Pleiotropy and Epistasis
in constraint-based models of microbial metabolism

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Inaugural dissertation

for the attainment of the title of doctor
in the Faculty of Mathematics and Natural Sciences
at the Heinrich Heine University Düsseldorf

Presented by
Deya Abdalla Ali Alzoubi
from Irbid, Jordan

Düsseldorf, 27. February 2019

from the Institute for Computer Science
at Heinrich Heine University Düsseldorf

Published by permission of the
Faculty of Mathematics and Natural Sciences at
Heinrich Heine University Düsseldorf

Supervisor: Prof. Dr. Martin J. Lercher
Co-supervisor: Prof. Dr. Oliver Ebenhöh

Date of the oral examination: 29.04.2019

Declaration

I declare under oath that I have compiled my dissertation independently and without any undue assistance by third parties under consideration of the ‘Principles for the Safeguarding of Good Scientific Practice at Heinrich Heine University Düsseldorf. I have not submitted the dissertation to any other institution in this or similar form. So far I have not made any unsuccessful or successful attempts to obtain a doctorate.

D sseldorf, 27.02.2019

Deya Abdalla Ali Alzoubi

To my loving parents
To my loving wife Ala and my loving daughters Rahaf and Tala
To all my loving brothers and sisters
For their endless support, encouragement and love

Contents

Abbreviations

ATP	Adenosine triphosphate
ccFBA	cost-constrained flux balance analysis
FBA	Flux Balance analysis
FVA	Flux variability analysis
GP	Genotype-phenotype
GPR	Gene to protein (enzyme) to reaction relationship
GSM	Genome-scale metabolic model
LP	Linear programming
MILP	Mixed integer linear programming
MOMA	Minimization of metabolic adjustment
MOMENT	Metabolic modeling with enzyme kinetics
MTF	Minimization of total flux
NADPH	Nicotinamide adenine dinucleotide phosphate
pFBA	Parsimonious flux balance analysis
ROOM	Regulatory on/off minimization of metabolic flux
SBML	Systems biology markup language

List of Figures

Preface

This doctoral thesis presents three manuscripts, along with additional chapters that relate them to the literature as well as to each other: Chapter 1 summarizes the background of my contributions, while Chapter 2 discusses the results. I generated and analyzed the results of all three manuscripts; detailed contributions are listed on the page preceding each manuscript. Manuscript 1 analyzed how the severity of a mutation affects pleiotropy in genome-scale metabolic networks. This work was published as: Alzoubi D., Desouki A. A., Lercher M. J., *Scientific Reports* 8: 17252, 22. November 2018, https://doi.org/10.1038/s41598-018-35092-1. Manuscript 2 presents the predictions of epistasis from flux balance analysis with molecular crowding; it is currently under review at *Scientific Reports* (Alzoubi D., Desouki A. A., Lercher M. J., Epistasis predictions from flux balance analysis with molecular crowding). Finally, Manuscript 3, which has not yet been submitted to a journal, presents the inability of constraint-based methods to make quantitative predictions for non-lethal metabolic gene knockouts in *E. coli* and *Saccharomyces cerevisiae*, a finding with major implications for the interpretation of Manuscripts 1 and 2 (Alzoubi D., Desouki A. A., Papp B., Lercher M. J., Flux balance analysis and other constraint-based methods fail to predict mutant fitness for non-lethal metabolic gene knockouts in *Escherichia coli* and *Saccharomyces cerevisiae*).

Summary

Understanding the relationships between genotypes and phenotypes remains a major challenge for biological research. Uncovering these relationships is hampered by the interconnectedness of biological systems, leading to non-independence of genes and of phenotypes. The most prominent emergent systems-level effects are summarized under the terms pleiotropy (one allele affecting multiple phenotypes) and epistasis (effects of one allele depend on the alleles of other genes). Metabolism is an ideal system to study pleiotropy and epistasis, as metabolic reactions can be studied in isolation. Constraint-based methods, in particular Flux Balance Analysis (FBA) represent the current state-of-the-art in genome-scale metabolic modelling. FBA has been successfully used to predict phenotypes such as growth rate, nutrient uptake rates, and gene essentiality(Edwards, Ibarra and Palsson 2001, Edwards and Palsson 2000, Famili et al. 2003, Forster et al. 2003, Ibarra, Edwards and Palsson 2002).

In Manuscript 1, we used constraint-based simulations of the metabolic models for the bacterium *Escherichia coli* and the Baker's yeast *Sacchormyces cerevisiae* to predict the pleiotropy of metabolic genes, allowing for mutations of variable severity. This work also represents the first analysis of how pleiotropy is associated with the generation of currency metabolites such as ATP and NADPH. We found that the knockout of a majority of genes that contribute to fitness has pleiotropic effects. For most of these genes, pleiotropy increases strongly with increasingly debilitating effects of mutations; in many cases, this was associated with increasing effects on currency metabolite production.

While standard FBA ignores the concentrations of enzymes catalyzing metabolic reactions, FBA with molecular crowding (ccFBA) accounts for the need to solve them in cellular volumes of limited capacity. In Manuscript 2, we tested if ccFBA can significantly improve the prediction of epistasis in yeast. The results indeed show that FBA with molecular crowding can predict some positive epistatic interactions not detectable with other constraint-based methods. However, the most important conclusion was that at least 70% of experimentally observed epistatic interactions are not detectable by any of the popular constraint-based methods. This result hinted at some fundamental problems of these methods.

These problems were addressed in Manuscript 3, where we found that all tested constraint-based methods are essentially useless when predicting the fitness effects of non-essential gene knockouts. If these models cannot quantify single gene knockout fitness reliably, it is no surprise that they fail to predict higher order effects (genetic interactions, *i.e.*, epistasis). More generally, these results show that one has to be careful when interpreting computational predictions of gene knockouts, such as done in our analysis of pleiotropy in Manuscript 1.

References

Edwards, J. S., R. U. Ibarra and B. O. Palsson (2001). In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol **19**(2): 125-130.

Edwards, J. S. and B. O. Palsson (2000). Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions. BMC Bioinformatics **1**: 1.

Famili, I., J. Förster, J. Nielsen and B. O. Palsson (2003). Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. Proceedings of the National Academy of Sciences **100**(23): 13134.

Forster, J., I. Famili, B. O. Palsson and J. Nielsen (2003). Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae. Omics **7**(2): 193-202.

Ibarra, R. U., J. S. Edwards and B. O. Palsson (2002). Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature **420**(6912): 186-189.

Chapter 1 - Background

Introduction: gene function and knockouts

“Ultimately, one wishes to determine how genes—and the proteins they encode—function in the intact organism. Although it may sound counterintuitive, one of the most direct ways to find out what a gene does is to see what happens to the organism when that gene is missing … Because mutations can interrupt cellular processes, mutants often hold the key to understanding gene function”(Alberts 2002).

A mutation can be beneficial, harmful, or neutral to its host, depending on the location and nature of the mutation and on the larger genomic and ecological context. A mutation in a gene introduces a new genotype. A genotype may be neutral and have no observable effects; but often it causes changes in the organism’s observable properties, and a new phenotype arises. The impacts of different mutations are crucial to our understanding of a biological system, and therefore researchers have devoted substantial efforts to the study of mutations. These efforts have led to the estimation of the distributions of mutational effects, where mostly the property of interest is fitness (Loewe 2008).

A major challenge of biological studies is to understand the relationship between genotypes and phenotypes, as this provides the key for identifying genetic variants responsible for organismal effects of interest. A common and conventional approach to connect genotypes to phenotypes is to perform *in vivo* gene deletion experiments and observe the subsequent phenotype changes. However, many phenotypic traits are complex traits affected by many genes. For example, Mendelian disorders, which are caused by mutations to single genes, account for only a small fraction of rare human diseases, while most common diseases appear to have more complex genetic causes that remain largely unknown (see OMIM; http://www.ncbi.nlm.nih.gov/omim). Prominent examples include Alzheimer’s disease or type 2 diabetes in humans (Plomin, Haworth and Davis 2009, Mackay, Stone and Ayroles 2009). In other cases, the disease-related genes affect multiple traits, and hence a single genotype is mapped to multiple unrelated phenotypes, an example of pleiotropy (Stearns 2011, Wagner and Zhang 2011). These pleiotropic genes mostly do not work alone, but instead they cooperate with other genes to control different target phenotypes; their effect on trait variation often depends on their interactions with other genes, giving rise to epistatic effects (Phillips 2008, de Visser, Cooper and Elena 2011).

Studies of the genotype-phenotype (GP) mapping have broad implications for our understanding of evolutionary biology, functional genomics, and disease (Segrè and Marx

2010). Figure 1 illustrates the concepts of pleiotropy and epistasis on the basis of the GP mapping. How pleiotropic and epistatic gene effects are organized, *i.e.*, the network of connections between genotypes and phenotypes, plays a major role in the capability of organisms to adapt and evolve (Wagner, Pavlicev and Cheverud 2007, Wagner and Altenberg 1996, Hansen 2006, Armbruster et al. 2014). Large-scale genetic interaction screens in yeast and other model systems have revealed common properties of genetic interaction networks, whose features appear to be maintained over extensive evolutionary distances (Botstein and Fink 2011).

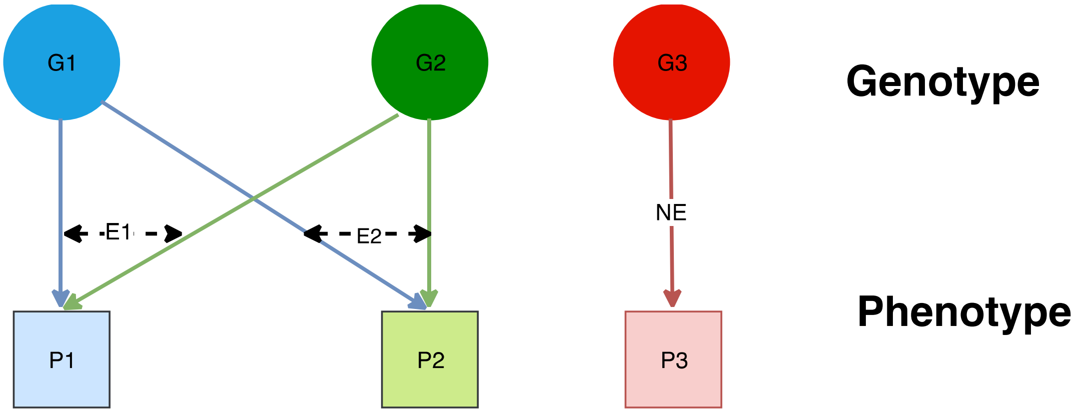


Figure 1: GP-map. P1, P2, and P3 are three phenotypes encoded by genes G1, G2, and G3. G3 encoding P3 has no pleiotropic effects and no epistatic interactions. G1 and G2 both affect P1 and P2, and thus both genes have pleiotropic effects. Conversely, both P1 and P2 are affected by G1 and G2, potentially giving rise to epistasis between G1 and G2 (dashed arrows E1, E2).

Systems biology models summarize our current understanding of metabolism

The aim of Systems Biology is to study “the structure and dynamics of cellular and organismal function”(Kitano 2002), typically utilizing different data types obtained from high-throughput measurements of cellular processes (Ideker, Galitski and Hood 2001, Kitano 2002). System biologists utilize mathematical modeling methods to analyze biological interactions represented by different types of networks, such as metabolic pathways, transcriptional regulation networks, or signal transduction networks, in order to understand the behavior of the cell as a whole. System models are constructed using powerful computational tools in order to interpret specific mechanisms and cellular phenotypes from a systems or network perspective (Ashburner et al. 2000, Cherry et al. 1998, Gasch et al. 2000, Harbison et al. 2004, Stark et al. 2006, Teixeira et al. 2006).

Metabolism is arguably the subcellular system most suited for systems level GP analyses. This is because metabolic features such as energy generation and amino acid synthesis are closely related to observable phenotypic traits (Nielsen 2017). For many years, cellular metabolism has been studied by biochemists, and a comprehensive collection of metabolic reactions has been characterized. Molecular biology techniques helped to uncover gene-to-enzyme-to-reaction (GPR) associations. Systems biology models of metabolism aim to summarize the current state of knowledge into a coherent, systems-level framework. While metabolism can be considered an extraordinarily well-characterized subcellular system, in many cases, phenotypes are still not predictable from genotypes, as changes in gene expression resulting from genetic changes are not well understood (Pavey et al. 2010). Accurate prediction of cellular phenotypes using genome-scale metabolic models is commonly confined to the prediction of gene essentiality (O'Brien, Monk and Palsson 2015).

Metabolic Network Models

A metabolic network connects metabolites and reactions. A node of the metabolic network represents a particular metabolite (chemical compound), and a link (or edge) between nodes represents a reaction that converts one set of metabolites into another. A reaction is catalysed by one or several enzymes, and thus can be linked to one or several genes encoding those enzymes. Through simulations, one can find out the flux of metabolites through the network; comparison of model predictions to experimental data often shows good agreement (Orth, Thiele and Palsson 2010).

Genome-scale metabolic models (GSMs) are useful tools to analyze the metabolism of an organism, which represents the complete set of reactions a cell can perform based on the enzymes and transporters encoded in the genome. GSMs are constructed based on the sequencing and annotation of an organism's genome (Thiele and Palsson 2010). To date, more than 100 manually curated GSMs for different organisms and strains, including humans, have been published (Aurich et al. 2015). Thiele *et al.* (Thiele and Palsson 2010) summarized the current paradigm in preparing GSM reconstructions, which involves five parts: (i) draft network reconstruction, (ii) refinement, (iii) conversion to model, (iv) evaluation, and (v) assembly. One example for models constructed in this way is the iJO1366 genome-scale metabolic model for *E. coli* (Orth et al. 2011), which contains 2583 reactions and 1805 metabolites and accounts for the functions of 1366 genes.

Constraint-based analysis of metabolic models

In recent years, many approaches have been introduced to predict gene function systemically. Constraint based methods (Edwards, Covert and Palsson 2002, Orth et al. 2010, Price et al. 2003, Lewis, Nagarajan and Palsson 2012) applied to GSMs are among the most important *in silico* methods that have been used to evaluate the effects of gene deletions under various environmental conditions (*e.g.,* (Harrison et al. 2007)). Constraint-based models formulate known (typically linear) constraints on a biological system, such as reaction stoichiometries and the directions of effectively irreversible reactions. They then typically try to find a cellular state (a distribution of reaction fluxes) that optimizes some objective, such as maximizing the biomass production rate of a microbial organism. Mathematical details of different constraint-based methods are given below.

Flux Balance Analysis (FBA)

Flux Balance analysis (FBA) (Orth et al. 2010) is a typical constraint-based modelling technique, which aims to predict the flux distribution of genome scale metabolic networks (Edwards et al. 2002). FBA was developed to predict the metabolic state of a strain optimized by natural selection or bioengineering for a particular function. Linear optimization is used to identify one or more optimal flux distribution. Metabolite balancing and stoichiometry of reactions in the system are used to construct constraints for the optimization under the steady-state assumption that all internal metabolites must be consumed at the same rate at which they are produced. Reaction directions (which represent a coarse-grained consideration of thermodynamics), maximal reaction capacities (if known), and the availability of nutrients in different environments impose further constraints on the linear system (Orth et al. 2010).

FBA applies these constraints in an optimization problem. The steady state assumption can be summarized as Sv=0. Here, v is a vector of fluxes, and S is the stoichiometric matrix, with each row representing a metabolite, each column representing a reaction, while each matrix entry is a stoichiometric coefficient s_{ij} , specifying the number of molecules of metabolite i produced (s_{ij} >0) or consumed (s_{ij} <0) in reaction j in a single reaction step (Orth et al. 2010). To find a solution, FBA solves the following linear programming (LP) problem:

$$\text{Maximize } \boldsymbol{c}^{\mathrm{T}}\,\boldsymbol{v}$$

$$\text{s.t} \qquad S\boldsymbol{v}=\mathbf{0}$$

$$\boldsymbol{v_{min}} \leq \boldsymbol{v} \leq \boldsymbol{v_{max}}$$

Here, $\boldsymbol{c}$ is a vector of constant weights; $\boldsymbol{v}$ is a vector of fluxes; S is the stoichiometric matrix; $\boldsymbol{v_{min}}$ and $\boldsymbol{v_{max}}$ are vectors of fixed lower and upper bounds, respectively, for every reaction, with the inequalities to be read component-wise. The solution of the LP is a flux distribution maximizing the objective function $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{v}$ under the assumption of steady state. As objective function, one frequently chooses an artificial reaction that simulates the accumulation of biomass; this is done under the assumption that natural selection acted to maximize the rate of biomass production of the metabolic system. The solution space is typically of high dimensions, *i.e.*, there are infinitely many solutions for *v*.

FBA can simulate the metabolic model under different environmental conditions. A variety of applications for such models were introduced, such as the the prediction of gene knockout effects, the identification of drug targets, the study of the evolution of metabolic systems, as well as the improved annotation of genomes (Raman and Chandra 2009).

FBA is one of the most important tools for analysis of the capabilities of a metabolic network (Varma and Palsson 1994, Teusink et al. 2009, Terzer et al. 2009, Schuster, Pfeiffer and Fell 2008, Price et al. 2003, Orth et al. 2010, Mahadevan and Schilling 2003, Durot, Bourguignon and Schachter 2009). To understand the complicated characteristics of metabolism in living cells, the repertoire of constraint-based analysis methods has been expanded continuously. One example of a variant of FBA is flux variability analysis (FVA), which evaluates the minimum and maximum flux for each reaction across the multiple optima of the FBA problem (or, if desired, across all flux distributions compatible with the underlying set of constraints) (Mahadevan and Schilling 2003). There are also many extension of FBA to predict flux states.

Applied to gene knockouts, FBA assumes optimality of growth, which might not accurately represent the behavior of real biological systems; hence, two widely recognized extensions of FBA were proposed for this case, namely MOMA (Segre, Vitkup and Church 2002) and ROOM (Shlomi, Berkman and Ruppin 2005). Both methods calculate the flux state of a knockout by minimizing a metric that estimates the distance between the wildtype and knockout flux distributions, assuming that gene regulation remains largely unchanged after the knockout.

The distance metric used by MOMA is the Euclidean distance of the two flux vectors. This can be formally expressed through the following quadratic programming problem:

$$\begin{gathered} \min \quad (\boldsymbol{w}-\boldsymbol{v})^{\mathbf{2}} \\ \text{s.t.} \quad S\mathrm{v}=\mathbf{0} \\ \boldsymbol{v_{min}} \leq \boldsymbol{v} \leq \boldsymbol{v_{max}} \end{gathered}$$

$$v_j = 0, j \in G$$

Here, $\boldsymbol{w}$ is the wildtype optimal flux vector obtained from standard FBA; $\boldsymbol{v}$ is a vector in the mutant flux space; and $\boldsymbol{G}$ is the set of reactions that are deactivated by the gene deletion (knockout constraints).

ROOM instead minimizes the total number of significant flux changes between the wildtype flux distribution (again estimated by FBA) and the mutant flux distribution. To solve this problem, ROOM is implemented using Mixed Integer Linear Programming (MILP):

$$Min \sum_{i=1}^{m} y_i$$

$$\text{s.t} \qquad S\boldsymbol{v}=\boldsymbol{0}$$

$$\boldsymbol{v_{min}} \leq \boldsymbol{v} \leq \boldsymbol{v_{max}}$$

$$v_j = 0, j \in \boldsymbol{G}$$

$$y_i \in \{0,1\}$$

$$v_i - y_i\,(v_i^{ub} - w_i^u) \leq w_i^u$$

$$v_i - y_i\left(v_i^{lb} - w_i^l\right) \geq w_i^l$$

$$w_i^u = w_i - \delta|w_i| + \varepsilon$$

$$w_i^l = w_i - \delta|w_i| - \varepsilon$$

where $\boldsymbol{y_i}$ is a Boolean auxiliary variable, which reflects the significant flux change between wildtype and mutant. $\boldsymbol{w_i^u}$ and $\boldsymbol{w_i^l}$ are used as a thresholds to distinguish significant from non-significant flux changes. The tolerance parameters δ and ε are specifying absolute and relative ranges in tolerance.

cost-constrained Flux Balance Analysis (ccFBA)

FBA ignores important biological constraints. In particular, these include molecular crowding, which refers to a cellular constraint on total macromolecular concentrations due to the limited solvent capacity of the cytosol (Beg et al. 2007). FBA does not require any kinetic information about the reactions, but the price to be paid for this simplicity is FBA's inability to model a range of metabolic phenomena such as overflow metabolism (Basan et al. 2015) or the evolution of cross-feeding in originally monoclonal bacterial populations grown on abundant glucose. Constraints from enzyme kinetics and cellular volume must be imposed in order to explain these phenomena (Pfeiffer and Bonhoeffer 2004).

cost-constrained FBA (ccFBA, available on CRAN) (Desouki 2016) is an improved general implementation of MetabOlic Modeling with ENzyme kineTics (MOMENT) (Adadi et al. 2012) that includes parameterizations for *E. coli* and *S. cerevisiae*. ccFBA uses enzyme molecular weights to constrain total cellular enzyme concentration, and enzyme kinetic data to constrain the fluxes catalyzed by these enzymes. ccFBA improves the original implementation of MOMENT by explicitly considering multifunctional enzymes.

ccFBA converts Boolean gene to protein (enzyme) to reaction mapping (GPR) rules into constraints as follows:

- For a reaction *j* catalyzed by single enzyme *i*, this equation is used:

$$\boldsymbol{v_i \leq k_{cat,j} * \ g_i}$$

- For a reaction *j* catalyzed by two isozymes *a* OR *b*, this equation is used:

$$\boldsymbol{v_i \leq k_{cat,j} * \ (g_a + \ g_b)}$$

- For a reaction *j* catalyzed by an enzyme complex consisting of gene products *a* AND *b*, this equation is used:

$$\boldsymbol{v_i \leq k_{cat,j} * \ min(g_a \, , g_b)}$$

where v_i is the metabolic flux of this reaction and $\boldsymbol{k_{cat,j}}$ is the corresponding (apparent) turnover number. g_a and g_b are the molar concentrations of protein copies of *a* and *b* respectively. In addition, ccFBA formulates a global constraint on the volume available for enzymes:

- Cost constraint (crowding) using molecular weights:

$$\boldsymbol{\sum g_i * \ MW_i \leq C * DW}$$

where g_i denotes molar enzyme concentration; MW_i denotes the (molar) molecular weight of the protein encoded by gene *i*; *C* denotes the fraction of dry weight accounted for by metabolic proteins, assumed to be known and constant (and assumed to be proportional to the volume available for enzymes); and *DW* is the cellular dry weight per volume of cytosol.

Minimization of Total Flux (MTF)

The solutions in FBA are not unique and can contain thermodynamically infeasible cycles (Price et al. 2003, Orth et al. 2010, De Martino et al. 2013). MTF was developed to solve these problems by finding a solution with minimum total flux among the alternative optima, while maintaining an optimal value of the objective function (Holzhutter 2004). This strategy is often applied unter the name “parsimonious FBA” (pFBA).

Chapter 2 – Results and Discussion

Pleiotropy and its importance in biology and medicine

One of the fundamental phenomena in studying gene mutations is pleiotropy, where a single gene is responsible for more than one phenotypic traits (Wagner and Zhang 2011, Stearns 2011). A mutation can result in different types of changes. It can cause a change to transcript regulation, a change in the amino acid sequence that affects its binding to a substrate, or – in the most extreme case – it can result in a gene deletion that abolishes expression completely (Alberts 2002). If a gene product participates in multiple cellular processes, its perturbation or deletion can affect some or all of its functions; the resulting effects can cascade down to all phenotypic traits affected by the gene.

Experiments (Johnsson et al. 2012, Wright 2015, Gratten and Visscher 2016, Burstin et al. 2007) performed on plants and animals in the past decade have shed important new light on pleiotropy. In particular, when selection is applied to one trait, the mean of other, pleiotropically related, traits also changes in the subsequent generations. If natural selection, sexual selection, or artificial selection on one trait favors one specific version of a gene, then pleiotropic effects may limit the rate of evolution. This is because while a mutation can have a positive effect on one trait, it can cause negative effects on other traits, thus counteracting the selected effects with negative fitness effects. This phenomenon is closely related to the Hill-Robertson effect of genetically linked genes (Hill and Robertson 1966), an effect that could be considered as due to pleiotropy of a haploblock rather than of a gene.

The presence of pleiotropy has important implications on genomic medicine, *i.e.*, the use of personalized medicine and genome editing. Because of the associated risks, we must thoroughly understand the implication of different effects of mutants of a gene, since specific genetic variants may show strong associations with multiple traits but in opposite directions (Parkes et al. 2013). This is important when we try to identify molecular targets for drug development (Sivakumaran et al. 2011), and when we try to "fix" mutations using genome editing approaches such as the CRISPR-Cas system (Gratten and Visscher 2016).

Measuring the extent of pleiotropy for a single gene requires substantial experimental effort. Thus, systematic, genome-wide screens for pleiotropy can best be performed *in silico.* Computer simulations of pleiotropy have previously been based on modified versions of FBA that assess the contribution of a reaction to different constituents of biomass (Szappanos et al. 2011). In other words, pleiotropy was measured as the number of biomass components whose maximal production is reduced by a given mutation, as shown in Figure 2.

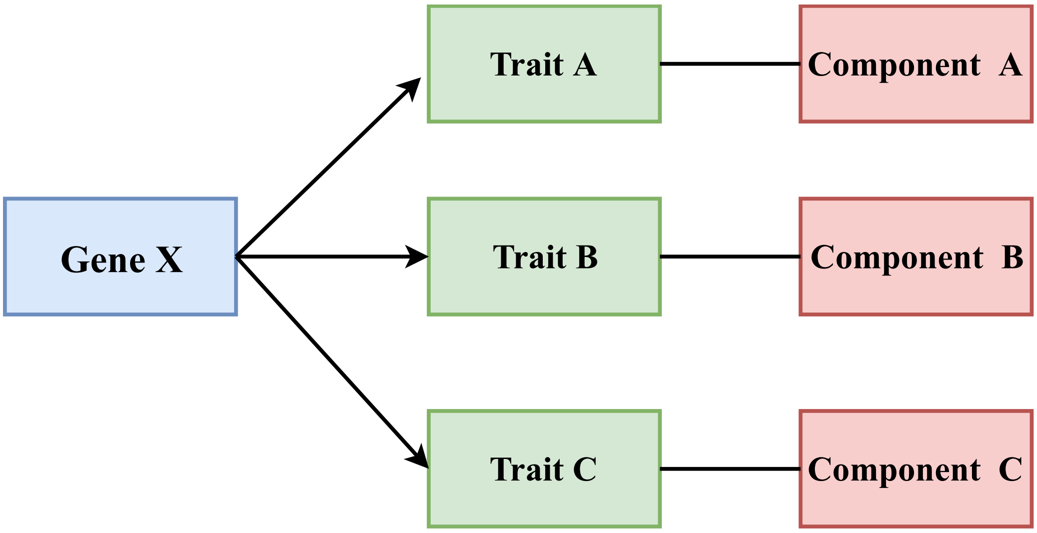


Figure 2. A graphical representation of pleiotropy. For three biomass components (A,B,C), maximal production is reduced by the mutation of gene X; thus, the degree of pleiotropy of gene X is 3.

Contribution 1: Alleles of a gene differ in pleiotropy, often mediated through currency metabolite production, in *E. coli* and yeast metabolic simulations

We used metabolic reconstructions of *Escherichia coli* (iJO1366 (Orth et al. 2011)) and *Saccharomyces cerevisiae* (Yeast v. 7.6, (https://sourceforge.net/projects/yeast (Aung, Henry and Walker 2013)) to study how the severity of a mutation affects pleiotropy in genome-scale metabolic networks. The production of so-called "currency metabolites", such as adenosine triphosphate (ATP) and nicotinamide adenine dinucleotide phosphate (NADPH), is essential to the functioning of many cellular processes, and hence it appears *a priori* likely that mutations affecting currency metabolite production will be highly pleiotropic. Thus, we specifically asked how pleiotropy is associated with the generation of currency metabolites in the studied networks.

Following earlier work, we measured pleiotropy as the number of biomass components whose maximal production is reduced by a given mutation (Szappanos et al. 2011). Our results indicate that many essential genes affect the production of multiple biomass components: in *E. coli*, essential genes affect on average 20% of biomass components, while the corresponding number for *S. cerevisiae* is 34%. We also found that pleiotropy strongly depends on the severity of the mutation. Pleiotropy typically increases with increasing mutation severity: often, small reductions in flux affect the production of only a small number of biomass components, while large flux reductions (or full knockouts) may affect a much larger number of biomass components. Our model also allows us to quantify the relative contributions of pleiotropy type I, which arises due to multiple molecular functions of a gene product and is responsible for about 40% of metabolic pleitoropy, and type II, which is caused by multiple physiological consequences of a single molecular function. We further find that metabolic pleiotropy networks are less modular than other pleiotropy networks. This suggests that the underlying metabolic network is more highly interconnected than the genetic networks underlying other types of traits, likely because of crosslinks provided by currency metabolites. As hypothesized, pleiotropy is indeed frequently mediated by currency metabolites: if we make currency metabolites (such as ATP or NADPH) freely available, 55.3% of essential genes in *E. coli* and 87.4% of essential genes in *S. cerevisiae* show reduced pleiotropy.

The importance of looking at flux reductions rather than full knockouts

Mutations of different severity to the same gene may reduce the metabolic fluxes of the associated reactions to different degrees, but how the severity of a mutation affects pleiotropy in genome-scale metabolic networks was previously largely unknown. Xu *et al.* (Xu, Barker and Gu 2012) used FBA to investigate the epistatic landscape of different mutant alleles in the same gene, and I here transferred that approach to the study of pleiotropy.

Pleiotropy plays an important role in many genetic diseases. Human disease-associated genes are typically not fully non-functional – at least not homozygously – and thus studying pleiotropy for different degrees of protein activity reduction is very important. Similarly, pleiotropy is important for genomic evolution (Stearns 2011); in long-term evolutionary processes, small effect mutations may be very abundant and may play a critical role in long-term evolutionary processes, and thus it is equally important to analyze the pleiotropy of small-effect mutations in this context (Rutter, Shaw and Fenster 2010).

The connection of currency metabolites to pleiotropy

We can divide internal metabolites into "currency" metabolites, defined as those that are involved in many reactions, typically to provide energy or redox equivalents, and "primary" metabolites (Fritzemeier et al. 2017). A mutated gene can directly affect a biomass component's production if that gene catalyzes a reaction in its pathway of production. Instead, the mutated gene may affect the production of currency metabolites utilized in the component's production rather than the primary metabolites. We found that the pleiotropy of many genes is indeed mediated through the generation of currency metabolites such as ATP, NADPH, or $FADH_2$. This is true for more than half of the pleiotropic genes in *E. coli*, and for 87% of pleiotropic genes in yeast. Two main factors appear to provide more potential for pleiotropic effects in yeast: the higher interconnectedness of the yeast pleiotropic network as well as the larger active metabolic network size. A generation of pleiotropy through effects on currency metabolite production was expected *a priori;* however, we were the first to systematically test this hypothesis explicitly, and the first to quantify its contribution to genomic pleiotropy.

Differences between *E. coli* & yeast

For our analyses, we chose the two best studied model organisms among eukaryotes and prokaryotes, respectively, the baker's yeast *Saccharomyces cerevisiae* and *E. coli*. Our analysis of pleiotropy reveals that yeast genes have generally higher pleiotropy than those of *E. coli*. As we considered only the active metabolic networks (755 reactions and 184 metabolites for yeast, 462 reactions and 103 metabolites for *E. coli*), it appears that yeast metabolism is more complex and yet less flexible. The average number of reactions per metabolite is for yeast 4.10, slightly lower than the corresponding number for *E. coli,* 4.49. This indicates that reactions often connect non-distant parts of the network.

