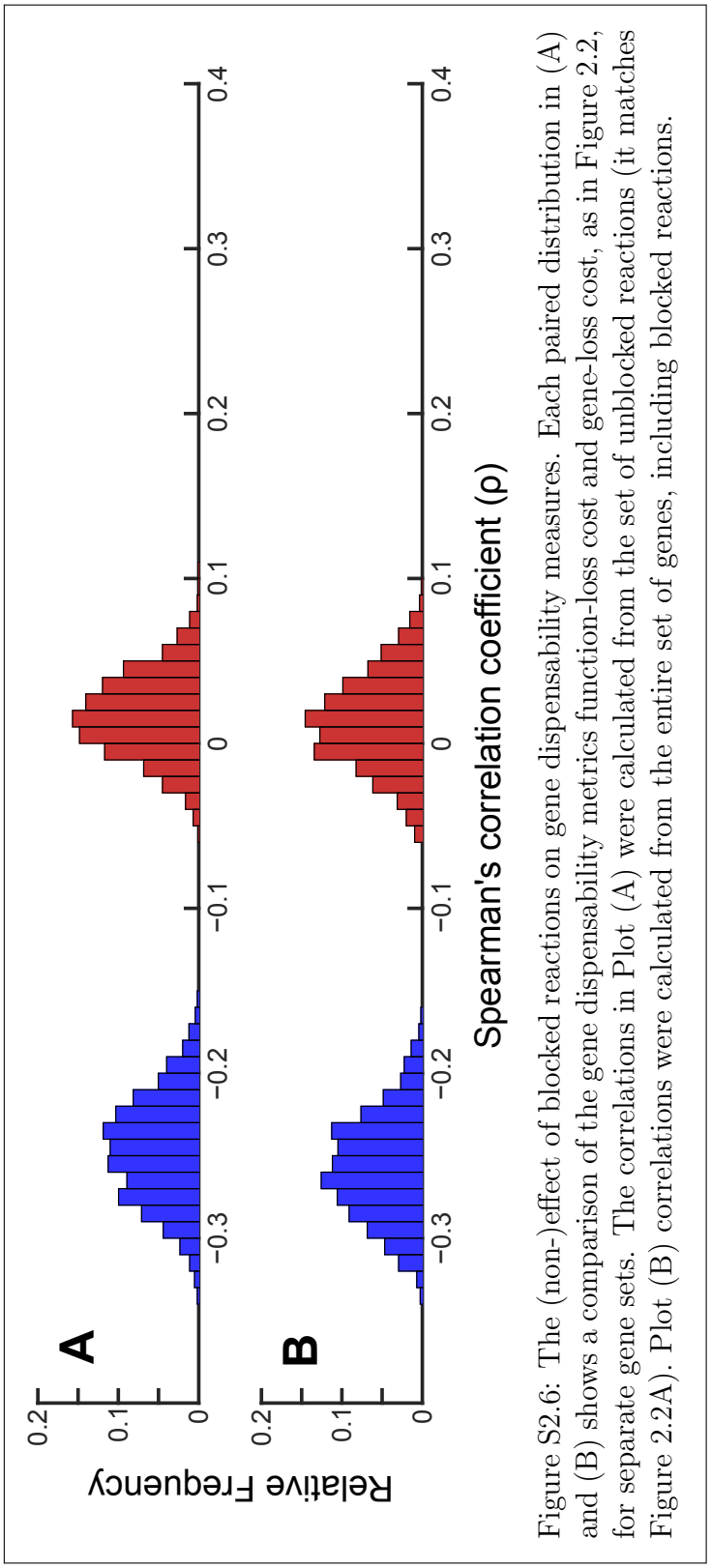

Figure S2.6: The (non-)effect of blocked reactions on gene dispensability measures. Each paired distribution in (A) and (B) shows a comparison of the gene dispensability metrics function-loss cost and gene-loss cost, as in Figure 2.2, for separate gene sets. The correlations in Plot (A) were calculated from the set of unblocked reactions (it matches Figure 2.2A). Plot (B) correlations were calculated from the entire set of genes, including blocked reactions.

		True Status: Essential	True Status: Non-Essential	Total
FBA Prediction	Essential	tp_1	fn_1	$tp_1 + fn_1$
		tp_2	fn_1	$tp_2 + fn_2$
	Non-Essential	fp_1	tn_1	$fp_1 + tn_1$
		fp_2	tn_2	$fp_2 + tn_2$
	Total	$tp_1 + fp_1$	$fn_1 + tn_1$	N_1
		$tp_2 + fp_2$	$fn_2 + tn_2$	N_2

Table S2.2: Comparison of the gene dispensability metrics in predicting gene essentiality. This table is a combined contingency table for function-loss cost (blue numbers) and gene-loss cost (red numbers). The TPR for predicting essential genes in *S. cerevisiae*

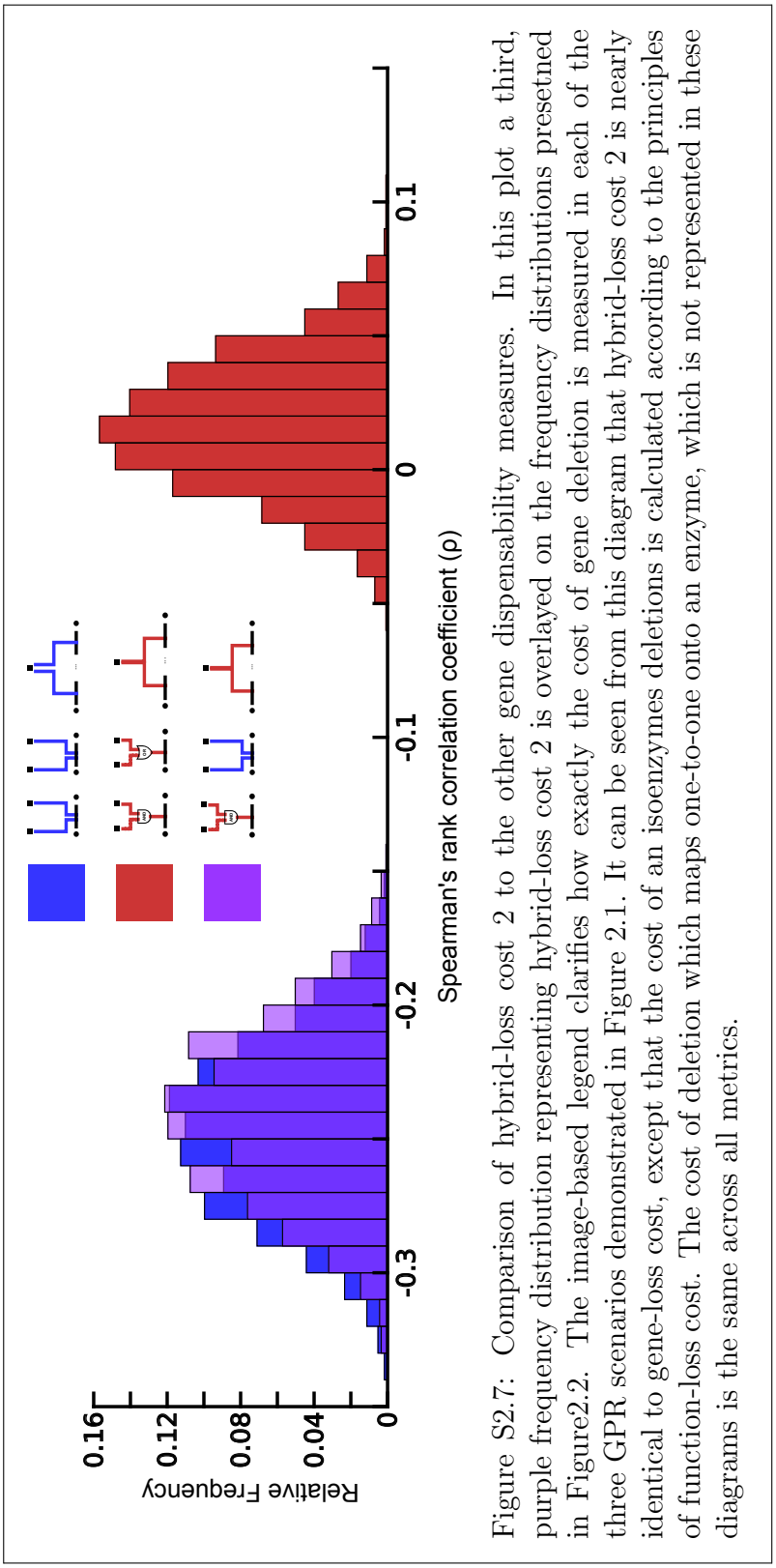

Figure S2.7: Comparison of hybrid-loss cost 2 to the other gene dispensability measures. In this plot a third, purple frequency distribution representing hybrid-loss cost 2 is overlayed on the frequency distributions presetned in Figure2.2. The image-based legend clarifies how exactly the cost of gene deletion is measured in each of the three GPR scenarios demonstrated in Figure 2.1. It can be seen from this diagram that hybrid-loss cost 2 is nearly identical to gene-loss cost, except that the cost of an isoenzymes deletions is calculated according to the principles of function-loss cost. The cost of deletion which maps one-to-one onto an enzyme, which is not represented in these diagrams is the same across all metrics.

Table S2.3: List of metabolites used to simulate media sets. The first row specifies the reference minimal media set that was used to generate all other minimal media sets. The next two rows, labeled "Carbons" and "Nitrogens", list all possible carbon and nitrogen sources that could substitute for D-glucose and ammonium (NH4+) in the reference minimal media set. The final columns provide a complete listing of the metabolites within the rich media sets tested.

Media Set	Metabolites
Reference minimal media	Glucose, ammonia, water, potassium, sodium, phosphate (Pi), sulfate, sodium, iron
Carbons	(1→3)-β-D-glcan, 2-hydroxybutyrate, 4-aminobutanoate, acetate, acetaldehyde, adenosine, 2-oxoglutarate, L-alanine, S-adenosyl-L-methionine, L-arginine, L-asparagine, L-aspartate, citrate, cytidine, ethanol, fuctose, fumarate, D-galactose, D-glucosamine 6-phosphate, L-glutamine, L-glutamate, glycine, glycerol, guanosine, inosine, D-lactose, L-lactose, L-malate, maltose, D-mannose, melibiose, oxaloacetate, ornithine, adenosine 3',5'-bisphosphate, L-proline, pyruvate, D-ribose, D-sorbitol, L-serine, succinate, sucrose, L-threonine, trehalose, uridine, xanthosine, D-xylose, xylitol

Table S2.3 continued on next page...

...Table S2.3 continuation from previous page.

Nitrogens	4-aminobutanoate, adenine, adenosine, L-alanine, allantoin, allantoate, S-adenosyl-L-methionine, L-arginine, L-asparagine, L-aspartate, cytosine, cytidine, deoxycytidine, D-glucosamine 6-phosphate, L-glutamine, L-glutamate, glycine, guanosine, guanine, L-isoleucine, L-leucine, ornithine, adenosine 3',5'-bisphosphate, L-phenylalanine, L-proline, putrescine, L-serine, spermidine, spermine, L-threonine, L-tryptophan, urea, L-valine
YPD	Oxygen, glucose, ammonia, Pi, sulfate, all 20 amino acids (L chiral configuration), potassium, sodium, biotin, choline, riboflavin, thiamine, inositol, thymidine, nicotinate, 4-aminobenzoate, (R)-pantothenate, pyridoxine, uracil
YPLactate	Same as YPD, less glucose, plus D-/L-lactate
SD	Oxygen, glucose, ammonia, Pi, sulfate, all 20 amino acids (L chiral configuration), potassium, sodium, biotin, choline, inositol, uracil
SD-His	Same as SD, less L-histidine