Epistasis and its importance in biology and medicine

Epistasis describes the situation that the consequences of a mutation in one gene depend on mutations at another gene (Harrison et al. 2007, He et al. 2010, Szappanos et al. 2011, Xu, Barker and Gu 2012). In complex genetic systems, understanding epistasis can help us to understand evolutionary dynamics, as epistasis plays a major role in shaping the fitness landscape (Hayden, Ferrada and Wagner 2011, Weinreich et al. 2006) and in maintaining sexual reproduction (Kondrashov 1988, Otto 2009), and as epistasis affects the speed of adaptation (Sanjuan et al. 2005, Kryazhimskiy, Tkacik and Plotkin 2009, Khan et al. 2011, Chiu, Marx and Segrè 2012).

More generally, the study of epistasis is relevant to different branches of biology and medicine (Churchill 2001). For instance, for diverse traits of medical importance evidence for epistasis has been reported, which includes cancer, hypertension, kidney disease, epilepsy, and alcoholism (Churchill 2001). Further, there are many genetic modifiers of disease phenotypes that alter the severity of a trait depending on the genetic background in which they occur. Geneticists have used epistasis to analyze functional relationships between genes, the genetic ordering of regulatory pathways, and quantitative differences of allele-specific effects (Phillips 2008). For example, epistasis has been shown to be of critical importance in mouse models of epilepsy (Browman and Crabbe 2001).

Mathematically, epistasis is defined based on the fitness of single and double mutants of two genes *X* and *Y* (Elena and Lenski 1997, Phillips 2000):

$$\varepsilon = W_{XY} - W_X W_Y$$

where W_X and W_Y represent the fitness values of single mutants and W_{XY} represents the fitness value of the corresponding double mutant. Based on the genetic interaction score ε, epistatic interactions can be broadly classified into three major classes, which we refer to as (1) negative epistasis, $\varepsilon < 0$; (2) positive epistasis, $\varepsilon > 0$; and (3) no epistasis, $\varepsilon \approx 0$.

An epistatic interaction is negative when the double mutant is less fit than the expectation based on the two single mutants (Figure 3). An extreme example of a negative genetic interaction is synthetic lethality, in which the combination of two mutations, each of which causes at most a slowing down of growth on its own, results in a fatal phenotype(Tong et al. 2001, Novick and Botstein 1985). A positive epistatic interaction is found when a double mutant's fitness is less than the expectation based on the two independent single mutants (Figure 2). One example is genetic suppression, which occurs when a mutation in one gene rescues the fitness defect associated with a mutation in another gene, such that the double mutant's fitness is greater than that of the worst single mutant(Baryshnikova et al. 2010).

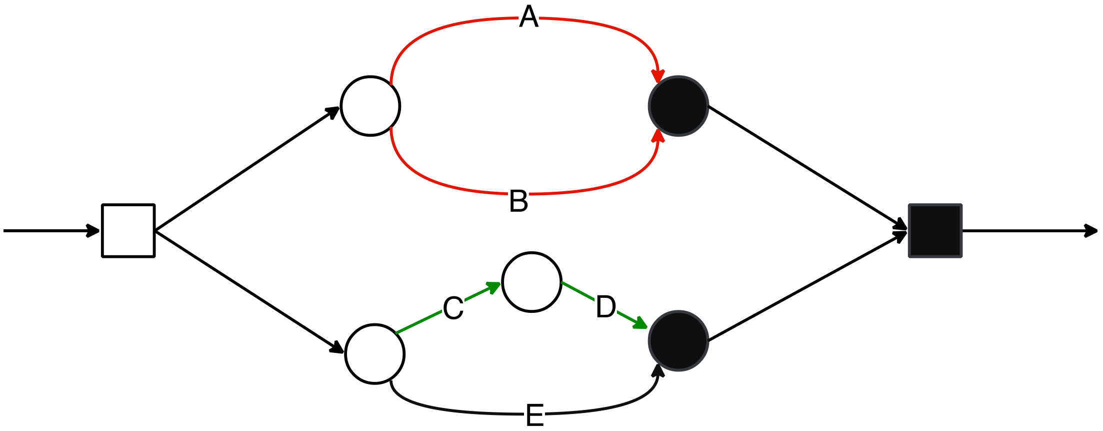


Figure 3. A graphical representation of the three main types of epistatic interactions; showing examples of negative epistatic interaction for genes A and B (red), positive epistatic interaction for genes C and D (green), and no epistatic interaction for genes E and B. Adapted from (Segre et al. 2005).

Contribution 2: Epistasis predictions from flux balance analysis with molecular crowding

Published analyses (Jacobs et al. 2017, Szappanos et al. 2011) show very low agreement between FBA predictions and experimental data on epistasis. In particular, Szappanos *et al.* (Szappanos et al. 2011) predicted negative and positive genetic interactions for genes in yeast metabolism based on FBA and MOMA. The results showed a very low agreement between FBA predictions and experimental data on epistasis. Only a small fraction of interacting genes were recovered *in silico*, with recall values of 2.8% and 12.9% for negative and positive interactions, respectively.

However, FBA predictions ignore important biological constraints; the most important of these may be macro-molecular crowding, arising through a limited solvent capacity of the cell and a corresponding maximal protein "budget". Thus, we hypothesized that constraint-based modeling approaches that consider the enzymatic costs of metabolic pathways may be able to improve the prediction of epistatic interactions. We tested if FBA with molecular crowding (ccFBA) can significantly improve the prediction of genetic interactions in yeast in comparison to other constraint-based methods, in particular standard FBA and a linearized version of Minimization of Metabolic Adjustment (MOMA).

FBA models with molecular crowding limit cell growth by imposing a maximal mass concentration of enzymes, which in turn limits the total flux through the reactions the enzymes catalyze. We calculated epistasis for all metabolic gene pairs in the yeast metabolic network (Yeast v. 7.6, https://sourceforge.net/projects/yeast (Aung et al. 2013)) at full gene knockouts based on the *in silico* biomass production rates of the single and double mutants, assumed to be proportional to fitness values. For this analysis, we used Cost Constrained Flux Balance Analysis (ccFBA) from the sybilccFBA R package (Desouki 2016), which represents an extension of the MOMENT model described in (Adadi et al. 2012). Our findings show that ccFBA was able to predict some positive epistatic interactions not detectable with other constraint-based methods. However, and more importantly, around two thirds of experimentally observed epistatic interactions are undetectable by any of the widely used constraint-based methods.

Very low recall values by all tested constraint-based methods

FBA appears to show the worst compromise between precision (fraction of predictions that are correct) and recall (fraction of interactions that are predicted correctly). However, although the recall increases when using ccFBA, there is also a rise in the number of false positives, especially for negative interactions. Regardless of what constraint-based method we used, the highest recall was 24% for negative and 30% for positive epistatic interactions, even with the most generous cutoffs. Accordingly, more that 70% of experimentally verified genetic interactions are not detectable, regardless of how many false positives we are willing to accept. In order to achieve 20% recall, the false positive rates are more than 10% for negative and 3% for positive interactions rates; due to the high number of comparisons made (71,994 experimentally verified genetic interactions in the dataset used here), such false positive rates are unacceptable for any practical purpose: true predictions of epistasis are drowned in a sea of false predictions. To achieve a recall value of around 12%, a reasonable false positive rate is 1% for both negative and positive interactions – but this means that almost 90% of interactions remain undetectable.

It is conceivable that the failure of all tested constraint-based methods to recover a majority of interactions points to a general flaw in these types of metabolic models. However, we need to emphasize that the ccFBA model contains known enzyme turnover numbers (kcat) for only 535 out of 4,594 protein-associated reactions, and an improved parameterization may well lead to a somewhat improved prediction accuracy in the future.

Calculation of epistasis relies on quantitative estimates of mutant fitnesses for single and double knockouts

Constraint-based methods show a very accurate prediction of gene essentiality (Orth et al. 2010, O'Brien et al. 2015, Hartleb, Jarre and Lercher 2016). In contrast, as explored in the third contribution of this thesis (see below), quantitative predictions of non-lethal gene knockout fitness values correlate only weakly with experimental observations (Papp, Szappanos and Notebaart 2011). As genetic interactions are deduced from predictions based on single mutant fitness values, predictions for genetic interactions with a metabolic model can only be as good as the predictions for single mutant fitness. Therefore, if our models cannot quantify single gene knockout fitness reliably, maybe it is no surprise that they fail to predict epistatic interactions. When predicting synthetic lethals, *i.e.*, gene pairs where the single mutants are viable while the double mutant is unviable, it is reasonable to expect that constraint-based methods also perform well – these models accurately predict gene essentiality, after all. However, all published results (Aziz et al. 2015, Harrison et al. 2007, Heavner and Price 2015) show very low agreement between FBA predictions and experimental data on synthetic lethal interactions in yeast and *E. coli.* In agreement with these earlier, small-scale analyses, we found that neither ccFBA nor any of the other tested methods could also predict synthetic lethality correctly.

Contribution 3: Constraint-based methods fail to predict mutant fitness for non-lethal gene knockouts

As already mentioned above, FBA and related constraint-based methods have been shown to predict gene essentiality with high accuracy, while it is not fully clear to which extent they are capable of predicting mutant physiology of non-essential genes. Therefore, we systematically analyzed the ability of multiple constraint-based modeling methods to predict growth features of *S. cerevisiae* and *E. coli*. As already hinted at by earlier work (Papp et al. 2011), we found that FBA and any of the alternative constraint-based methods, including ccFBA, fail to quantitatively predict knockout effects of non-essential genes. For a given metabolic model and environment, the predictions of biomass fluxes of different non-essential gene knockouts give a limited number of distinct values in contrast to the observations of growth rate, fitness, or biomass yield. Even in the best cases (dataset/model combinations), model-based predictions are only barely better than predictions by a trivial "prediction" model that assumes identical fitness of all knockouts. The models perform slightly better when attempting to classify non-essential gene knockouts into those with and without fitness effects. However, the best performing methods, linear and quadratic MOMA, still only predict between 20% and 40% of experimentally observed deleterious fitness effects. Even these methods never reach recall values above 0.5 for *E. coli* and 0.25 for yeast in any of the datasets tested, indicating that the majority of deleterious knockout effects are unpredictable by current constraint-based methodologies for non-essential genes.

Why is knockout fitness prediction for non-essential genes so much harder than for essential genes?

FBA has been shown to successfully predict gene essentiality, with reported accuracy values between 91% and 95% for the bacterium *E. coli* (Hartleb, Jarre and Lercher 2016a) and between 83% and 90% for different models of the yeast *S. cerevisiae* (Duarte, Herrgard and Palsson 2004, Forster et al. 2003, Kuepfer, Sauer and Blank 2005). The reason behind this high accuracy is likely that gene essentiality is largely a consequence of network topology only, and is independent of kinetic parameters and of regulatory circuits. In contrast, these details may strongly influence the physiological effects of non-essential gene knockouts. For example, consider a genome encoding two homologs of an enzyme, with different kinetics of the two isoforms. If only the catalytically more efficient enzyme is utilized in a given environment, its knockout will always reduce the growth rate, while the knockout of its homolog will have no effect. As FBA only sees the same stoichiometries of the reactions catalyzed by the homologs, it considers them as redundant and predicts that both knockouts have no physiological effects.

More generally, the deletion of a non-essential gene encoding an enzyme active in the wildtype requires a global re-routing of reaction fluxes to re-establish a steady state. This will not happen "automatically", as implicitly assumed by FBA. Rather, the knockout of a non-essential enzyme that is active in the wildtype will result in a concentration increase of the enzyme's substrates as well as a concentration reduction of its products. These concentration changes will likely cascade through the system, perturbing a large number of metabolite concentrations. Such an avalanche of concentration changes may be sensed by regulatory circuits that misinterpret its details as resulting from environmental changes, leading to non-optimal (and often even non-sensical) regulatory responses. Thus, it appears *a priori* likely that the effects of non-essential gene knockouts are much harder to predict by constraint-based methods than those of essential gene knockouts.

On the applicability of constraint-based model techniques on non-essential gene knockouts

That all tested constraint-based methods fail when trying to predict the fitness of gene knockouts throws strong shadows of doubt on the applicability of such methods to all types of phenomena that are based on knockout analysis. In particular, calculations of epistasis are based on the difference between the fitness of single and double-mutant phenotypes. Therefore, it may not be surprising that only a small percentage of genetic interactions can be predicted successfully in the 2nd contribution of this thesis.

In the 1st contribution, we analyzed how the severity of a mutation affects pleiotropy in genome-scale metabolic networks. As these analyses were based on optimality assumptions, they show how pleiotropic genes tend to be. They further show how pleiotropy would change with reduced gene function *if* optimal cellular physiology would be upheld, *e.g.*, by allowing the corresponding strains to evolutionarily adapt to the changed enzyme expression/efficiency. Thus, these calculations – and our accompanying analysis of the role of currency metabolites in pleiotropy generation – certainly have theoretical merit. However, based on our 3rd contribution, it appears highly doubtful that these *optimality-based* predictions would apply to experimental data from enzyme knock-downs. Thus, while our 2nd contribution can elucidate the fundamental role and the characteristics of pleiotropy in the wildtype, they probably say little about pleiotropic effects of gene knock-downs or knockouts. However, no genome-scale experimental data of that type is currently available, and thus a final conclusion on whether the patterns we observe *in silico* are recovered at least qualitatively by such data is not currently possible.

Conclusion

In this thesis, I used various constraint-based methods for modeling genome-scale metabolism to assess the effects of genetic perturbations on microbial physiology, focusing on the two most fundamental systems-level characteristics, pleiotropy and epistasis. While the pleiotropy estimates are of theoretical utility for understanding the systems-level organization of wild type cells, it is not clear if they can explain observations based on experimental genetic perturbations. In contrast, for epistasis, it is rather clear that none of the constraint-based models in widespread use can account for the majority of experimentally observed genetic interactions.

The latter observation prompted us to examine in detail the ability of these methods to predict non-essential gene knockouts, with very discouraging results. Based on the observation that there is practically no biologically meaningful correlation between *in silico* predictions of non-essential knockout fitness and experimental observations, we have to conclude that *all* tested methods are unsuitable for the description of microbial physiology after genetic perturbations – except when looking at lethal phenotypes.

In contrast, constraint-based methods such as ccFBA have been shown to predict growth rates of wildtype cells on a variety of carbon sources (Sanchez et al. 2017), indicating that such methods are capable of predicting wildtype physiology. This points to the conclusion that the phenotypic effects of genetic perturbations are fundamentally different from expectations for wildtype cells, probably because of the perturbations' interference with regulatory systems that evolved in the wildtype, not the perturbed system.

References

Adadi, R., B. Volkmer, R. Milo, M. Heinemann & T. Shlomi (2012) Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS Comput Biol,* 8**,** e1002575.

Alberts, B. 2002. *Molecular biology of the cell*. New York: Garland Science.

Armbruster, W. S., C. Pelabon, G. H. Bolstad & T. F. Hansen (2014) Integrated phenotypes: understanding trait covariation in plants and animals. *Philos Trans R Soc Lond B Biol Sci,* 369**,** 20130245.

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

Aung, H. W., S. A. Henry & L. P. Walker (2013) Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism. *Ind Biotechnol (New Rochelle N Y),* 9**,** 215-228.

Aurich, M. K., G. Paglia, O. Rolfsson, S. Hrafnsdottir, M. Magnusdottir, M. M. Stefaniak, B. O. Palsson, R. M. Fleming & I. Thiele (2015) Prediction of intracellular metabolic states from extracellular metabolomic data. *Metabolomics,* 11**,** 603-619.

Aziz, R. K., J. M. Monk, R. M. Lewis, S. In Loh, A. Mishra, A. Abhay Nagle, C. Satyanarayana, S. Dhakshinamoorthy, M. Luche, D. B. Kitchen, K. A. Andrews, N. L. Fong, H. J. Li, B. O. Palsson & P. Charusanti (2015) Systems biology-guided identification of synthetic lethal gene pairs and its potential use to discover antibiotic combinations. *Sci Rep,* 5**,** 16025.

Baryshnikova, A., M. Costanzo, Y. Kim, H. Ding, J. Koh, K. Toufighi, J.-Y. Youn, J. Ou, B.-J. San Luis, S. Bandyopadhyay, M. Hibbs, D. Hess, A.-C. Gingras, G. D. Bader, O. G. Troyanskaya, G. W. Brown, B. Andrews, C. Boone & C. L. Myers (2010) Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nature Methods,* 7**,** 1017.

Basan, M., S. Hui, H. Okano, Z. Zhang, Y. Shen, J. R. Williamson & T. Hwa (2015) Overflow metabolism in Escherichia coli results from efficient proteome allocation. *Nature,* 528**,** 99-104.

Beg, Q. K., A. Vazquez, J. Ernst, M. A. de Menezes, Z. Bar-Joseph, A. L. Barabasi & Z. N. Oltvai (2007) Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. *Proc Natl Acad Sci U S A,* 104**,** 12663-8.

Botstein, D. & G. R. Fink (2011) Yeast: An Experimental Organism for 21st Century Biology. *Genetics,* 189**,** 695.

Browman, K. E. & J. C. Crabbe. 2001. Alcoholism: Genetic Aspects. In *International Encyclopedia of the Social & Behavioral Sciences,* eds. N. J. Smelser & P. B. Baltes, 371-378. Oxford: Pergamon.

Burstin, J., P. Marget, M. Huart, A. Moessner, B. Mangin, C. Duchene, B. Desprez, N. Munier-Jolain & G. Duc (2007) Developmental Genes Have Pleiotropic Effects on Plant Morphology and Source Capacity, Eventually Impacting on Seed Protein Content and Productivity in Pea. *Plant Physiology,* 144**,** 768-781.

Cherry, J. M., C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng & D. Botstein (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res,* 26**,** 73-9.

Chiu, H.-C., C. J. Marx & D. Segrè (2012) Epistasis from functional dependence of fitness on underlying traits. *Proceedings. Biological sciences,* 279**,** 4156-4164.

Churchill, G. A. 2001. Epistasis. In *Encyclopedia of Genetics,* eds. S. Brenner & J. H. Miller, 638-641. New York: Academic Press.

De Martino, D., F. Capuani, M. Mori, A. De Martino & E. Marinari (2013) Counting and correcting thermodynamically infeasible flux cycles in genome-scale metabolic networks. *Metabolites,* 3**,** 946-66.

de Visser, J. A. G. M., T. F. Cooper & S. F. Elena (2011) The causes of epistasis. *Proceedings of the Royal Society B-Biological Sciences,* 278**,** 3617-3624.

Desouki, A. A. 2016. Algorithms for improving the predictive power of flux balance analysis. In *Institute for Computer Science*. Heinrich Heine University Duesseldorf.

Duarte, N. C., M. J. Herrgard & B. O. Palsson (2004) Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res,* 14**,** 1298-309.

Durot, M., P. Y. Bourguignon & V. Schachter (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. *Fems Microbiology Reviews,* 33**,** 164-190.

Edwards, J. S., M. Covert & B. Palsson (2002) Metabolic modelling of microbes: the flux-balance approach. *Environ Microbiol,* 4**,** 133-40.

Edwards, J. S., R. U. Ibarra & B. O. Palsson (2001) In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat Biotechnol,* 19**,** 125-30.

Edwards, J. S. & B. O. Palsson (2000) Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions. *BMC Bioinformatics,* 1**,** 1.

Elena, S. F. & R. E. Lenski (1997) Test of synergistic interactions among deleterious mutations in bacteria. *Nature,* 390**,** 395-8.

Famili, I., J. Förster, J. Nielsen & B. O. Palsson (2003) Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proceedings of the National Academy of Sciences,* 100**,** 13134.

Forster, J., I. Famili, B. O. Palsson & J. Nielsen (2003) Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae. *Omics,* 7**,** 193-202.

Fritzemeier, C. J., D. Hartleb, B. Szappanos, B. Papp & M. J. Lercher (2017) Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. *PLOS Computational Biology,* 13**,** e1005494.

Gasch, A. P., P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein & P. O. Brown (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell,* 11**,** 4241-57.

Gratten, J. & P. M. Visscher (2016) Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Med,* 8**,** 78.

Hansen, T. F. (2006) The Evolution of Genetic Architecture. *Annual Review of Ecology, Evolution, and Systematics,* 37**,** 123-157.

Harbison, C. T., D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel & R. A. Young (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature,* 431**,** 99-104.

Harrison, R., B. Papp, C. Pal, S. G. Oliver & D. Delneri (2007) Plasticity of genetic interactions in metabolic networks of yeast. *Proc Natl Acad Sci U S A,* 104**,** 2307-12.

Hartleb, D., F. Jarre & M. J. Lercher (2016) Improved Metabolic Models for E. coli and Mycoplasma genitalium from GlobalFit, an Algorithm That Simultaneously Matches Growth and Non-Growth Data Sets. *PLoS Comput Biol,* 12**,** e1005036.

Hayden, E. J., E. Ferrada & A. Wagner (2011) Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature,* 474**,** 92-5.

He, X., W. Qian, Z. Wang, Y. Li & J. Zhang (2010) Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. *Nat Genet,* 42**,** 272-6.

Heavner, B. D. & N. D. Price (2015) Comparative Analysis of Yeast Metabolic Network Models Highlights Progress, Opportunities for Metabolic Reconstruction. *PLOS Computational Biology,* 11**,** e1004530.

Hill, W. G. & A. Robertson (1966) The effect of linkage on limits to artificial selection. *Genet Res,* 8**,** 269-94.

Holzhutter, H. G. (2004) The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur J Biochem,* 271**,** 2905-22.

Ibarra, R. U., J. S. Edwards & B. O. Palsson (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature,* 420**,** 186-9.

Ideker, T., T. Galitski & L. Hood (2001) A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet,* 2**,** 343-72.

Jacobs, C., L. Lambourne, Y. Xia & D. Segre (2017) Upon Accounting for the Impact of Isoenzyme Loss, Gene Deletion Costs Anticorrelate with Their Evolutionary Rates. *PLoS One,* 12**,** e0170164.

Johnsson, M., I. Gustafson, C. J. Rubin, A. S. Sahlqvist, K. B. Jonsson, S. Kerje, O. Ekwall, O. Kampe, L. Andersson, P. Jensen & D. Wright (2012) A sexual ornament in chickens is affected by pleiotropic alleles at HAO1 and BMP2, selected during domestication. *PLoS Genet,* 8**,** e1002914.

Khan, A. I., D. M. Dinh, D. Schneider, R. E. Lenski & T. F. Cooper (2011) Negative epistasis between beneficial mutations in an evolving bacterial population. *Science,* 332**,** 1193-6.

Kitano, H. (2002) Systems biology: a brief overview. *Science,* 295**,** 1662-4.

Kondrashov, A. S. (1988) Deleterious mutations and the evolution of sexual reproduction. *Nature,* 336**,** 435-40.

Kryazhimskiy, S., G. Tkacik & J. B. Plotkin (2009) The dynamics of adaptation on correlated fitness landscapes. *Proc Natl Acad Sci U S A,* 106**,** 18638-43.

Kuepfer, L., U. Sauer & L. M. Blank (2005) Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome Res,* 15**,** 1421-30.

Lewis, N. E., H. Nagarajan & B. O. Palsson (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol,* 10**,** 291-305.

Loewe, L. (2008) Genetic mutation. *Nature Education,* 1(1):113.

Mackay, T. F., E. A. Stone & J. F. Ayroles (2009) The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet,* 10**,** 565-77.

Mahadevan, R. & C. H. Schilling (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng,* 5**,** 264-76.

Nielsen, J. (2017) Systems Biology of Metabolism: A Driver for Developing Personalized and Precision Medicine. *Cell Metabolism,* 25**,** 572-579.

Novick, P. & D. Botstein (1985) Phenotypic analysis of temperature-sensitive yeast actin mutants. *Cell,* 40**,** 405-416.

O'Brien, E. J., J. M. Monk & B. O. Palsson (2015) Using Genome-scale Models to Predict Biological Capabilities. *Cell,* 161**,** 971-987.

Orth, J. D., T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist & B. O. Palsson (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol Syst Biol,* 7**,** 535.

Orth, J. D., I. Thiele & B. O. Palsson (2010) What is flux balance analysis? *Nature Biotechnology,* 28**,** 245-248.

Otto, S. P. (2009) The evolutionary enigma of sex. *Am Nat,* 174 Suppl 1**,** S1-s14.

Papp, B., B. Szappanos & R. A. Notebaart (2011) Use of genome-scale metabolic models in evolutionary systems biology. *Methods Mol Biol,* 759**,** 483-97.

Parkes, M., A. Cortes, D. A. van Heel & M. A. Brown (2013) Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet,* 14**,** 661-73.

Pavey, S. A., H. Collin, P. Nosil & S. M. Rogers (2010) The role of gene expression in ecological speciation. *Annals of the New York Academy of Sciences,* 1206**,** 110-129.

Pfeiffer, T. & S. Bonhoeffer (2004) Evolution of cross-feeding in microbial populations. *Am Nat,* 163**,** E126-35.

Phillips, P. C. (2008) Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet,* 9**,** 855-67.

Phillips, P. C., Otto, S.P. & Whitlock, M.C. 2000. Beyond the Average: The Evolutionary Importance of Gene Interactions and Variablity of Epistatic Effects. In *Epistasis and the Evolutionary Process,* ed. J. B. Wolf, Brodie III, E.D. & Wade, M.J., 20-38. New York: Oxford University Press.

Plomin, R., C. M. Haworth & O. S. Davis (2009) Common disorders are quantitative traits. *Nat Rev Genet,* 10**,** 872-8.

Price, N. D., J. A. Papin, C. H. Schilling & B. O. Palsson (2003) Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol,* 21**,** 162-9.

Raman, K. & N. Chandra (2009) Flux balance analysis of biological systems: applications and challenges. *Brief Bioinform,* 10**,** 435-49.

Rutter, M. T., F. H. Shaw & C. B. Fenster (2010) SPONTANEOUS MUTATION PARAMETERS FOR ARABIDOPSIS THALIANA MEASURED IN THE WILD. *Evolution,* 64**,** 1825-1835.

Sanchez, B. J., C. Zhang, A. Nilsson, P. J. Lahtvee, E. J. Kerkhoven & J. Nielsen (2017) Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol Syst Biol,* 13**,** 935.

Sanjuan, R., J. M. Cuevas, A. Moya & S. F. Elena (2005) Epistasis and the adaptability of an RNA virus. *Genetics,* 170**,** 1001-8.

Schuster, S., T. Pfeiffer & D. A. Fell (2008) Is maximization of molar yield in metabolic networks favoured by evolution? *Journal of Theoretical Biology,* 252**,** 497-504.

Segre, D., A. Deluna, G. M. Church & R. Kishony (2005) Modular epistasis in yeast metabolism. *Nat Genet,* 37**,** 77-83.

Segrè, D. & C. J. Marx (2010) Introduction to Focus Issue: Genetic Interactions. *Chaos: An Interdisciplinary Journal of Nonlinear Science,* 20**,** 026101.

Segre, D., D. Vitkup & G. M. Church (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 15112-15117.

Shlomi, T., O. Berkman & E. Ruppin (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A,* 102**,** 7695-700.

Sivakumaran, S., F. Agakov, E. Theodoratou, J. G. Prendergast, L. Zgaga, T. Manolio, I. Rudan, P. McKeigue, J. F. Wilson & H. Campbell (2011) Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet,* 89**,** 607-18.

Stark, C., B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz & M. Tyers (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res,* 34**,** D535-9.

Stearns, F. W. (2011) One Hundred Years of Pleiotropy: A Retrospective (vol 186, pg 767, 2010). *Genetics,* 187**,** 355-355.

Szappanos, B., K. Kovacs, B. Szamecz, F. Honti, M. Costanzo, A. Baryshnikova, G. Gelius-Dietrich, M. J. Lercher, M. Jelasity, C. L. Myers, B. J. Andrews, C. Boone, S. G. Oliver, C. Pal & B. Papp (2011) An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature Genetics,* 43**,** 656-U182.

Teixeira, M. C., P. Monteiro, P. Jain, S. Tenreiro, A. R. Fernandes, N. P. Mira, M. Alenquer, A. T. Freitas, A. L. Oliveira & I. Sa-Correia (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res,* 34**,** D446-51.

Terzer, M., N. D. Maynard, M. W. Covert & J. Stelling (2009) Genome-scale metabolic networks. *Wiley Interdisciplinary Reviews-Systems Biology and Medicine,* 1**,** 285-297.

Teusink, B., A. Wiersma, L. Jacobs, R. A. Notebaart & E. J. Smid (2009) Understanding the Adaptive Growth Strategy of Lactobacillus plantarum by In Silico Optimisation. *Plos Computational Biology,* 5.

Thiele, I. & B. O. Palsson (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc,* 5**,** 93-121.

Tong, A. H., M. Evangelista, A. B. Parsons, H. Xu, G. D. Bader, N. Page, M. Robinson, S. Raghibizadeh, C. W. Hogue, H. Bussey, B. Andrews, M. Tyers & C. Boone (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science,* 294**,** 2364-8.

Varma, A. & B. O. Palsson (1994) Metabolic Flux Balancing - Basic Concepts, Scientific and Practical Use. *Bio-Technology,* 12**,** 994-998.

Wagner, G. P. & L. Altenberg (1996) PERSPECTIVE: COMPLEX ADAPTATIONS AND THE EVOLUTION OF EVOLVABILITY. *Evolution,* 50**,** 967-976.

Wagner, G. P., M. Pavlicev & J. M. Cheverud (2007) The road to modularity. *Nat Rev Genet,* 8**,** 921-31.

Wagner, G. P. & J. Zhang (2011) The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat Rev Genet,* 12**,** 204-13.

Weinreich, D. M., N. F. Delaney, M. A. Depristo & D. L. Hartl (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science,* 312**,** 111-4.

Wright, D. (2015) The Genetic Architecture of Domestication in Animals. *Bioinform Biol Insights,* 9**,** 11-20.

Xu, L., B. Barker & Z. L. Gu (2012) Dynamic epistasis for different alleles of the same gene. *Proceedings of the National Academy of Sciences of the United States of America,* 109**,** 10420-10425.0

Manuscripts

Manuscript 1: Alleles of a gene differ in pleiotropy, often mediated through currency metabolite production, in *E. coli* and yeast metabolic simulations

Status

This manuscript was published as:

Deya Alzoubi, Abdelmoneim Amer Desouki & Martin J. Lercher (2018). Alleles of a gene differ in pleiotropy, often mediated through currency metabolite production, in *E. coli* and yeast metabolic simulations. *Scientific Reports* **8**:10.1038/s41598-018-35092-1.

Contributions

Together with Martin Lercher, I developed the concept and methodology of this work and wrote the manuscript. I wrote the software for simulations and analyses and performed all analyses.

Manuscript

(see next page)

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Alleles of a gene differ in pleiotropy, often mediated through currency metabolite production, in *E. coli* and yeast metabolic simulations

Received: 27 June 2018

Accepted: 22 October 2018

Published online: 22 November 2018

Deya Alzoubi, Abdelmoneim Amer Desouki & Martin J. Lercher

A major obstacle to the mapping of genotype-phenotype relationships is pleiotropy, the tendency of mutations to affect seemingly unrelated traits. Pleiotropy has major implications for evolution, development, ageing, and disease. Except for disease data, pleiotropy is almost exclusively estimated from full gene knockouts. However, most deleterious alleles segregating in natural populations do not fully abolish gene function, and the degree to which a polymorphism reduces protein function may influence the number of traits it affects. Utilizing genome-scale metabolic models for *Escherichia coli* and *Saccharomyces cerevisiae*, we show that most fitness-reducing full gene knockouts of metabolic genes in these fast-growing microbes have pleiotropic effects, *i.e.*, they compromise the production of multiple biomass components. Alleles of the same metabolic enzyme-encoding gene with increasingly reduced enzymatic function typically affect an increasing number of biomass components. This increasing pleiotropy is often mediated through effects on the generation of currency metabolites such as ATP or NADPH. We conclude that the physiological effects observed in full gene knockouts of metabolic genes will in most cases not be representative for alleles with only partially reduced enzyme capacity or expression level.

A gene is pleiotropic if it affects more than one phenotypic trait[1,2]. A classic example is phenylketonuria, a human disease that is caused by a single gene defect but which affects multiple systems, with symptoms ranging from lighter skin color to mental disorders[3]. Pleiotropic effects can cause alleles to affect fitness differentially at different ages, a phenomenon believed to be a major cause of aging[4–6]; indeed, alleles contributing to increased longevity often show reduced fertility and stress tolerance[7]. Similar antagonistic epistasis may underlie other important biological phenomena such as speciation[8] and cooperation[9]. Understanding the factors that contribute to pleiotropy is of fundamental importance in genetics[10–12], evolution[13–16], development[17,18], as well as in disease[19,20] and ageing[4]. In comparison to its fundamental importance, empirical knowledge of the prevalence and especially on the causal mechanisms of pleiotropy is scarce[2,21].

Pleiotropy may be classified according to the types of traits considered[22]. Molecular gene pleiotropy refers to the number of functions of a gene and its products, *e.g.*, the number of reactions catalyzed by a single enzyme. Developmental pleiotropy describes the genetic and evolutionary interdependence of phenotypic aspects. Finally, selectional pleiotropy refers to the number of separate fitness components affected by mutations to a gene. In this study, we focus on the latter type of pleiotropy.

Experimental studies generally assess pleiotropy through the analysis of gene knockouts[23–25]. The degree of pleiotropy is then defined as the number of traits affected when a gene becomes fully non-functional. Wang *et al.*[25] analyzed phenotypes of large numbers of yeast, nematode, and mouse mutants. They found that pleiotropy is widespread: on average, yeast gene knockouts affect 8% of the examined traits; for the nematode, the corresponding number is 10%, for the mouse 3% (see also[24]). The distributions of the degree of pleiotropy appear rather similar across very different study systems, from the skeletal features of mice[25] to metabolic systems[26–28]. Moreover, pleiotropy was found to be modular, such that sets of genes tend to affect the same sets of traits[25].

Institute for Computer Science and Department of Biology, Heinrich Heine University, Universitätsstraße 1, Düsseldorf, D-40221, Germany. Correspondence and requests for materials should be addressed to M.J.L. (email: martin.lercher@hhu.de)

In *E. coli*, 36% of metabolic reactions are catalyzed by enzymes also involved in other reactions; the same is true for 27% of metabolic reactions in the yeast *Saccharomyces cerevisiae*[28]. Pleiotropic effects of mutations that affect enzyme activity can be simulated from genome-scale metabolic models using constraint-based modeling techniques such as flux balance analysis (FBA)[29,30]. The functional pleiotropy of a metabolic gene can then be defined as the number of biomass components whose maximal production is affected by the gene's knockout[26]. Previous studies using this definition found that a metabolic gene's functional pleiotropy is related to its propensity to form negative epistatic interactions with other metabolic genes[26,27].

While full gene knockouts are easily examined experimentally, they may not be representative of the effects of deleterious alleles segregating in natural populations: individual mutations may affect only a subset of all traits influenced by the gene[31]. Thus, it is important to distinguish between the pleiotropy of the gene and the pleiotropy of individual mutations, especially in evolutionary and clinical contexts. For example, while 4.6% of human SNPs implicated in complex non-Mendelian phenotypes show pleiotropic effects, most of these do not fully abolish protein function[32]. Experimental studies indicate that mutational pleiotropy tends to be smaller than gene pleiotropy[31].

Genome-scale metabolic models allow us to dissect the relationship between gene and mutational pleiotropy in quantitative detail, without being hampered by the detection limits of experimental assays. Does the degree of pleiotropy depend on how severely a given allele of a metabolic gene reduces protein activity, *i.e.*, are the same number of functions affected when protein function or expression is reduced only partially? How modular is metabolic pleiotropy? Currency metabolites, such as ATP and NADPH, are used as cofactors in many otherwise unrelated reactions; it thus appears highly likely that a substantial fraction of metabolic pleiotropy is due to effects on the production of currency metabolites. Is such an effect of currency metabolites on patterns of pleiotropy confirmed by simulated data?

Below, we address these questions by analyzing the metabolic networks of a representative bacterial model system, *Escherichia coli*, and a corresponding eukaryotic system, the baker's yeast *Saccharomyces cerevisiae*. We find that most gene knockouts that impact fitness do so by affecting the production of multiple biomass components, and that the number of affected biomass components typically increases with increasing mutation severity. Pleiotropy is rarely a consequence of multiple molecular gene functions, but is an emergent property of the metabolic network. For many genes, pleiotropy is indeed mediated through their involvement in the generation of currency metabolites.

Results

Estimating pleiotropy from contributions to biomass components within the wildtype flux distribution.

We first estimated wildtype flux distributions in the default growth condition for the genome-scale metabolic model of *E. coli*[33] and the yeast *S. cerevisiae*[34] (obtained from https://sourceforge.net/projects/yeast/files/). The maximal biomass production rates were estimated using flux balance analysis (FBA)[29,30]. For both model systems, we identified the flux distribution compatible with maximal biomass production that had the smallest sum of absolute fluxes, a strategy often termed parsimonious FBA (pFBA), which approximates optimal utilization of limited cellular protein resources[35].

To simulate mutations that cause different reductions of protein function or expression and correspond to different deleterious alleles of a metabolic gene, we restricted the maximal flux through all reactions requiring this gene to a fixed percentage of the estimated wildtype flux[36], starting from 100% (the wildtype) down to 0% (a full gene knockout) in steps of 0.5%. For each flux reduction, we defined the degree of pleiotropy (referred to simply as "pleiotropy" below) as the number of biomass components whose production was reduced by at least 0.01% compared to the maximal (wildtype) production. Note that with this definition, only genes with pleiotropy ≥ 2 are pleiotropic, while genes with pleiotropy 0 (no affected biomass component) or pleiotropy 1 (one affected biomass component) are non-pleiotropic.

Flux distributions at maximal biomass production rate are usually not unique[35], and so in many cases a flux restriction through one reaction may be compensated by a rerouting of fluxes through alternative pathways. Such rerouting would require the upregulation of the corresponding genes. While it has been observed experimentally that cells can survive many gene deletions in central metabolism without drastic changes in gene expression[37], the necessary upregulation of protein expression will not occur spontaneously at least for some pathways[38]. More importantly, if we are interested in the *de facto* contribution of a given gene to the production of biomass components, then it is of no consequence if alternative pathways *could* take over part of this functionality. Thus, when calculating the maximal (wildtype) production rate of individual biomass components as well as when simulating the effects of mutations to a given metabolic gene, we did not allow the redistribution of fluxes to alternative pathways: we allowed only decreases, not increases, of the absolute value of any flux compared to the wildtype flux distribution obtained with pFBA.

Note that experimental studies often employ a pragmatic working definition of pleiotropy that lies somewhere between the definitions of pleiotropy proposed here based on the wild type flux distribution on the one hand and a quantitative measure of essentiality based on an analogous calculation that allows the free redistribution of fluxes. In these studies, pleiotropy is typically estimated as the number of traits with observable phenotypic changes after the gene knockout, but before allowing the strain to adapt to its new genotype. In this case, some fluxes may be rerouted due to enzymes and transporters that are expressed either spuriously or because of other roles they play in wildtype physiology, while other fluxes that require the upregulation of the corresponding enzymes and transporters will not yet be active. Thus, our definition of pleiotropy describes a "worst case scenario", providing an upper limit on experimentally measured pleiotropy.

Many genes affect the production of multiple biomass components.

Pleiotropy varies widely between different genes. Mutations to the majority of genes affect no biomass components in the minimal growth

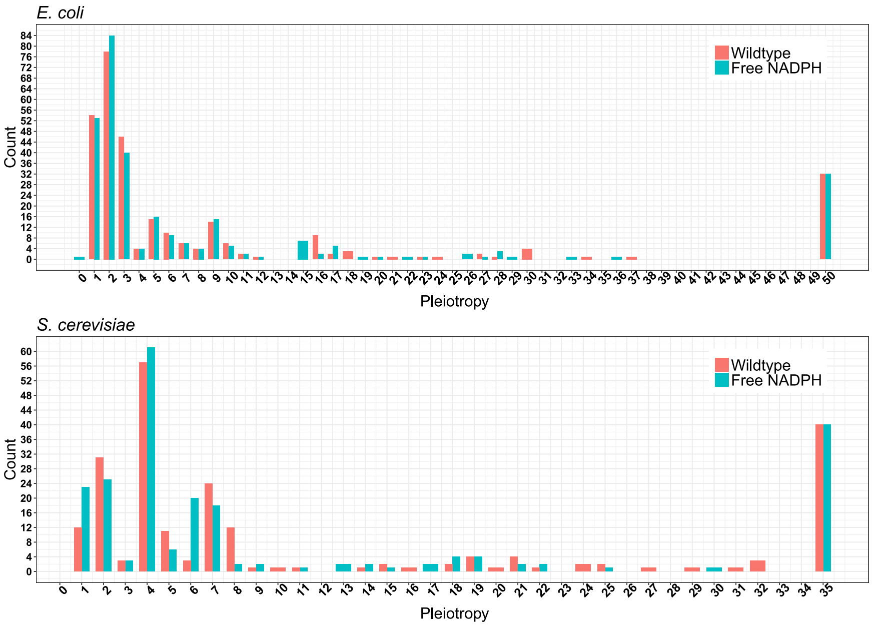


Figure 1. Most complete gene knockouts of fitness-relevant genes have pleiotropic effects, *i.e.*, they affect the production of multiple biomass components. For some genes, pleiotropy is reduced when NADPH is made freely available (cyan bars). For other freely available currency metabolites, see Supplementary Figure S2.

		E. coli[c]	*S. cerevisiae*[c]
Number of biomass components		**50**	**35**
Standard model	Pleiotropy[a]	10.0 (3)	12.1 (5)
	Essentiality[a]	4.6 (2)	9.3 (2)
Free NADPH[b]	Pleiotropy[a]	9.9 (3)	11.1 (4.5)
	Essentiality[a]	4.3 (2)	8.0 (1)

Table 1. Average number of biomass components whose production is affected by a full gene knockout. [a]In the FBA calculations, fluxes are either constrained to not exceed the wildtype (WT) fluxes to estimate the *de facto* contribution of gene products to biomass production (Pleiotropy), or they are allowed to vary freely to assess the number of biomass components for which gene products are essential even after allowing the mutant strain to adapt (Essentiality). [b]Solution when allowing unlimited conversion of NADPH to NADP+ [c]Mean (median) number of affected biomass components.

medium assayed, independent of mutation severity (*E. coli*: 1,067 genes or 78.1%; *S. cerevisiae*: 687 genes or 75.6%). Among genes contributing to biomass production—and thus fitness—in the wildtype, non-pleiotropic cases are rare: in *E. coli*, only 54 full-gene knockouts (out of 299 knockouts with fitness contributions, 18.1%) affect exactly one biomass component, while the same is true for only 12 knockouts (out of 222 knockouts with fitness contributions, 5.4%) in *S. cerevisiae*. Conversely, knockouts of 32 genes in *E. coli* (10.7% of knockouts with fitness contributions) and 40 genes in *S. cerevisiae* (18.1% of knockouts with fitness contributions) affect the production of *all* biomass components. Many of the remaining genes show low degrees of pleiotropy, affecting the production of only a few biomass components; on average, full gene knockouts of fitness-relevant genes affect the production of 20% of biomass components in *E. coli* and 34% of biomass components in *S. cerevisiae* (Fig. 1, Table 1).

These percentages reflect functional pleiotropy, the *de facto* contribution of gene products to biomass component production. If, instead, we are interested in the phenotypic effects of gene knockouts after allowing the mutant strain to adapt its physiology to its altered gene content, we must allow free redistributions of fluxes after the gene knockouts. Corresponding simulations show that after adaptation, genes with fitness contributions are, on average, essential for the production of 9.2% of *E. coli* biomass components and of 26.6% of *S. cerevisiae* biomass components (Table 1, Supplementary Figure S1). The degree of gene pleiotropy for yeast is substantially higher than previous experimental estimates, which are around 2 (corresponding to 2–11% of considered traits depending on the types of traits analyzed)[25]; however, experimental estimates of gene pleiotropy tend be downwardly biased due to experimental detection limits[2,22].

Pleiotropy is an emergent property of the metabolic network. Pleiotropy can be classified by its origin into type I pleiotropy, caused by multiple molecular functions of a gene product, and type II pleiotropy, caused by multiple physiological consequences of a single molecular function[2]. Similar distinctions have been made previously using the terms “horizontal” vs. “vertical”[10] and “mosaic” vs. “relational”[39] pleiotropy. Our model allows us to quantify the relative contributions of these two pleiotropy types. 41.7% of *E.coli* genes and 40.5% of yeast genes in our metabolic models catalyze multiple reactions. To what extent does this functional diversity cause functional pleiotropy as measured in the number of biomass components affected by a gene knockout? To answer this question, we compared the gene pleiotropy (Fig. 1) to the pleiotropy of individual reactions catalyzed by the gene product. For example, fully abolishing all functions of the purB (b1131) gene, whose gene product catalyzes two distinct biochemical reactions, reduced the production of 18 biomass components. In contrast, blocking only one of the catalyzed reactions results in a pleiotropy estimate of 16, while blocking only the other reaction results in a pleiotropy of 10. Thus, the pleiotropy of the b1131 gene is largely of type II, and is only in small part due to its multiple molecular functions.

This pattern is typical: the maximal pleiotropy arising from blocking only a single out of several reactions catalyzed by the same protein accounts for over 97% of the gene pleiotropy (*E. coli* 97.4%, yeast 97.6%). These numbers drop only marginally when we consider only gene products that are essential for multiple reactions, to 92.2% in *E. coli* and to 94.5% in yeast (Supplementary Figure S3). We conclude that the vast majority of metabolic epistasis is of type II, *i.e*., is an emergent property of the metabolic network rather than a consequence of multiple molecular functions. This finding is consistent with the previous observation that the degree of pleiotropy in yeast is not significantly correlated with the number of molecular gene functions[40].

Metabolic networks show significant but low modularity. The relationship between genes and biomass components (traits) can be represented as a bipartite graph, with links connecting genes with affected biomass components. Modules are defined as sets of genes and traits with significantly more within-module than between-module links[25]. A high degree of modularity thus indicates that pleiotropic genes tend to affect groups of related traits (*e.g*., chemically related biomass components) rather than random sets of traits. Supplementary Figure S4 shows heatmaps that illustrate the modularity of both metabolic pleiotropy networks. To quantitatively assess the modularity, we used the LP&BRIM algorithm[41], resulting in raw modularities of $Q = 0.235$ for *E. coli* and $Q = 0.197$ for *S. cerevisiae*. Both networks show highly statistically significant modularity: in each case, the modularity of 10,000 randomly rewired networks was always lower than observed for the real pleiotropy network (*i.e*., $P < 0.0001$; Supplementary Figure S5).

Following ref.[25], we then defined a *z*-score for modularity (or “scaled modularity”)[42] as the difference between the observed modularity and the mean modularity of randomly rewired networks, measured in number of standard deviations. The *E. coli* pleiotropy network exhibits a scaled modularity of 9.1, while the *S. cerevisiae* network has a scaled modularity of 4.9, *i.e*., the modularity of metabolic pleiotropy is about 9 and 5 standard deviations higher than for corresponding random gene-trait networks. These values are surprisingly low: for five different experimental study systems and trait definitions, Wang *et al.* found a median scaled modularity of 37 (range 34–238). Thus, metabolic pleiotropy networks are less modular than other pleiotropy networks, suggesting that the underlying metabolic network shows more interconnections between the pathways producing different sets of biomass components than the genetic networks underlying other types of traits. Our findings on modularity may be related to the role of currency metabolites, which crosslink the diverse metabolic pathways (see below).

Pleiotropy typically increases with increasing mutation severity. We next examined the pleiotropy of alleles with small-effect mutations, *i.e*., mutations that reduce enzyme capacity without fully abolishing enzyme function. About 20% of *E. coli* genes with fitness contributions have constant pleiotropy: small-effect mutations of these genes affect the same number of biomass components as full gene knockouts. In comparison, only 7.7% of yeast genes contributing to fitness exhibit constant pleiotropy.

All other genes contributing to fitness affect an increasing number of biomass components for increasingly deleterious alleles. Figure 2 shows this stepwise increase in pleiotropy for the example of *Lipoamide dehydrogenase* (gene names: *E. coli* b0116, *S. cerevisiae* YFL018C; for additional examples, see Supplementary Figure S6. In both organisms, pleiotropy typically increases in about a dozen steps from weakly to strongly deleterious alleles (Fig. 3; mean number of steps: *E. coli* 11.6, *S. cerevisiae* 12.6).

The pleiotropy of the full gene knockout constitutes an upper limit to the number of stepwise increases in pleiotropy. If there was otherwise no systematic relationship between maximal pleiotropy and the number of steps, we would expect the numbers of steps to be uniformly distributed between zero and the pleiotropy of the full knockout. However, the correlation between the number of steps and pleiotropy at full knockout was much stronger than expected from such a relationship (Supplementary Figure S7, Spearman’s $\rho = 0.926$ (*E. coli*) and $\rho = 0.986$ (*S. cerevisiae*), $P < 10^{-6}$ in each case from randomizations; see Methods). Thus, genes whose full knockout showed higher metabolic pleiotropy also showed more stepwise increases in pleiotropy for increasingly debilitating mutations.

All genes whose mutations affect the production of at least one biomass component must also affect the overall production of biomass (*i.e*., in the common interpretation of FBA, fitness). The reverse is not true: a mutation to a gene may affect the maximal production of biomass, but not the production of any individual biomass component. This is a consequence of the algorithm employed to estimate production capabilities for individual biomass components. If we maximize the production of a single compound, then pathways usually concerned with the production of other biomass components can be diverted to the production of this compound. While we find no such genes for *S. cerevisiae*, this is indeed the case for 3 essential *E. coli* genes, which encode transporters for acetate (b4067), magnesium/nickel/cobalt (b3816), and calcium/sodium (b3196, an antiporter).

The pleiotropy of most genes is mediated by currency metabolites. We can conceptually partition internal metabolites into currency metabolites—those involved in many reactions, *e.g*., to provide energy or redox equivalents[43]—and primary metabolites. A deleterious allele may affect the production of a given biomass

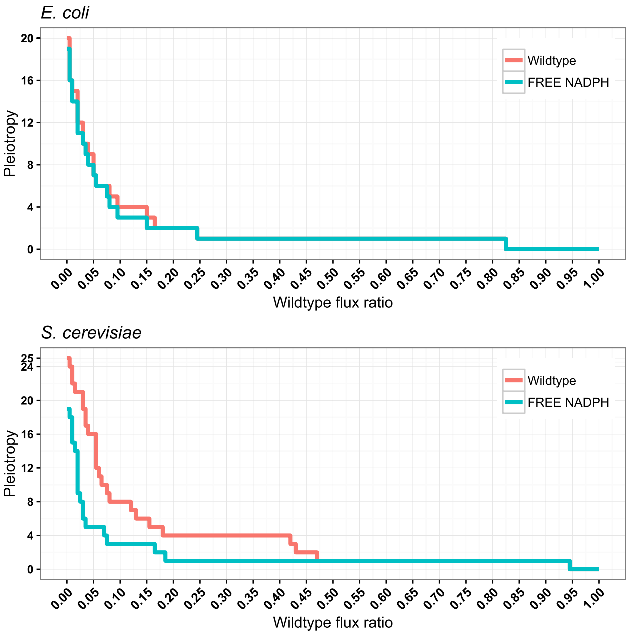


Figure 2. Pleiotropy for the *Lipoamide dehydrogenase* gene increases for increasingly deleterious alleles. Pleiotropy is reduced when NADPH is made freely available (cyan curves). For additional examples, see Supplementary Figure S6.

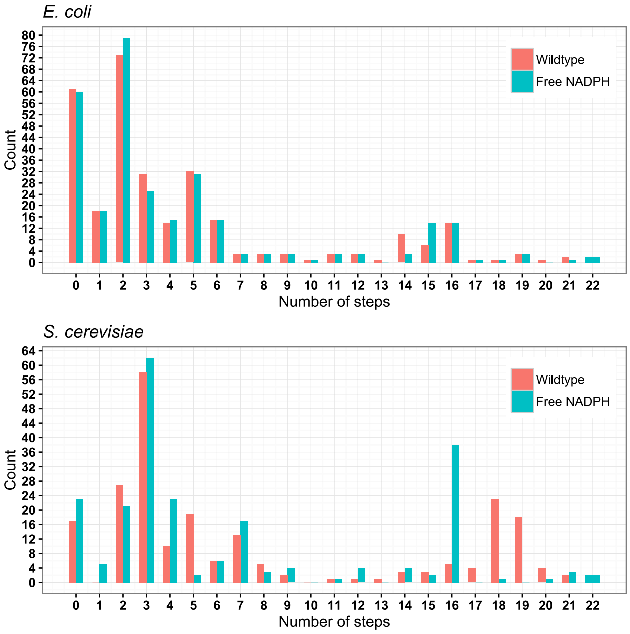


Figure 3. For the majority of genes contributing to biomass production, pleiotropy increases for increasingly deleterious alleles in multiple steps. Histograms for the number of pleiotropy steps in *E. coli* and the yeast *S. cerevisiae*. Cyan bars reflect the reduced numbers of pleiotropy increases when making NADPH freely available.

component because the mutated gene catalyzes a reaction in a pathway of primary metabolites that directly leads to the component's production. Conversely, a deleterious allele may affect not the primary metabolites, but the currency metabolites utilized in the component's production. A list of 14 currency metabolites was obtained from ref.[43]. Excluding exchange reactions, 753 out of a total of 2,251 reactions (33.5%) in *E. coli* and 310 out of 3,324 reactions (9.3%) in *S. cerevisiae* involved at least one of these metabolites.

A substantial fraction of pleiotropy is indeed associated with the generation of currency metabolites: 87.4% of previously pleiotropic genes show reduced pleiotropy when we make metabolites such as ATP, UTP, or NADPH freely available in yeast (Fig. 4). The free availability of ATP alone reduces the degree of pleiotropy of over half

35

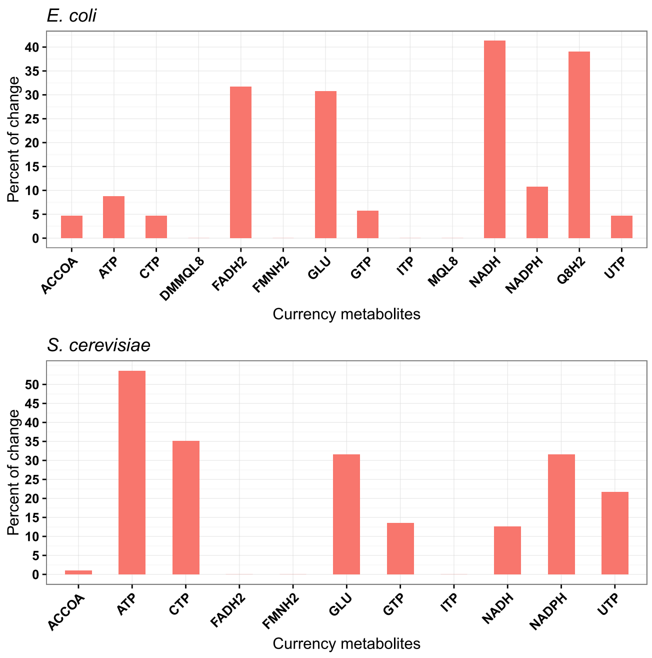


Figure 4. Many genes show reduced pleiotropy when currency metabolites are made freely available. The bar chart shows the percentage of previously pleiotropic genes with reduced pleiotropy in response to the free availability of different currency metabolites. Abbreviations: Adenosine triphosphate (ATP); Cytidine triphosphate (CTP); Guanosine triphosphate (GTP); Uridine triphosphate (UTP); Inosine triphosphate (ITP); Nicotinamide adenine dinucleotide (NADH); Nicotinamide adenine dinucleotide phosphate (NADPH); Flavin adenine dinucleotide reduced (FADH2); Reduced flavin mononucleotide (FMNH2); Ubiquinol-8 (Q8H2); Menaquinol 8 (MQL8); 2-Demethylmenaquinol 8 (DMMQL8); Acetyl-CoA (ACCOA); L-Glutamate (GLU).

of pleiotropic yeast genes. The influence of currency metabolite production on pleiotropy is weaker, yet still substantial in *E. coli*: here, 55.3% of pleiotropic genes are affected, with NADH making the biggest contribution (over 40%) (Fig. 4).

Involvement in currency metabolite production is an important determinant of the number of biomass components for which a gene knockout is essential even after allowing the mutant strain to adapt its protein expression to the altered gene content of its genome. This contribution is particularly striking in yeast: for over half of the tested currency metabolites, free availability reduces the number of biomass components for which a gene is essential for almost half of the genes (Supplementary Figure S8).

Discussion

Using constraint-based simulations of the metabolic models for *E. coli* and the yeast *S. cerevisiae*, we have characterized the distributions of pleiotropy. Consistent with earlier computational[26–28] and experimental[23–25] studies, we found that the knockout of a majority of genes that contribute to fitness has pleiotropic effects. The vast majority of this gene pleiotropy is not caused by multiple molecular functions of the gene product (type I), but is an emergent property of the metabolic network (type II). Pleiotropy is modular, but to a lower degree than estimated experimentally for non-metabolic systems[25].

For most pleiotropic genes, pleiotropy increases strongly for alleles with increasingly debilitating effects. Thus, standard measures of pleiotropy based on gene knockout studies are more likely to reflect the maximal degree of mutational pleiotropy of a given gene[2,31]. Alleles that only knock down protein activity (by reducing enzyme/transporter function or expression level) often affect only a subset of phenotypic traits, with additional traits affected progressively as alleles become more deleterious. Thus, the physiological effect of the full gene knockout will in most cases not be representative for the effects of deleterious alleles that retain some level of enzyme function. This type of effect is also evident from individual medical observations of pleiotropy. For example, some small-effect mutations affecting human SOX9 expression lead to minor skeletal malformations, while the consequences of large-effect mutations can include sex reversals[44].

How can we understand the dependence of pleiotropy on the degree to which an allele reduces protein activity? For increasingly deleterious alleles, more and more metabolic resources must be channeled into the compensation of the compromised pathway; as a consequence of this increasing drain of resources, more and more other pathways are affected. Not surprisingly[27], we found that the pleiotropy of many genes is mediated through the generation of currency metabolites such as ATP, NADPH, or $FADH_2$. This is true for more than half of the pleiotropic genes in *E. coli*, and for 87% of pleiotropic genes in yeast.

While the overall patterns of pleiotropy appear qualitatively similar between *E. coli* and yeast, we found a number of quantitative differences. Compared to *E. coli* genes, yeast genes (i) showed generally higher pleiotropy and were rarely of pleiotropy 1; (ii) were less likely to have constant pleiotropy; and (iii) were more likely to show reduced pleiotropy when supplied with currency metabolites. Moreover, (iv) the yeast pleiotropy network exhibited lower modularity. In part, these differences may be related to network size. The *E. coli* metabolic network encompasses substantially more genes overall than the yeast network. However, we constrained network usage to reactions active in the wildtype. In contrast to total network sizes, the *active* metabolic network of yeast (755 reactions and 184 metabolites) is substantially larger than the *active* metabolic network of *E. coli* (462 reactions and 103 metabolites); this difference is consistent with the notion that yeast metabolism is more complex, yet less flexible than *E. coli* metabolism. While the average number of reactions per metabolite is similar between *E. coli* (4.49) and yeast (4.10), the lower yeast modularity indicates that reactions more often connect otherwise distant network parts. A role of currency metabolites in such connecting reactions would be consistent with the larger effect of currency metabolite supply on pleiotropy in yeast. In sum, the higher interconnectedness of the yeast pleiotropic network, combined with the larger active metabolic network size, appears to provide more potential for pleiotropic effects.

Pleiotropy is a complex phenomenon: it is not constant, but varies between different alleles of the same gene, and its causes are often indirect. Thus, experimental as well as computational analyses of pleiotropy should move away from focusing on full gene knockouts, and instead consider explicitly the degree to which mutations reduce protein activity. The necessity of a corresponding nuanced view of pleiotropy may be particularly evident in studies of medically relevant mutations, where full knockouts are often lethal, while small-effect mutations may segregate at appreciable frequencies in the human population[45].

Materials and Methods

Metabolic models. To simulate *Escherichia coli* metabolism, we used the metabolic reconstruction iJO1366[33], encompassing 1,366 metabolic genes associated with 2,251 reactions. For the yeast *S. cerevisiae*, we used the yeast7.6 model (https://sourceforge.net/projects/yeast)[34], accounting for 909 metabolic genes associated with 3,324 reactions. The published models were used without any modifications. The *E. coli* model contains a growth-independent maintenance energy consumption term (the ATPM reaction), which enforcies a minimal ATPase activity of 3.15 mmol/gDW/h. We utilized the default biomass reactions for *E. coli* (Ec_biomass_iJO1366_core_53p95M), which comprises 50 essential biomass components (Supplementary Table S1), considering only "substrates" of the biomass reactions and excluding inorganic ions and H_2O. For *S. cerevisiae*, we used the "yeast 5 biomass pseudoreaction", which comprises 35 essential biomass components (Supplementary Table S2), again considering only "substrates" and excluding inorganic ions and H_2O.

Flux distribution constraints derived from wildtype simulations. In order to approximate the *de facto* contribution of individual metabolic proteins to the production of individual biomass components *in vivo*, we should only consider flux distributions that are naturally active during growth (biomass production) in the nutritional environment studied, and fluxes should not exceed these wildtype fluxes. We thus first estimate the wildtype flux distribution $\mathbf{v}^{WT}$, by running a flux balance analysis (FBA) with the biomass reaction as the objective function, followed by a minimization of the sum of absolute fluxes at the previously determined maximal biomass production rate (parsimonious FBA[35]).

When simulating the production of individual biomass components, we constrained all fluxes v_i to values between zero and the wildtype flux v_i^{WT} for this reaction, *i.e.*,

$$\begin{aligned} 0 \le v_i \le v_i^{\mathrm{WT}} \text{ for } v_i^{\mathrm{WT}} \ge 0 \\ 0 \ge v_i \ge v_i^{\mathrm{WT}} \text{ for } v_i^{\mathrm{WT}} < 0 \end{aligned} \tag{1}$$

Estimating pleiotropy. For each essential biomass component (Supplementary Tables S1 and S2, respectively), we added a new exchange reaction representing its secretion[46]. As some biomass components may be coupled through the biomass reaction, we allowed the free excretion of all other biomass components when maximizing the production of one selected biomass component (*i.e.*, $v_j \ge 0$ for all added exchange reactions *j*).

We then calculated the maximum production of each biomass component by maximizing its exchange reaction flux while enforcing the wildtype flux distribution constraints (Eq. 1). For each metabolic gene, we compared this unperturbed maximal production with the maximal production rate of alleles with increasingly reduced protein activity, simulated by restricting the flux through all reactions catalyzed by the gene to a fixed fraction of the wildtype flux[36], which we reduced from 100% to 0% in steps of 0.5%. The flux through a specific reaction was constrained in this way only if the gene-protein-reaction (GPR) mapping contained the affected gene either alone or only in an "AND" relationship (*i.e.*, as an essential part of a protein complex); if the GPR listed the affected gene in an "OR" relationship (*i.e.*, as one of multiple isoenzymes or alternative transporters), the reaction was not affected.