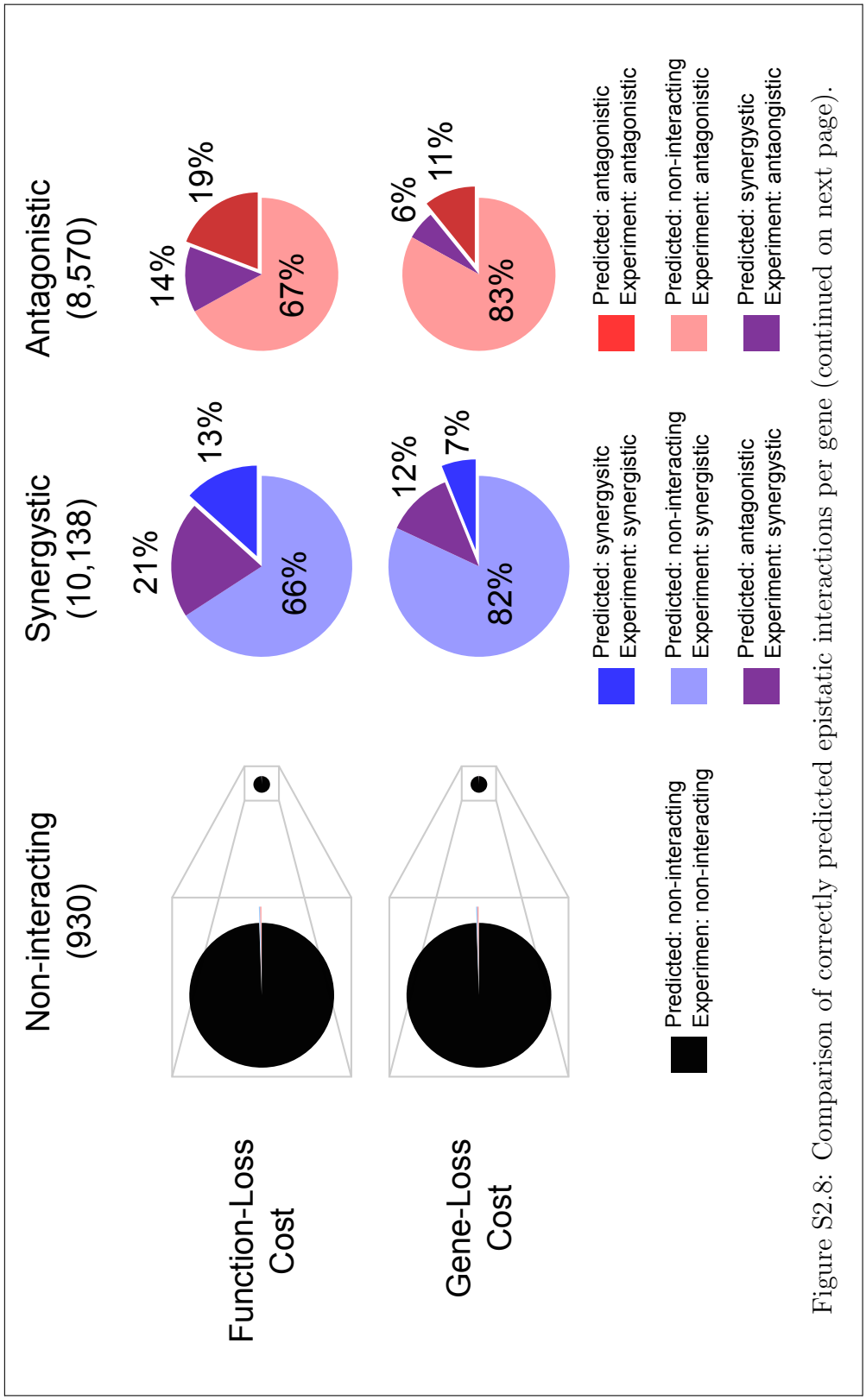

Figure S2.8: Comparison of correctly predicted epistatic interactions per gene (continued on next page).

Figure S2.8: Comparison of correctly predicted epistatic interactions per gene (continuation from previous page). Each row represents one of the two gene dispensability measures in comparison: either function-loss cost (top) or standard gene-loss cost (bottom). Columns represent experimental classification of an epistatic interaction: synergistic (left) or antagonistic (right). The total number of these interactions are listed at the top just below the headers. The pie chart in each quadrant represents the capacity of FBA to predict these interactions. In each pie, the offset slices show the true positives, while the other two slices show the false negatives (and which category they were incorrectly assigned). The size of each chart is determined by the number of interactions in each whole pie relative to the sum total number of interactions in all both pies in the row.

CHAPTER 3

ORGANIZATION PRINCIPLES IN GENETIC INTERACTION NETWORKS

The majority of text in this chapter has been adapted from sections within:

> Jacobs C, Segrè D. Organization principles in genetic interaction networks. In: Soyer OS, editor. Evolutionary systems biology. No. 751 in Advances in Experimental Medicine and Biology. New York: Springer New York; 2012. p. 53–78. EISBN: 978-1-4614-3567-9

I wish to emphasize from the start that this chapter is not meant to be a comprehensive overview of the history and importance of the concept of epistasis in biology. For this purpose, the reader could consult several recent review articles [111, 39, 40, 112, 113] and books [32], in addition to classical textbooks and literature. Rather, I will focus on a specific, relatively recent direction, namely the interplay between the concept of epistasis and the approaches and viewpoints of systems biology.

In the upcoming sections, I will explore in detail some of the concepts I outlined in the introductory section on epistasis. First, I will provide an overview of how epistasis may substantially differ depending on the types of perturbations performed, on the phenotype observed, and on the environmental conditions of the experiment. Next, I will illustrate a standard definition of epistasis in systems biology and the ensuing types of interactions typically encountered. I will spend then a good portion of this chapter describing how the organization of epistatic interaction networks relates to functional classification of cellular components, and how this organization varies as one monitors different phenotypes, with potential evolutionary implications. Finally,

drawing from recent reports of epistasis in laboratory evolution, I will discuss how one might bridge the gap between fitness-level epistasis and epistasis at lower trait levels, perhaps heading toward a global view of the genotype–phenotype mapping and its implications to evolutionary and systems biology.

3.1 Epistasis and Systems Biology

While the central concept of epistasis in systems biology — perturbations combining in unexpected ways — is common to several studies (Figure 3.1), the embedding of this concept in specific biological systems can take many different shapes. First of all, a genetic perturbation may range from a single nucleotide polymorphism (SNP) in the coding or regulatory region of a gene, to a complete deletion of the gene, or its substitution by a different allele. Also, one can focus on either naturally occurring mutations (e.g., beneficial mutations in evolutionary experiments or natural genetic variation in a population) or artificially imposed genetic modifications (such as the systematic deletion of individual genes in an organism or engineered point mutations within a protein [28]). In systems biology, epistasis is typically assessed concurrently for multiple pairs of alleles or perturbations, or, ideally, for all possible perturbations of a certain type in a given system, e.g., the deletion of all gene pairs in a microbial species. Hence, the study of genetic interactions often entails performing high-throughput experiments or computer simulations. In turn, the type of data generated with these approaches can be effectively visualized in the form of a network, where epistatic interactions of a certain type and/or above a certain threshold can be represented as links between nodes associated with individual genes.

It is important to emphasize that the response of an organism to individual perturbations carries in itself abundant biological information, e.g., about essentiality of

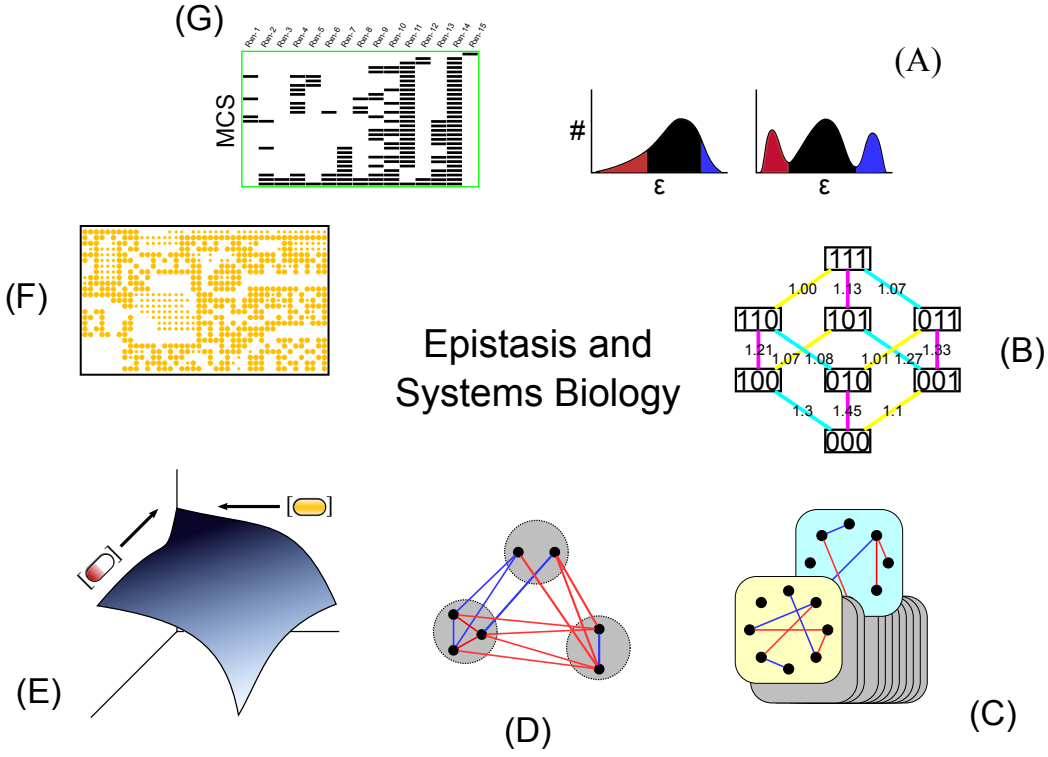

Figure 3.1: Highlights of some of the research areas and questions at the interface between epistasis and systems biology (continued on next page).

Figure 3.1: Highlights of some of the research areas and questions at the interface between epistasis and systems biology (continuation from previous page): (a) The distribution of genetic interactions between several alleles has been the subject of substantial research, largely due to its possible evolutionary implications. The definition and quantification of non-epistatic (black), synergistic (red), and antagonistic (blue) effects depends in general on the null model used (e.g. multiplicative), on the type of mutations (beneficial/deleterious) and on cutoffs in the distribution. (b) Laboratory evolution experiments allow one to identify beneficial mutations occurring during adaptation. Epistasis (in this case antagonistic, or diminishing returns) can then be estimated by measuring fitness for all possible combinations of alleles (represented here as 3-letter strings). (c) Epistasis can be measured or predicted relative to any measurable trait. Hence, one can talk about multi-phenotype epistatic networks. Networks obtained relative to different phenotypes can show different patterns of antagonistic (blue) and synergistic (red) interactions. (d) Epistatic networks can be analyzed using unsupervised clustering into monochromatically interacting modules, i.e. such that all edges between any two clusters are all of the same color. (e) Epistasis can be studied between drugs, in addition to genetic perturbations. Combinations of drugs in different doses give rise to drug–drug interaction landscapes. (f) Epistasis can be measured through high throughput assays, such as epistatic miniarrays, through which vast numbers of single- and double-deletion mutant strains can be grown in parallel, and assayed for colony size (yellow dots). (g) The approach of minimal cut sets (MCS) can be used to find sets (rows) of metabolic network reactions (columns) whose concurrent deletion will cause a drastic change in a specific metabolic flux phenotype, giving rise to what has been named deep epistasis

genes under specific conditions [77, 76]. In order to estimate epistasis, it is necessary to perform all single and all double perturbations of the alleles under study, so that the deviation between the behavior expected from two individual perturbations and the phenotype of the double perturbation can be appropriately quantified. In addition to the most elementary instance of epistasis — pair-wise interactions between perturbations — one could quantify epistasis for all possible sets of three, four, or k perturbations. Even for small genomes, though, this quickly expands to a massive undertaking. For example, to test all the possible pair-wise interactions between deletions of the approximately 6,275 genes in yeast, even assuming that a pair-wise interaction is not dependent on the order of perturbation, one would need to carry out over 19.5 million knockout experiments. Extending such a study to include all possible triplets would need on the order of 1.0×10^{11} knockout experiments.

Another crucial parameter in the definition and quantification of epistasis is the phenotype relative to which an interaction is detected. Classical work on gene deletions, as described below, focuses on growth rate phenotype, partly because it is easily measurable, and partly because of its close relationship to evolutionary fitness in microbial systems. However, this choice is somehow arbitrary, and it is legitimate to ask whether two genes interact epistatically relative to any alternative, non-fitness phenotype. Mapping genetic influences relative to alternative phenotypes is especially important for the study of human disease, where the reduced fitness of an individual is often not readily apparent and/or is directly relatable to the expression of the alternative phenotype. For example, the aberrant phenotype of Alzheimer's disease, a neurodegenerative disease causing dementia, usually only manifests in the elderly, thus its impact on human fitness is not readily apparent until beyond the ages of reproduction. Nevertheless, Combarros et al. were able to statistically investigate

100 potential gene-pair epistatic interactions related to sporadic (i.e., non-Mendelian) Alzheimer's, eventually finding that 27 of these interactions were significantly related to Alzheimer's, including a few pairs which helped reduce the risk of onset of the disease [114]. Such studies may prove to be extremely important to human health in the future, as most traits are not under the control of a single locus [115], and epistatic interactions contributing to susceptibility and resistance seem ubiquitous throughout human disease [116].

In addition to considering multiple perturbations and multiple phenotypes, one can ask how epistasis varies for multiple environmental conditions. Though the environmental impact on human disease phenotypes has been studied for a long time [116, 117, 114], only recently has the idea of environment dependency migrated to epistatic networks in computational simulations and other investigations [118, 76, 119]. Most work in this area focuses on how epistasis depends on only one of the three key variables mentioned (perturbations, phenotype, and environment), largely because of the combinatorial explosion of possibilities, though some examples exist of studies that address the interplay between different variables, e.g., perturbations and environment [120], or perturbations and phenotypes [121]. The evolutionary implications of the environmental dependence of mutational effects and epistasis are in themselves a topic of high importance, recently addressed in RNA enzyme adaptation experiments [122].