We defined pleiotropy as the number of biomass components whose maximal production was reduced by at least 0.01% compared to the unperturbed state (WT) for the allele considered[26]. Thus, an allele not involved in the maximal production of any essential biomass component is considered to have pleiotropy 0; an allele that affects the production of exactly one essential biomass component has pleiotropy 1.

Our estimate of pleiotropy reflects the actual contribution of a gene product to biomass formation, based on estimated enzyme and transporter activities in the wildtype. If instead, one is interested in a quantitative measure of essentiality, defined as the number of biomass components affected by a deleterious allele *after* the mutant strain has been allowed to adapt its physiology to the gene deletion, a different algorithm is more appropriate. In this case, one needs to allow the free redistribution of fluxes after the simulated activity reduction of the protein encoded by the gene in question.

Statistical test for the relationship between pleiotropy at full gene knockout and number of steps. The full knockout (maximal) pleiotropy sets an upper limit to the possible number of pleiotropy steps at decreasing enzyme activity. The null hypothesis is that the number of steps is uniformly distributed between zero and maximal pleiotropy; *i.e.*, the null hypothesis assumes that apart from the upper limit, there is no systematic relationship between maximal pleiotropy and number of steps. We tested this through a randomization protocol, where we constructed 10^6 datasets with the same maximal pleiotropies, but with step numbers drawn from the corresponding uniform distributions. All random datasets for both the *E. coli* and the yeast data had Spearman rank correlation coefficients between pleiotropy at full knockout and number of steps that were lower than the observed correlation coefficients. Thus, the empirical *P*-value was $<10^{-6}$ for both data sets.

Currency metabolites. In additional analyses, we made several cofactors freely available to study how pleiotropy is associated with the generation of currency metabolites. We did this by adding a balanced biochemical reaction that interconverts the activated and inactivated versions of the cofactor and allowing unlimited flux of this reaction in both directions. For example, to simulate free NADPH, we added the following reversible reaction:

$$\mathrm{NADPH} \rightleftharpoons \mathrm{NADP}^{+} + \mathrm{H}^{+} + 2\mathrm{e}^{-}$$

A list of currency metabolites was obtained from ref.[43]. Supplementary Table S3 lists the currency metabolites and the corresponding exchange reactions as well as the number of reactions utilizing each currency metabolite.

Software used. All simulations were performed in R[47] using *sybil*, a computer library optimized for efficient constraint-based modeling of metabolic networks[48]. We used IBM ILOG CPLEX as the linear solver, connected to sybil via the cplexAPI R package.

To calculate network modularities, we used the LP&BRIM algorithm (Label Propagation with Bipartite Recursively Induced Modules) implemented in Matlab[49].

Data Availability

All input files, R scripts, and raw data used to generate the results and figures can be found on github at https://github.com/deyazoubi/pleiotropy-.git. An overview over the individual files is given in the Readme file.

References

1. Stearns, F. W. One Hundred Years of Pleiotropy: A Retrospective. *Genetics* **187**, 355–355 (2011).
2. Wagner, G. P. & Zhang, J. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat Rev Genet* **12**, 204–213 (2011).
3. Paul, D. A double-edged sword. *Nature* **405**, 515–515 (2000).
4. Williams, G. C. Pleiotropy, Natural-Selection, and the Evolution of Senescence. *Evolution* **11**, 398–411 (1957).
5. Zwaan, B. J. The evolutionary genetics of ageing and longevity. *Heredity* **82**, 589–597 (1999).
6. Moorad, J. A. & Promislow, D. E. L. What can genetic variation tell us about the evolution of senescence? *Proceedings of the Royal Society B-Biological Sciences* **276**, 2271–2278 (2009).
7. Paaby, A. B. & Schmidt, P. S. Dissecting the genetics of longevity in Drosophila melanogaster. *Fly (Austin)* **3**, 29–38 (2009).
8. Slatkin, M. Pleiotropy and Parapatric Speciation. *Evolution* **36**, 263–270 (1982).
9. Foster, K. R., Shaulsky, G., Strassmann, J. E., Queller, D. C. & Thompson, C. R. Pleiotropy as a mechanism to stabilize cooperation. *Nature* **431**, 693–696 (2004).
10. Tyler, A. L., Asselbergs, F. W., Williams, S. M. & Moore, J. H. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *Bioessays* **31**, 220–227 (2009).
11. Wright, S. *Evolution and the genetics of populations: a treatise in four volumes*. (University of Chicago Press, 1968).
12. Barton, N. H. Pleiotropic Models of Quantitative Variation. *Genetics* **124**, 773–782 (1990).
13. Fisher, R. A. *The genetical theory of natural selection*. (Clarendon Press, 1930).
14. Orr, H. A. Adaptation and the cost of complexity. *Evolution* **54**, 13–20 (2000).
15. Waxman, D. & Peck, J. R. Pleiotropy and the preservation of perfection. *Science (New York, N.Y.)* **279**, 1210–1213 (1998).
16. Otto, S. P. Two steps forward, one step back: the pleiotropic effects of favoured alleles. *Proc Biol Sci* **271**, 705–714 (2004).
17. Hodgkin, J. Seven types of pleiotropy. *Int J Dev Biol* **42**, 501–505 (1998).
18. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).
19. Albin, R. L. Antagonistic pleiotropy, mutation accumulation, and human genetic disease. *Genetica* **91**, 279–286 (1993).
20. Brunner, H. G. & van Driel, M. A. From syndrome families to functional genomics. *Nat Rev Genet* **5**, 545–551 (2004).
21. Zhang, J. & Wagner, G. P. On the definition and measurement of pleiotropy. *Trends Genet* **29**, 383–384 (2013).
22. Paaby, A. B. & Rockman, M. V. The many faces of pleiotropy. *Trends Genet* **29**, 66–73 (2013).
23. Dudley, A. M., Janse, D. M., Tanay, A., Shamir, R. & Church, G. M. A global view of pleiotropy and phenotypically derived gene function in yeast. *Molecular Systems Biology* **1**, 2005.0001 (2005).
24. Wagner, G. P. *et al.* Pleiotropic scaling of gene effects and the 'cost of complexity'. *Nature* **452**, 470–472 (2008).
25. Wang, Z., Liao, B. Y. & Zhang, J. Z. Genomic patterns of pleiotropy and the evolution of complexity. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 18034–18039 (2010).
26. Szappanos, B. *et al.* An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat Genet* **43**, 656–662 (2011).
27. Bajic, D., Moreno-Fenoll, C. & Poyatos, J. F. Rewiring of genetic networks in response to modification of genetic background. *Genome Biol Evol* **6**, 3267–3280 (2014).
28. Wang, Z. & Zhang, J. Z. Abundant Indispensable Redundancies in Cellular Metabolic Networks. *Genome Biology and Evolution* **1**, 23–33 (2009).
29. Watson, M. R. Metabolic Maps for the Apple-II. *Biochem Soc T* **12**, 1093–1094 (1984).
30. Orth, J. D., Thiele, I. & Palsson, B. O. What is flux balance analysis? *Nature Biotechnology* **28**, 245–248 (2010).
31. Stern, D. L. Evolutionary developmental biology and the problem of variation. *Evolution* **54**, 1079–1091 (2000).
32. Sivakumaran, S. *et al.* Abundant Pleiotropy in Human Complex Diseases and Traits. *American Journal of Human Genetics* **89**, 607–618 (2011).
33. Orth, J. D. *et al.* A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. *Mol Syst Biol* **7**, 535 (2011).

34. Aung, H. W., Henry, S. A. & Walker, L. P. Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism. *Ind Biotechnol (New Rochelle N Y)* **9**, 215–228 (2013).
35. Holzhutter, H. G. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *European Journal of Biochemistry* **271**, 2905–2922 (2004).
36. Xu, L., Barker, B. & Gu, Z. L. Dynamic epistasis for different alleles of the same gene. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 10420–10425 (2012).
37. Ishii, N. *et al*. Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science (New York, N.Y.)* **316**, 593–597 (2007).
38. Shlomi, T., Berkman, O. & Ruppin, E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci USA* **102**, 7695–7700 (2005).
39. Hadorn, E. *Developmental genetics and lethal factors*. (Methuen; Wiley, 1961).
40. He, X. & Zhang, J. Toward a molecular understanding of pleiotropy. *Genetics* **173**, 1885–1891 (2006).
41. Liu, X. & Murata, T. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* (IEEE, Milan, Italy, 2009).
42. Wang, Z. & Zhang, J. Z. In search of the biological significance of modular structures in protein networks. *Plos Computational Biology* **3**, 1011–1021 (2007).
43. Fritzemeier, C. J., Hartleb, D., Szappanos, B., Papp, B. & Lercher, M. J. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. *PLoS Comput Biol* **13**, e1005494 (2017).
44. Cameron, F. J. & Sinclair, A. H. Mutations in SRY and SOX9: Testis-determining genes. *Human Mutation* **9**, 388–395 (1997).
45. McKusick-Nathans Institute of Genetic Medicine - Johns Hopkins University (Baltimore MD). *Online Mendelian Inheritance in Man (OMIM)*, http://www.ncbi.nlm.nih.gov/omim/.
46. Shlomi, T. *et al*. Systematic condition-dependent annotation of metabolic genes. *Genome Research* **17**, 1626–1633 (2007).
47. The R Foundation. The R Project for Statistical Computing, https://www.r-project.org.
48. Gelius-Dietrich, G., Desouki, A. A., Fritzemeier, C. J. & Lercher, M. J. Sybil–efficient constraint-based modelling in R. *BMC Syst Biol* **7**, 125 (2013).
49. Flores, C. O., Poisot, T., Valverde, S. & Weitz, J. S. BiMat: a MATLAB package to facilitate the analysis of bipartite networks. *Methods Ecol Evol* **7**, 127–132 (2016).

Acknowledgements

We thank Balazs Papp and Jonathan Fritzemeier for helpful discussions. This work was supported financially through fellowships by the German Academic Exchange Service (DAAD) to DA and AAD, and through German Research Foundation (DFG) grants to MJL (CRC 680 and CRC 1310).

Author Contributions

D.A. and M.J.L.: conceptualization, methodology, writing. D.A.: investigation, software, validation. A.A.D.: assistance with metabolic modeling and software implementation. M.J.L.: funding acquisition, supervision.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-35092-1.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Information:

Alleles of a gene differ in pleiotropy, often mediated through currency metabolite production, in *E. coli* and yeast metabolic simulations

Deya Alzoubi, Abdelmoneim Amer Desouki, Martin J. Lercher

Contents

Supplementary Figures

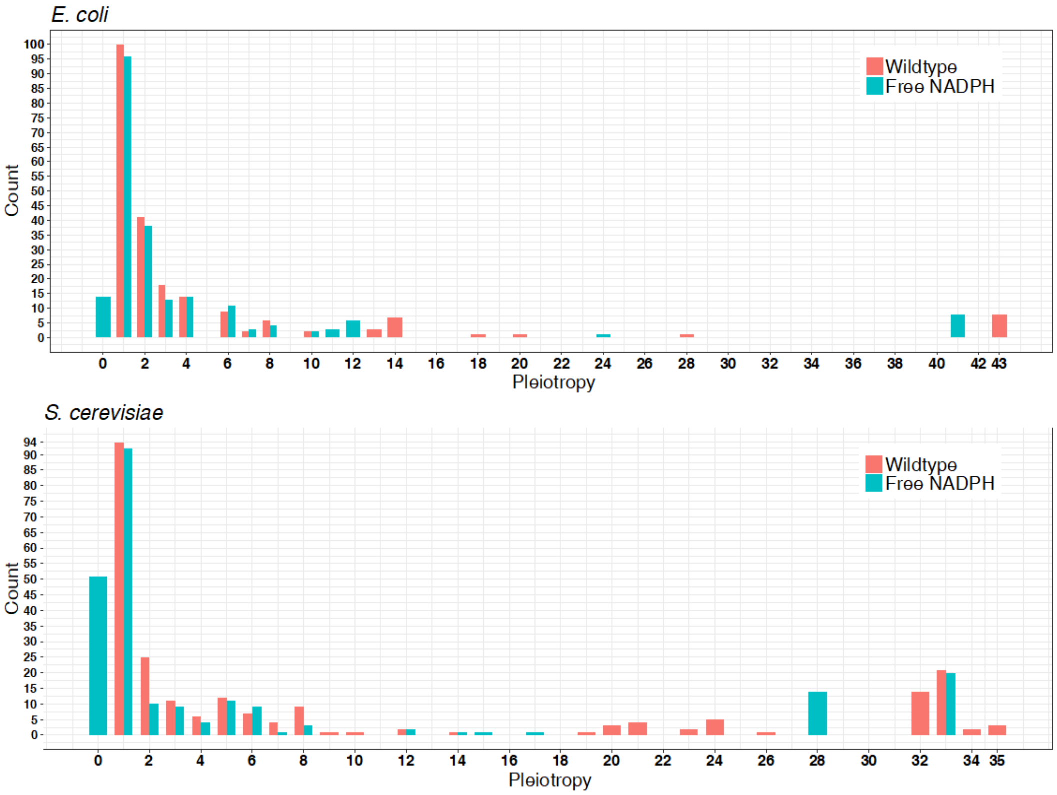
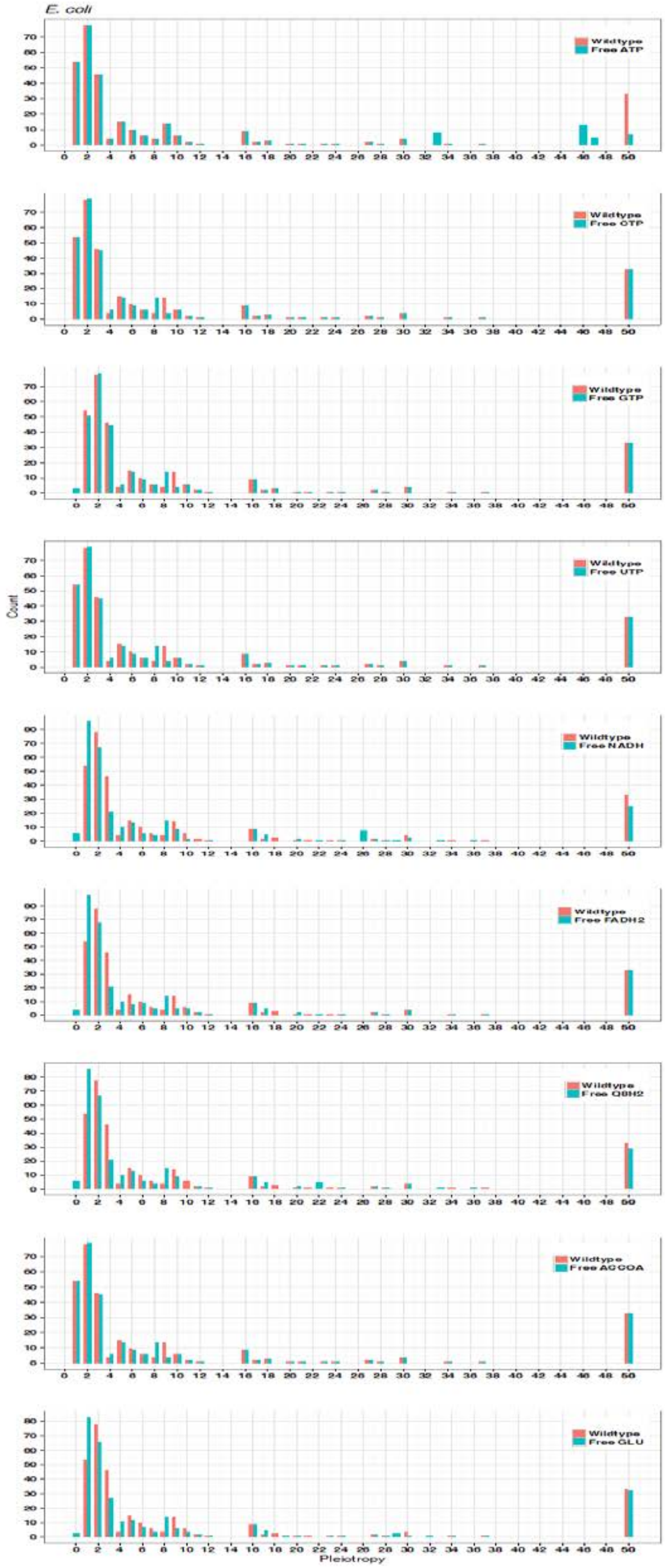


Figure S1: Distribution of the number of biomass components for which genes are essential even after allowing the mutant strain to adapt its protein expression (excluding genes with no effect on biomass production). The number of biomass components for which a given gene is essential is often reduced when NADPH is made freely available (cyan bars).


E. coli
Wildtype
Free ATP
Wildtype
Free CTP
Wildtype
Free GTP
Wildtype
Free UTP
Wildtype
Free NADH
Wildtype
Free FADH2
Wildtype
Free Q8H2
Wildtype
Free ACCOA
Wildtype
Free GLU
Count
0 10 20 30 40 50 60 70 80
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50
Pleiotropy

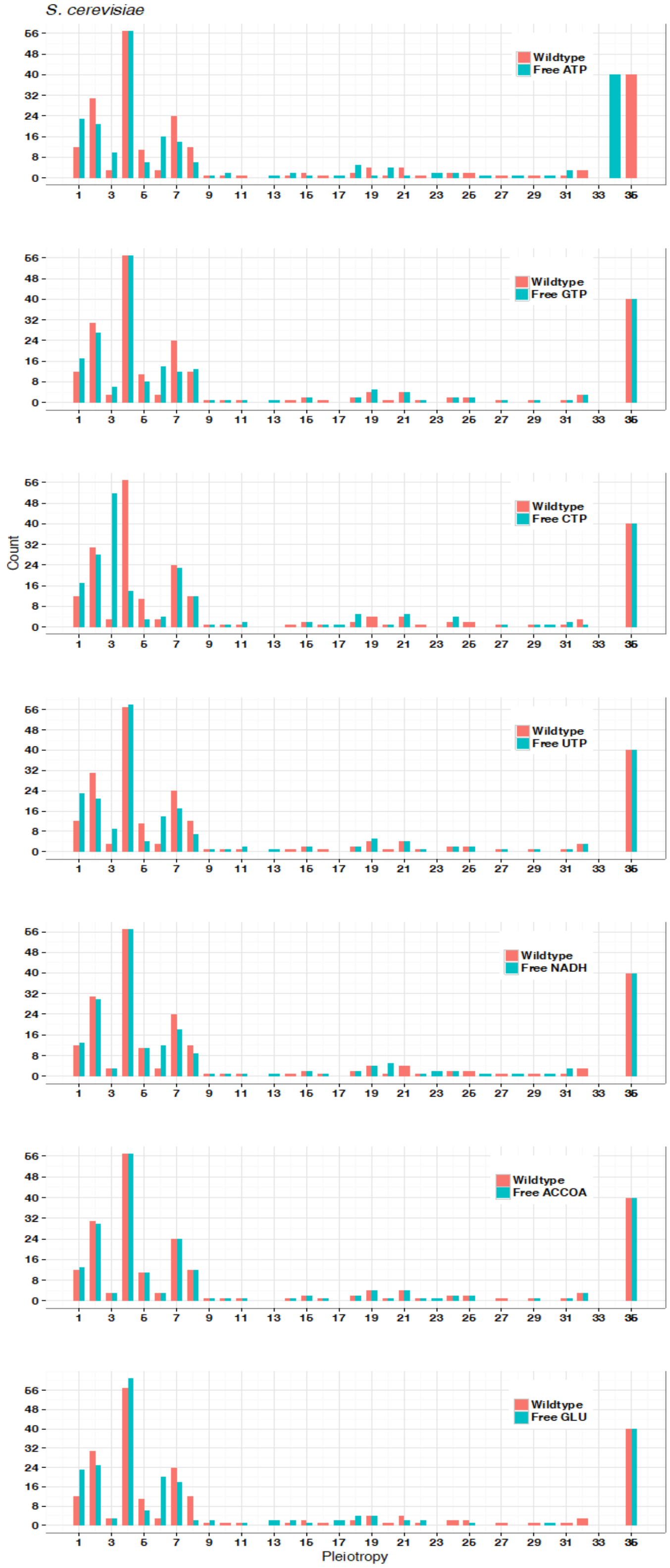


Figure S2. Distribution of pleiotropy for full gene knockouts. Pleiotropy is reduced for many *E. coli* and *S. cerevisiae* genes when different currency metabolites (one per panel) are made freely available.

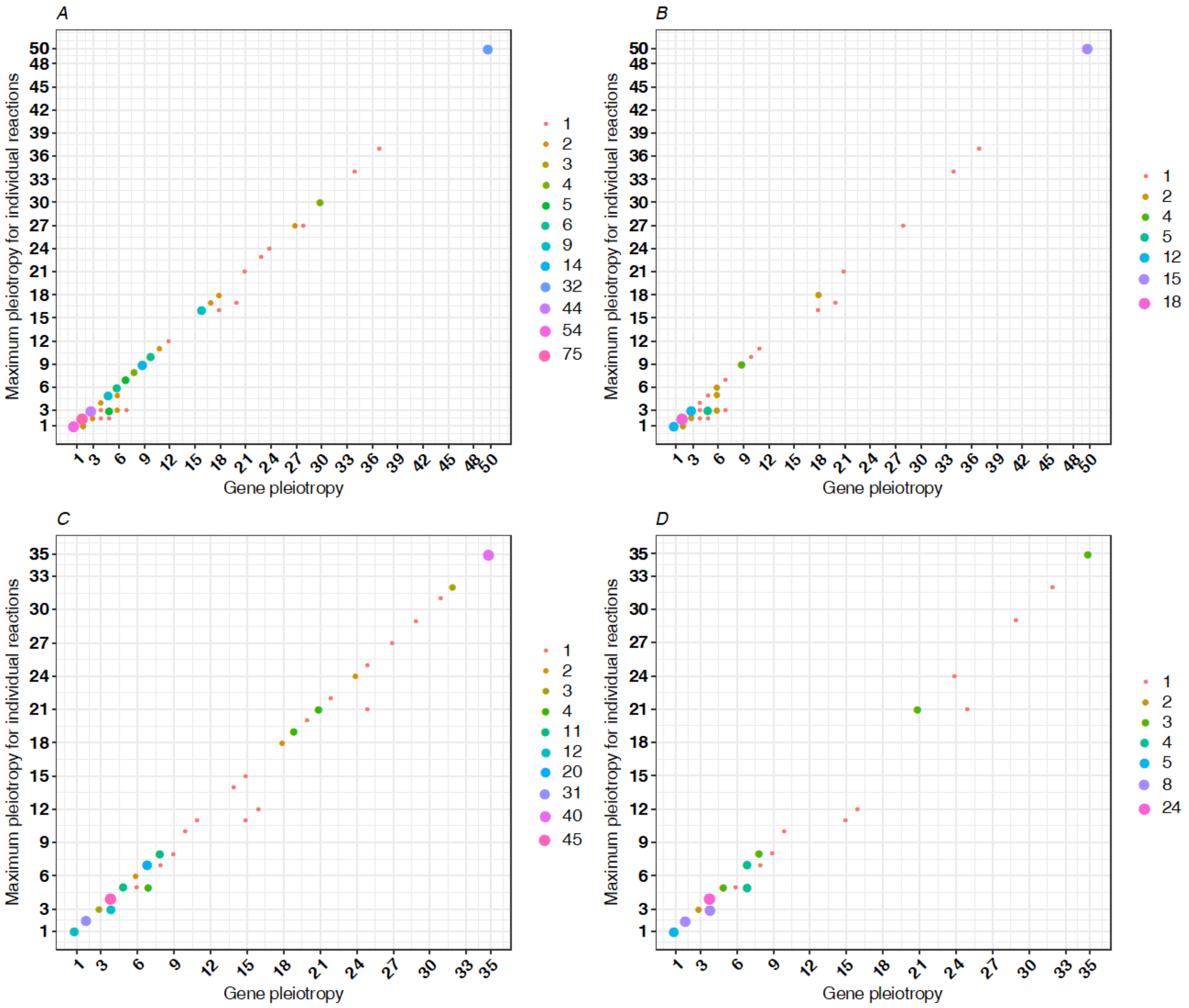


Figure S3. Comparison of gene pleiotropy (the number of biomass components affected by a full gene knockout) to the maximal pleiotropy observed when blocking the individual reactions for which this gene is essential. (A) *E. coli*, all genes; (B) *E. coli*, only genes essential for ≥2 reactions; (C) *S. cerevisiae*, all genes; (D) *S. cerevisiae*, only genes essential for ≥2 reactions. Dot size and color indicates the number of genes represented by each dot.

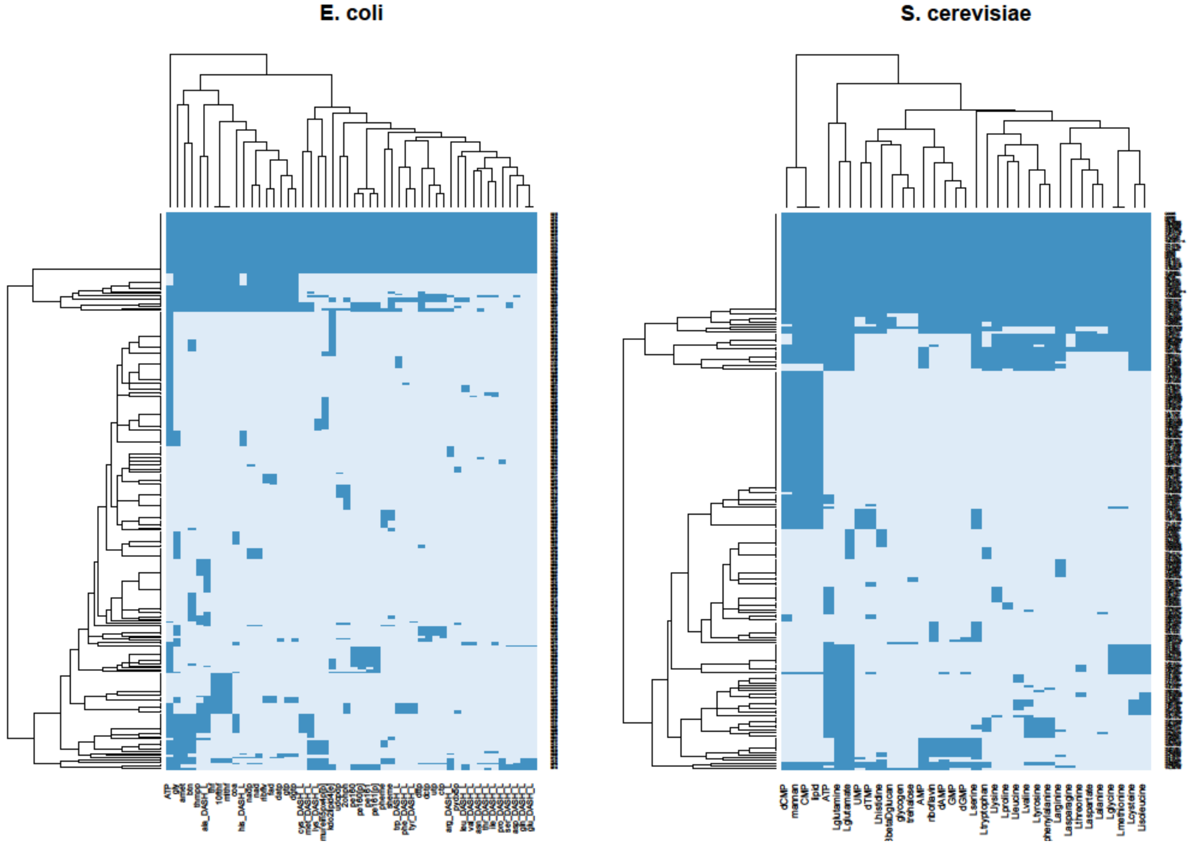


Figure S4. Graphical representation of the modularity of the bipartite pleiotropy networks for *E. coli* and *S. cerevisiae*. The figures show heatmaps produced with the R function of the same name, which order biomass components (columns) and genes (rows) through hierarchical clustering. Dark blue means that the biomass component's production is decreased through a knockout of the gene, light blue means it is not. Modularity shows up as blocks of dark blue, where several rows of gene affect the same neighbouring columns of traits.

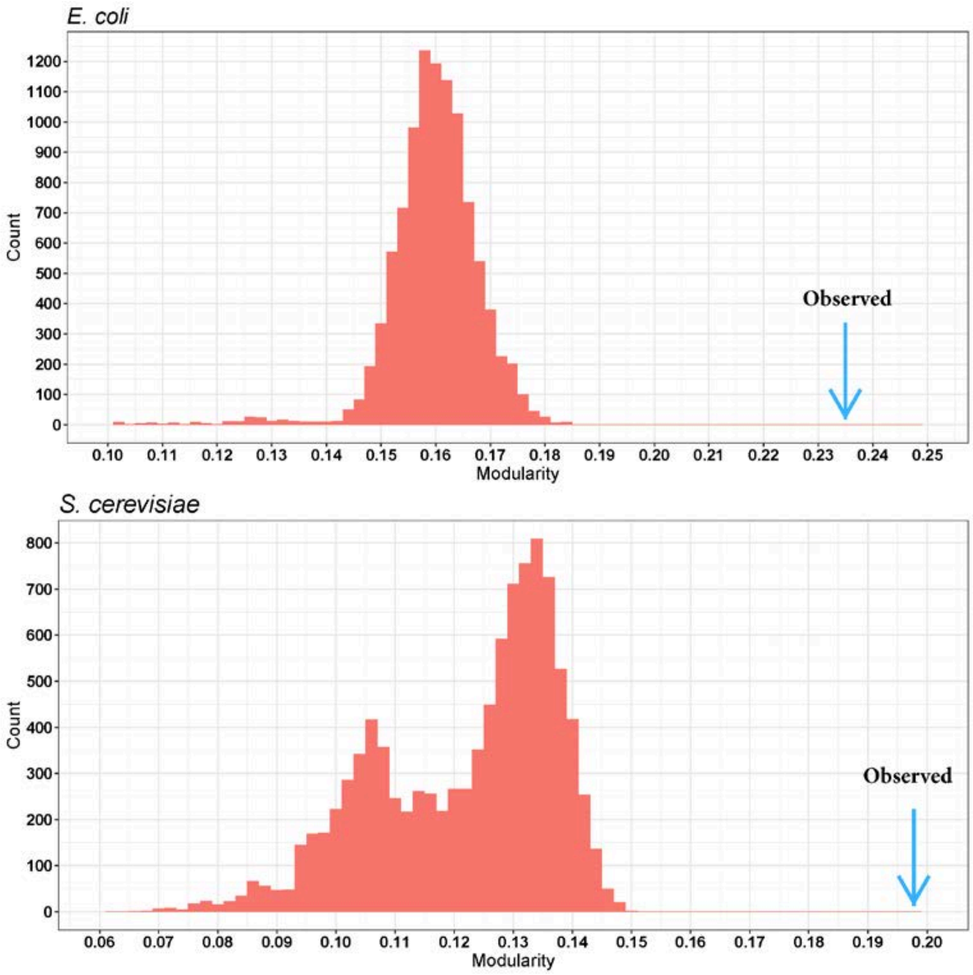
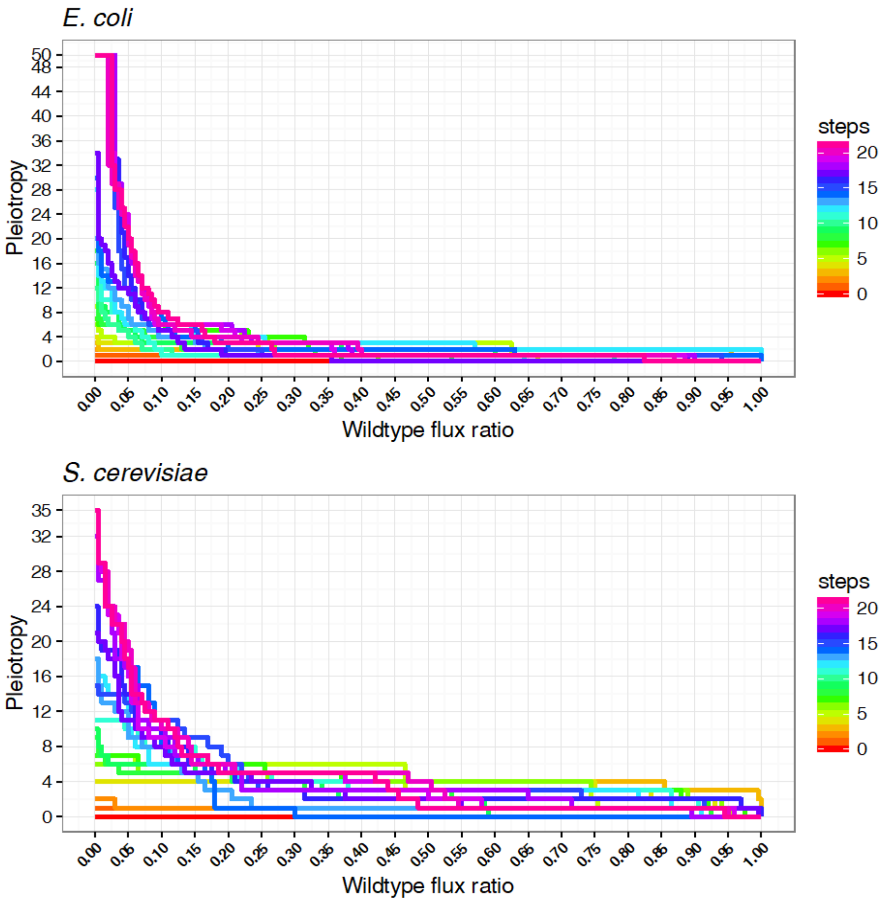


Figure S5. Observed modularity *Q* (as calculated with LP&BRIM, blue arrow) compared to the distribution of modularities obtained for pleiotropy networks with randomized links between genes and biomass components (red).