3.1.1 Measuring and Predicting Epistasis

For the majority of the lifetime of the term, epistasis was quantitatively deduced by deviations from the expected relative frequencies of phenotype expression [40, 39, 28, 112]. A gene **X** would be epistatic to a gene **Y** if, the presence of the dominant

allele of **X** (*X* written in italics) masked the effect of both alleles of gene **Y** (*Y* / *y*), that is, the phenotypic expression of either *Y* or *y* is not observable in the presence of dominant allele *X*, but is observable with allele *x* (*xx* only, in diploid organisms). This was the definition of an epistatic interaction first described by Bateson and Mendel [123]. Though Bateson's definition of "epistatic" was unidirectional, it was soon after modified slightly, to lose this constraint, such that two genes could be epistatic to each other [39].

For the purpose of quantitative assessment and modeling of epistasis, it is essential to define epistasis in a more formal way, beyond the identification of phenotype masking effects. In particular, this is important for many modeling applications, including epistasis in human disease where different alleles often lead not directly to disease or immunity, but rather to increased susceptibility or resistance to the disease. This requires agreeing on a definition of what it means for a gene to have an effect on a particular trait and on assumptions about gene independence.

For quantitative traits, various mathematical/statistical models of epistasis have been developed [124, 112]. As mentioned above, I will focus here on recent definitions used in functional genomics, rather than other classical definitions found in the population genetics literature. Epistasis, in this context, can be defined as the deviation from a null model, corresponding to a multiplicative law for the combination of individual effects. In mathematical terms words, epistasis is defined as:

$$\varepsilon_{ij} = W_{ij} - W_i \cdot W_j \tag{3.1}$$

where W_{ij} is a measure of the phenotype under consideration, typically fitness, and the null expectation is then given by $W_i \cdot W_j$. All values are expressed assuming a normalized wild-type fitness $W_0 = 1$. A number of alternative metrics for measuring

have been used throughout the literature, including (most notably) additive models where the null expectation matches $(W_i + W_j - 1)$, models of "minimal mathematical function" where the expectation of the double mutant is equal to the minimally "fit" of the single mutants, according to some measure (usually fitness) [124, 112], as well as many variations on the above, including heterogeneity models [39], and scaled measures of ε [110] to name only a couple of examples (more examples may be found in [112]).

An epistatic interaction may be classified as either synergistic or antagonistic. Synergistic epistasis (sometimes aggravating epistasis) describes an interaction which is more severe, i.e., larger in magnitude, than expected. For a combination of beneficial mutations, this would mean that ε has a positive sign, i.e., the double mutant is more fit than expected. However, combinations of deleterious mutations would have negative ε: the double mutant is less fit than expected. Antagonistic epistasis (sometimes buffering epistasis) describes the diminished effects of a genetic interaction, with an opposite trend relative to synergistic effects. One should be aware that the terms positive and negative epistasis can be used with different meanings in the literature. In some papers (mainly dealing with deleterious mutations), positive and negative are used to indicate respectively antagonistic and synergistic epistasis [36, 125], while others (considering mostly beneficial mutations) use positive and negative in the opposite way [126, 127]. In other works positive vs. negative epistasis refers to the sign of ε, as defined in Equation 3.1, where negative ε would imply antagonistic epistasis between beneficial mutations and synergistic epistasis between deleterious ones. Due to this potential ambiguity, I will avoid as much as possible the use of "positive" or "negative" epistasis throughout this chapter.

In addition to synergistic and antagonistic epistasis, it is possible to encounter

cases in which not only the magnitude, but the sign (beneficial/deleterious) of a mutation changes based on the genetic background. For example, one could have deleterious effects for individual mutations ($W_i < W_0$, $W_j < W_0$), but a beneficial effect for the double mutation ($W_{ij} > W_0$). This type of epistasis, which has been named sign epistasis [34], may play a particularly significant role in adaptation, because it is a necessary precondition to the multi-peaked fitness landscapes [33], which force organisms to potentially go through decreased fitness (or wait for alternative phenotype-altering environmental conditions) in order to reach higher peaks.

The availability of robotics and parallelization of experimental assays has made it possible to measure the phenotypic effects for even larger numbers of perturbations. Charles Boone's group began the daunting task of mapping complete epistatic interaction networks for an organism by focusing on a particular form of extreme synergistic deleterious epistasis known as synthetic sick/lethal, or SSL in the model organism *Saccharomyces cerevisiae* (baker's yeast) [31]. SSL double mutants are dead/nongrowing mutants resulting from the crossing of relatively healthy single mutants. Tong et al. introduced a new experimental methodology called the synthetic genetic array (SGA) to test SSL double mutants in a high-throughput manner in their yeast strains [29, 30]. The SGA method was later expanded upon to form E-MAPs (Figure 3.1F), epistatic miniarray profiles [41]. E-MAPs are advantageous because they provide quantitative data on growth rate differences (based on colony size), which in turn allow both antagonistic and synergistic interactions to be observed, using a metric analogous to Equation 3.1.

In parallel to experimental high-throughput technologies for generating genenetic modifications and measuring their phenotypic impact, computational biology has been used to explore the patterns and nature of such perturbations using large-scale

models of biological systems. Often, these endeavors venture into *in silico* studies at the edge of, or beyond, experimental feasibility. A particular example of this type of computational model, which has been used extensively for precisely these types of simulated experiments is provided in Subsection 1.2.1 — flux balance analysis (FBA), a mathematical model of organism metabolism. I will not reiterate the specific history and applications of FBA here, but I do wish to emphasize its particular importance as an efficient computational tool for perfoming large-scale perturbation studies in metabolism such as with the generation of epistatic interaction maps [110, 128, 129, 121].

It is important to mention that while both experimental and computational studies can evaluate growth rates and epistasis based on the multiplicative null model, a potentially thorny issue is the definition of the point beyond which a genetic interaction deviates far enough from the null model to be classified as an epistatic interaction. I will not delve into this issue in this chapter, but point the reader to relevant discussions [124, 112].

3.2 Modularity in Interaction Networks

As is often the case, the analysis of complex biological networks poses difficult computational and interpretational challenges. Genetic networks are no exception: they form graphs containing hundreds or thousands of nodes (genes) and interactions (epistatic links). One useful approach for understanding the biological significance of complex networks has been to organize the nodes into appropriately defined modules — self-contained units sharing common attributes — which underlie the functional hierarchies of biology [130]. Note that a distinction has been suggested between pathways, a (usually linear) chain of information flow through a network, and mod-

ules, which do not necessarily imply a notion of information flow [131]. Despite the name, genetic interactions are not real physical interactions between genes, but rather conceptual links related to the way the system responds to their joint perturbation. Hence, according to the above definition, we expect genetic networks to form modules rather than pathways.

Functional gene modules (or simply modules) in epistatic networks arise from the idea that nodes (i.e. genes) have some functional relationship to one another not only if they are directly interacting, but also if their patterns of interactions with other genes display certain regularities (Figure 3.2), e.g. if they share common neighbors. In this sense, epistatic networks can be clustered into modules using criteria and approaches similar to those implemented for clustering protein–protein interaction networks [132, 133]. Most notably, modules may be defined either as a result of enrichment of edges between the member nodes (within-module) or as a consequence of shared interactions between member nodes and nodes of distant modules (between-module).

Based on the two principles of within-group and between-group clustering, several researchers have proposed clustering schemes and applied them to different datasets in order to understand the nature of modules in epistatic networks. The SSL interaction networks generated by Tong et al. were first clustered within-group by the overlap of interactions between the first and second gene deletions [29, 30]. Segrè et al. found that FBA-generated epistasis data formed hierarchies of pathway-related modules when clustered with respect to their between-group connectivity and monochromaticity, a concept will be explored further in the next section [110]. Costanzo et al. expanded the work of the previous two studies by describing multiple types of monochromaticity in the largest yeast epistasis dataset available so far [31, 134].

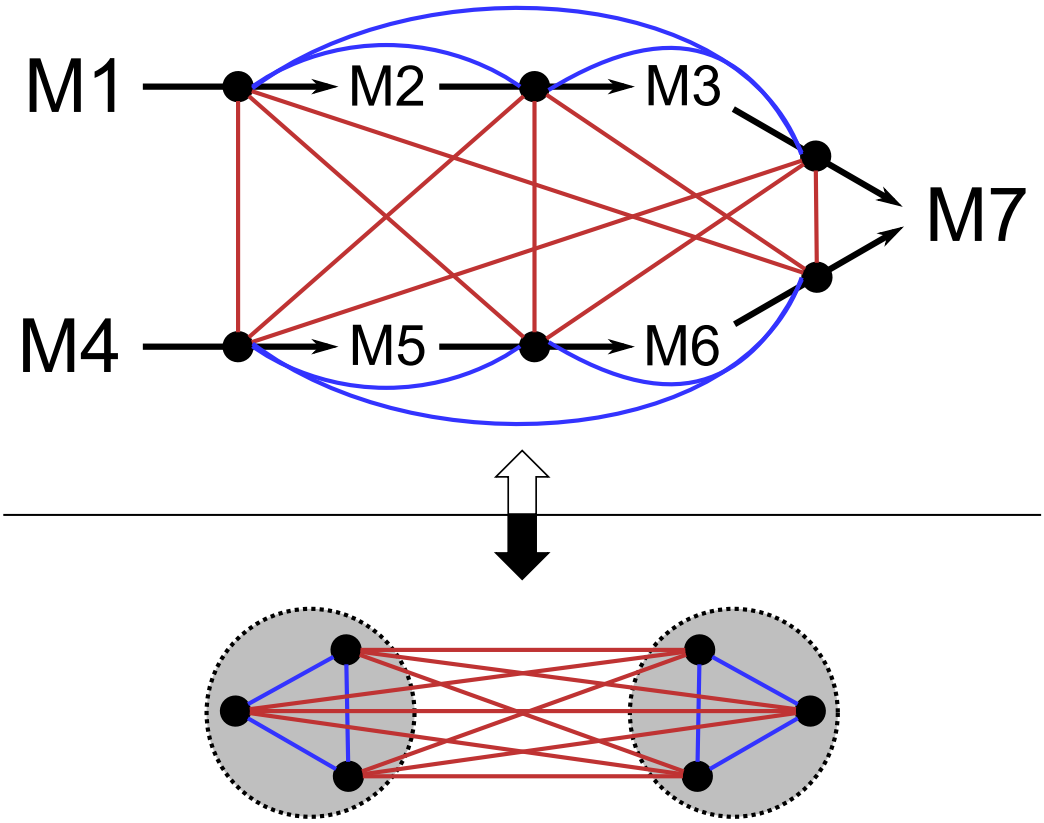

Figure 3.2: From metabolic pathways to epistatic modules. In this toy example we show how genetic interaction modules (bottom) can be related to a segment of a typical metabolic network (top), where either of two precursor metabolites (M1 and M4), but not both, are required for the production of a subsequent essential metabolite (M7). Interactions within either of the parallel pathways are antagonistic (blue), because the loss of a single edge along the pathway is sufficient for the entire pathway to become defunct, thus subsequent deletions have no further impact on fitness. Interactions between pathways are synergistic (red) because even though single deletions in either pathway may be only mildly deleterious, the loss of both genes is lethal. Given that all edges between modules are of a single color, this type of hierarchical organization is named "monochromatic clustering" (see the section on monochromaticity for more details).

Lehár et al. investigated the role of monochromaticity as an agent for selectivity within drug–drug interaction networks [135]. Guo et al. combined previous data on gene–gene interactions with gene–environment and gene–drug interaction data in their description of a recursive expectation-maximization clustering algorithm they ultimately use as a hypothesis-generating tool for investigations into the nature of robustness in cellular processes [136]. In this section, we will first describe in some detail the idea of monochromaticity in genetic networks, and then summarize some large-scale epistasis measurement efforts that corroborated the relevance and utility of this concept.

3.2.1 Hierarchies of Monochromatic Modules

One of the most surprising outcomes of the analysis of the genetic interaction networks predicted with flux balance modeling for yeast metabolism was the discovery of monochromaticity [110] (see example in Figure 3.2). To understand the concept of monochromaticity, it is useful to recall some aspects of how epistatic interaction networks are computed through FBA. A very general property of the solutions to an FBA problem (upon maximization of the biomass production flux) is that any new constraint can only decrease the predicted growth rate. Hence, in FBA calculations, all epistatic effects occur necessarily among deleterious mutations, and synergism/antagonism refers to growth rates that are respectively smaller or larger than expected based on individual perturbations. Hence, if we draw links between epistatic gene pairs in a metabolic network and color-code them according to their class (synergistic/antagonistic), the result is a network connected by edges of two colors (conventionally red for synergistic, green for antagonistic). Upon performing a standard agglomerative hierarchical clustering algorithm, the color of the edges

can be taken into account by requiring that, at every step in the clustering process, two genes (or sets of genes) can combine into a new set only if they do not interact in different colors with any other node or sets of nodes (Figure 3.3). If this property was satisfied for a genetic interaction network, this would imply that, at any level in the hierarchy, modules would interact with each other with only one color. Indeed it was found that for the metabolic network of *S. cerevisiae* [65], the FBA-computed genetic network satisfies the property of monochromatic clusterability [110]. This coherence (or monochromaticity) of interactions between modules allows one to define epistasis as a property of modules, in addition to a property of genes. Modules in metabolic networks display stronger coherent types when epistatic interactions match well against known metabolic pathways. For example, numerous genes belonging to the fermentatory pathway interact synergistically with genes belonging to respiration. The interpretation, in this case, is that these two major energy-transducing pathways play related functional roles and cannot be simultaneously impaired without serious consequences for the cell.

It is interesting to observe that the monochromatic clusterability of the FBA-produced genetic network is not easily satisfied by random networks. In fact, the odds that a random network would be monochromatically clusterable are extremely small. In a small network, it is enough to swap a single edge color to change a monochromatically clusterable network into a non-clusterable network.

From this example of hierarchical modularity in yeast metabolism, we can see how system level properties may arise naturally from interactions at the gene level, which will be an important concept in the next sections.

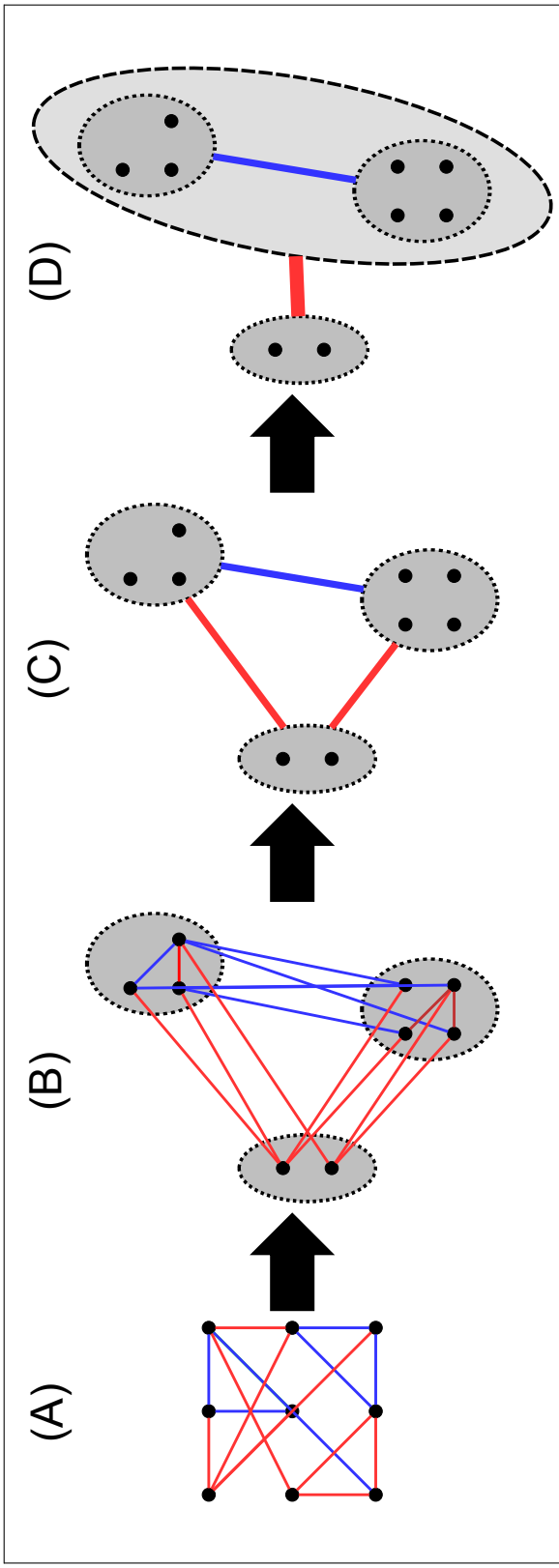

Figure 3.3: Building hierarchies of monochromatic modules. First (a), the epistatic interaction network is expressed as a bi-color graph according to interaction type: red for synergistic interactions, green for antagonistic. Next (b), the network is clustered into monochromatically pure modules using the Prism algorithm [110] (hierarchical agglomerative algorithm accounting for edge color), and groups of nodes are replaced as meta-nodes representing the entire module. Then (c), the links between module-nodes are consolidated into a single representative edge between meta-nodes. Finally (d), the Prism algorithm completes the hierarchy of modules by iterating previous steps (b) and (c) over meta-nodes until no more feasible groupings can be made.

3.2.2 Modularity and Monochromaticity in Experimental Data

While FBA-based phenotype predictions for single gene deletions can reach surprising accuracy, it is not obvious, *a priori*, whether properties of genetic networks discovered in silico should be expected to hold also for experimentally measured networks. In other words, is monochromatic modularity simply a theoretical construct? The idea that clustering methods would be useful to define modules of functionally related genes was already present in the early work on mapping SSL interactions in yeast [29]. The subsequent papers on SGA analysis and E-MAPs by Tong et al. [30], Schuldiner et al. [41], and Collins et al. [137] had increased focus on clustering the interaction networks resultant from their high-throughput experiments. These works mostly focused on clustering around enrichment of epistatic interactions within group. Beginning with the E-MAP data, the Boone and Weissmann groups and others have increasingly examined the role of between-group interactions, including a search for monochromaticity. Constazo et al. observed monochromatic modules of interactions across several cellular processes [31, 134], e.g., metabolism and posttranscriptional modifications, and based on their observations, were able to suggest novel functional annotations for some genes (e.g., for PAR32 and SGT2) and to explain the relationship between the urmylation pathway (posttranslational modification) and elongator complex (transcription). More recently, Szappanos et al. imposed novel experimental knowledge on-top of FBA-derived epistatic interaction predictions, whereupon they found that gene dispensability can be related to degree of synergistic deleterious interactions participated in a property which itself is driven by pleiotropy [107].

The broad concept of monochromatic clustering of genetic interactions is becoming increasingly valuable as a tool for refining our understanding of cellular organiza-

tion. For example, Bandyopadhyay et al. combined E-MAP data and computational predictions of epistasis with TAP-MS (tandem affinity purification followed by mass spectrometry) data, identifying proteins acting within complexes [138]. By doing so they were able to improve predictions of functionally related proteins and protein subunits, which they used to construct a functional map of 91 protein complexes involved in chromosomal architecture. This map led to the discovery of several previously uncharacterized complexes and complex subunits.

Hierarchical modularity has also been applied to classifying drug–drug interactions. Yeh et al. have applied the concept of hierarchical monochromatic clustering to epistatic networks between pairs of drugs [43, 139]. These clusters also map well into classes based on their putative functions, with the exception of drugs affecting the two subunits of ribosomes, which form two classes of protein synthesis inhibitors (PSIs). The separation of PSIs between functional classes was not something which had been noted before, and indeed many of the class–class interactions between drugs had not been well characterized. In related drug–drug interaction screens and clustering, Lehár et al. showed how some combinations of drugs may increase their selectivity [135], a reversal of what is commonly feared by prescribing multiple drugs.

These examples demonstrate how epistasis constitutes an organizing principle for the hierarchy of biological networks, with important practical applications. A fascinating, mostly unanswered question is how evolutionary adaptation gives rise to this unique architecture, and — conversely — whether and how this hierarchical modular organization imposes constraints on evolutionary trajectories.

3.3 Epistasis and Robustness Relative to Multiple Quantitative Traits

Epistasis, in the context of systems biology and evolutionary biology of populations, is often interpreted as the mutual dependence of genetic modifications in their impact on fitness. Interestingly, however, in other contexts — most notably in the study of human disease — researchers care about epistasis insofar as it affects alternative (i.e. non-fitness) measurable traits, such as the predisposition to a genetic disease [117, 140], or the level of metabolites in the blood, bone, etc. [141]. The effect of epistasis on non-fitness phenotypes plays also an important role in metabolic engineering, where the concurrent tinkering with multiple genes is aimed at increasing a practically important phenotype, typically the production of specific industrially or medically important molecules [142, 143, 144]. Might non-fitness phenotypes play an important role also in systems and evolutionary biology?

Genes, and thus epistasis, ultimately act upon fitness by acting on the intermediate phenotypes which comprise fitness. Hence, there are several reasons why alternative phenotypes are relevant to systems and evolutionary biology: (1) Even if one observes epistasis relative to fitness, it is unclear whether this is the result of epistasis relative to some specific trait (e.g. nutrient uptake rate) propagating all the way to fitness, or the outcome of interference amongst several traits; (2) Knowing that two genes are interacting relative to fitness does not provide much information on the underlying molecular mechanism for this interaction; (3) The existence of epistasis relative to various intracellular traits (e.g. size of a given metabolite pool) would imply that simultaneous changes in multiple genes could nonlinearly alter cellular dynamics, posing new questions on the evolutionary and regulatory constraints on cellular organization.

Research on polygenic quantitative trait loci (QTLs) has been concerned with

epistasis relative to non-fitness phenotypes for many years. Such alternative phenotypes may include any quantifiable trait, including metabolic abundance [145], penetrance for disease [146, 147], and several plant-related traits including the two just mentioned [148, 149, 150]. Most relevant to systems biology, largely because of the high-throughput nature, are gene expression QTLs, also referred to as eQTLs. Mapping eQTLs in clonal yeast populations has removed some of the complexity in identifying causal loci, allowing Brem et al. to trace the global expression patterns of over 1,500 yeast genes to causative loci [151, 115]. Epistasis plays a major role in this study as we will see below in the hyperref[subsection:epistasis:phenotyped:robustness]section on on robustness.

Taking a system-level perspective, gene expression quantitative traits are one of many possible phenotypes quantitatively measurable in the cell. However, outside of fitness and expression, large datasets suitable for assessing the degree and nature of epistasis relative to multiple quantitative phenotypes are not readily available. This is why genome-scale models of biological networks can be helpful for a preliminary assessment of such multi-phenotype maps.

3.3.1 Phenotype-Specific Epistasis in Metabolic Networks

In flux balance modeling, each calculation produces, in addition to growth rate, a prediction of all the metabolic fluxes in the cell. This fact offers the opportunity to utilize these fluxes as quantitative traits relative to which epistasis can be estimated. Snitkin and Segrè used flux balance modeling (specifically, MOMA) to compute the entire genetic interaction map for all double mutants in the yeast model with respect to all metabolic flux phenotypes [121]. As before, interactions could be largely classified into antagonistic and synergistic relationships between gene pairs. It is worth

mentioning that, in this case, sign epistasis could occur as well, due to the fact that flux phenotypes may increase or decrease upon perturbations, whereas, in growth-optimized FBA simulations, the growth rate can only decrease upon perturbation.

A key question one can ask about these genetic interaction networks is how similar their connectivity is relative to different flux phenotypes. The model calculations predict that these networks can be quite different, reflecting the fact that different fluxes highlight different regions of the metabolic chart (see below). This can also be expressed in terms of the number of new interactions that each phenotype highlights relative to other phenotypes. Across all phenotypes, more than 2,200 unique epistatic interactions were observed, far more than can be found for fitness or any of the alternative phenotypes alone (see Figure 4 in [121]). Approximately 80 out of 300 different phenotypes are required to capture all unique epistatic interactions. One should keep in mind that these numbers depend on the statistical cutoff used to determine epistasis, and should not be interpreted necessarily as universal quantities.

A specific consequence of the diversity of epistatic maps relative to different phenotypes is that genes can change the sign of interaction depending on the phenotype monitored. Similar to Equation 3.1, for a phenotype k, epistasis can be defined as follows:

$$\varepsilon_{ij}^{k} = W_{ij}^{k} - W_{i}^{k} \cdot W_{j}^{k} \tag{3.2}$$

The phenotype-dependence of the sign of epistasis could then be expressed by saying that a pair of gene knockouts (i, j) could have synergistic epistasis relative to phenotypes $\{k_1,\ k_2,\ \ldots,\ k_q\}$, and antagonistic epistasis relative to phenotypes $\{k_{q+1},\ k_{q+2},\ \ldots,\ k_h\}$. This is indeed abundantly observed in the computationally generated flux balance predictions (see Figure 3.3 in [121]). These predicted mixed interactions indicate that epistasis is not an absolute characteristic of gene-pairs,

but should be contextualized by the phenotype being examined. To our knowledge specific instances of this phenomenon have not been documented experimentally yet. Since several metabolic fluxes (in particular uptake and secretion rates) are experimentally measurable, it should be possible to directly test many of these predictions in the future.