E. coli

S. cerevisiae

Figure S6. The pleiotropy of genes contributing to biomass production typically increases for increasingly debilitating mutations. Shown are curves for one randomly chosen gene for each value of the number of pleiotropy increases (steps).

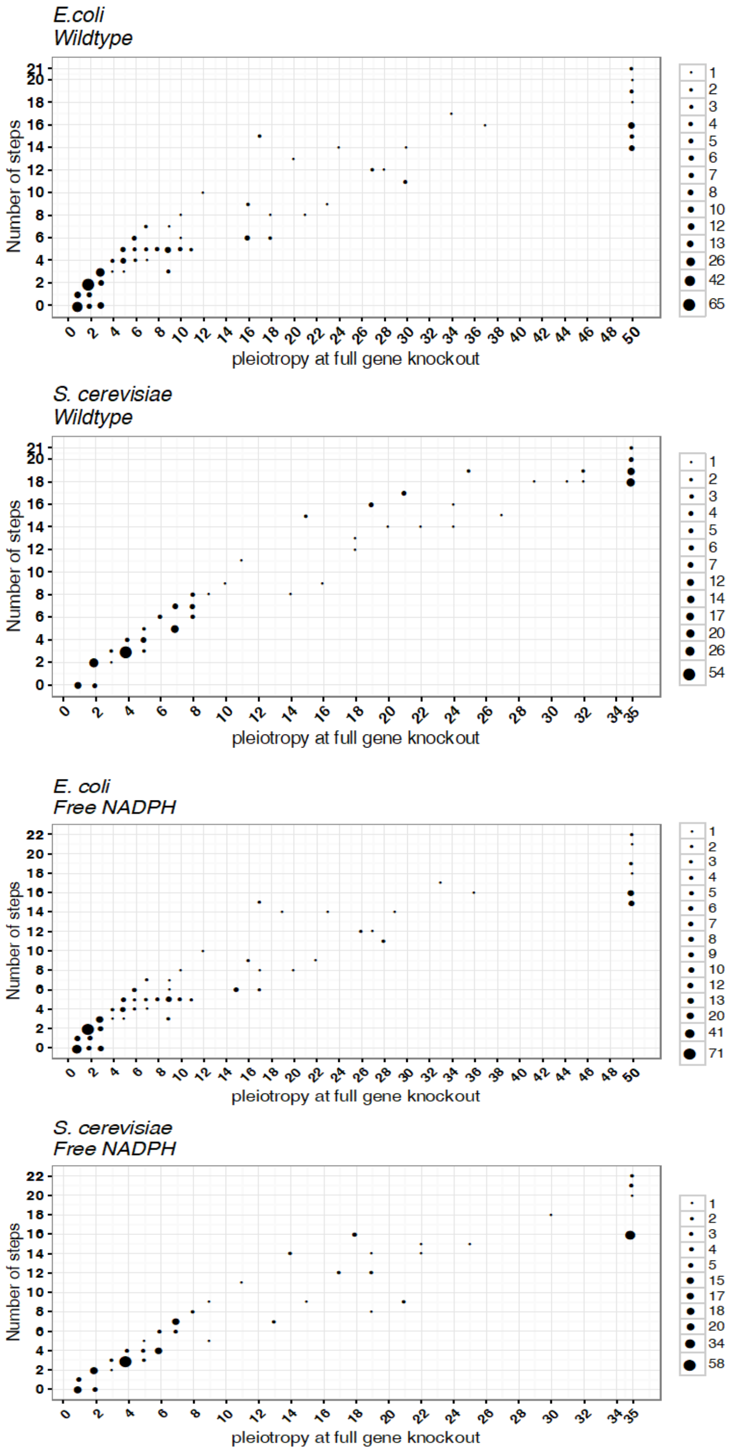


Figure S7. The number of step-wise increases in pleiotropy for increasingly debilitating mutations and the pleiotropy at full gene knockout are strongly correlated (Spearman's rank correlation when considering only genes contributing to biomass production, wildtype: *ρ*=0.926 (*E. coli*) and *ρ*=0.986 (*S. cerevisiae*); when making NADPH freely available: *ρ*=0.922 (*E. coli*) and *ρ*=0.986 (*S. cerevisiae*)). The diameter of each point is proportional to the number of genes with this combination of pleiotropy and step number.

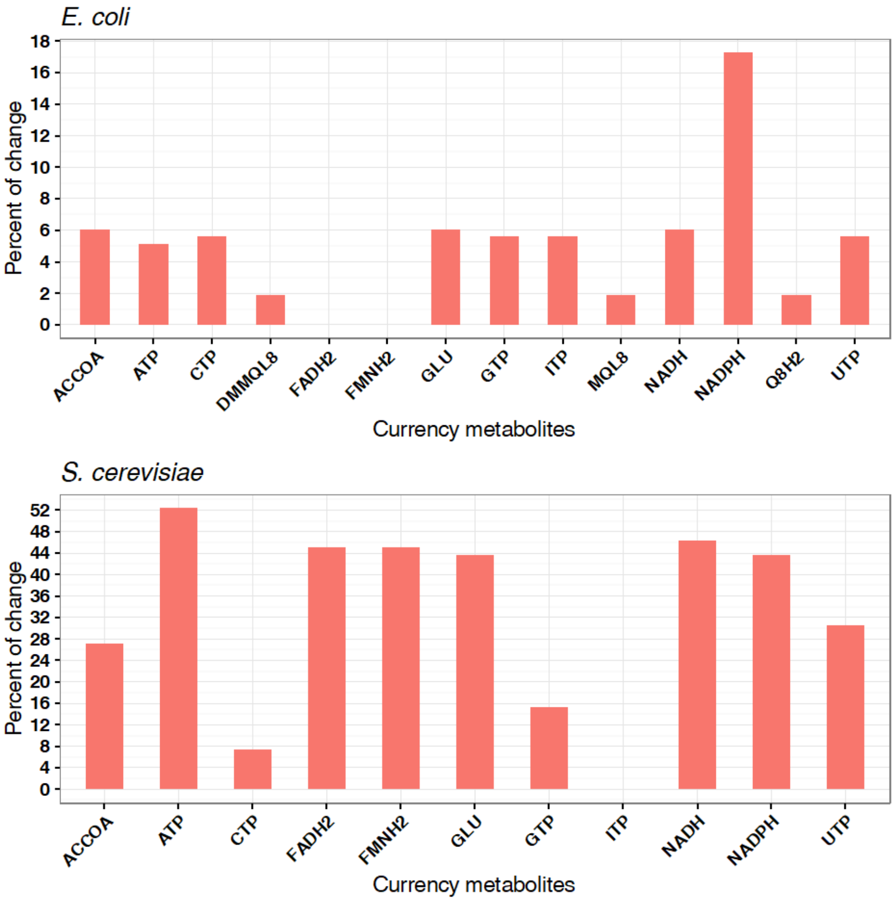


Figure S8. Percentage of genes for which free availability of a given currency metabolite reduces the number of biomass components for which this gene is essential even after allowing the mutant strain to adapt its protein expression to the altered gene content of its genome. Abbreviations: Adenosine triphosphate (ATP); Cytidine triphosphate (CTP); Guanosine triphosphate (GTP); Uridine triphosphate (UTP); Inosine triphosphate (ITP); Nicotinamide adenine dinucleotide (NADH); Nicotinamide adenine dinucleotide phosphate (NADPH); Flavin adenine dinucleotide reduced (FADH2); Reduced flavin mononucleotide (FMNH2); Ubiquinol-8 (Q8H2); Menaquinol 8 (MQL8); 2-Demethylmenaquinol 8 (DMMQL8); Acetyl-CoA (ACCOA); L-Glutamate (GLU).

Supplementary Tables

Table S1. Essential Biomass components in *E. coli* (excluding inorganic ions, H_2O, and products of the biomass reaction)

Components	
ala_DASH_L[c]	gtp[c]
arg_DASH_L[c]	utp[c]
asn_DASH_L[c]	murein5px4p[p]
asp_DASH_L[c]	kdo2lipid4[e]
cys_DASH_L[c]	pe160[c]
gln_DASH_L[c]	pe160[p]
glu_DASH_L[c]	pe161[c]
gly[c]	pe161[p]
his_DASH_L[c]	10fthf[c]
ile_DASH_L[c]	2ohph[c]
leu_DASH_L[c]	amet[c]
lys_DASH_L[c]	btn[c]
met_DASH_L[c]	coa[c]
phe_DASH_L[c]	fad[c]
pro_DASH_L[c]	mlthf[c]
ser_DASH_L[c]	nad[c]
thr_DASH_L[c]	nadp[c]
trp_DASH_L[c]	pheme[c]
tyr_DASH_L[c]	pydx5p[c]
val_DASH_L[c]	ribflv[c]
datp[c]	sheme[c]
dctp[c]	thf[c]
dgtp[c]	thmpp[c]
dttp[c]	udcpdp[c]
ctp[c]	atp[c]

Table S2. Essential Biomass components in *S. cerevisiae* (excluding inorganic ions, H_2O, and products of the biomass reaction)

Components	
ATP [cytoplasm]	L-proline [cytoplasm]
(1->3)-beta-D-glucan [cytoplasm]	L-phenylalanine [cytoplasm]
lipid [cytoplasm]	L-tyrosine [cytoplasm]
mannan [cytoplasm]	L-histidine [cytoplasm]
glycogen [cytoplasm]	UMP [cytoplasm]
L-alanine [cytoplasm]	AMP [cytoplasm]
L-glycine [cytoplasm]	GMP [cytoplasm]
L-glutamate [cytoplasm]	CMP [cytoplasm]
L-glutamine [cytoplasm]	L-methionine [cytoplasm]
L-valine [cytoplasm]	L-cysteine [cytoplasm]
L-serine [cytoplasm]	L-tryptophan [cytoplasm]
L-leucine [cytoplasm]	trehalose [cytoplasm]
L-lysine [cytoplasm]	dAMP [cytoplasm]
L-threonine [cytoplasm]	dTMP [cytoplasm]
L-asparagine [cytoplasm]	dCMP [cytoplasm]
L-aspartate [cytoplasm]	dGMP [cytoplasm]
L-isoleucine [cytoplasm]	riboflavin [cytoplasm]
L-arginine [cytoplasm]	

Table S3. Currency metabolites and the corresponding supply reactions

Name	Abbrev.	Chemical equation	*E. coli* model equation	#Reactions in *E. coli*	*S. cerevisiae* model equation	#Reactions in *S. cerevisiae*
Adenosine triphosphate	ATP	ATP + H2O --> ADP + H(+) + Phosphate	atp[c] + h2o[c] --> adp[c] + h(+)[c] + pi[c]	359	0434 +0803 --> 0394 + 0794 + 1322	158
Cytidine triphosphate	CTP	CTP + H2O --> CDP + H(+) + Phosphate	ctp[c] + h2o[c] --> cdp[c] + h(+)[c] + pi[c]	18	0539 +0803 --> 0467 + 0794 + 1322	13
Guanosine triphosphate	GTP	GTP + H2O --> GDP + H(+) + Phosphate	gtp[c] + h2o[c] --> gdp[c] + h(+)[c] + pi[c]	25	0785 + 0803 --> 0739 +0794 + 1322	13
Uridine triphosphate	UTP	UTP + H2O --> UDP + H(+) + Phosphate	utp[c] + h2o[c] --> udp[c] + h(+)[c] + pi[c]	8	1559 + 0803 --> 1538 + 0794 + 1322	11
Inosine triphosphate	ITP	ITP + H2O --> IDP + H(+) + Phosphate	itp[c] + h2o[c] --> idp[c] + h(+)[c] + pi[c]	4	0950 + 0803 --> 0846 + 0794 + 1322	3
Nicotinamide adenine dinucleotide	NADH	Nicotinamide adenine dinucleotide - reduced --> H(+) + Nicotinamide adenine dinucleotide	nadh[c] --> h(+)[c] + nad[c]	119	1203 --> 0794 + 1198	36
Nicotinamide adenine dinucleotide phosphate	NADPH	Nicotinamide adenine dinucleotide phosphate – reduced --> H(+) + Nicotinamide adenine dinucleotide phosphate	nadph[c] --> h(+)[c] + nadp[c]	97	1212 --> 0794 + 1207	58
Flavin adenine dinucleotide reduced	FADH2	Flavin adenine dinucleotide reduced --> 2 H(+) + Flavin adenine dinucleotide oxidized	fadh2[c] --> 2 h(+)[c] +fad[c]	25	0689 --> 2 (0794) + 0687	2
Reduced flavin mononucleotide	FMNH2	Reduced FMN --> 2 H(+) + FMN	fmnh2[c] --> 2 h(+)[c] + fmn[c]	13	0717 --> 2 (0794) + 0714	3
Ubiquinol-8	Q8H2	Ubiquinol-8 --> 2 H(+) + Ubiquinone-8	q8h2[c] --> 2 h(+)[c] + q8[c]	24	NA	NA
Menaquinol 8	MQL8	Menaquinol 8 --> 2 H(+) + Menaquinone 8	mql8[c] --> 2 h(+)[c] + mqn8[c]	25	NA	NA
2-Demethylmenaquinol 8	DMMQL8	2-Demethylmenaquinol 8 --> 2 H(+) + 2-Demethylmenaquinone 8	2dmmql8[c] --> 2 h(+)[c] + 2dmmq8[c]	14	NA	NA
Acetyl-CoA	ACCOA	H2O + Acetyl-CoA --> H(+) + Acetate + Coenzyme A	h2o[c] + accoa[c] --> h(+)[c] + ac[c] + coa[c]	37	0803 + 0373 --> 0794 + 0362 + 0529	44
L-Glutamate	GLU	L-Glutamate + H2O --> 2-Oxoglutarate + Ammonium + 2 H(+)	glu_dash_l[c] + h2o[c] --> akg[c] +nh4[c] + 2h(+)[c]	49	0991 + 0803 --> 0180 + 0419 + 2 (0794)	44

Manuscript 2: Epistasis predictions from flux balance analysis with molecular crowding

Status

This manuscript is currently under evaluation at the journal *Scientific Reports.*

Authors: Deya Alzoubi, Abdelmoneim Amer Desouki & Martin J. Lercher

Title: Epistasis predictions from flux balance analysis with molecular crowding

Contributions

Together with Martin Lercher, I developed the concept and methodology of this work and wrote the manuscript. I wrote the software for simulations and analyses and performed all analyses.

Manuscript

(see next page)

Epistasis predictions from flux balance analysis with molecular crowding

Deya Alzoubi[1], Abdelmoneim Amer Desouki[1], Martin J. Lercher[1,*]

[1] Institute for Computer Science and Department of Biology,

Heinrich Heine University, Universitätsstraße 1, Düsseldorf D-40221, Germany

* To whom correspondence should be addressed

Corresponding author: Martin Lercher, Institute for Computer Science,

Heinrich Heine University, D-40225 Düsseldorf, Germany, martin.lercher@hhu.de,

+49 151 22964073

Email addresses: DA deya.alzoubi@hhu.de; AAD abdelmoneim.desouki@hhu.de;

MJL martin.lercher@hhu.de

Abstract

The computational prediction of double gene knockout effects by flux balance analysis (FBA) has been used widely to characterize genome-wide patterns of epistasis in microorganisms. However, is it unclear how *in silico* epistasis predictions are related to *in vivo* epistasis, as FBA was unable to predict the vast majority of observed genetic interactions in a high-throughput experiment that generated double knockouts of non-essential metabolic genes in yeast. It has been proposed that FBA predictions, which are based purely on metabolic network stoichiometry and on the assumption of maximal biomass yield, can be improved by incorporating approximate enzyme kinetics and a constraint on macromolecular crowding. Here, we test if FBA with molecular crowding (ccFBA) can predict previously unexplained epistatic interactions in yeast. We find that while FBA with molecular crowding predicts some positive epistatic interactions not detectable with alternative constraint-based methods, more than 70% of epistatic interactions are undetectable by any of the widely used constraint-based methods.

Introduction

Epistasis measures the extent to which the consequences of a mutation in one gene depend on mutations in another gene[1]. Epistasis is said to be negative (aggravating) if the double mutant has lower fitness than expected, *i.e.*, if its fitness is lower than the product of the single-mutant fitnesses; epistasis is called positive (alleviating) if the double mutant has higher fitness. Understanding the distribution of epistasis is fundamental to our understanding of gene function and interaction[2-4]. Epistasis is important for a wide range of theoretical issues in biology, including the evolution of sex[5,6], speciation[7], ploidy[8], mutation load[9], and genetic buffering[10]; epistasis is also fundamental to our understanding of as human disease[11,12] and drug resistance[13].

Epistasis can be assayed experimentally through the analyis of double gene knockouts[14-23]. However, such experiments are technically demanding, and the number of possible interactions grows quadratically with genome size. An attractive alternative to the generation of experimental knockouts for all possible gene combinations is the *in silico* prediction of double gene knockout effects. One approach towards the computational prediction of epistasis uses machine learning based on various experimentally observed gene and gene pair properties; Table 1 of Ref.[24] provides on overview over such predictions.

Here, we will focus instead on prediction methods based on *in silico* models of gene function, which are inherently more suited to generate increased biological understanding. Epistasis is a property of functional links between genes, not of individual genes. Thus, large-scale predictions of epistasis from first principles are only possible with computational models that account for functional connections between gene products. The best-studied complex biological system is metabolism. Excellent representations of metabolic networks have been compiled for several unicellular organisms such as *E. coli*[25] and the baker's yeast *Saccharomyces cerevisiae*[26]. So far, all attempts at genome-scale *in silico* epistasis prediction[27-34] have used flux balance analysis (FBA), which maximizes the yield of biomass production in the wild-type and in the mutants[35,36], or a variant of FBA that attempts to minimize the difference between wild-type and knockout distributions of metabolic reaction rates (minimization of metabolic adjustment, MOMA[37]).

Several studies used these simulation methods to perform large-scale characterizations of epistasis *in silico.* Segrè et al. first used FBA to study the spectrum of epistatic interactions between metabolic genes in *S. cerevisiae*[27]. These authors introduced a new concept of epistasis between functional modules rather than between individual genes, intended to describe functional relationships among metabolic pathways. They found that modules interact with each other 'monochromatically', *i.e.*, epistatic interactions between two specific modules are either largely positive or largely negative[27]. Examining the metabolic networks of

E. coli and *S. cerevisiae*, He *et al.*[28] found negative epistatic interactions largely among nonessential reactions with overlapping functions; in contrast, positive interactions were found predominantly between reactions without overlapping functions, and these were frequently essential[28].

Snitkin *et al.*[29] studied epistatic interactions between yeast gene deletions based on their influence on the reaction rates of individual enzymatic reactions. They found that gene pairs interact incoherently relative to different phenotypes, and that genes involved in many genetic interactions across multiple phenotypes tend to be highly expressed, to evolve slowly, and to be associated with human diseases[29].

Xu *et al.*[30] compared epistatic interactions for different alleles of the same gene; alleles of different enzymatic activities were simulated by reducing the admissible flux (reaction rate) relative to the wild-type by a given percentage. They found that different alleles of the same gene typically interact with very different gene sets *in silico*; they argued that the distribution of the sign of epistasis in their simulations can speed up the purging of deleterious mutations in eukaryotes[30].

Finally, Barker *et al.*[31] studied epistatic relationships between genes under various environments, finding that epistatic interactions can differ substantially between growth conditions and that the epistasis network structure differs fundamentally between condition-independent (stable) and condition-dependent interactions[31].

While these *in silico* analyses of epistatic landscapes purport to fundamentally advance our understanding of epistasis in nature, it is not clear that *in silico* and *in vivo* epistasis are correlated sufficiently on the genome-scale to allow such conclusions. Several experimental platforms for the high-throughput detection of epistasis have been developed, among them synthetic genetic arrays (SGA)[15,23] , diploid-based synthetic lethality analyses with microarrays[16,19], synthetic dosage-suppression and lethality screens[14,17,18], and epistatic miniarray profiles[20-22]. The most comprehensive estimates of epistasis are available for the baker's yeast *Saccharomyces cerevisiae*[23,33], obtained through SGA.

Synthetic lethality – an extreme case of epistasis – was successfully predicted for some genes using FBA already in 2007; however, these authors could correctly predict only 7 out of 29 previously described synthetic lethals, corresponding to a recall of only 24%[32]. Two further studies in 2015 compared FBA predictions of synthetic lethality to experimental observations in yeast[38] and *E. coli*[39], confirming that only a minority of observed synthetic lethal interactions can be predicted successfully.

Szappanos *et al.*[33] were the first to compare quantitative epistasis predictions from FBA and MOMA with high-throughput experimental data, examining 67,517 pairs of non-essential yeast genes (high-confidence empirical interactions from SGA). They also found that only a minority of empirically observed interactions can be successfully predicted. For negative epistatic interactions, at 45% precision (percentage of predicted interactions that are indeed experimentally observed), they obtained a recall (percentage of observed interactions that are correctly predicted) of 2.8%. While the recall can be increased to slightly above 4% by lowering the prediction threshold, this comes at the cost of many false positive predictions, associated with a drastic reduction of precision to below 6%. For positive interactions, Szappanos *et al.* obtained a recall of 12.9% at a precision of around 10%, which could not be improved much further by lowering the prediction threshold. The quality of predictions could only be improved marginally by an automated model refinement procedure[33]. These results suggest that the physiological responses of yeast to double gene knockouts are not sufficiently captured by computational methods based on yield maximization such as FBA and MOMA. A later study that calculated epistasis from a new "function-loss cost" metric did not result in significantly improved predictions of the same data[34].

Why do the methods tested – FBA and MOMA – perform so poorly when predicting epistatic interactions? FBA captures epistasis based on the maximal biomass yield of the single and double mutants. MOMA assumes that the redistribution of reaction fluxes relative to the FBA wild-type solution is minimized upon the genetic perturbation[37]. Both FBA and MOMA predictions ignore the protein cost of enzymatic reactions, which arises from the necessary investment of cellular currencies, such as ATP and carbon, into enzyme production. Furthermore, it has been suggested that enzymes and the protein translation apparatus compete for the limited intracellular concentration space, a suggestion consistent with the observation that total cellular protein concentrations appear to be approximately constant across conditions[40]. In particular the latter constraint, summarized under the term (macro-) molecular crowding, has been explored in detail in the literature[41-43]. Instead of a largely arbitrary constraint on the uptake of a limiting nutrient, FBA models with molecular crowding limit cell growth by imposing a maximal mass concentration of enzymes, which in turn limits the total flux through the reactions the enzymes catalyze. Note that FBA and related constraint-based models do not consider internal metabolite concentrations explicitly, and thus FBA with molecular crowding methods calculate the enzyme concentration necessary for a given reaction flux v as $[E] = k_{eff}\ v$, with a ocnstant effective rate constant that is often approximated through the enzyme turnover number k_{cat}[41,43].

Could molecular crowding be responsible for epistatic interactions? FBA considers different yields of pathways, but pathways also differ in their kinetics, such that the same overall flux may require much more protein investment in one pathway compared to an alternative pathway; such differences in pathway costs of fluxes are believed to be the origin of overflow metabolism[44,45]. Accordingly, the fitness effect of a non-essential enzyme knockout will depend not only on the stoichiometry of the catalyzed reaction (which is what FBA considers), but also on the enzyme's kinetics.

Two toy examples for positive and negative epistasis are given in Fig. 1. If multiple isoenzymes or pathways can convert metabolite A into B (Fig. 1a), then FBA will predict that the corresponding single and double knockouts are all without fitness effect. However, if the isoenzymes and pathways differ in the protein cost per catalyzed flux, then a double knockout involving the most efficient enzymes will result in a reduced total flux, unless protein investment into the remaining pathway is increased at the cost of reduced investment into other pathways that contribute to biomas. The least effective pathway is utilized only in the double knockout, and this will result in negative epistasis. Positive epistasis may arise, *e.g.*, if two pathways are coupled by a downstream enzyme that jointly uses the products of both pathways as substrates (Fig. 1b). If there exists a catalytically less efficient alternative pathway for each of the two inputs, then the double knockout of the two efficient pathways will result – at identical protein investment – in a flux that is identical to the lower flux of the two single knockouts.

Previous applications of MOMA[37] suffer from a second problem. FBA solutions are generally redundant, *i.e.*, multiple flux distributions lead to the same biomass yield. Thus, the distance of the MOMA to the FBA flux distribution may depend strongly on the particular FBA solution returned by the numerical solver of the wild-type optimization problem. A straightforward possibility to rectify this problem is to use the wild-type flux distribution returned by parsimonious FBA (pFBA), which attempts to minimize protein investment at a given biomass yield[46] and has been shown to perform well in predicting the effects of single gene knockouts[47].

Here, to test if the poor performance of previous *in silico* predictions of epistasis[33,34] can be improved by correcting the shortcomings discussed above, we compare epistasis predictions from (i) FBA with molecular crowding and (ii) MOMA starting from the pFBA solution to the double gene knockout data for yeast in Ref.[33].

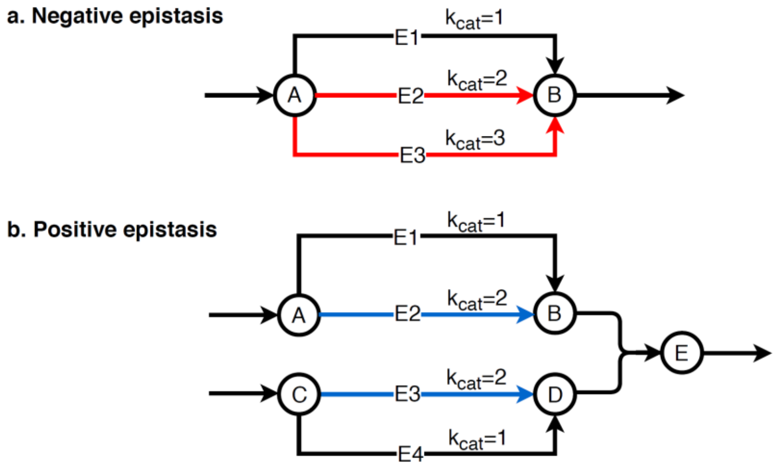


Figure 1. Toy examples of epistatic interactions that arise because of different enzymatic costs of pathways. **a**. Negative epistasis between E2 and E3. b. Positive epistasis between E2 and E3. The example assumes equal protein costs for all enzymes.

Materials and methods

Experimental data

We used a high-confidence subset of *S. cerevisiae* epistasis data for metabolic genes identified in Szappanos *et al.*[33]. This data was generated using synthetic genetic array (SGA) screens. We excluded genes deemed to be essential by the metabolic model or that are blocked in the model. This resulted in 291 negative and 123 positive interactions among 71,994 non-essential gene pairs.

Metabolic models and Media

To model *S. cerevisiae* metabolism, we used the metabolic reconstruction yeast7.6 (https://sourceforge.net/projects/yeast)[48]. Following the authors of Ref.[33], we removed a set of genes from the metabolic model (CAN1, LYP1, URA3, LEU2, MET17) to mimic the strain background used in the experiments; we also used the same definition of the growth medium, which mimics the experimental conditions[33]. The resulting, strain-specific model encompasses 904 metabolic genes associated with 3,326 reactions.

We performed all simulations using *sybil*, a computer library for efficient modelling of metabolic networks[49] in R[50]. Among other methods, sybil implements FBA, pFBA (minimization of total flux, MTF), and diverse methods for genome-scale simulation of genetic perturbations.

Flux balance analysis (FBA)

FBA identifies a flux distribution across the metabolic network that maximizes biomass yield under the constraints given by (i) the stoichiometry of enzymatic and transport reactions and (ii) lower and upper bounds on individual fluxes. The upper bounds on individual enzymatic fluxes are meant to reflect maximal enzyme capacity, and hence FBA could in principle also take enzyme kinetics into account; however, as enzyme capacities are generally unknown, the upper bounds are typically set to a value that is effectively infinite. Lower bounds on individual enzymatic reactions are set to zero for reactions deemed irreversible, and are (effectively) set to negative infinity for reversible reactions. Bounds on exchange reactions reflect maximal nutrient uptake or excretion rates. To estimate epistasis with FBA, we need to calculate the maximal biomass production yield of the double gene knockout, v_{12} , and the two single gene knockouts, v_1 and v_2 ; in each case, all fluxes through reactions for which one of the knockouts is essential are forced to zero. We convert the biomass yield values to fitness estimates by dividing them by the wild-type biomass yield, v_{WT}: $W_i = v_i / v_{WT}$. The fitness of the single and double mutants then allows the calculation of epistasis as[33]:

$$\varepsilon \coloneqq W_{12} - W_1 \times W_2 \tag{1}$$

Minimization of metabolic adjustment (MOMA)

MOMA is an extension of FBA for the prediction of gene knockouts. MOMA employs quadratic programming to identify the closest point (in terms of its Euclidean distance) in the permissible flux space of the knockout to the wild-type flux[37]. Previous applications of MOMA to epistasis predictions minimized the distance to an arbitrary FBA solution returned by the linear solver of the FBA problem[33]. As FBA flux distributions are highly degenerate, we instead use the parsimonious FBA (pFBA or MTF) solution to the wild-type problem[46], which should lead to biologically more relevant results[47]. Following previous applications[33], we minimize the Manhattan rather than Euclidean distance between wild-type and knockout flux distributions, which results in a linear optimization problem (lMOMA). As for FBA, epistasis was then estimated from the difference between the double knockout fitness and the product of the single knockout fitnesses (Eq. (1)).