So far, we have mostly discussed the connectivity and phenotype-dependent sign of epistasis in multi-phenotype interaction networks. Next, we want to illustrate the biological insight that multiple phenotype maps can provide. One concept emerging from flux balance predictions of these maps is that different phenotypic readouts provide useful mechanistic insight about the interacting genes or processes, much more than growth rate alone would do. While in growth-based interaction maps the only way to relate genes to function is through clustering and modular organization (two genes interacting may be inferred to have related functions, but there is no information on what that function is), in multi-phenotype maps, knowing that two gene interact relative to a specific metabolic phenotype is in itself informative about the functional relationship between those genes. Two examples of predicted epistatic interactions not visible relative to growth rate, reported by Snitkin et al., illustrate this point. The first example, a synergistic relationship between serine biosynthesis and the genes comprising electron transport chain complex II, results in unexpectedly large secretions of succinate (which in this case can be considered as the observed phenotype). This occurs because the alternate predicted pathway for serine biosynthesis includes succinate as an additional byproduct. A further synergistic relationship between glutamate synthase and the electron transport chain results in surprisingly large secretions of glycerol. Hence, similar to monochromatic modules [110], and to environment-dependent perturbations [108], also multi-phenotype

interaction maps can in principle help annotate genes with unknown functions, and infer relationships between processes.

The predicted existence of several epistatic interactions between different cellular processes relative to a multitude of metabolic phenotypes is yet to be directly tested experimentally. However, it was found that genes highly interacting with other genes through antagonistic interactions relative to multiple phenotypes tend to evolve slower, providing indirect evidence for the value of these predictions, and the importance of these networks from the adaptive standpoint [121].

3.3.2 Multi-Phenotype k-Robustness in Metabolism

One of the consequences of epistasis measured across multiple traits is evolved robustness of cellular systems due to availability of alternative routes to many destinations. Here we use the term robustness to indicate the constancy of a particular (quantitative) trait in the face of large numbers of genetic perturbations. For example, one can think of the entire metabolic network of yeast as being robust under rich growth medium, because less than 20% of genes are essential for growth in YPD (yeast peptone dextrose, a common growth medium) [17]. Such robustness is common across several cellular subsystems [152, 153, 154, 155, 128]. It has been argued that this type of robustness may be largely due to the existence of modules whose genes are linked to each other by synergistic (i.e. aggravating) interactions [80, 155].

While throughout this chapter we have so far only dealt with pair-wise genetic interactions, it has been shown that it is not uncommon for a larger number of genes to be engaged in a single k-wise epistatic relationship. The manifestation of this phenomenon, also known as deep epistasis, gives rise to k-robustness, where multiple genes have to be deleted for a phenotypic change to be detectable [80]. One of the

problems with investigating k-robustness, is that one needs to perform all combinations of k knockouts for large networks per phenotype examined. Although flux balance modeling is very useful in this context, performing exhaustive calculations beyond $k = 4$ becomes prohibitive, requiring other types of approaches to reveal the abundant k-robustness shown to exist above this k value [80, 128, 129]. In particular, the identification of k-robust models for larger values of k has been approached using minimal cut sets (MCSs, Figure 3.1G). The idea of MCSs is to search efficiently for gene sets of arbitrary size k whose removal will result in phenotype loss, while the removal of any subset of such set would not. Initially applied to small biochemical networks [156], this approach has been adapted to genome-scale metabolic networks of *E. coli* [128] and human [129], relative to several different metabolic flux phenotypes. Notably, in these investigations, and in similar studies using *in silico* yeast models [80], the vast majority of k-robust modules discovered are of the highest cardinality investigated: for example, in the work of Imielinski and Belta [129], over 80% of (approximately 33,000 human) essential sets contain 9–10 redundant genes. This general trend of several traits having a high cardinality of epistasis matches well to experimental data in yeast [151, 115].

Deep epistasis and MCSs are another way in which modularity in genetic networks can be used to infer the function of genes where single knockouts fail [157, 158]. The removal of the an individual gene from a k-robust module provides context to the role that gene plays in the overall network, both because of the functional annotation of the other $k-1$ genes within the same module, and because the phenotype relative to which it was observed is potentially informative. Another practical use proposed for deep epistasis and robustness measures is the prediction of gene targets in pathogens, especially multidrug resistant bacteria [159].

3.4 Epistasis as an Organizing Principle

Computational predictions and analyses of epistasis using genome-scale models of metabolism, as well as high throughput experiments, such as SGA and E-MAP have provided snapshots of specific features of genetic interaction networks: hierarchical modularity, monochromaticity, phenotype-dependence, k-robustness, just to mention the ones discussed at length throughout this chapter. Several fundamental questions, however, are still open. One very important challenge is the pursuit of further understanding of the role of epistasis in evolution. While a lot of the high throughput work has been focused on the effects of epistasis between gene deletions, evolution typically involves many different scales of perturbations, from single base mutations, to whole chromosome duplication events. Another related challenge is piecing together these snapshots into a coherent view of the genotype–phenotype map, and on how evolution may have influenced (and be influenced by) its architecture and nonlinearities. In this section we will summarize some recent evidence of epistasis in evolutionary adaptation experiments, and describe how some of the conclusions drawn from these studies may suggest avenues for building an integrated model of epistasis in biological networks.

3.4.1 Epistasis in Evolutionary Adaptation

The recent availability of inexpensive sequencing technologies makes it possible to explore the outcome of adaptation in natural or laboratory evolution experiments. Several authors have now documented in detail the occurrence of epistasis in different settings, ranging from RNA viruses [160, 161], ribozymes [162, 122], to individual proteins [163, 164][9, 111] and whole organisms [165, 166, 127, 167].

Two recent adaptive evolution experiments using *Methylobacterium extorquens*

and *E. coli* demonstrated the emergence of antagonistic (diminishing returns) epistasis between beneficial mutations arising during laboratory evolution [127, 167]. One of these two works, by Chou et al., analyzed evolution of a metabolically impaired *M. extorquens* strain, and identified four major beneficial mutations that provided improved fitness in the evolved strain. By introducing all possible combinations of beneficial mutations onto the ancestor's background and measuring fitness of the ensuing strains, Chou et al. were able to obtain a complete map of the fitness increase of each mutation on the any background of any possible combination of the other alleles [167]. This analysis highlighted an overall general trend of diminishing returns epistasis, a form of antagonistic epistasis whereby the fitness advantage of a beneficial mutation decreases on top of successively more fit backgrounds (Figure 3.1b), which is well in agreement with analogous studies [168, 169, 127]. An intriguing theoretical consideration that emerged from this study is that such diminishing returns epistasis at the level of fitness could be explained by expressing fitness ($f = W^{\text{growth}}$) as the difference between two other traits, a benefit (b) and a cost (c) [170]. For the unperturbed system, fitness is then expressed as:

$$f_0 = b_0 - c_0 \tag{3.3}$$

The decomposition of fitness into benefit and cost in the Chou et al. system was largely motivated by the observation that changes in enzyme levels could tune fluxes affecting metabolic efficiency (benefit), and also alter the degree of morphological defects caused by excessive protein expression (cost). The model proposed to explain the diminishing returns trend assumes that any given mutation may independently alter both the benefit and the cost. If, for a mutation i the benefit and the cost are respectively modified by coefficients θ_i and λ_i (irrespective of previous mutations),

then fitness upon an arbitrary number n of mutations can be expressed through the following generalized equation:

$$f_{i,j,\ldots,n} = \theta_i\theta_j \ldots \theta_n b_0 - \lambda_i\lambda_j \ldots \lambda_n c_0 \tag{3.4}$$

Once b_0, c_0, and each θ and λ are inferred from the experimental data, this Equation 3.3 provides an excellent fit to all the fitness values for all possible combinations of mutations, and recapitulates the experimentally observed diminishing returns effect. Importantly, this antagonistic epistasis emerges at the level of fitness, despite the assumption that, relative to the benefit and cost traits, mutations combine multiplicatively, i.e. non-epistatically. This result underpins a fundamental property of epistatic networks, i.e. that epistasis at "high-order" phenotypes could result naturally from the interrelationship between two "low-order" phenotypes, in turn affected non-epistatically by multiple mutations [171].

While in the work by Chou et al. the decomposition of fitness into simpler traits takes the specific shape of a benefit-cost function, one should not necessarily expect that the relationships between different phenotypic traits will be obvious or intuitive. However, as explored next, we maintain that a hierarchical relationship between traits, and the emergence of epistasis when transitioning from one level of description to the one above, fit nicely with several other observations on genetic networks discussed in the previous sections.

3.4.2 Towards a Hierarchical Genotype–Phenotype Map

Three main principles of organization can be distilled out of the above discussions: (i) Monochromaticity: genetic interactions within and between modules tend to display coherent patterns of synergistic/antagonistic links; (ii) Phenotype-specificity: the

same pair of genes may interact with different types of interactions depending on the phenotype or trait relative to which epistasis is evaluated; (iii) Emergence of epistasis from coupling of traits: genes may display no epistasis relative to two simple traits, but could become interacting relative to a more complex trait that can be expressed as a function of the simpler traits [171]. In this final subsection we ask whether these three principles fit into a coherent view of how epistatic networks are organized.

In Figure 3.4, we propose a possible connection between these three principles that we think captures some important aspects of genetic network organization. The two bottom panels of Figure 3.4 display two very different genetic interaction networks resulting from measuring the two phenotypes $\mathbf{X}$ and $\mathbf{Y}$, highlighting the phenotype-specificity of epistasis (principle (ii)). Fitness in this toy model is an arbitrary function f of the two traits $\mathbf{X}$ and $\mathbf{Y}$. Principle (iii) suggests that it is possible for two genes to have no interaction relative to individual traits (e.g. two genes from sets S_{X} and S_{Y}), but become epistatic relative to fitness, due to the dependence of fitness on such traits, giving rise to the links between sets in the top panel. In general, the transition from low to high level could also cause the disappearance of specific epistatic links. Finally, genes that belong to sets highlighted by specific phenotypes in the lower levels will tend to cluster monochromatically (principle (i)), i.e. interact in a coherent fashion with genes that were responsive relative to a different phenotype.

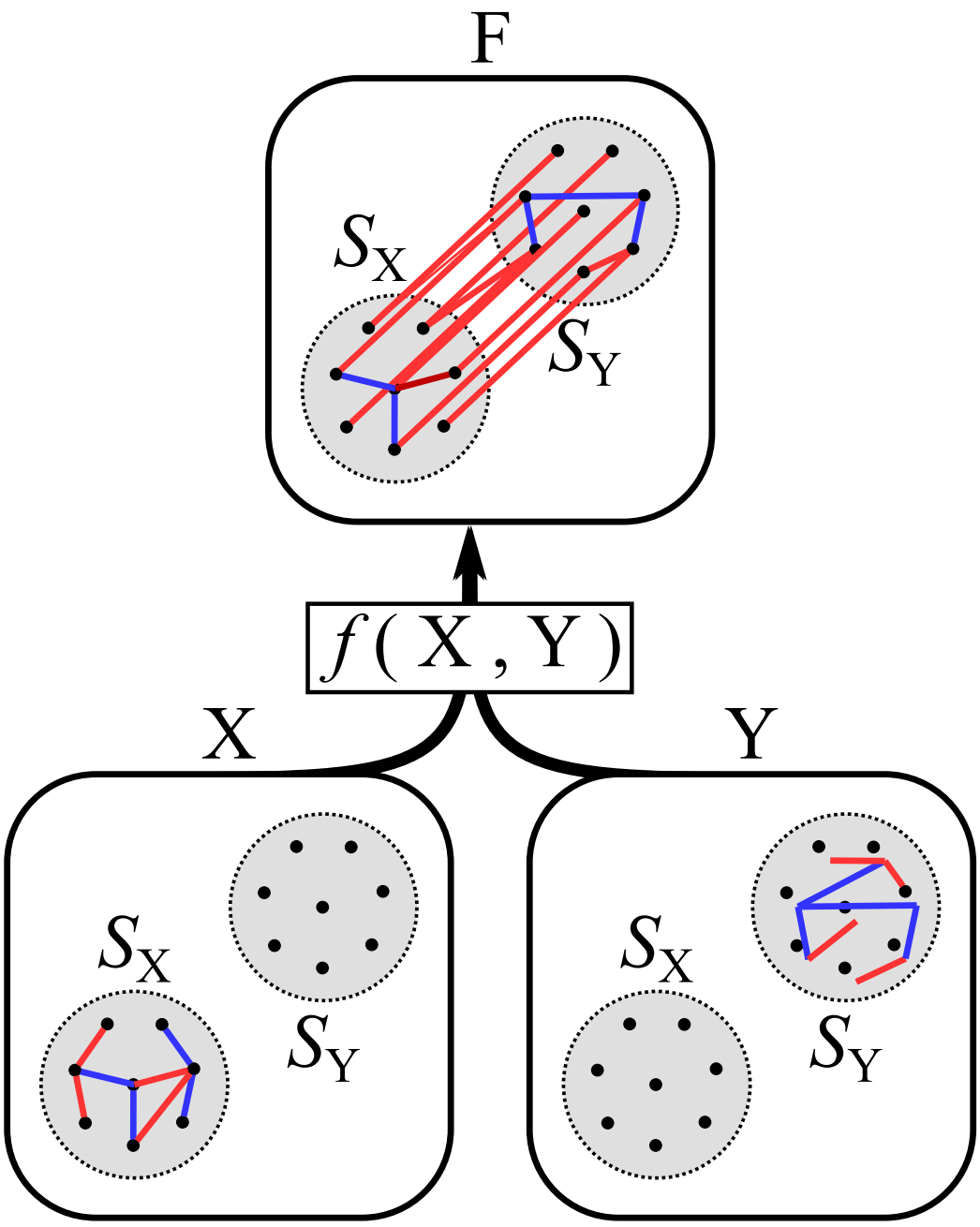

Figure 3.4: Organizing principles of epistasis (continued on next page).

Figure 3.4: Organizing principles of epistasis (continuation from previous page). Each panel represents the complete epistasis interaction map for a toy genome relative to the phenotypes **X**, **Y** and **F**. The set of genes S_{X} is associated with phenotype **X** and similarly S_{Y} are those genes associated with phenotype **Y**. The fitness phenotype, **F**, is dependent on the phenotypes **X** and **Y** through a function $\mathrm{F} = f(\mathrm{X}, \mathrm{Y})$. The genetic interaction map of **F** includes monochromatic epistasis between the sets S_{X} and S_{Y} , which could not be detected relative to either **X** or **Y**, and informs the functional relationship between **X** and **Y**.

CHAPTER 4

CONCLUSIONS AND OUTLOOK

4.1 Review and Additional Discussion Points

The major work of this dissertation focused on the implications of gene deletion studies in the model organism *Saccharomyces cerevisiae*. I used computational biology methods to systematically estimate the cost effect of a gene loss event for both single gene deletion and double gene deletion studies in yeast metabolism. I developed a new quantitative measure of a gene loss event and showed that this new metric was capable of bridging a conceptual gap between the functional importance of a gene "here and now" and the same gene's historical importance. I then demonstrated how this metric could be expanded beyond studies in a single gene to the study of epistasis between two genes. Finally, I explored the implications of the study epistatic interactions in systems biology in general, and distilled from this examination a general formulation for epistasis as an organizing principle in biological networks.

4.1.1 Open Questions in the Domain of Epistasis

The subtle complexity of the multilevel relationships between different proposed organizing principles of genetic networks leaves a lot of questions unanswered. First, much of the evidence for these principles is based on partially tested computational predictions. Known limitations of flux balance methods may influence our perspective of epistasis between metabolic enzyme genes. For example, predictions of phenotypic traits and genetic interactions may be affected by the choice of the objective function [172, 76], by the presence of alternative optima in flux balance calculations

[71, 173] or by the lack of explicit regulatory dynamics. Hence, we still do not really know how pervasive epistasis may end up being in real metabolic networks when measured relative to different phenotypes. Given that several genetic diseases involve the manifestation of aberrant phenotypes (typically other than fitness), the prevalence of epistasis relative to such phenotype could have important consequences on the study of human biology and diseases. In addition to the potential relevance of epistasis in genetic studies, a notable recent example of how epistasis can play a role in fighting diseases is the model-mediated discovery of a cancer-specific gene deletion, whose synthetic lethal interaction with a second perturbation makes it possible to selectively target cancer cells without affecting normal ones [174, 175].

Second, if indeed so many internal degrees of freedom of a cell can be nonlinearly affected by multiple minor-effect perturbation of other variables, how does the cell cope? Have cells evolved, as part of their regulatory wiring, the capacity to dampen these effects, avoiding uncontrollable chaos? Or, could biological systems have embraced these epistatic effects, and learned to master them in order to control some portions of the network through subtle manipulation of more easily tunable parameters or genes? Third, it will be interesting to think whether it is possible to explain the whole hierarchy of cellular functions through multi-level traits connected by a complex, but structured web of genetic links. The existence of k-robustness points to the necessity of expanding genetic interaction networks from pair-wise graphs to more complex hypergraphs [176]. Particularly important will be to try and understand how these networks have evolved, and, in turn, how they affect the rate and possibilities of evolutionary adaptation. For example, it would be interesting to explore the relationship between the robustness of metabolism relative to genetic perturbations and its robustness upon changes in environmental parameters, such as

the availability of different nutrients. It is possible that the evolution of a network towards robustness to environmental uncertainty also provides robustness to single and multiple genetic perturbations under certain conditions.

Future research on epistasis will address some of the issues mentioned above through increased computational power and enhanced high-throughput experimental technologies. However, novel insights in the study of genetic interaction networks will likely stem from newly rising research directions in systems biology as well. For example, it will be interesting to explore whether nonlinearities detected at the level of population averages hold also at the single cell level, where gene expression and metabolism can be modulated by stochastic effects and cell individuality. From the mathematical perspective, several groups have started looking beyond current genome scale modeling methods, trying to incorporate thermodynamic constraints (e.g. energy balance analysis [177]), or formulate detailed mass balance models that take explicitly into account all possible macromolecules. Finally, both in the study of human biology and of microbial dynamics and evolution, we expect that a lot of new insight will come from studying the interplay of multiple cell types and microbial species. There is no reason why the synergistic and antagonistic interactions observed between genes and modules should not extend beyond the whole organism level. Stoichiometric flux balance models are already being extended from genome-scale to whole organism [178] and ecosystem level [97, 179, 180], suggesting indeed that metabolic cross-talk may play an important role in the evolution and dynamics of microbial diversity and multicellularity.

4.2 Future Directions and Prospects

4.2.1 The Future Fuzzification of Flux Balance Analysis

In the discussion section at end of the second chapter, I highlighted a particular issue within the GPR mapping of flux balance models, namely the Boolean logic that governs the gene-to-reaction map (GPR set). This simplified logic, specifically the OR gates used to approximate isoenzymatic function, essentially lay at the root of the major problem that was addressed by the function-loss cost metric in that chapter. However, as I noted in that discussion, my proposed solution essentially overcorrects for this issue, effectively ignoring OR gates (isoenzymatic activity) entirely. Unfortunately, the issue is not capable of being fixed by simply updating the each isoenzymatic set to use the most appropriate logical function. The reality of when and how isoenzymes are capable of providing backup for one another is a complex function that depends on many complicating factors, including the extent of their functional overlap, the current cellular regulatory schema, and even environmental condition. This incredible complexity makes it difficult to assign any single Boolean logic function to a set of isoenzymes to describe their behaviors, because even a "best fit" approach will be in error a large percentage of the time. Even if one were to step beyond the GPR set, to my knowledge there does not currently exist a formulation of FBA which is capable of dealing with all of these complexities in a satisfactory way.

The area of fuzzy logic may offer a way in which to address this problem or at least make it more manageable. Fuzzy logic operates similarly to Boolean logic, except that it is capable of working with partial truths or truths to which there may be a degree function applied. A fuzzy GPR would start with a base Boolean logic map and add to that a notion of gene levels in the inputs and the possibly in the

protein output. One could also add inputs describing the environmental condition or any other aspect you wish. In this way, the inputs in to the GPR can be gradated to arbitrary degree, while the protein outputs may still assume a Boolean logic property where desired (e.g. backup, no-backup, some backup). Compensatory fuzzy logic allows for joining fuzzy logic systems using conjunction and/or disjunctions, which also allows the GPR set to make use of rules subsets defined within itself to define further mappings.

Fuzzy logic systems are already being put to practical use every day. Common examples for this include washing machines, where the "dirtiness" of a load is generally not a Boolean value, and subway trains, where ride comfort may be controlled by fuzzy logic inference on a plethora of information such as current train speed and quality of the track. A fuzzy flux balance model would offer a few advantages over FBA formulations that attempt to take advantage of expression data in order to inform bounds on reaction fluxes. Most significantly, such models would not require experimental measurements of expression levels, although they could also take advantage of this data in the inference step. In addition, there is the potential for the speed of calculation to be close to or on-par with standard FBA linear programming.

4.2.2 Environment-Specific Epistasis Relative to Multiple Phenotypes

Chapter three devoted an entire section to the discussion of epistasis in non-fitness phenotypes. However, I was only able to scratch the surface of what it is possible to do with non-fitness epsitasis. Flux balance methods are high-throughput enough that it is not infeasible to imagine computing entire genetic interaction maps for all double mutants in a genome-scale model with respect to all metabolic flux phenotypes. In the current version of the yeast consensus model [67], this would translate to over

1 billion individual interaction scores. If we extrapolate from previous explorations in this area using an older model [121] and assume an increased ability to predict epistasis on par with that gained from using function-loss cost, performing such a study in the yeast model alone would possibly result in up to an 8-fold increase in the number of interactions identified, most of which would not be relative to the fitess phenotype.

Provided with such a plethora of data, the nautral question arises what specifically should be done with it and how it should be organized such that these ideas are best realized. In general, it is possible to describe any epistatic interaction map (ε) as a vector where every possible genetic interaction between pairs of n genes (gene pair $\langle 2,1\rangle$, gene pair $\langle 3,2\rangle$, $\langle 3,1\rangle$, ..., $\langle n, n-1\rangle$) represents a feature for which we can calculate a specific value:

$$\boldsymbol{\varepsilon} = \langle \varepsilon_{2,1}, \varepsilon_{3,2}, \varepsilon_{3,1}, \ldots, \varepsilon_{n,n-1} \rangle \tag{4.1}$$

This formulation is simply and naturally expandable. A phenotype-specific map would be expressed, thusly:

$$\boldsymbol{\varphi}_k = \langle \varepsilon_{2,1,k}, \varepsilon_{3,2,k}, \varepsilon_{3,1,k}, \ldots, \varepsilon_{n,n-1,k} \rangle \tag{4.2}$$

where k represents the phenotype of interest. Of course, one of the major points of emphasis in chapter three was that the measurable or predictable affect of a genetic perturbation is altered by the environmental condition in which it was performed. Similar to phenotype-specific epistasis, we can account for epistatic variation due to

environmental conditions with:

$$\boldsymbol{\vartheta}_k = \langle \varepsilon_{2,1,k}, \varepsilon_{3,2,k}, \varepsilon_{3,1,k}, \ldots, \varepsilon_{n,n-1,k} \rangle \tag{4.3}$$

It is perhaps obvious that these metrics may be combined to produce genetic interaction maps relative to environment-phenotype pairs.

Using these mathematical definitions of epistatic maps would allow us to perform comparative analysis on such data sets with relative ease. A metric to measure the distance between phenotype- or environment-specific epistatic maps would be both novel and incredibly useful. The obvious use for such a metric would be to relate phenotypes to one another simply through their genetic interaction profiles. We know that genes can be clustered into fucntionally-related modules based on their interaction profiles relative to the fitness wildtype, but it remains to be seen to what extent these modules persist across non-growth phenotypes. We can even expand this idea to the level of epistatic profiles and ask the question: do phenotype-specific genetic interaction maps cluster together to form modules of related phenotypes? One could even ask whether or not certain phenotypic interaction maps resemble certain environmental conditions. All of this eventually leads us back to the idea of defining a functional relationship between epistatic profiles, further expounding on one of the organization principles in genetic interaction networks (Figure 3.4).

4.3 Closing Remarks

In this dissertation I explored how genetic perturbations — a fundamental tool across the field biology and a fundamentally simple concept — could be used in novel ways to help the bridge a gap between some some aspects of biological study. I hope the reader comes away from this with a greater understanding of how the tools of

computational and systems biology can be used to address similar concerns in their particular field of study.

BIBLIOGRAPHY

[1] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Studying gene expression and function. In: Molecular biology of the cell. 2nd ed. New York: Garland Science; 2002. p. 525–545. EISBN: 978-0-8153-4072-0. Available from: `http://www.ncbi.nlm.nih.gov/books/NBK26818/`.

[2] Benner SA, Sismour AM. Synthetic biology. Nature Reviews Genetics. 2005 Jul;6(7):533–543.

[3] Andrianantoandro E, Basu S, Karig DK, Weiss R. Synthetic biology: new engineering rules for an emerging discipline. Molecular Systems Biology. 2006;2:2006.0028.

[4] Mahner M, Kary M. What exactly are genomes, genotypes and phenotypes? And what about phenomes? Journal of Theoretical Biology. 1997 May;186(1):55–63.

[5] Pigliucci M. Genotype–phenotype mapping and the end of the 'genes as blueprint' metaphor. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2010 Feb;365(1540):557–566.

[6] Alberch P. From genes to phenotype: dynamical systems and evolvability. Genetica. 1991;84(1):5–11.

[7] Landry CR, Rifkin SA. The genotype-phenotype maps of systems biology and quantitative genetics: distinct and complementary. In: Soyer OS, editor. Evolutionary systems biology. No. 751 in Advances in experimental medicine and biology. New York: Springer New York; 2012. p. 371–398. EISBN: 978-1-4614-3567-9.

[8] Darwin C. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. 1st ed. London: John Murray; 1859.