Cost-constrained FBA (ccFBA)

ccFBA is a general implementation of FBA with molecular crowding[33,41-43], which comes with an existing parameterization for yeast[51]. It is implemented in R[50] and builds on the sybil package[49]. ccFBA improves on the MOMENT methodology for molecular modeling with enzyme kinetics[43] by explicitly considering multifunctional enzymes. Put simply, ccFBA extends FBA by adding a global constraint on the total mass concentration (assumed to be proportional to volume concentration) of enzymes:

$$\textstyle\sum_i [E_i]\, m_i \leq C \tag{2}$$

where the sum runs over all enzymes (or enzyme complexes) i, $[E_i]$ is the molar concentration of enzyme i per gram dry weight, m_i is the molar mass of the enzyme, and C is an upper limit on the total enzyme mass per gram dry weight. As the maximal flux through a reaction is constrained by the corresponding enzyme concentration and turnover number, $v_i \leq k_{cat,i} E_i$, Eq. (2) represents an additional linear constraint on the modeled fluxes. This constraint replaces the constraint on nutrient uptake rates imposed by FBA as the limiting factor for biomass production. We modified the yeast model distributed with ccFBA, which is based on the iMM904 model, matching it to the yeast 7.6 adapted to the experimental data (see above).

The resulting ccFBA model contains experimental k_{cat} values for 535 enzymes; for the remaining enzymes, we use the median of the 535 known values, $k_{cat,med}$=11.5[51]. The proportion of biomass devoted to metabolic enzymes was set to C=0.27.

Epistasis is again calculated according to Eq. (1) from the single and double gene knockout fitness estimates; in ccFBA, the maximal biomass fluxes v_{WT}, v_1, v_2, and v_{12} represent maximal growth rates rather than yields.

Data availability

The empirical data, the modified yeast7.6 metabolic model, the turnover numbers (k_{cat}) and molecular weigths used as input to ccFBA, and the raw data summarized in our results and figures can be found on github at https://github.com/deyazoubi/Epistasis-. An overview over the individual files is given in the Readme files.

Results

Comparing the predictions of different methods

For each pair of non-essential genes contained in the metabolic model, we calculated Epistasis (Eq. (1)) based on three methods: flux balance analysis (FBA); a linear version of minimization of metabolic adjustment (lMOMA) that finds the knockout flux distribution most similar to the pFBA solution; and a recent implementation of FBA with molecular crowding (ccFBA). To obtain an overview over the differences between the three tested methods, we first classified gene pairs into those showing negative epistasis ($\varepsilon \leq -0.0001$), positive epistasis ($\varepsilon \geq +0.0001$), or no epistasis ($|\varepsilon| < 0.0001$).

The Venn diagrams in Fig. 2a and 2b summarize the sets of gene pairs that show negative and positive epistasis, respectively, according to the three methods. 49 negative and 151 positive interactions are predicted jointly by all three methods. 48.8% of negative and 82.8% of positive epistasis predictions with FBA are also predicted by lMOMA. The numbers of interactions predicted uniquely by one of the methods differ substantially: While FBA predicts only 217 genetic interactions not predicted by any of the two other methods, ccFBA makes 640 unique predictions and lMOMA makes 1116 unique predictions. If ccFBA and/or lMOMA starting from the pFBA solution are better at capturing physiological knockout effects than FBA, or if they capture other types of physiological responses, then the additional predictions might improve the unsatisfactory recall of epistasis predictions by FBA.

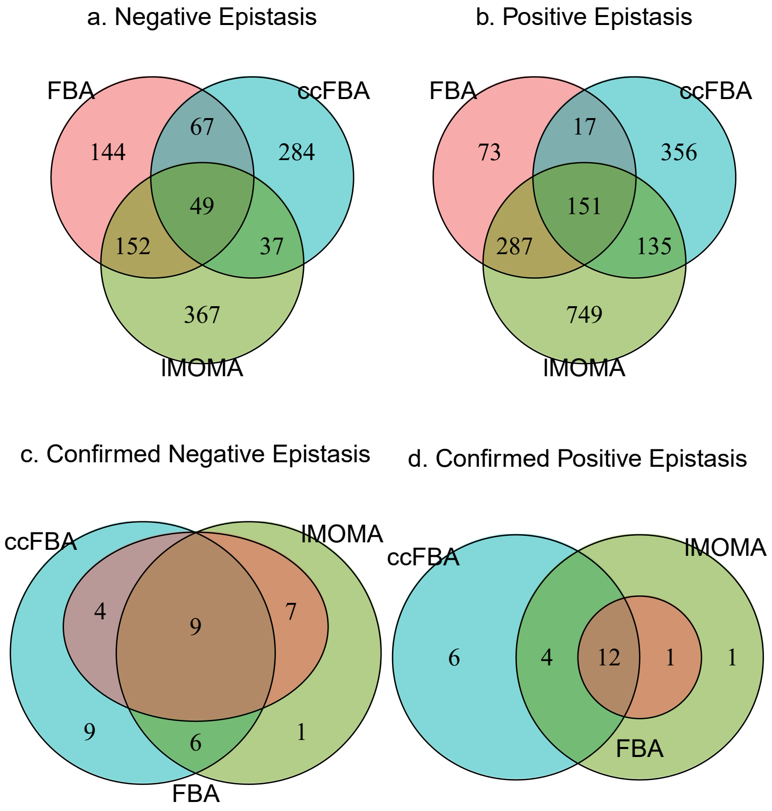


Figure 2. Venn diagrams showing the overlap of negative (a,c) and positive (b,d) epistasis predictions by the three methods. Panels a and b show total predictions. Panels c and d show only those predictions confirmed by the high-confidence set of experimental epistasis estimates.

Comparing predictions to experiments

To test this possibility, we compared the epistasis predictions by the three methods to the high-confidence experimental epistasis data set provided by[33]. Figures 2c and 2d show Venn diagrams that compare the numbers of correctly predicted experimentally observed epistatic interactions between the three metabolic simulation methods. Only a small fraction of the predicted interactions are confirmed by the data in each case. Not surprisingly, the most reliable predictions are those that are jointly made by all three simulation methods (9 correct out of 49 joint predictions of negative epistasis, *i.e.*, a precision of 9/49=18.4%; and 12 correct out of 151 joint predictions of positive epistasis, *i.e.*, a precision of 12/151=7.9%). In contrast, genetic interactions uniquely predicted by one of the three methods are confirmed only rarely. Strikingly, none of the 217 interactions uniquely predicted by FBA were confirmed by the data. Unique ccFBA predictions were confirmed in 9/284=3.2% of cases for negative epistasis and in 6/356=1.7% of cases for positive epistasis. The very many unique predictions by lMOMA were confirmed only once each for negative epsistasis (1/367=0.3%) and for positive epistasis (1/749=0.1%).

Both lMOMA and ccFBA recovered some epistatic interactions not detected with FBA, where those predicted by lMOMA – with only two exceptions – form a subset of those predicted by ccFBA. Accordingly, the precision of ccFBA (50/1096=4.6%) exceeds that of FBA (33/940=3.5%), which exceeds that of lMOMA (18/1322=2.1%). Thus, the consideration of molecular crowding in ccFBA indeed leads to an improvement of the recall achievable in epistasis predictions, while lMOMA does not.

The cutoff of $|\varepsilon| = 0.0001$ for epistasis used to select the predicted interacting gene pair sets in Figure 2 was chosen largely arbitrarily. Figure 3 shows the influence of other cutoffs for the simulated epistasis scores ε on prediction accuracy. Overall, FBA seems to show the worst compromise between precision (fraction of predictions that are correct) and recall (fraction of interactions that are predicted correctly). ccFBA allows the highest recall; however, for negative interactions, high recalls come at the price of high false positive rates.

Importantly, even with the most generous cutoffs, the highest recall reachable by any of the three methods is 24% for negative and 30% for positive epistatic interactions. Thus, at least 70% of experimentally observed epistatic interactions are not detectable by any of the constraint-based methods tested, regardless of how many false positives we are willing to accept. To achieve recall values above 20%, we have to accept false positive rates of more than 10% for negative and 3% for positive interactions; given the high number of comparisons made (71,994 in the dataset used here), this means that true predictions of epistasis are drowned in a sea of false predictions. At a more reasonable false positive rate of 1%, the highest achievable recall values are around 12%.

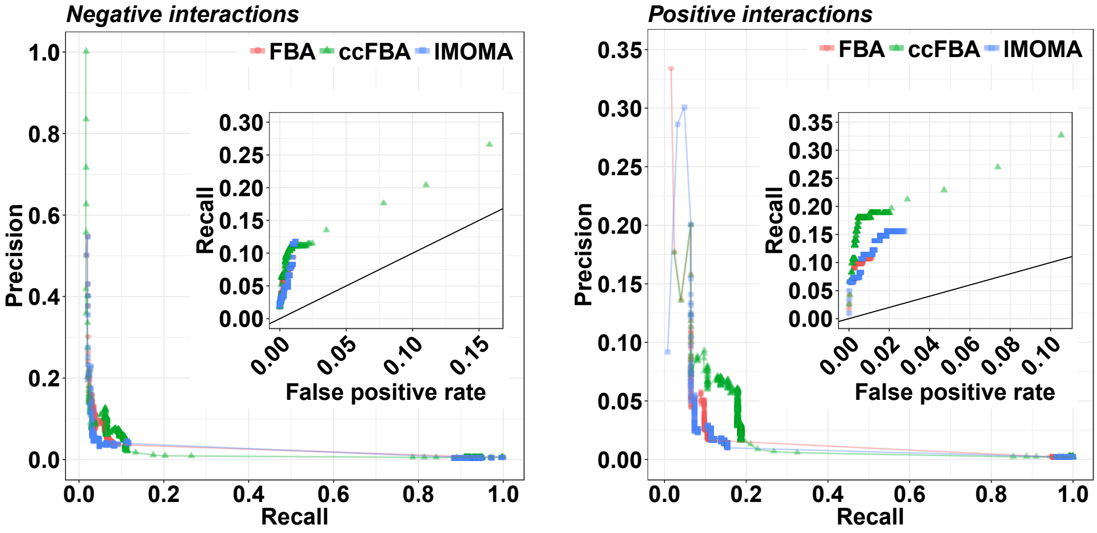


Figure 3. The accuracy of the three prediction methods for negative and positive epistatic interactions. The outer panels show precision (fraction of predictions that are correct) vs. recall (fraction of interactions that are predicted correctly), while the insets show a detail of the receiver operator characteristic (ROC) curve, tracing the dependence of recall on the false positive prediction rate (= 1 – specificity).

Synthetic lethals

To predict epistasis scores for viable double mutants, we need to calculate fitness values quantitatively for the single and double knockouts. It is conceivable that the underwhelming performance of constraint-based methods to predict genetic interactions (Figs. 2, 3) is due to this necessity. However, the strength of constraint-based methods may lie more in qualitative predictions: FBA has been demonstrated to accurately predict gene essentiality, i.e., genes whose knockout is lethal[52,53]. The likely reason is that knockout lethality often arises from the inability to produce a biomass component without the knocked out reaction, *i.e.*, from an effect of the knockout on metabolic network topology rather than on kinetics, regulation, or biomass yield. Thus, it might be reasonable to expect that constraint-based methods also perform well when predicting synthetic lethals, *i.e.*, gene pairs where the single mutants are viable but the double mutant is not. Previous studies showed recall values below 25% for the FBA prediction of synthetic lethals[32,38]. However, these observations were based on the analysis of small numbers of experimentally confirmed synthetic lethals drawn from diverse studies, and thus it seems advisable to compare model predictions of synthetic lethality to a systematic, genome-wide screen of metabolic genes.

To identify pairs of synthetic lethal genes in the raw data from Ref.[33], we selected non-essential gene pairs with experimentally confirmed negative epistasis ($\varepsilon < -0.08$, see Ref.[23]) and with very low double mutant fitness ($f < 0.2$). Only 146 out of a total of 207,060 non-essential gene pairs represented in the model and assayed by Szappanos *et al.*[33] were labeled as synthetic lethal according to this definition.

When using the same cutoffs ($\varepsilon < -0.08$ and $f < 0.2$) for the computational epistasis predictions, we recover only 4 (FBA), 0 (ccFBA), and 4 (lMOMA) of the experimentally confirmed synthetic lethal pairs, corresponding to recall values of less than 3%. For FBA and lMOMA, recall cannot be improved by choosing less stringent cutoffs, as long as we require negative epistasis. For ccFBA, we can obtain 6 true positive predictions if we relax the double mutant fitness cutoff to $f < 0.55$ while requiring negative epistasis ($\varepsilon < -0.0001$). These findings confirm the earlier results on smaller datasets of synthetic lethals[32,38]: constrained based methods appear no better at predicting synthetic lethality than at predicting epistasis in general.

Discussion

The essence of ccFBA is the incorporation of a tradeoff, where the expression of one pathway reduces the cellular resources available for other pathways. This interdependence between pathways in terms of available resources may underlie at least some epistatic interactions, and may hence contribute to explaining the slightly better performance of ccFBA compared to the two alternatives at methods.

While ccFBA added a small number of correct epistasis predictions, the most important conclusion that can be drawn from the above analyses is a sobering one: We still fail to predict 70% of epistatic interactions, regardless of the constraint-based method and the cutoffs used. Neither the inclusion of molecular crowding, as in ccFBA, nor the use of a more realistic wild-type flux distribution in lMOMA led to a substantial improvement over the previously reported failure of FBA to predict a majority of experimentally observed interactions[33].

It is conceivable that a substantial fraction of observed epistatic interactions can only be understood through the consideration of detailed reaction kinetics and the associated cellular investment into enzymes. While ccFBA approximately accounts for enzyme kinetics and the corresponding cellular investment, we need to emphasize that the ccFBA model contains known enzyme turnover numbers (k_{cat}) for only 535 out of 4,594 protein-associated reactions, and an improved parameterization may well lead to improved prediction accuracy in the future. However, ccFBA could also not correctly predict synthetic lethal interactions, which in most cases probably arise from changes in network topology rather than from enzyme kinetics; this failure suggests that the problem is more fundamental.

A second potential explanation for the observed underperformance of constraint-based methods is the influence of regulatory feedbacks. Regulatory interactions evolved in the ancestors of the wild-type strain as responses to environmental conditions. Changes in metabolite concentrations resulting from the knockouts may be mis-interpreted by the cell's regulatory system as environmental cues, and may thus lead to regulatory responses that cause suboptimal metabolic network usage. Such inappropriate cellular regulatory responses might lead to large discrepancies between mutant physiology and predictions by optimization-based methods.

Such discrepancies should be evident not only for double knockouts, but also for single gene knockouts. FBA is highly accurate in the prediction of gene essentiality[36,52,53]. In contrast, quantitative predictions of non-lethal gene knockout fitness values correlate only weakly with experimental observations[54]. If our models cannot quantify single gene knockout fitness reliably, maybe it is no surprise that they fail to predict epistasis scores, which result from a comparison of single and double mutant fitness values.

References

1 de Visser, J. A. G. M., Cooper, T. F. & Elena, S. F. The causes of epistasis. *P Roy Soc B-Biol Sci* **278**, 3617-3624, doi:10.1098/rspb.2011.1537 (2011).

2 Hartman, J. L. t., Garvik, B. & Hartwell, L. Principles for the buffering of genetic variation. *Science (New York, N.Y.)* **291**, 1001-1004 (2001).

3 Boone, C., Bussey, H. & Andrews, B. J. Exploring genetic interactions and networks with yeast. *Nature reviews. Genetics* **8**, 437-449, doi:10.1038/nrg2085 (2007).

4 Phillips, P. C. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nature reviews. Genetics* **9**, 855-867, doi:10.1038/nrg2452 (2008).

5 Kondrashov, A. S. Selection against harmful mutations in large sexual and asexual populations. *Genetical research* **40**, 325-332 (1982).

6 Otto, S. P. Unravelling the evolutionary advantage of sex: a commentary on 'Mutation-selection balance and the evolutionary advantage of sex and recombination' by Brian Charlesworth. *Genetical research* **89**, 447-449, doi:10.1017/s001667230800966x (2007).

7 Presgraves, D. C. Speciation genetics: epistasis, conflict and the origin of species. *Curr Biol* **17**, R125-127, doi:10.1016/j.cub.2006.12.030 (2007).

8 Kondrashov, A. S. & Crow, J. F. Haploidy or diploidy: which is better? *Nature* **351**, 314-315, doi:10.1038/351314a0 (1991).

9 Crow, J. F. & Kimura, M. Efficiency of truncation selection. *Proc Natl Acad Sci U S A* **76**, 396-399 (1979).

10 Jasnos, L. & Korona, R. Epistatic buffering of fitness loss in yeast double deletion strains. *Nat Genet* **39**, 550-554, doi:10.1038/ng1986 (2007).

11 Cordell, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* **11**, 2463-2468 (2002).

12 Moore, J. H. & Williams, S. M. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays : news and reviews in molecular, cellular and developmental biology* **27**, 637-646, doi:10.1002/bies.20236 (2005).

13 Trindade, S. *et al.* Positive epistasis drives the acquisition of multidrug resistance. *PLoS genetics* **5**, e1000578, doi:10.1371/journal.pgen.1000578 (2009).

14 Measday, V. & Hieter, P. Synthetic dosage lethality. *Methods Enzymol* **350**, 316-326 (2002).

15 Tong, A. H. *et al.* Global mapping of the yeast genetic interaction network. *Science (New York, N.Y.)* **303**, 808-813, doi:10.1126/science.1091317 (2004).

16 Pan, X. *et al.* A robust toolkit for functional profiling of the yeast genome. *Mol Cell* **16**, 487-496, doi:10.1016/j.molcel.2004.09.035 (2004).

17 Measday, V. *et al.* Systematic yeast synthetic lethal and synthetic dosage lethal screens identify genes required for chromosome segregation. *Proc Natl Acad Sci U S A* **102**, 13956-13961, doi:10.1073/pnas.0503504102 (2005).

18 Sopko, R. *et al.* Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell* **21**, 319-330, doi:10.1016/j.molcel.2005.12.011 (2006).

19 Pan, X. *et al.* A DNA integrity network in the yeast Saccharomyces cerevisiae. *Cell* **124**, 1069-1081, doi:10.1016/j.cell.2005.12.036 (2006).

20 Collins, S. R. *et al.* Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446**, 806-810, doi:10.1038/nature05649 (2007).

21 Fiedler, D. *et al.* Functional organization of the S. cerevisiae phosphorylation network. *Cell* **136**, 952-963, doi:10.1016/j.cell.2008.12.039 (2009).

22 Kornmann, B. *et al.* An ER-mitochondria tethering complex revealed by a synthetic biology screen. *Science (New York, N.Y.)* **325**, 477-481, doi:10.1126/science.1175088 (2009).

23 Costanzo, M. *et al.* The genetic landscape of a cell. *Science (New York, N.Y.)* **327**, 425-431, doi:10.1126/science.1180823 (2010).

24 Boucher, B. & Jenna, S. Genetic interaction networks: better understand to better predict. *Frontiers in genetics* **4**, 290, doi:10.3389/fgene.2013.00290 (2013).

25 Orth, J. D. *et al.* A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Molecular systems biology* **7**, 535, doi:10.1038/msb.2011.65 (2011).

26 Schellenberger, J., Park, J. O., Conrad, T. M. & Palsson, B. O. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* **11**, 213, doi:10.1186/1471-2105-11-213 (2010).

27 Segre, D., Deluna, A., Church, G. M. & Kishony, R. Modular epistasis in yeast metabolism. *Nat Genet* **37**, 77-83, doi:10.1038/ng1489 (2005).

28 He, X., Qian, W., Wang, Z., Li, Y. & Zhang, J. Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. *Nat Genet* **42**, 272-276, doi:10.1038/ng.524 (2010).

29 Snitkin, E. S. & Segre, D. Epistatic interaction maps relative to multiple metabolic phenotypes. *PLoS genetics* **7**, e1001294, doi:10.1371/journal.pgen.1001294 (2011).

30 Xu, L., Barker, B. & Gu, Z. Dynamic epistasis for different alleles of the same gene. *Proc Natl Acad Sci U S A* **109**, 10420-10425, doi:10.1073/pnas.1121507109 (2012).

31 Barker, B., Xu, L. & Gu, Z. Dynamic Epistasis under Varying Environmental Perturbations. *PloS one* **10**, e0114911 (2015).

32 Harrison, R., Papp, B., Pal, C., Oliver, S. G. & Delneri, D. Plasticity of genetic interactions in metabolic networks of yeast. *Proc Natl Acad Sci U S A* **104**, 2307-2312, doi:10.1073/pnas.0607153104 (2007).

33 Szappanos, B. *et al.* An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat Genet* **43**, 656-662, doi:10.1038/ng.846 (2011).

34 Jacobs, C., Lambourne, L., Xia, Y. & Segre, D. Upon Accounting for the Impact of Isoenzyme Loss, Gene Deletion Costs Anticorrelate with Their Evolutionary Rates. *PloS one* **12**, e0170164, doi:10.1371/journal.pone.0170164 (2017).

35 Watson, M. R. Metabolic Maps for the Apple-II. *Biochem Soc T* **12**, 1093-1094, doi:DOI 10.1042/bst0121093 (1984).

36 Orth, J. D., Thiele, I. & Palsson, B. O. What is flux balance analysis? *Nat Biotechnol* **28**, 245-248, doi:10.1038/nbt.1614 (2010).

37 Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *P Natl Acad Sci USA* **99**, 15112-15117, doi:10.1073/pnas.232349399 (2002).

38 Heavner, B. D. & Price, N. D. Comparative Analysis of Yeast Metabolic Network Models Highlights Progress, Opportunities for Metabolic Reconstruction. *PLoS computational biology* **11**, e1004530, doi:10.1371/journal.pcbi.1004530 (2015).

39 Aziz, R. K. *et al.* Systems biology-guided identification of synthetic lethal gene pairs and its potential use to discover antibiotic combinations. *Scientific reports* **5**, 16025, doi:10.1038/srep16025 (2015).

40 Basan, M. *et al.* Inflating bacterial cells by increased protein synthesis. *Molecular systems biology* **11**, 836, doi:10.15252/msb.20156178 (2015).

41 Beg, Q. K. *et al.* Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. *Proc Natl Acad Sci U S A* **104**, 12663-12668, doi:10.1073/pnas.0609845104 (2007).

42 Goelzer, A., Fromion, V. & Scorletti, G. Cell design in bacteria as a convex optimization problem. *Automatica* **47**, 1210-1218, doi:10.1016/j.automatica.2011.02.038 (2011).

43 Adadi, R., Volkmer, B., Milo, R., Heinemann, M. & Shlomi, T. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS computational biology* **8**, e1002575, doi:10.1371/journal.pcbi.1002575 (2012).

44 Schuster, S., Boley, D., Moller, P., Stark, H. & Kaleta, C. Mathematical models for explaining the Warburg effect: a review focussed on ATP and biomass production. *Biochem Soc Trans* **43**, 1187-1194, doi:10.1042/BST20150153 (2015).

45 Basan, M. *et al.* Overflow metabolism in Escherichia coli results from efficient proteome allocation. *Nature* **528**, 99-104, doi:10.1038/nature15765 (2015).

46 Holzhutter, H. G. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur J Biochem* **271**, 2905-2922, doi:10.1111/j.1432-1033.2004.04213.x (2004).

47 Machado, D. & Herrgard, M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS computational biology* **10**, e1003580, doi:10.1371/journal.pcbi.1003580 (2014).

48 Aung, H. W., Henry, S. A. & Walker, L. P. Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism. *Industrial biotechnology (New Rochelle, N.Y.)* **9**, 215-228, doi:10.1089/ind.2013.0013 (2013).

49 Gelius-Dietrich, G., Desouki, A. A., Fritzemeier, C. J. & Lercher, M. J. Sybil--efficient constraint-based modelling in R. *BMC systems biology* **7**, 125, doi:10.1186/1752-0509-7-125 (2013).

50 R Development Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, <http://www.R-project.org.> (2008).

51 Desouki, A. A. *Algorithms for improving the predictive power of flux balance analysis*, Heinrich Heine University Duesseldorf, (2016).

52 O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using Genome-scale Models to Predict Biological Capabilities. *Cell* **161**, 971-987, doi:10.1016/j.cell.2015.05.019 (2015).

53 Hartleb, D., Jarre, F. & Lercher, M. J. Improved Metabolic Models for E. coli and Mycoplasma genitalium from GlobalFit, an Algorithm That Simultaneously Matches Growth and Non-Growth Data Sets. *PLoS computational biology* **12**, e1005036, doi:10.1371/journal.pcbi.1005036 (2016).

54 Papp, B., Szappanos, B. & Notebaart, R. A. Use of genome-scale metabolic models in evolutionary systems biology. *Methods in molecular biology (Clifton, N.J.)* **759**, 483-497, doi:10.1007/978-1-61779-173-4_27 (2011).

Acknowledgements

We thank Balazs Papp and Jonathan Fritzemeier for helpful discussions. This work was supported financially through fellowships by the German Academic Exchange Service (DAAD) to DA and AAD, and through German Research Foundation (DFG) grants to MJL (CRC 680 and CRC 1310).

Supplementary Information:

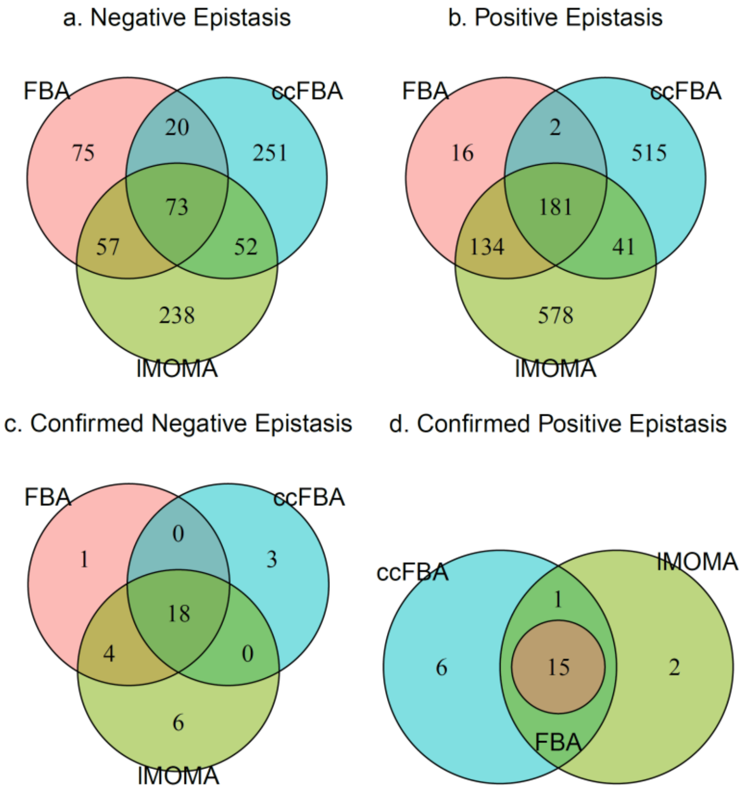


Figure S1. Venn diagrams showing the overlap of negative (a,c) and positive (b,d) epistasis predictions by the three methods based on the iMM904 yeast model. Panels (a) and (b) show total predictions. Panels (c) and (d) show only those predictions confirmed by the high-confidence set of experimental epistasis estimates. See Fig. 2 of the main text for the corresponding figure based on the newer yeast7.6 model.

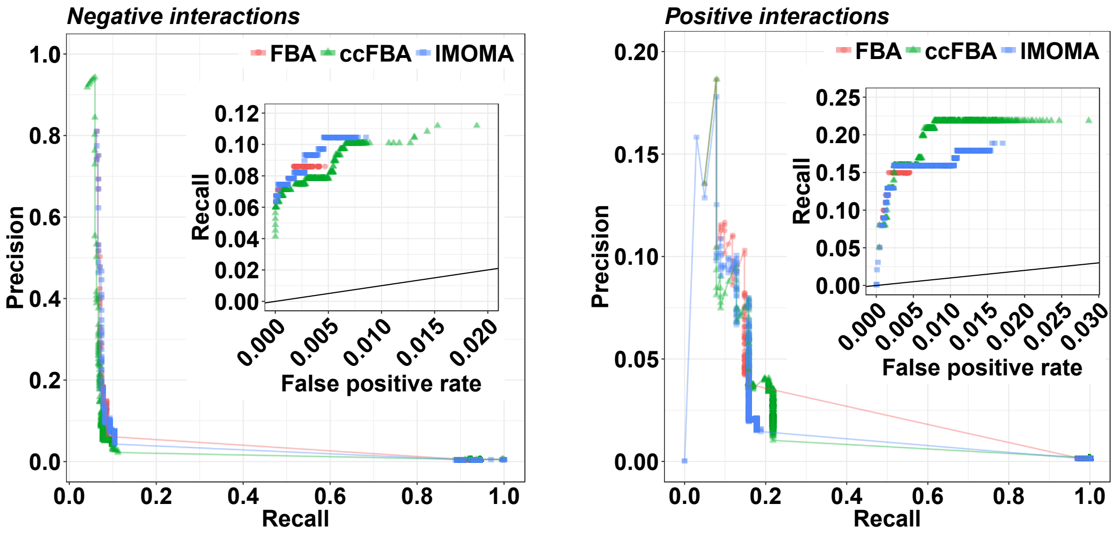


Figure S2. The accuracy of the three prediction methods for negative (left) and positive (right) epistatic interactions based on the iMM904 yeast model. The outer panels show precision (fraction of predictions that are correct) *vs.* recall (fraction of interactions that are predicted correctly), while the insets show a detail of the receiver operator characteristic (ROC) curve, tracing the dependence of recall on the false positive prediction rate (= 1 – specificity). See Fig. 3 of the main text for the corresponding figure based on the newer yeast7.6 model.

Manuscript 3: Flux balance analysis and other constraint-based methods fail to predict mutant fitness for non-lethal metabolic gene knockouts in *E. coli* and yeast

Status

This manuscript is in preparation for submission.

Authors: Deya Alzoubi, Abdelmoneim Amer Desouki, Balázs Papp & Martin J. Lercher

Contributions

Together with Martin Lercher, I developed the concept and methodology of this work and wrote the manuscript. I wrote the software for simulations and analyses and performed all analyses.

Manuscript

(see next page)

Flux balance analysis and other constraint-based methods fail to predict mutant fitness for non-lethal metabolic gene knockouts in *E. coli* and yeast

Deya Alzoubi[1], Abdelmoneim Amer Desouki[1], Balázs Papp[2], Martin J. Lercher[1,*]

[1] Institute for Computer Science and Department of Biology,
Heinrich Heine University, Universitätsstraße 1, Düsseldorf D-40221, Germany

[2] Synthetic and Systems Biology Unit, Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences, Szeged, Hungary

* To whom correspondence should be addressed

Abstract

FBA and related constraint-based methods are considered state-of-the-art for genome-scale metabolic modeling. These methods have been shown to predict gene essentiality with high accuracy. However, it is not clear how reliably they predict mutant physiology for non-essential gene knockouts, as systematic analyses that apply the different constraint-based methods proposed for this problem to genome-scale data for multiple organisms. Here, we apply FBA and its popular extensions, including methods specifically developed for non-essential gene knockout predictions (MOMA, ROOM) and a method accounting for macro-molecular crowding (ccFBA), to data on knockout effects of non-essential genes in *E. coli* and *Saccharomyces cerevisiae*. For a given metabolic model, simulation method, and environment, we find that predicted biomass fluxes across non-essential gene knockouts are restricted to a small number of distinct values. In each case analyzed, predictions explain only a small fraction of the observed variance in growth rate, fitness, or biomass yield. Even in the best cases, model-based predictions lead to coefficients of determination that are barely better than those of a trivial "model" assuming identical fitness of all knockouts. The constraint-based models perform slightly better when attempting to qualitatively distinguish between non-essential gene knockouts with and without fitness effects. However, even the best-performing methods – linear or quadratic MOMA – predict only between 20% and 40% of experimentally observed deleterious fitness effects. What could be the reason for this poor performance of non-lethal knockout predictions? The physiological response of microbes to gene knockouts is based on regulatory systems that evolved in the wildtype. The flux changes resulting from metabolic gene knockouts lead to changes of metabolite concentrations both upstream and downstream of the knocked-out reaction. We speculate that these concentration changes cause misguided regulatory responses that are impossible to predict by optimization-based methods agnostic of regulatory interactions.