[9] Kimura M. Evolutionary rate at the molecular level. Nature. 1968 Feb;217(5129):624–626.

[10] King JL, Jukes TH. Non-Darwinian evolution. Science. 1969 May;164(3881):788–798.

[11] Arigoni F, Talabot F, Peitsch M, Edgerton MD, Meldrum E, Allet E, et al. A genome-based approach for the identification of essential bacterial genes. Nature Biotechnology. 1998 Sep;16(9):851–856.

[12] Lazebnik Y. Can a biologist fix a radio?—or, what I learned while studying apoptosis. Cancer Cell. 2002 Sep;2(3):179–182.

[13] Duina AA, Miller ME, Keeney JB. Budding yeast for budding geneticists: a primer on the Saccharomyces cerevisiae model system. Genetics. 2014 May;197(1):33–48.

[14] Rothstein RJ. One-step gene disruption in yeast. Methods in Enzymology. 1983;101:202–211.

[15] Forsburg SL. The art and design of genetic screens: yeast. Nature Reviews Genetics. 2001 Sep;2(9):659–668.

[16] Giaever G, Nislow C. The yeast deletion collection: a decade of functional genomics. Genetics. 2014 Jun;197(2):451–465.

[17] Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science. 1999 Aug;285(5429):901–906.

[18] Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, et al. Functional profiling of the Saccharomyces cerevisiae genome. Nature. 2002 Jul;418(6896):387–391.

[19] Mohr SE, Smith JA, Shamu CE, Neumüller RA, Perrimon N. RNAi screening comes of age: improved techniques and complementary approaches. Nature Reviews Molecular Cell Biology. 2014 Sep;15(9):591–600.

[20] Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, et al. SGD: Saccharomyces Genome Database. Nucleic Acids Research. 1998 Jan;26(1):73–79.

[21] Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Research. 2012 Jan;40(Database issue):D700–705.

[22] Botstein D, Chervitz SA, Cherry JM. Yeast as a model organism. Science. 1997 Aug;277(5330):1259–1260.

[23] Botstein D, Fink GR. Yeast: an experimental organism for 21st Century biology. Genetics. 2011 Nov;189(3):695–704.

[24] Kachroo AH, Laurent JM, Yellman CM, Meyer AG, Wilke CO, Marcotte EM. Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science. 2015 May;348(6237):921–925.

[25] McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM. Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proceedings of the National Academy of Sciences of the United States of America. 2010 Apr;107(14):6544–6549.

[26] Woods JO, Singh-Blom UM, Laurent JM, McGary KL, Marcotte EM. Prediction of gene-phenotype associations in humans, mice, and plants using phenologs. BMC bioinformatics. 2013;14:203.

[27] Greenspan RJ. The flexible genome. Nature Reviews Genetics. 2001 May;2(5):383–387.

[28] Moore JH. A global view of epistasis. Nature Genetics. 2005 Jan;37(1):13–14.

[29] Tong AHY, Evangelista M, Parsons AB, Xu H, Bader GD, Pagé N, et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science. 2001 Dec;294(5550):2364–2368.

[30] Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, et al. Global mapping of the yeast genetic interaction network. Science. 2004 Feb;303(5659):808–813.

[31] Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, et al. The genetic landscape of a cell. Science. 2010 Jan;327(5964):425–431.

[32] Wolf JB, Edmund D Brodie III, Wade MJ. Epistasis and the evolutionary process. Oxford: Oxford University Press; 2000.

[33] Poelwijk FJ, Tănase-Nicola S, Kiviet DJ, Tans SJ. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. Journal of Theoretical Biology. 2011 Mar;272(1):141–144.

[34] Weinreich DM, Watson RA, Chao L. Perspective: sign epistasis and genetic constraint on evolutionary trajectories. Evolution. 2005 Jun;59(6):1165–1174.

[35] Kondrashov AS. Deleterious mutations and the evolution of sexual reproduction. Nature. 1988 Dec;336(6198):435–440.

[36] Azevedo RBR, Lohaus R, Srinivasan S, Dang KK, Burch CL. Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. Nature. 2006 Mar;440(7080):87–90.

[37] Kondrashov FA, Kondrashov AS. Multidimensional epistasis and the disadvantage of sex. Proceedings of the National Academy of Sciences. 2001 Oct;98(21):12089–12092.

[38] MacCarthy T, Bergman A. Coevolution of robustness, epistasis, and recombination favors asexual reproduction. Proceedings of the National Academy of Sciences. 2007 Jul;104(31):12801–12806.

[39] Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molecular Genetics. 2002 Oct;11(20):2463–2468.

[40] Phillips PC. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics. 2008 Nov;9(11):855–867.

[41] Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, et al. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. Cell. 2005 Nov;123(3):507–519.

[42] Parsons AB, Brost RL, Ding H, Li Z, Zhang C, Sheikh B, et al. Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. Nature Biotechnology. 2004 Jan;22(1):62–69.

[43] Yeh PJ, Tschumi AI, Kishony R. Functional classification of drugs by properties of their pairwise interactions. Nature Genetics. 2006 Apr;38(4):489–494.

[44] Chait R, Craney A, Kishony R. Antibiotic interactions that select against resistance. Nature. 2007 Apr;446(7136):668–671.

[45] Penrod NM, Greene CS, Granizo-MacKenzie D, Moore JH. Artificial immune systems for epistasis analysis in human genetics. In: Pizzuti C, Ritchie MD, Giacobini M, editors. Evolutionary computation, machine learning and data mining in bioinformatics. No. 6023 in Lecture Notes in Computer Science. Berlin, Germany: Springer Berlin Heidelberg; 2010. p. 194–204. EISBN: 978-3-642-12211-8.

[46] Mitchison NA, Rose AM. Epistasis: the key to understanding immunological disease? European Journal of Immunology. 2011 Aug;41(8):2152–2154.

[47] Williams HS. A history of science. vol. 4. New York: Harper & Brothers; 1904.

[48] Buchner E. Alkoholische Gährung ohne Hefezellen. Berichte der deutschen chemischen Gesellschaft. 1897 Jan;30(1):117–124.

[49] Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. Nucleic Acids Research. 2002 Jan;30(1):47–49.

[50] Chang A, Schomburg I, Placzek S, Jeske L, Ulbrich M, Xiao M, et al. BRENDA in 2015: exciting developments in its 25th year of existence. Nucleic Acids Research. 2015 Jan;43(Database issue):D439–446.

[51] Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A. The EcoCyc and MetaCyc databases. Nucleic Acids Research. 2000 Jan;28(1):56–59.

[52] Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Research. 2015 Nov;.

[53] Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, et al. Reactome: a knowledge base of biologic pathways and processes. Genome Biology. 2007;8(3):R39.

[54] Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Research. 2014 Jan;42(Database issue):D472–477.

[55] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 2000 Jan;28(1):27–30.

[56] Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Research. 2014 Jan;42(Database issue):D199–205.

[57] Peregrin-Alvarez JM, Tsoka S, Ouzounis CA. The phylogenetic extent of metabolic enzymes and pathways. Genome Research. 2003 Mar;13(3):422–427.

[58] Peregrín-Alvarez JM, Sanford C, Parkinson J. The conservation and evolutionary modularity of metabolism. Genome Biology. 2009;10(6):R63.

[59] He X, Zhang J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics. 2005 Feb;169(2):1157–1164.

[60] Ohno S. Evolution by gene duplication. Berlin: Springer Berlin Heidelberg; 1970. EISBN: 978-3-642-86659-3.

[61] Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999 Apr;151(4):1531–1545.

[62] Edwards JS, Palsson BØ. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proceedings of the National Academy of Sciences. 2000 May;97(10):5528–5533.

[63] Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. Molecular Systems Biology. 2011;7:535.

[64] Smallbone K. Standardized network reconstruction of E. coli metabolism; 2013. Available from: `http://arxiv.org/abs/1304.2960`.

[65] Förster J, Famili I, Fu P, Palsson BØ, Nielsen J. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Research. 2003 Feb;13(2):244–253.

[66] Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nature Biotechnology. 2008 Oct;26(10):1155–1160.

[67] Aung HW, Henry SA, Walker LP. Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism. Industrial Biotechnology. 2013 Aug;9(4):215–228.

[68] Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nature Biotechnology. 2010 Mar;28(3):245–248.

[69] Edwards JS, Covert MW, Palsson BØ. Metabolic modelling of microbes: the flux-balance approach. Environmental Microbiology. 2002 Mar;4(3):133–140.

[70] Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. Current Opinion in Biotechnology. 2003 Oct;14(5):491–496.

[71] Segrè D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. Proceedings of the National Academy of Sciences. 2002 Nov;99(23):15112–15117.

[72] Varma A, Palsson BØ. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Applied and Environmental Microbiology. 1994 Oct;60(10):3724–3731.

[73] Edwards JS, Ibarra RU, Palsson BØ. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nature Biotechnology. 2001 Feb;19(2):125–130.

[74] Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, et al. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Molecular Systems Biology. 2007;3:121.

[75] Kuepfer L, Sauer U, Blank LM. Metabolic functions of duplicate genes in Saccharomyces cerevisiae. Genome Research. 2005 Oct;15(10):1421–1430.

[76] Snitkin ES, Dudley AM, Janse DM, Wong K, Church GM, Segrè D. Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions. Genome Biology. 2008;9(9):R140.

[77] Papp B, Pál C, Hurst LD. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. Nature. 2004 Jun;429(6992):661–664.

[78] Harrison R, Papp B, Pál C, Oliver SG, Delneri D. Plasticity of genetic interactions in metabolic networks of yeast. Proceedings of the National Academy of Sciences. 2007 Feb;104(7):2307–2312.

[79] Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. Briefings in Bioinformatics. 2009 Jul;10(4):435–449.

[80] Deutscher D, Meilijson I, Kupiec M, Ruppin E. Multiple knockout analysis of genetic robustness in the yeast metabolic network. Nature Genetics. 2006 Sep;38(9):993–998.

[81] Wang F, Yang W. Structural insight into translesion synthesis by DNA Pol II. Cell. 2009 Dec;139(7):1279–1289.

[82] Duarte NC, Herrgård MJ, Palsson BØ. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Research. 2004 Jul;14(7):1298–1309.

[83] Mo ML, Palsson BØ, Herrgård MJ. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Systems Biology. 2009;3:37.

[84] Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Research. 2012 Jan;40(Database issue):D54–56.

[85] Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLoS computational biology. 2010 Feb;6(2):e1000667.

[86] Davis C, Kota K, Baldhandapani V, Gong W, Abubucker S, Becker E, et al. mBLAST: Keeping up with the sequencing explosion for (meta)genome analysis. Journal of Data Mining in Genomics & Proteomics. 2015 Aug;4(3).

[87] Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature Biotechnology. 2010 Sep;28(9):977–982.

[88] Orr HA. Fitness and its role in evolutionary genetics. Nature Reviews Genetics. 2009 Aug;10(8):531–539.

[89] Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution. 2000 Dec;15(12):496–503.

[90] Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular Biology and Evolution. 2000 Jan;17(1):32–43.

[91] Hurst LD. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends in Genetics. 2002 Sep;18(9):486.

[92] Hirsh AE, Fraser HB, Wall DP. Adjusting for selection on synonymous sites in estimates of evolutionary distance. Molecular Biology and Evolution. 2005 Jan;22(1):174–177.

[93] Hirsh AE, Fraser HB. Protein dispensability and rate of evolution. Nature. 2001 Jun;411(6841):1046–1049.

[94] Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, et al. Systematic screen for human disease genes in yeast. Nature Genetics. 2002 Aug;31(4):400–404.

[95] Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, et al. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science. 2008 Apr;320(5874):362–365.

[96] Heavner BD, Smallbone K, Price ND, Walker LP. Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. Database. 2013;2013:bat059.

[97] Klitgord N, Segrè D. Environments that induce synthetic microbial ecosystems. PLOS Computational Biology. 2010;6(11):e1001002.

[98] Covert MW, Schilling CH, Palsson BØ. Regulation of gene expression in flux balance models of metabolism. Journal of Theoretical Biology. 2001 Nov;213(1):73–88.

[99] Xia Y, Franzosa EA, Gerstein MB. Integrated assessment of genomic correlates of protein evolutionary rate. PLOS Computational Biology. 2009 Jun;5(6):e1000413.

[100] Ihmels J, Levy R, Barkai N. Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. Nature Biotechnology. 2004 Jan;22(1):86–92.

[101] Kafri R, Bar-Even A, Pilpel Y. Transcription control reprogramming in genetic backup circuits. Nature Genetics. 2005 Mar;37(3):295–299.

[102] DeLuna A, Springer M, Kirschner MW, Kishony R. Need-based up-regulation of protein levels in response to deletion of their duplicate genes. PLOS Biology. 2010 Mar;8(3):e1000347.

[103] Ihmels J, Collins SR, Schuldiner M, Krogan NJ, Weissman JS. Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. Molecular Systems Biology. 2007;3:86.