Introduction

Constraint-based methods such as flux balance analysis (FBA) [1-3] and its derivatives [4] are considered state-of-the-art methods for the genome-scale modeling of microbial physiology. These methods start from the known stoichiometry of metabolic reactions encoded in an organism's genome. Assuming a steady state and constraints on the reversibility of reactions, FBA and related methods then impose the requirement that the fluxes producing and consuming each internal metabolite must be balanced. An optimal state (typically assumed to result from natural selection) is found, *e.g.*, by maximizing the biomass production rate under these constraints. FBA solves a linear optimization problem and is thus computationally highly efficient. However, this efficiency comes at the price of ignoring many mechanistic details, including reaction kinetics and regulatory interactions [5].

FBA models accurately predict gene essentiality, with reported accuracy values between 91% and 95% for the bacterium *Escherichia coli* [6] and between 83% and 90% for different models of the yeast *Saccharomyces* cerevisiae [7-9]. Gene essentiality refers to the inability of a gene knockout strain to produce biomass. If we assume that typically, each metabolic gene is expressed at least marginally in some cells of a microbial population, all metabolic genes encoded in the genome can contribute to alleviating deleterious knockout effects. Thus, gene essentiality is independent of kinetic parameters and of regulatory circuits, which may explain why FBA is able to predict gene essentiality despite ignoring such mechanistic details.

However, the same details may strongly influence the physiological effects of non-essential gene knockouts. As a first example demonstrating such effects, consider a genome encoding two homologs of an enzyme with different kinetics. If only the catalytically more efficient enzyme is utilized in a given environment, then its knockout will reduce the growth rate, while the knockout of its homolog will have no effect. As FBA only sees the identical stoichiometries of the reactions catalyzed by the homologs, it considers them as redundant and predicts that both knockouts have no physiological effects.

More generally, the deletion of a non-essential gene encoding an enzyme active in the wildtype requires a global re-routing of reaction fluxes to re-establish a steady state. The knockout of a non-essential gene encoding an enzyme that is active in the wildtype will result in a concentration increase of the enzyme's substrates as well as a concentration reduction of its products. These changes will in turn affect concentrations further upstream and downstream of the reaction, which may be sensed by regulatory circuits that misinterpret them as resulting from environmental changes, leading to non-optimal (and potentially hard to predict) regulatory responses. Thus, it appears *a priori* likely that the effects of non-essential gene knockouts are much harder to predict by constraint-based methods than those of essential gene knockouts.

Indications that this may indeed be the case come from two previous studies. Following up on earlier small-scale analyses [10,11], Papp *et al.* analyzed observed competitive fitness, growth rates, and growth efficiencies (a proxy for yield) of a *S. cerevisiae* genome-wide gene knockout collection on two different media (minimal, SD; and rich, YPD) [5]. They predicted biomass production rates using MOMA (Minimization Of Metabolic Adjustment) [12], a variant of FBA specifically designed for gene knockouts, which finds the permissible flux distribution in the knockout that is most similar to the wildtype flux distribution. Papp *et al.* found only weak correlations between the predicted biomass production rates and the observed data (Spearman rank correlation coefficients: ρ=0.46 for competitive fitness on SD, ρ=0.26 for competitive fitness on YPD, ρ=0.14 for growth rate on SD, and ρ=0.05 for growth efficiency on SD). While all correlations were statistically significant, the model was able to explain only between 0.25% and 21% of the observed variation across metabolic knockouts.

In a later study, Vandersluis *et al.* analyzed growth rates of a genome-wide collection of prototropic gene knockout strains of *S. cerevisiae* across 28 metabolic environments [13]. Instead of directly comparing growth rates with predicted biomass production rates, these authors calculated *z*-scores quantifying the deviation between growth in a given condition and a reference condition for experimental growth rates and biomass production rates predicted with FBA and MOMA, using two different metabolic models (iMM940 [14] and yeast5

[15]). In many growth conditions, they found that FBA and MOMA predicted only a few discrete levels of biomass production, with the mode accounting for between 39% and 95% of predictions. Comparing *z*-scores between experiments and predictions, they found that Spearman's rank correlations were statistically significant (P<0.05) in between 40% (iMM904 with MOMA) and 70% (yeast5 with FBA) of growth conditions. Labeling gene knockouts as with and without fitness effects, they found average recall values (correct predictions among all knockouts with observed fitness effects) across growth conditions between 14% and 25%, at precision values (true positives among all predictions with fitness effects) between 49% and 18%, respectively (see Fig. 4 of Ref. [13]). Note that in a given condition, only a small minority of knockouts show observable fitness effects. Thus, for relative changes in biomass production abilities, even modest recall values can be achieved only at the price of a massive number of false positive predictions (precision < 50%).

In sum, the two previous studies that compared large numbers of predictions by FBA and/or MOMA with observed yeast growth data indicate that neither method is capable of predicting gene knockout effects quantitatively, and that at best a minority of fitness effects can be predicted even qualitatively. If these conclusions would apply more generally across organisms and constraint-based metabolic modeling methods, than these would be sobering conclusions for a suite of methods that is widely considered to be the state of the art [4]. On one hand, it is conceivable that low predictive abilities are restricted to yeast (or to eukaryotes in general). On the other hand, several modifications aimed at improving the predictive abilities of constraint-based methods have been proposed [4,10,16-18]. In particular, a minimization of the number of required regulatory changes in the knockout [10] or a consideration of enzyme kinetics and a limited enzyme "budget" [16-18] might alleviate some of the problems faced by the methods tested previously. Here, we systematically analyze the ability of multiple constraint-based modeling methods to predict growth features of the two best-studied unicellular model organisms, the yeast *S. cerevisiae* and the bacterium *Escherichia coli*.

Methods

Flux Balance Analysis (FBA) [19]

Flux balance analysis (FBA) is a mathematical approach for analyzing the flow of metabolites through a metabolic network [1]. The central object in an FBA formulation (as in other constraint-based methods) is a stoichiometric matrix *S*. Its columns correspond to reactions R_j (with metabolite fluxes v_j), while its rows correspond to individual metabolites *i*; the entries of S are the stoichiometric coefficients of metabolite *i* in reaction R_j, with negative signs for substrates and positive signs for products of the reaction. FBA then solves the following linear optimization problem [1]:

$$\max v_{bio} \tag{1}$$

Subject to:

$$S\boldsymbol{v} = 0 \tag{2}$$

$$v_i^{min} \leq v_i \leq v_i^{max} \tag{2}$$

$v_j = 0$ for any reaction that requires the product of a knocked-out gene for its activity (3)

Here, $\boldsymbol{v}$ is a vector of steady-state metabolic fluxes, where $S\boldsymbol{v} = 0$ enforces the steady state assumption (each internal metabolite is produced and consumed at the same rate); v_{bio} is the rate of biomass production. v_i^{min} and v_i^{max} are lower and upper bounds for the flux through reaction *i;* these bounds are typically enforced only for irreversible reactions ($v_i^{min} = 0$) and to constrain the uptake rate for a limiting nutrient. Note that modeling gene knockout effects with FBA assumes that the biomass production rate in the knockout mutant is maximal (under the given constraints).

Minimization of Metabolic Adjustment (MOMA) [12]

MOMA is a variant of FBA for modeling gene knockouts. Its basic idea is that due to a gene regulatory system that evolved in (and is presumably optimized for) the wildtype, the flux distribution is not optimized for maximal biomass production, but is instead maximally similar to the wildtype flux distribution given the constraints. However, the wildtype flux distribution is not uniquely defined by FBA, as typically many flux distributions with the same maximal biomass production rate exist. Previous authors have often ignored this redundancy, and thus used a "random" maximal solution for the calculation of the MOMA flux distribution (*e.g.*, [5]). Here, we generally use the parsimonious FBA (pFBA) solution for the wildtype; the pFBA solution corresponds to the FBA solution with a minimal sum of absolute fluxes [20], attempting to approximately minimize enzyme usage. pFBA has been shown to provide a reasonable approximation to biological reality [21].

Below, we use three versions of MOMA, each of which replaces Eq. (1) with a different optimization:

(i) **qMOMA** – This version was originally described by Segre *et al.* and minimizes the (squared) Euclidean distance between the two flux vectors, $\left\|\boldsymbol{v}^{knockout}-\boldsymbol{v}^{WT}\right\|_2^2=\sum_i\left(v_i^{knockout}-v_i^{WT}\right)^2$ [22].

(ii) **lMOMA** – A linearized version that instead minimizes the Manhattan distance $\left\|\boldsymbol{v}^{knockout}-\boldsymbol{v}^{WT}\right\|_1 = \sum_i\left|v_i^{knockout}-v_i^{WT}\right|$; this linear MOMA method has been used by several later authors because of its computational efficiency (*e.g.*, [5]).

(iii) **sqMOMA** – The distances calculated by MOMA and linear MOMA weigh each flux equally. Thus, a 1% change of a wildtype flux $v_j^{WT}=100$ (in arbitrary units) is penalized exactly in the same way as the doubling of a much smaller flux $v_{j'}^{WT}=1$. This may not reflect biological reality: a flux change by 1% may easily be accommodated by an unchanged enzyme concentration, while a flux doubling will typically require upregulating the catalyzing enzyme. Thus, we additionally use a scaled version of MOMA that penalizes fractional flux changes rather than absolute flux changes. sqMOMA minimizes the weighted Euclidean distance between flux distributions, where each flux is scaled by dividing through the corresponding wildtype flux $\left|v_j^{WT}\right|$ (or, if $\left|v_j^{WT}\right|=0$, through the maximal absolute value of possible loopless fluxes v_j^{max} under maximal biomass production, calculated using cycleFreeFVA [23]; in the implementation, we replace $|v|=0$ with $|v|<0.0001$). Thus, the optimization problem solved by sqMOMA (and replacing Eq. (1)) is:

$$\min \sum_i \frac{\left(v_i^{knockout}-v_i^{WT}\right)^2}{v_i^*}$$

$$\text{with} \quad v_i^* = \begin{cases} \left|v_i^{WT}\right| & \text{if } v_i^{WT}\neq 0 \\ v_i^{max} & \text{if } v_i^{WT}=0 \text{ and } v_i^{max}\neq 0 \\ 1 & \text{else} \end{cases},$$

$$v_i^{max} = \max \quad \left\{\left|\min v_i^{ccFVA}\right|, \left|\max v_i^{ccFVA}\right|\right\}.$$

Regulatory on/off minimization of metabolic flux (ROOM) [10]

ROOM minimizes the total number of significant flux changes in the knockout relative to the wild type flux distribution (pFBA) [20]. We implemented two versions of ROOM: the published version [10]; and a new version of ROOM (ROOMw), which penalizes only flux increases, but not flux decreases. The logic of ROOMw is that flux decreases do not require changes in enzyme expression, while flux increases require upregulation of enzyme expression. To implement ROOMw, first all reversible reactions are split into two independent forward and backward reactions. Since all reactions in the modified model can only carry positive fluxes, we now minimize only the number of reactions with flux increases:

$$n_{increases} \coloneqq \min \quad \sum_i \ y_i$$

Subject to:

$$v_i^{min} \leq v_i \leq v_i^{max}$$

$v_j = 0$ for any reaction that requires the product of a knocked-out gene for its activity

$$y_i \ \in \{0,1\}$$

$$v_i - y_i\ (v_{\mathrm{i}}^{\max} - v_i^{WTu}) \leq v_i^{WTu}$$

$$v_i^{WTu} = v_i^{WT} + \delta\left|v_i^{WT}\right| + \varepsilon$$

Here, $\delta = 0.03$ and $\varepsilon = 0.0001$ are relative and absolute tolerances, respectively, for "significant" flux changes. Note that a solution to this minimization problem is always given by $\boldsymbol{v} = 0$. To avoid this trivial solution, we first enforce a minimal amount of biomass production, $v_{bio} \geq 0.05\, v_{bio}^{WT}$ to obtain a lower bound $n_{increases} = n^*$ on the necessary number of flux increases. In a second step, we again solve the above optimization problem, now enforcing this lower bound: $n_{increases} \geq n^*$.

FBA with molecular crowding [18]

The constraint-based methods listed above ignore important cellular constraints beyond the stoichiometry of biochemical reactions. In particular, the enzymes that catalyze biochemical fluxes must be produced from limited cellular resources, and the total volume of macromolecules solved in a microliter of cytosol cannot become arbitrarily high: molecular crowding of enzymes and other macromolecules may hinder the efficient diffusion of proteins and metabolites {Atkinson}. Accordingly, Beg *et al.* introduced an upper limit on total enzymatic capacity into FBA (FBA with molecular crowding, FBAwMC) [17]. They added a constraint on the volume available for enzymes, showing that this extension of FBA improved the prediction of phenotypes for *E. coli*. MetabOlic Modeling with ENzyme kineTics (MOMENT) [16] extended this approach by including detailed gene-protein-reaction (GPR) associations, and cost-constrained FBA (ccFBA) further improved this framework and added an explicit model for *S. cerevisiae* [18]. Here, we use the ccFBA implementations for *E. coli* and *S. cerevisiae* (available on CRAN). ccFBA uses enzyme molecular weights to constrain total cellular enzyme concentration, and enzyme kinetic data (k_{cat}) to constrain the fluxes catalyzed by these enzymes [18]

Numerical calculations

All calculations were performed using sybil 24, an efficient computational framework for constraint-based analyses in the R environment for statistical computing 25. As external optimizer, we used IBM ILOG CEPLEX. We considered predicted changes in biomass production rates to be biologically significant if they exceeded 0.15% of the wildtype value, except for ROOM, where we set this "tolerance" to 4.5%; the higher value for ROOM reflects the high relative tolerance δ=0.03 in the determination of "significant" flux changes 10. As the calculations for cycleFreeFVA did not converge for many genes in the yeast7.6 model, we used standard flux variability analysis 26as implemented in sybil for sqMOMA for this model instead of cycleFreeFVA.

S. cerevisiae data

We analysed growth data for non-essential metabolic gene knockout strains from *S. cerevisiae* from three different sources. Genes assayed experimentally were mapped to the gene-protein-reaction associations of the yeast7.6 model (https://sourceforge.net/projects/yeast)) [27].

Dataset 1: Szappanos *et al.* 2011 [28]

The first dataset was obtained from Szappanos *et al.*, who performed single and double gene knockouts of 613 metabolic genes [28]; only the single gene knockouts were considered here. Strains were grown on a nutrient-rich synthetic medium, and fitness was assessed quantitatively by measuring colony size [29].

We removed CAN1, LYP1, URA3, LEU2, and MET17 from the yeast7.6 reconstruction to mimic the strain background used in the experiments [28]. We defined the model growth medium as in [5], listed in **Suppl. Table S1**. Genes essential in the resulting model were excluded from further analysis, resulting in 796 non-essential genes covered by the reduced model. 532 genes were contained in both the experimental dataset and the model.

We classified gene knockouts as having deleterious fitness effects by defining a z-score, $:= \frac{1-f}{SD}$, where *f* is the mean fitness of the single gene mutant across replicates, and *SD* is the corresponding standard deviation. We considered knockouts with z>2 as having significant deleterious fitness effects.

To examine if the details of the metabolic model have a major influence on the prediction accuracy, we analyzed Dataset 1 non only with the yeast7.6 model, but also with the older iMM904 model [14] after a correction of NAD metabolism [28]. Again, we used the media definition utilized by Szappanos *et al.* and removed CAN1, LYP1, URA3, LEU2, and MET17 to mimic the strain background [28]. The merged experimental/model dataset contained 563 genes.

Dataset 2: Deutschbauer *et al.* 2005 [30]

The second dataset was obtained from Deutschbauer *et al.*, who examined 5922 single gene knockouts in a glucose-limited aerobic minimal medium and in YPD, estimating fitness through parallel fitness profiling [30]. The minimal medium corresponds to the default medium of the yeast7.6 model. To simulate YPD, we used the same complex medium as for dataset 1, but allowing uptake of the three amino acids histidine, arginine and lysine at a maximal rate of 0.36; we additionally allowed unlimited uptake of Ammonium, Sodium, Choline, Inosine, pyridoxine (see **Suppl. Table S2** for the medium definition). After removing essential genes, the merged experiment/model datasets comprised 762 genes for YPD and 690 genes for the minimal medium, respectively.

As Deutschbauer *et al.* examined only 2 replicates per gene, we could not define a meaningful *z*-score. Based on the approximately normal distribution of experimental fitness values *f* around 1.0 across knockouts, we classified genes as with fitness effect if *f*<0.97. This resulted in a total of 194 non-essential gene knockouts with experimentally observed deleterious fitness effects on YPD and a total of 222 non-essential gene knockouts with experimentally observed deleterious fitness effects on the minimal medium.

Dataset 3: Breslow *et al.* 2008 [31]

The final *S. cerevisiae* dataset was obtained from Beslow *et al.,* who assayed 4204 single gene knockouts in a glucose-limited minimal medium, assessing competitive fitness using flow cytometry [31]. The minimal medium was simulated as the default medium of the yeast7.6 model. After removing essential genes, the merged experiment/model datasets covered 550 genes.

As for Dataset 2, genes did not consistently have >2 replicates, so we again considered genes with *f* < 0.97 to have a fitness effect. This resulted in a total of 130 non-essential gene knockouts with experimentally observed deleterious fitness effects.

E. coli data

For *E. coli*, we analysed growth data for non-essential metabolic gene knockout strains from two different sources. Genes assayed experimentally were mapped to the gene-protein-reaction associations of the iJO1366 model [32].

Dataset 1: Fuhrer *et al.* 2016 [33]

Fuhrer *et al.* assayed growth rates of 3901 gene knockouts on a minimal medium with glucose and casein hydrolysate [34]. For 3631 genes, we were able to obtain mean and standard deviation (SD) of the knockout growth rate across replicates. We estimated the distribution of wildtype growth rates by fitting a normal distribution around the mode of the knockout growth rate distribution, resulting in a wildtype growth rate of (0.809 ± 0.126) h^{-1}. We divided the mean growth rate across replicates for each gene knockout by the wildtype growth rate to estimate observed fitness values of gene knockouts.

We simulated the growth medium using the default glucose-limited minimal medium definition with maximal glucose uptake rate -10, additionally allowing uptake of the 20 amino acids. We mapped all other compounds and set the lower bounds of the associated exchange reactions to -1000. Allowing unlimited influx of all amino acids results in some internal reactions reaching their upper bounds in parsimonious FBA (pFBA) [20], indicating unrealistically high uptake rates. We therefore constrained all amino acid uptake rates to the same value -0.5, such that the predicted growth rate including the 20 amino acids (the simulated casein hydrolysate) was approximately twice the wildtype growth rate on glucose alone (μ=1.76 instead of μ= 0.98). The simulated medium definition is listed in **Suppl. Table S3**.

Excluding essential genes resulted in a model that covers 1212 non-essential metabolic *E. coli* genes. Merging this model with the experimental data resulted in 1064 non-essential gene knockouts covered by both datasets.

As for yeast Dataset 1, we classified gene knockouts as having deleterious fitness effects by defining a z-score, $\coloneqq \frac{1-f}{SD}$, where f is the mean fitness of the single gene mutant across replicates, and SD is the corresponding standard deviation. We considered the 168 knockouts with $z > 2$ as having significant deleterious fitness effects.

Dataset 2: Takeuchi *et al.* 2014 [35]

Takeuchi *et al.* assayed growth curves for 4105 knockout strains grown on LB media [35]. For 3631 gene knockouts, we could obtain mean and SD across replicates for maximal growth rate and maximal optical density (KO_maxgrowth and KO_saturation in the Supplementary Information of Ref. [35], respectively); optical density provides a proxy for biomass yield. We fitted normal distributions around the modes of the two distributions to obtain wildtype mean and SD, estimated to be (0.981±0.094) for growth rate and (0.982±0.082) for yield (both in units of the data provided). For each knockout strain, we obtained two independent fitness estimates by dividing (i) its mean growth rate and (ii) its yield by the corresponding wildtype values.

To simulate the growth medium, we started from the default minimal medium with glucose as the carbon source (max. uptake rate -10) and with Cob(I)alamin uptake limited to -0.01. We added all 20 amino acids, limiting the uptake of all carbon sources other than glucose to -0.5, such that the predicted wildtype growth rate was approximately twice that on glucose alone (μ=1.97 instead of μ= 0.98). The resulting media definition is listed in **Suppl. Table S4**. Merging the resulting model with the experimental data results, after exclusion of essential genes, in 1175 non-essential gene knockouts covered by both datasets.

As before, we calculated *z*-scores from the observed means and standard deviations. Knockouts of 88 genes have $z > 2$ for growth rate, while knockouts of 84 genes have $z > 2$ for biomass yield; these knockouts are considered to have deleterious fitness effects with respect to growth rate and yield, respectively.

Results and Discussion

As detailed in the Methods section, we predicted biomass production rates with seven different constraint-based methods:

(i) standard FBA;
(ii) minimization of metabolic adjustment (qMOMA);
(iii) a linear version of MOMA (lMOMA);
(iv) a scaled version of MOMA (sqMOMA) introduced here, which minimizes relative rather than absolute flux changes between wildtype and knockout strain;
(v) regulatory on-off minimization (ROOM);
(vi) a version of ROOM introduced here that only penalizes flux increases, not flux decreases in the knockout relative to the wildtype (ROOMw);
(vii) FBA with molecular crowding (ccFBA)

Only weak correlations between predicted biomass production rates and observed fitness for non-essential S. cerevisiae metabolic gene knockouts

All methods predict only a small number of distinct biomass production rates, as has been reported previously for FBA [13]. For the 532 non-essential metabolic knockouts in the Szappanos *et al.* data [28], the number of distinct values ranges between 10 (ROOM) and 34 (sqMOMA and ROOMw) (**Suppl. Table S5**). **Figure 1** shows predicted biomass production rates *vs*. observed growth rate or fitness for the seven constraint-based methods across metabolic gene knockouts from four different experimental datasets. Visual inspection does not indicate any strong relationship between predictions and observations. This impression is confirmed by statistical analysis. While Spearman rank correlation coefficients are positive and are in almost all cases statistically significant (P<0.05), their values are always below ρ=0.3 (**Suppl. Tables S5-S8**).

To estimate what proportion of the variation in observed fitness is explained by the model, we calculated coefficients of determination, defined as $COD \coloneqq 1 - SS_{residual}/SS_{total}$, where $SS_{residual}$ is the residual sum of squares (a measure of unexplained variation), and SS_{total} is the total sum of squares (a measure of total variation). Almost all *COD* values are negative (**Suppl. Tables S5-S8**), indicating that the mean of the observed data provides a better description of the observations than does any of the constraint-based models. FBA with molecular crowding performs slightly better than the other methods according to this measure, and is the only method that provides a positive *COD* for at least one data set (**Suppl. Table S5**).

Figure 1 and **Suppl. Tables S5-S8** are based on the yeast7.6 model; qualitatively very similar results are obtained with the NAD-modified iMM904 model [28] (**Suppl. Figures S1-S2** and **Suppl. Tables S9-S10**).

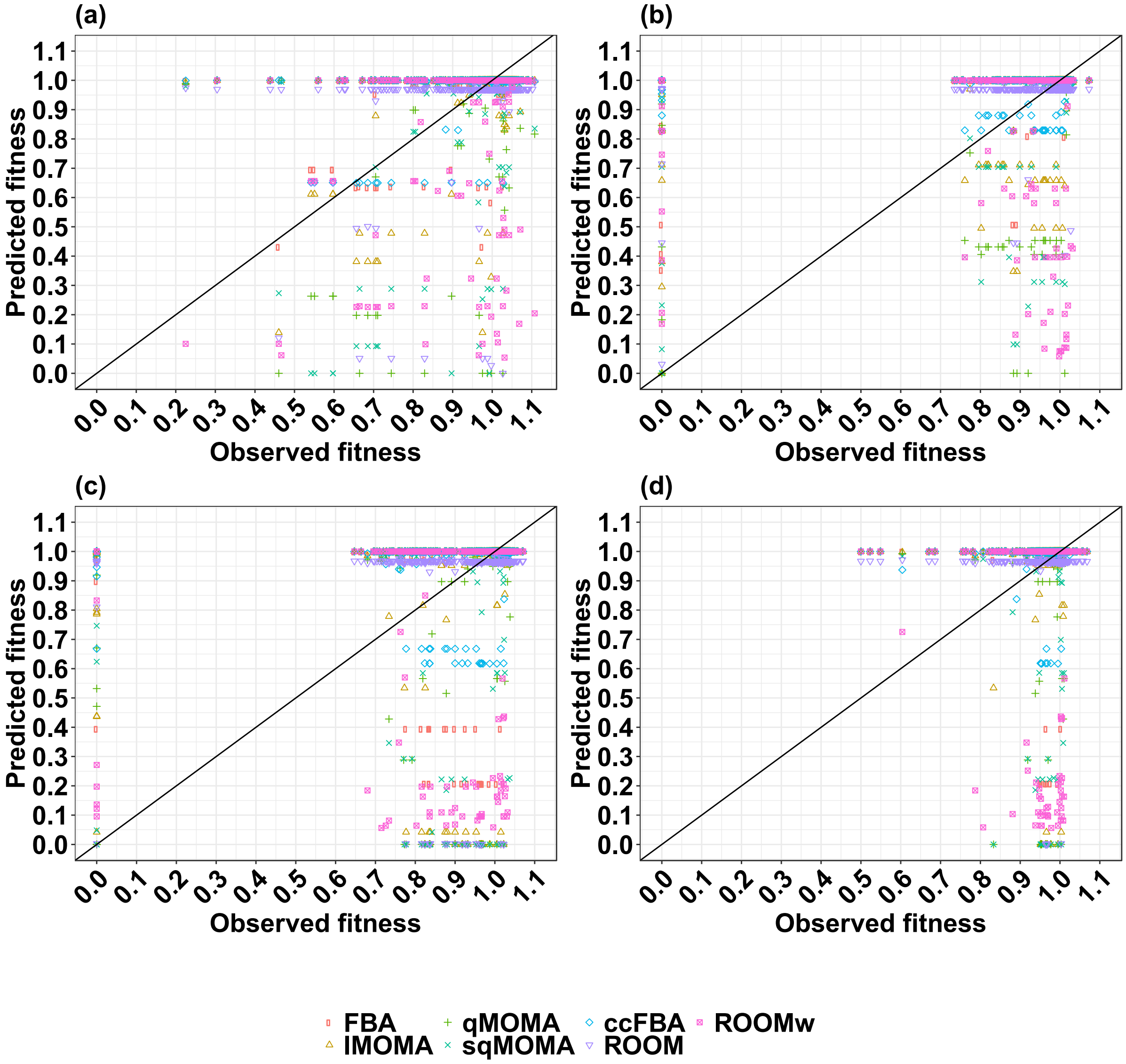


Figure 1. The (absence of) correlation between predicted and observed single-mutant growth rate and fitness for non-essential genes in the yeast *S. cerevisiae* for different knockout datasets. **(a)** Colony size (growth rate) data from Szappanos *et al.* [28]. **(b)** Parallel fitness profiling data from Deutschbauer *et al.* [30] on YPD. **(c)** Parallel fitness profiling data from Deutschbauer *et al.* [30] on minimal medium. **(d)** Competitive fitness data from Breslow *et al.* [31]. Note that the ROOM results are calculated with a relative tolerance δ=3%; this explains the large number of predicted fitness values at 0.97, which are indistinguishable from wildtype fitness within this tolerance.

Qualitative prediction of non-essential gene knockout fitness effects for S. cerevisiae

The negative *COD* values indicate a failure of all tested methods to predict fitness effects of gene knockouts in *S. cerevisiae* quantitatively. At the same time, statistically significant positive Spearman rank correlation coefficients for all methods across all datasets (with only one exception, **Suppl. Tables S5-S8**) indicate that the models capture fitness effects of some knockouts at least qualitatively. To further explore the qualitative prediction capabilities of constraint-based simulation methods, we classified non-essential gene knockouts into those with and without observed fitness effects (Methods). This resulted in a total of between 130 and 222 non-essential gene knockouts with experimentally observed deleterious fitness effects for the different datasets (**Table 1**). **Suppl. Tables S11-S14** list the numbers of true and false positives and negatives of the individual methods and the resulting values for recall (sensitivity), precision, specificity, and accuracy across *S. cervisiae* datasets. We scored any detectable reduction in predicted biomass production rate *f* as a fitness reduction (*i.e.*, f<0.9985 for all methods except ROOM, and f<0.0955 for ROOM; see Methods).

Specificities are above 0.9 for all methods across all datasets and are above 0.95 in most cases (**Suppl. Tables S11-S14**). Thus, all methods correctly predict wildtype biomass production rates for more than 95% of the knockouts without experimentally observable fitness defects. In contrast, no method is able to predict more than 25% of confidently observed deleterious fitness effects, *i.e.*, recall (sensitivity) is always below 0.25 (**Table 1**). We have to conclude that more than 75% of the fitness effects of metabolic gene knockouts in *S. cerevisiae* are not detectable with any of the tested methods. It is noteworthy that this low recall is still associated with precision values of around 0.6 (**Suppl. Tables S11-S14**), *i.e.*, around 60% of the fitness defects predicted by any of the methods are not confirmed by data. The only exception is ROOM, which – at very low numbers of positive predictions – has precision values between 0.8 and 0.9 in three of the four datasets.

As seen from **Table 1**, predictions with either the quadratic (qMOMA) or the linear (lMOMA) variant of MOMA appear to perform best in terms of their ability to predict deleterious knockout effects qualitatively, outperforming not only FBA but also FBA with molecular crowding (ccFBA) in terms of recall values (sensitivity). In contrast, the second method specifically developed to predict non-optimal physiological responses in non-essential gene knockouts, ROOM, exhibits the lowest recall values of all tested methods, less than half of those achieved with the two MOMA variants; note, however, that this is associated with a lower fraction of false positives (*i.e.,* a higher precision).

The results in **Table 1** and in **Suppl. Tables S11-S14** were obtained in an attempt to maximize recall by scoring the tiniest detectable reduction in predicted biomass production rate as a fitness defect. This was associated with high false positive rates (low precision, **Suppl. Tables S11-S14**), and it is generally more desirable to strike a balance between recall and precision. This can be achieved by scoring only fitness reductions larger than an appropriate cutoff *c* as deleterious. Each data point in **Figure 2** corresponds to a possible value for *c*; the resulting receiver operator characteristic (ROC) curve maps the possible combinations of recall (true positive rate) and 1–specificity (false positive rate). The diagonal of the figure signifies random expectations, and data points far away from the diagonal are thus preferable; the area under the curve is a measure of prediction accuracy.

Regardless of how many false positives are accepted (how low we allow specificity to drop), no individual method achieves recall values exceeding 0.25 in any of the datasets tested, consistent with what is seen in **Table 1**. When pooling all methods by labeling any knockout as deleterious for which at least one of the seven methods predicts reduced biomass production rates ("Min" in **Figure 2**), recall can be increased slightly, up to at most 0.31. Consistent with the resultsin **Table 1**, the best recall/specificity tradeoffs overall seem to be provided by the two MOMA variants (lMOMA and qMOMA, **Figure 2**). Thus, while ROOM was shown to perform better than qMOMA when tested on a small number of knockouts [10], this performance advantage cannot be replicated on large-scale datasets of *S. cerevisiae* metabolic knockouts.