[104] Sonderegger P, Christen P. Comparison of the evolution rates of cytosolic and mitochondrial aspartate aminotransferase. Nature. 1978 Sep;275(5676):157–159.

[105] Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature. 2004 Apr;428(6983):617–624.

[106] Jacobs C, Segrè D. Organization principles in genetic interaction networks. In: Soyer OS, editor. Evolutionary systems biology. No. 751 in Advances in Experimental Medicine and Biology. New York: Springer New York; 2012. p. 53–78. EISBN: 978-1-4614-3567-9.

[107] Szappanos B, Kovács K, Szamecz B, Honti F, Costanzo M, Baryshnikova A, et al. An integrated approach to characterize genetic interaction networks in yeast metabolism. Nature Genetics. 2011 Jul;43(7):656–662.

[108] Shlomi T, Herrgård MJ, Portnoy V, Naim E, Palsson BØ, Sharan R, et al. Systematic condition-dependent annotation of metabolic genes. Genome Research. 2007 Nov;17(11):1626–1633.

[109] Burgard AP, Nikolaev EV, Schilling CH, Maranas CD. Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Research. 2004 Feb;14(2):301–312.

[110] Segrè D, Deluna A, Church GM, Kishony R. Modular epistasis in yeast metabolism. Nature Genetics. 2005 Jan;37(1):77–83.

[111] Phillips PC. The language of gene interaction. Genetics. 1998 Jul;149(3):1167–1171.

[112] Gao H, Granka JM, Feldman MW. On the classification of epistatic interactions. Genetics. 2010 Mar;184(3):827–837.

[113] de Visser JAGM, Cooper TF, Elena SF. The causes of epistasis. Proceedings of the Royal Society of London B: Biological Sciences. 2011 Dec;278(1725):3617–3624.

[114] Combarros O, Cortina-Borja M, Smith AD, Lehmann DJ. Epistasis in sporadic Alzheimer's disease. Neurobiology of Aging. 2009 Sep;30(9):1333–1349.

[115] Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proceedings of the National Academy of Sciences. 2005 Feb;102(5):1572–1577.

[116] Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human Heredity. 2003;56(1-3):73–82.

[117] Carlborg Ö, Haley CS. Epistasis: too often neglected in complex trait studies? Nature Reviews Genetics. 2004 Aug;5(8):618–625.

[118] You L, Yin J. Dependence of epistasis on environment and mutation severity as revealed by in silico mutagenesis of phage t7. Genetics. 2002 Apr;160(4):1273–1281.

[119] Arnqvist G, Dowling DK, Eady P, Gay L, Tregenza T, Tuda M, et al. Genetic architecture of metabolic rate: environment specific epistasis between mitochondrial and nuclear genes in an insect. Evolution. 2010 Dec;64(12):3354–3363.

[120] Kishony R, Leibler S. Environmental stresses can alleviate the average deleterious effect of mutations. Journal of Biology. 2003;2(2):14.

[121] Snitkin ES, Segrè D. Epistatic interaction maps relative to multiple metabolic phenotypes. PLOS Genetics. 2011;7(2):e1001294.

[122] Hayden EJ, Ferrada E, Wagner A. Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. Nature. 2011 Jun;474(7349):92–95.

[123] Bateson W, Mendel G. Mendel's principles of heredity. 1st ed. Cambridge Library Collection. Cambridge: Cambridge University Press; 2009. EISBN: 978-0-511-69446-2.

[124] Mani R, St Onge RP, Hartman JL IV, Giaever G, Roth FP. Defining genetic interaction. Proceedings of the National Academy of Sciences. 2008 Mar;105(9):3461–3466.

[125] Trindade S, Sousa A, Xavier KB, Dionisio F, Ferreira MG, Gordo I. Positive epistasis drives the acquisition of multidrug resistance. PLOS Genetics. 2009 Jul;5(7):e1000578.

[126] Eronen VP, Lindén RO, Lindroos A, Kanerva M, Aittokallio T. Genome-wide scoring of positive and negative epistasis through decomposition of quantitative genetic interaction fitness matrices. PLOS ONE. 2010;5(7):e11611.

[127] Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. Negative epistasis between beneficial mutations in an evolving bacterial population. Science. 2011 Jun;332(6034):1193–1196.

[128] Imielinski M, Belta C. Exploiting the pathway structure of metabolism to reveal high-order epistasis. BMC Systems Biology. 2008;2:40.

[129] Imielinski M, Belta C. Deep epistasis in human metabolism. Chaos. 2010 Jun;20(2):026104.

[130] Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999 Dec;402(6761supp):C47–52.

[131] Carter GW, Rush CG, Uygun F, Sakhanenko NA, Galas DJ, Galitski T. A systems-biology approach to modular genetic complexity. Chaos. 2010 Jun;20(2):026102.

[132] Kelley R, Ideker T. Systematic interpretation of genetic interactions using protein networks. Nature Biotechnology. 2005 May;23(5):561–566.

[133] Michaut M, Baryshnikova A, Costanzo M, Myers CL, Andrews BJ, Boone C, et al. Protein complexes are central in the yeast genetic landscape. PLOS Computational Biology. 2011 Feb;7(2):e1001092.

[134] Costanzo M, Baryshnikova A, Myers CL, Andrews BJ, Boone C. Charting the genetic interaction map of a cell. Current Opinion in Biotechnology. 2011 Feb;22(1):66–74.

[135] Lehár J, Krueger AS, Avery W, Heilbut AM, Johansen LM, Price ER, et al. Synergistic drug combinations tend to improve therapeutically relevant selectivity. Nature Biotechnology. 2009 Jul;27(7):659–666.

[136] Guo J, Tian D, McKinney BA, Hartman JL IV. Recursive expectation-maximization clustering: a method for identifying buffering mechanisms composed of phenomic modules. Chaos. 2010 Jun;20(2):026103.

[137] Collins SR, Schuldiner M, Krogan NJ, Weissman JS. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. Genome Biology. 2006;7(7):R63.

[138] Bandyopadhyay S, Kelley R, Krogan NJ, Ideker T. Functional maps of protein complexes from quantitative genetic interaction data. PLOS Computational Biology. 2008 Apr;4(4):e1000065.

[139] Yeh PJ, Hegreness MJ, Aiden AP, Kishony R. Drug interactions and the evolution of antibiotic resistance. Nature Reviews Microbiology. 2009 Jun;7(6):460–466.

[140] Eleftherohorinou H, Wright V, Hoggart C, Hartikainen AL, Jarvelin MR, Balding D, et al. Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. PLOS ONE. 2009;4(11):e8068.

[141] Shao H, Burrage LC, Sinasac DS, Hill AE, Ernest SR, O'Brien W, et al. Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. Proceedings of the National Academy of Sciences. 2008 Dec;105(50):19910–19914.

[142] Koffas M, Roberge C, Lee K, Stephanopoulos G. Metabolic engineering. Annual Review of Biomedical Engineering. 1999;1:535–557.

[143] Khosla C, Keasling JD. Metabolic engineering for drug discovery and development. Nature Reviews Drug Discovery. 2003 Dec;2(12):1019–1025.

[144] Keasling JD. Manufacturing molecules through metabolic engineering. Science. 2010 Dec;330(6009):1355–1358.

[145] Dumas ME, Wilder SP, Bihoreau MT, Barton RH, Fearnside JF, Argoud K, et al. Direct quantitative trait locus mapping of mammalian metabolic phenotypes in diabetic and normoglycemic rat models. Nature Genetics. 2007 May;39(5):666–672.

[146] Xu C, Li Z, Xu S. Joint mapping of quantitative trait Loci for multiple binary characters. Genetics. 2005 Feb;169(2):1045–1059.

[147] Hunter KW, Crawford NPS. The future of mouse QTL mapping to diagnose disease in mice in the age of whole-genome association studies. Annual Review of Genetics. 2008;42:131–141.

[148] Young ND. QTL mapping and quantitative disease resistance in plants. Annual Review of Phytopathology. 1996;34:479–501.

[149] Lisec J, Meyer RC, Steinfath M, Redestig H, Becher M, Witucka-Wall H, et al. Identification of metabolic and biomass QTL in Arabidopsis thaliana in a parallel analysis of RIL and IL populations. The Plant Journal. 2008 Mar;53(6):960–972.

[150] Kliebenstein D. Advancing genetic theory and application by metabolic quantitative trait loci analysis. The Plant Cell. 2009 Jun;21(6):1637–1646.

[151] Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002 Apr;296(5568):752–755.

[152] de Visser JAGM, Hermisson J, Wagner GP, Ancel Meyers L, Bagheri-Chaichian H, Blanchard JL, et al. Perspective: evolution and detection of genetic robustness. Evolution. 2003 Sep;57(9):1959–1972.

[153] Li F, Long T, Lu Y, Ouyang Q, Tang C. The yeast cell-cycle network is robustly designed. Proceedings of the National Academy of Sciences. 2004 Apr;101(14):4781–4786.

[154] Stelling J, Sauer U, Szallasi Z, Doyle FJ III, Doyle J. Robustness of cellular functions. Cell. 2004 Sep;118(6):675–685.

[155] Wagner A. Distributed robustness versus redundancy as causes of mutational robustness. BioEssays. 2005 Feb;27(2):176–188.

[156] Klamt S, Gilles ED. Minimal cut sets in biochemical reaction networks. Bioinformatics. 2004 Jan;20(2):226–234.

[157] Kaufman A, Keinan A, Meilijson I, Kupiec M, Ruppin E. Quantitative analysis of genetic and neuronal multi-perturbation experiments. PLOS Computational Biology. 2005 Nov;1(6):e64.

[158] Deutscher D, Meilijson I, Schuster S, Ruppin E. Can single knockouts accurately single out gene functions? BMC Systems Biology. 2008;2:50.

[159] Lehár J, Krueger AS, Zimmermann GR, Borisy A. High-order combination effects and biological robustness. Molecular Systems Biology. 2008;4:215.

[160] Burch CL, Turner PE, Hanley KA. Patterns of epistasis in RNA viruses: a review of the evidence from vaccine design. Journal of Evolutionary Biology. 2003 Nov;16(6):1223–1235.

[161] Elena SF, Solé RV, Sardanyés J. Simple genomes, complex interactions: epistasis in RNA virus. Chaos. 2010 Jun;20(2):026106.

[162] Lehman N. The molecular underpinnings of genetic phenomena. Heredity. 2008 Jan;100(1):6–12.

[163] Weinreich DM, Delaney NF, Depristo MA, Hartl DL. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science. 2006 Apr;312(5770):111–114.

[164] Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. Nature. 2006 Dec;444(7121):929–932.

[165] Cooper TF, Remold SK, Lenski RE, Schneider D. Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in Escherichia coli. PLOS Genetics. 2008 Feb;4(2):e35.

[166] Draghi JA, Parsons TL, Plotkin JB. Epistasis increases the rate of conditionally neutral substitution in an adapting population. Genetics. 2011 Apr;187(4):1139–1152.

[167] Chou HH, Chiu HC, Delaney NF, Segrè D, Marx CJ. Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science. 2011 Jun;332(6034):1190–1192.

[168] Martin G, Elena SF, Lenormand T. Distributions of epistasis in microbes fit predictions from a fitness landscape model. Nature Genetics. 2007 Apr;39(4):555–560.

[169] Kvitek DJ, Sherlock G. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. PLOS Genetics. 2011 Apr;7(4):e1002056.

[170] Dekel E, Alon U. Optimality and evolutionary tuning of the expression level of a protein. Nature. 2005 Jul;436(7050):588–592.

[171] Chiu HC, Marx CJ, Segrè D. Epistasis from functional dependence of fitness on underlying traits. Proceedings of the Royal Society of London B: Biological Sciences. 2012 Oct;279(1745):4156–4164.

[172] Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Molecular Systems Biology. 2007;3:119.

[173] Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic Engineering. 2003 Oct;5(4):264–276.

[174] Folger O, Jerby L, Frezza C, Gottlieb E, Ruppin E, Shlomi T. Predicting selective drug targets in cancer through metabolic networks. Molecular Systems Biology. 2011;7:501.

[175] Frezza C, Zheng L, Folger O, Rajagopalan KN, MacKenzie ED, Jerby L, et al. Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. Nature. 2011 Sep;477(7363):225–228.

[176] Klamt S, Haus UU, Theis F. Hypergraphs and cellular networks. PLOS Computational Biology. 2009 May;5(5):e1000385.

[177] Beard DA, Babson E, Curtis E, Qian H. Thermodynamic constraints for biochemical networks. Journal of Theoretical Biology. 2004 Jun;228(3):327–333.

[178] Bordbar A, Feist AM, Usaite-Black R, Woodcock J, Palsson BØ, Famili I. A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. BMC Systems Biology. 2011;5:180.

[179] Wintermute EH, Silver PA. Emergent cooperation in microbial metabolism. Molecular Systems Biology. 2010 Sep;6:407.

[180] Klitgord N, Segrè D. Ecosystems biology of microbial metabolism. Current Opinion in Biotechnology. 2011 Aug;22(4):541–546.

CURRICULUM VITAE