Table 1. Recall (sensitivity) values for the seven tested methods for the prediction of negative fitness effects in four *S. cerevisiae* datasets of metabolic gene knockouts

	N [1]	**FBA**	**lMOMA**	**qMOMA**	**sqMOMA**	**ccFBA**	**ROOM**	**ROOMw**
Szappanos 2011	133	0.195	0.218	0.241	0.248	0.180	0.113	0.233
Deutschbauer 2005, YPD	194	0.067	0.211	0.216	0.216	0.134	0.062	0.155
Deutschbauer 2005, minimal	222	0.180	0.198	0.221	0.221	0.185	0.086	0.162
Breslow 2008	130	0.115	0.146	0.154	0.154	0.131	0.031	0.146

[1] Number of positives (strains with experimentally observed fitness at least 2 SD below the wildtype)

Only weak correlations between predicted biomass production rates and observed fitness for non-essential E. coli metabolic gene knockouts

We repeated the same analyses for metabolic non-essential gene knockout strains of *E. coli*. Biomass production rate predictions varied slightly more for the *E. coli* model than for the yeast model, with between 12 (ROOM) and 58 (ROOMw) distinct values for the 1175 non-essential metabolic genes covered by the Takeuchi *et al.* data [35] (**Suppl**. **Table S16**). **Figure 3** shows the biomass production rates predicted with the seven methods *vs*. observed growth rate [34,35] and biomass yield [35]. Similar to the yeast data, there is no obvious correlation between predictions and observations. Spearman's rank correlation coefficients are at best marginally statistically significant for the Fuhrer *et al.* growth rate data (**Suppl. Table S15**). However, they are positive and highly statistically significant for both growth rate and yield observed by Takeuchi *et al.*, with correlation coefficients around 0.17 for growth rate (**Suppl. Table S16**) and around 0.21 for yield (**Suppl. Table S17**). The only exception is ccFBA, for which the correlation is only marginally statistically significant also in the Takeuchi *et al.* data.

As before, we calculated coefficients of determination, $COD \coloneqq 1 - SS_{residual}/SS_{total}$ to estimate what proportion of the variation in observed growth rate or yield is explained by the model. For the growth rate data of Fuhrer *et al.*, all *COD* values are again negative (**Suppl. Table S15**), indicating that the mean of the observed data provides a better quantitative description of the observations than does any of the constraint-based models. The situation is slightly improved for the Takeuchi *et al.* data. Standard FBA and ROOM provide at least slightly positive *COD* values for both growth rate and yield, while ccFBA provides a marginally positive value at least for yield (**Suppl. Tables S16-S17**). The highest predictive power is observed for standard FBA and the experimental yield data, with *COD*=0.136. Thus, the best-case scenario appears to be that constraint-based methods are able to explain around 14% of the experimentally observed variation in biomass yield in *E. coli*.

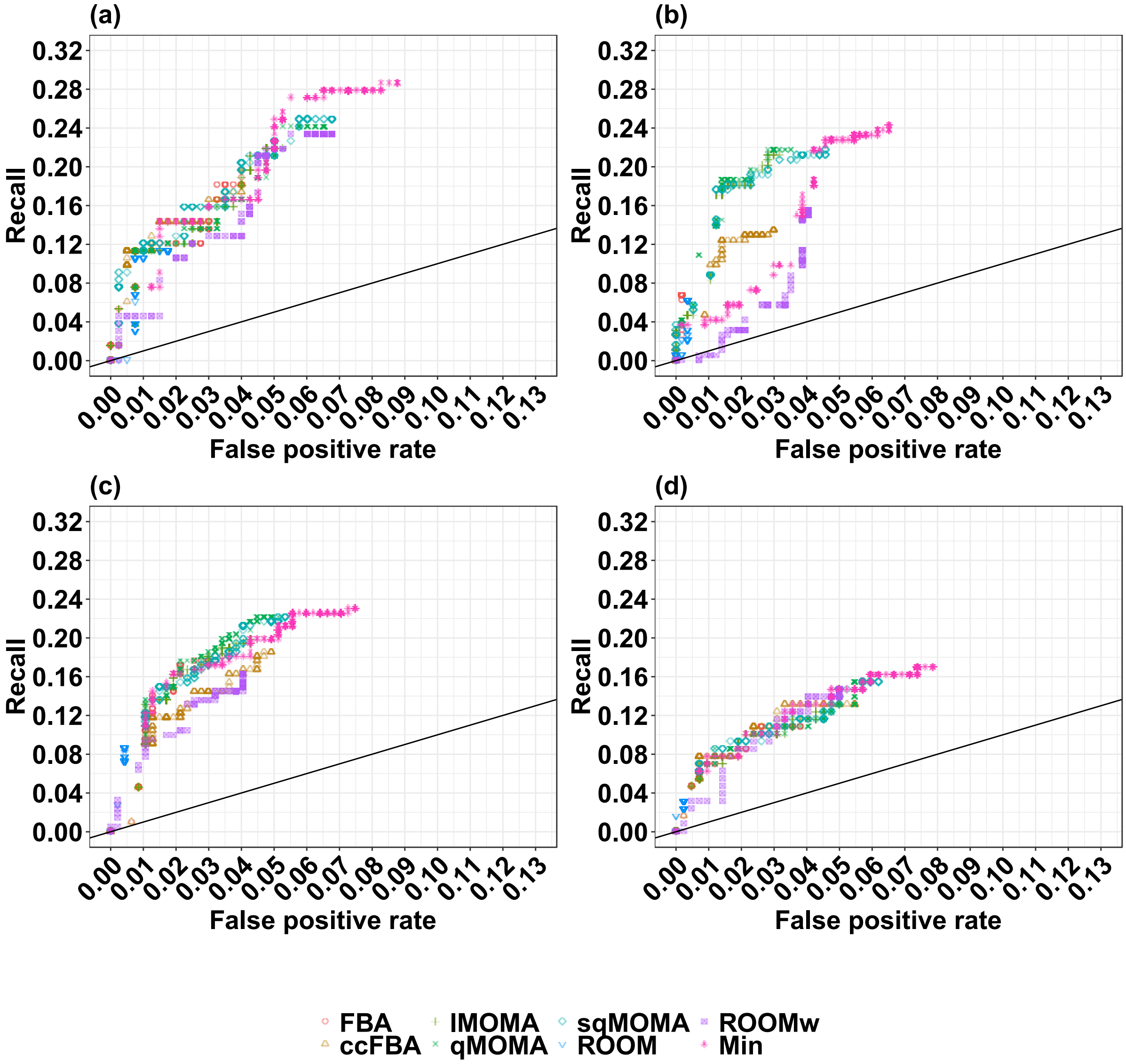


Figure 2. ROC curves for the qualitative prediction of negative fitness effects of non-essential gene knockouts in *S. cerevisiae*. “Min” uses the minimal fitness prediction for each knockout across all seven constraint-based methods, *i.e.*, it designates a knockout as deleterious if at least one method predicts a reduced biomass production rate. **(a)** Colony size (growth rate) data from Szappanos *et al.* [28]. **(b)** Parallel fitness profiling data from Deutschbauer *et al.* [30] on YPD. **(c)** Parallel fitness profiling data from Deutschbauer *et al.* [30] on minimal medium. **(d)** Competitive fitness data from Breslow *et al.* [30]. “Positives” are gene knockouts whose experimentally observed fitness is at least 2 standard deviations below the mean (z<2). See **Suppl. Figure S3** for individual panels per method.

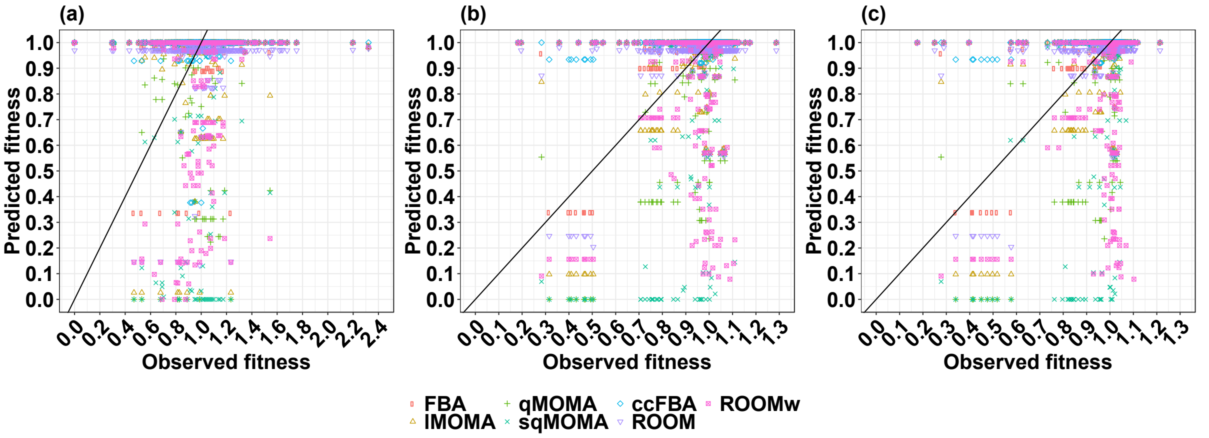


Figure 3. The (absence of) correlation between predicted and observed single-mutant growth rates for non-essential genes in *E. coli* for different knockout datasets. **(a)** Growth rate data from Fuhrer *et al.* [34]. **(b)** Growth rate data from Takeuchi *et al.* [35]. **(c)** Yield data from Takeuchi *et al.* [35]. Note that the ROOM results are calculated with a relative tolerance δ=3%; this explains the large number of predicted fitness values at 0.97, which are indistinguishable from wildtype fitness within this tolerance.

Qualitative prediction of non-essential gene knockout fitness effects for E. coli

As for the yeast data, the statistically significant rank correlation between predicted biomass production rates and the growth rate and yield data of Takeuchi *et al.* (**Suppl. Tables S16-S17**) indicates some ability of the models to predict deleterious fitness effects at least qualitatively. We thus again classified non-essential gene knockouts into those with and without observed fitness effects based on *z*-scores. This resulted in a total of between 84 and 168 non-essential gene knockouts with experimentally observed deleterious effects for the different *E. coli* datasets (**Table 2**).

Specificities are above 0.9 for all methods also across the *E. coli* datasets (**Suppl. Tables S18-S20**). As shown in **Table 2**, recall is extremely low for the Fuhrer growth rate data (<0.12). In contrast, for all methods except ROOM and ccFBA, recall is >0.31 for growth rate and >0.4 for yield data from Takeuchi *et al.*, with only minor differences between methods. Note, however, that these seemingly good performances have their downside in very low precision values around 0.3 (**Suppl. Tables S18-S20**).

The ROC curves (**Figure 4, Supplementary Figure S4**) confirm these impressions: predictions for the growth rate data of Fuhrer *et al.* (**Figure 4a**) are barely better than random guesses at any cutoff. In contrast, predictions by the merged classifier (“Min”) and by all MOMA variants perform favorably on the data from Takeuch *et al.*; they quickly reach recall values around 0.3 for growth rate (**Figure 4b, Supplementary Figure S4b**) and around 0.4 for yield (**Figure 4c, Supplementary Figure S4c**), at false positive rates of only a few percent. Interestingly, this is not only true for the original (quadratic, qMOMA) and linear (lMOMA) variants of MOMA, but especially for the newly introduced scaled MOMA version (sqMOMA). FBA and ROOMw reach similar recall values as the MOMA variants, but only catch up at higher false positive rates (**Figure 4**). ROOM and ccFBA perform considerably worse on the Takeuchi *et al.* data, strengthening the above conclusion that the previously reported superior performance of ROOM compared to MOMA and FBA [10] cannot be replicated on large-scale datasets.

Thus, the different MOMA versions appear to provide reasonable qualitative predictions for deleterious knockout effects on *E. coli* growth rate and yield. However, even these methods do not reach recall values above 0.5, indicating that also for *E. coli*, about half of deleterious knockout effects are unpredictable by current constraint-based methodologies.

Table 2. Recall (sensitivity) values for the seven tested methods for the prediction of negative fitness effects in three *E. coli* datasets of metabolic gene knockouts

	N [1]	**FBA**	**lMOMA**	**qMOMA**	**sqMOMA**	**ccFBA**	**ROOM**	**ROOMw**
Fuhrer growth rate	168	0.083	0.083	0.101	0.101	0.036	0.018	0.113
Takeuchi growth rate	88	0.318	0.318	0.341	0.341	0.114	0.159	0. 352
Takeuchi yield	84	0.405	0.417	0.440	0.440	0.167	0.214	0.440

[1] Number of positives (knockout strains whose experimentally observed fitness / growth rate / yield is at least 2 SD below the wildtype)

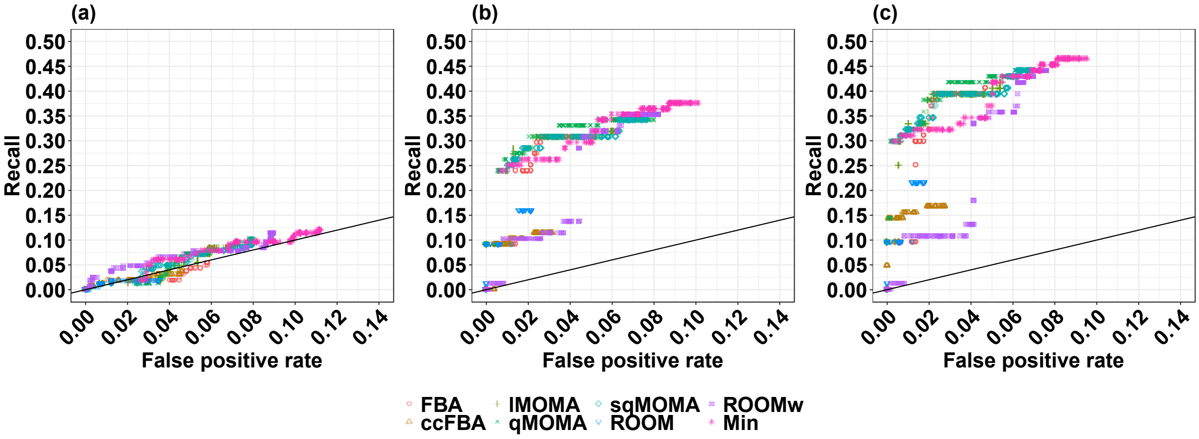


Figure 4. ROC curves for the qualitative prediction of negative fitness effects of non-essential gene knockouts in *E. coli*. “Min” uses the minimal fitness prediction for each knockout across all seven constraint-based methods, *i.e.*, it designates a knockout as deleterious if at least one method predicts a reduced biomass production rate. **(a)** Growth rate data from Fuhrer *et al.* 2014 [34]. **(b)** Growth rate data from Takeuchi *et al.* 2014 [35]. **(c)** Yield data from Takeuchi *et al.* 2014 [35]. “Positives” are gene knockouts whose experimentally observed fitness is at least 2 standard deviations below the mean (*z*<2). See **Suppl. Figure S4** for individual panels per method.

Conclusions

We found that neither FBA nor any of the six examined alternative constraint-based methods is able to predict gene knockout fitness in *S. cerevisiae* or *E. coli* quantitatively to any reasonable degree of accuracy (**Figure 1** and **Figure 3; Suppl. Tables S5-S8** and **S15-17**). This sobering conclusion holds not only for methods that assume metabolic optimality (FBA, ccFBA), but also for methods specifically developed to predict the deleterious effects of non-essential metabolic gene knockouts (the different variants of MOMA and ROOM). The quantitative estimates provided by the methods tested are not (substantially) better than a trivial "model" that predicts the same biomass production rate for all knockouts, as evidenced by the negative or very small *COD* values.

Predicting epistasis relies on a comparison of estimated double-mutant and single-mutant fitness. Thus, predictions for genetic interactions with a metabolic model can only be as good as the quantitative predictions for single mutants. Given our above observations, it appears no wonder that only a small percentage of genetic interactions can be predicted successfully by FBA or MOMA [28].

FBA with molecular crowding, as implemented in ccFBA, has been shown to predict maximal yeast growth rates across different media with an average relative error of only 8% [36]. Thus, the failure of ccFBA to predict growth rates of yeast knockout mutants is likely not rooted in an erroneous wildtype model. Instead, it appears likely that non-essential metabolic gene knockouts result in changes in internal metabolite concentrations, which are misinterpreted by the cellular signalling systems as the result of external conditions. Accordingly, the resulting regulatory responses are misguided, and are thus not predictable by models ignorant of regulatory feedbacks.

It has previously been argued that the physiology of knockout strains has not been optimized by natural selection unless sufficient time for evolutionary adaptation is provided, and thus that methods such as FBA and ccFBA that rely on the assumption of metabolic optimality may not be suited for the prediction of knockout effects [10,12]. Our results demonstrate that methods such as MOMA and ROOM, which minimize the difference between the flux distributions of wildtype and knockout mutant, also do not provide a satisfactory approximation to the regulatory feedbacks.

While quantitative growth predictions for metabolic gene knockouts appear to be outside of the scope reachable without explicit models of regulation, **Figure 2** and in particular **Figure 4** show that qualitative predictions are indeed possible at least for some knockout mutants. At false positive rates of a few percent, around 20% of knockouts in *S. cerevisiae* with deleterious fitness effects and up to 40% of knockouts in *E. coli* with yield reductions are predictable by MOMA; interestingly, this is true both for the original (quadratic) as well as the linearized version of MOMA.

References

1 Orth, J. D., Thiele, I. & Palsson, B. O. What is flux balance analysis? *Nat Biotechnol* **28**, 245-248, doi:10.1038/nbt.1614 (2010).

2 WATSON, M. R. Metabolic maps for the Apple II. *Biochemical Society Transactions* **12**, 1093-1094, doi:10.1042/bst0121093 (1984).

3 Fell, D. A. & Small, J. R. Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochem J* **238**, 781-786 (1986).

4 Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nature reviews. Microbiology* **10**, 291-305, doi:10.1038/nrmicro2737 (2012).

5 Papp, B., Szappanos, B. & Notebaart, R. A. Use of genome-scale metabolic models in evolutionary systems biology. *Methods in molecular biology (Clifton, N.J.)* **759**, 483-497, doi:10.1007/978-1-61779-173-4_27 (2011).

6 Hartleb, D., Jarre, F. & Lercher, M. J. Improved Metabolic Models for E. coli and Mycoplasma genitalium from GlobalFit, an Algorithm That Simultaneously Matches Growth and Non-Growth Data Sets. *PLoS computational biology* **12**, e1005036, doi:10.1371/journal.pcbi.1005036 (2016).

7 Duarte, N. C., Herrgard, M. J. & Palsson, B. O. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res* **14**, 1298-1309, doi:10.1101/gr.2250904 (2004).

8 Forster, J., Famili, I., Palsson, B. O. & Nielsen, J. Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae. *OMICS* **7**, 193-202, doi:10.1089/153623103322246584 (2003).

9 Kuepfer, L., Sauer, U. & Blank, L. M. Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome Res* **15**, 1421-1430, doi:10.1101/gr.3992505 (2005).

10 Shlomi, T., Berkman, O. & Ruppin, E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A* **102**, 7695-7700, doi:10.1073/pnas.0406346102 (2005).

11 Snitkin, E. S. & Segre, D. Optimality criteria for the prediction of metabolic fluxes in yeast mutants. *Genome informatics. International Conference on Genome Informatics* **20**, 123-134 (2008).

12 Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* **99**, 15112-15117, doi:10.1073/pnas.232349399 (2002).

13 VanderSluis, B. *et al.* Broad metabolic sensitivity profiling of a prototrophic yeast deletion collection. *Genome Biology* **15**, R64, doi:10.1186/gb-2014-15-4-r64 (2014).

14 Mo, M. L., Palsson, B. Ø. & Herrgård, M. J. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Systems Biology* **3**, 37, doi:10.1186/1752-0509-3-37 (2009).

15 Heavner, B. D., Smallbone, K., Barker, B., Mendes, P. & Walker, L. P. Yeast 5 – an expanded reconstruction of the Saccharomyces cerevisiae metabolic network. *BMC Systems Biology* **6**, 55, doi:10.1186/1752-0509-6-55 (2012).

16 Adadi, R., Volkmer, B., Milo, R., Heinemann, M. & Shlomi, T. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS computational biology* **8**, e1002575, doi:10.1371/journal.pcbi.1002575 (2012).

17 Beg, Q. K. *et al.* Intracellular crowding defines the mode and sequence of substrate uptake by <em>Escherichia coli</em> and constrains its metabolic activity. *Proceedings of the National Academy of Sciences* **104**, 12663-12668, doi:10.1073/pnas.0609845104 (2007).

18 Desouki, A. A. *Algorithms for improving the predictive power of flux balance analysis*, Heinrich Heine University Duesseldorf, (2016).

19 Watson, M. R. Metabolic Maps for the Apple-II. *Biochem Soc T* **12**, 1093-1094, doi:DOI 10.1042/bst0121093 (1984).

20 Holzhutter, H. G. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur J Biochem* **271**, 2905-2922, doi:10.1111/j.1432-1033.2004.04213.x (2004).

21 Machado, D. & Herrgård, M. Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. *PLoS computational biology* **10**, e1003580, doi:10.1371/journal.pcbi.1003580 (2014).

22 Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *P Natl Acad Sci USA* **99**, 15112-15117, doi:10.1073/pnas.232349399 (2002).

23 Desouki, A. A., Jarre, F., Gelius-Dietrich, G. & Lercher, M. J. CycleFreeFlux: efficient removal of thermodynamically infeasible loops from flux distributions. *Bioinformatics (Oxford, England)* **31**, 2159-2165, doi:10.1093/bioinformatics/btv096 (2015).

24 Gelius-Dietrich, G., Desouki, A. A., Fritzemeier, C. J. & Lercher, M. J. Sybil--efficient constraint-based modelling in R. *BMC Syst Biol* **7**, 125, doi:10.1186/1752-0509-7-125 (2013).

25 R Development Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, <http://www.R-project.org.> (2008).

26 Gudmundsson, S. & Thiele, I. Computationally efficient flux variability analysis. *BMC Bioinformatics* **11**, 489, doi:10.1186/1471-2105-11-489 (2010).

27 Aung, H. W., Henry, S. A. & Walker, L. P. Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism. *Industrial biotechnology (New Rochelle, N.Y.)* **9**, 215-228, doi:10.1089/ind.2013.0013 (2013).

28 Szappanos, B. *et al.* An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat Genet* **43**, 656-662, doi:10.1038/ng.846 (2011).

29 Baryshnikova, A. *et al.* Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nature Methods* **7**, 1017, doi:10.1038/nmeth.1534 https://www.nature.com/articles/nmeth.1534 - supplementary-information (2010).

30 Deutschbauer, A. M. *et al.* Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* **169**, 1915-1925, doi:10.1534/genetics.104.036871 (2005).

31 Breslow, D. K. *et al.* A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nat Methods* **5**, 711-718, doi:10.1038/nmeth.1234 (2008).

32 Orth, J. D. *et al.* A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Molecular systems biology* **7**, 535, doi:10.1038/msb.2011.65 (2011).

33 Fuhrer, T., Zampieri, M., Sevin, D. C., Sauer, U. & Zamboni, N. Genomewide landscape of gene-metabolome associations in Escherichia coli. *Molecular systems biology* **13**, 907, doi:10.15252/msb.20167150 (2017).

34 Fuhrer, T., Zampieri, M., Sévin, D. C., Sauer, U. & Zamboni, N. Genomewide landscape of gene–metabolome associations in <em>Escherichia coli</em>. *Molecular systems biology* **13**, 907, doi:10.15252/msb.20167150 (2017).

35 Takeuchi, R. *et al.* Colony-live — a high-throughput method for measuring microbial colony growth kinetics— reveals diverse growth effects of gene knockouts in Escherichia coli. *BMC Microbiology* **14**, 171, doi:10.1186/1471-2180-14-171 (2014).

36 Sanchez, B. J. *et al.* Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular systems biology* **13**, 935, doi:10.15252/msb.20167411 (2017).

Acknowledgements

We thank Jonathan Fritzemeier for helpful discussions. This work was supported financially through fellowships by the German Academic Exchange Service (DAAD) to DA and AAD, and through German Research Foundation (DFG) grants to MJL (CRC 680 and CRC 1310).

Supplemental Figures

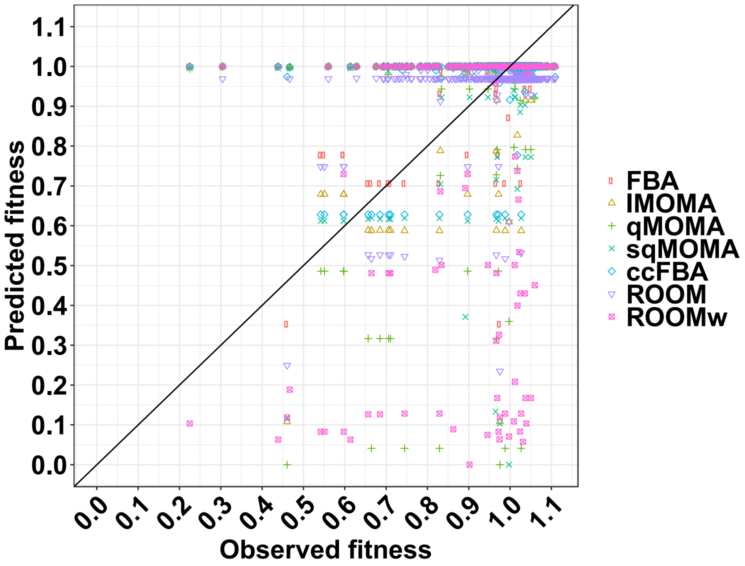


Figure S1. The (absence of) correlation observed single-mutant growth rates for non-essential genes in the yeast S. cerevisiae and the corresponding predictions based on the NAD-modified iMM904 model. Note that the ROOM results are calculated with a relative tolerance δ=3%; this explains the large number of predicted fitness values at 0.97, which are indistinguishable from wildtype fitness within this tolerance.

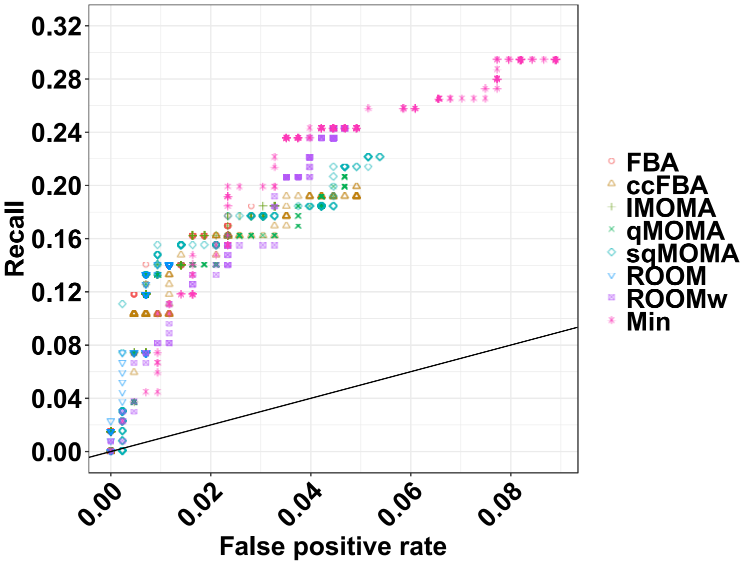


Figure S2. ROC curves for the qualitative prediction of negative fitness effects of non-essential gene knockouts in yeast based on the NAD-modified iMM904 model. “Min” uses the minimal fitness prediction for each knockout across all seven constraint-based methods. “Positives” are gene knockouts whose experimentally observed fitness is at least 2 standard deviations below the mean.

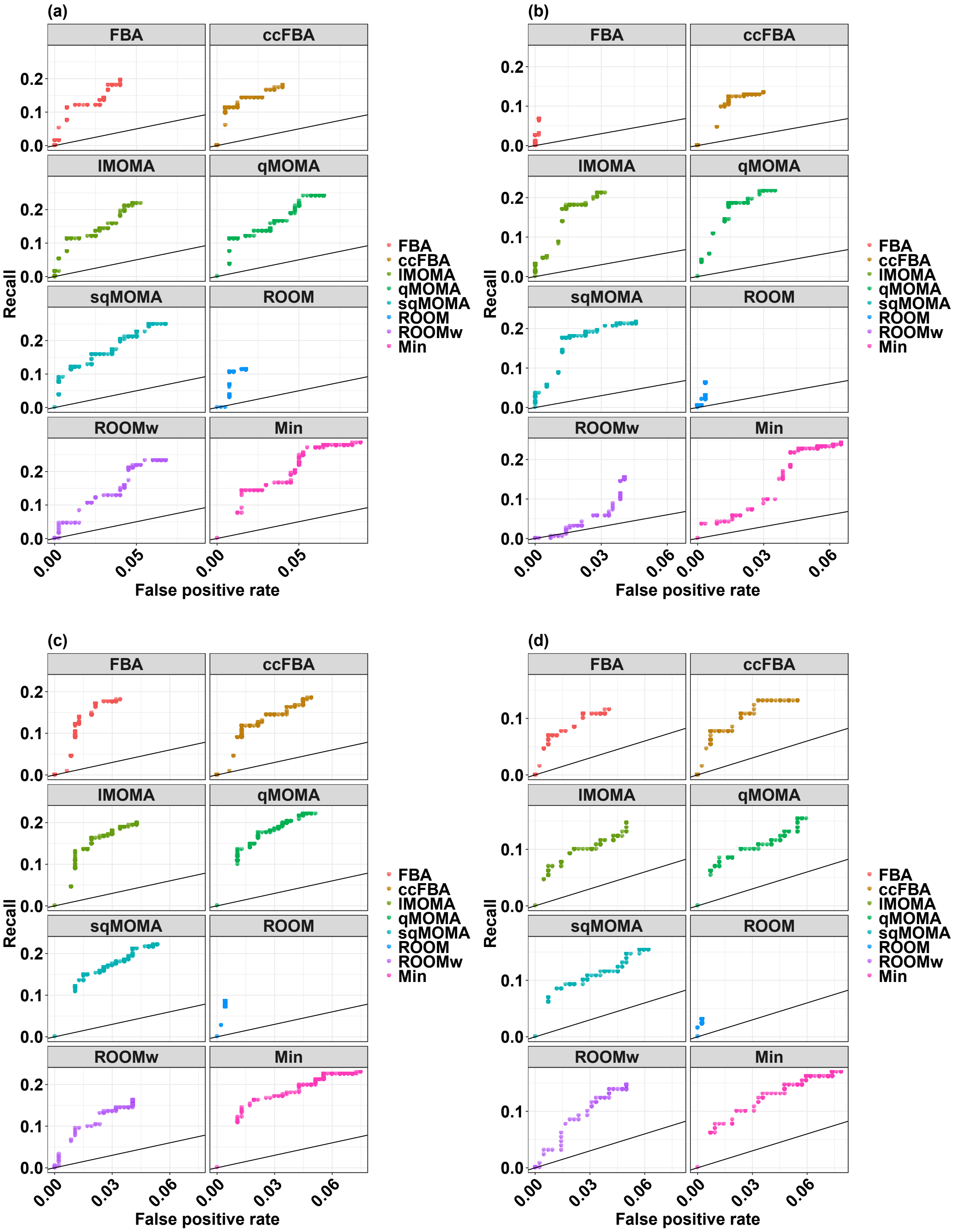


Figure S3. Individual panels for the *S. cerevisiae* ROC curves shown in **Figure 2**.
(a) Colony size data from Szappanos *et al.* (2011).
(b) Parallel fitness profiling data from Deutschbauer *et al.* (2005) on YPD.
(c) Parallel fitness profiling data from Deutschbauer *et al.* (2005) on minimal medium.
(d) Competitive fitness data from Breslow *et al.* (2008).

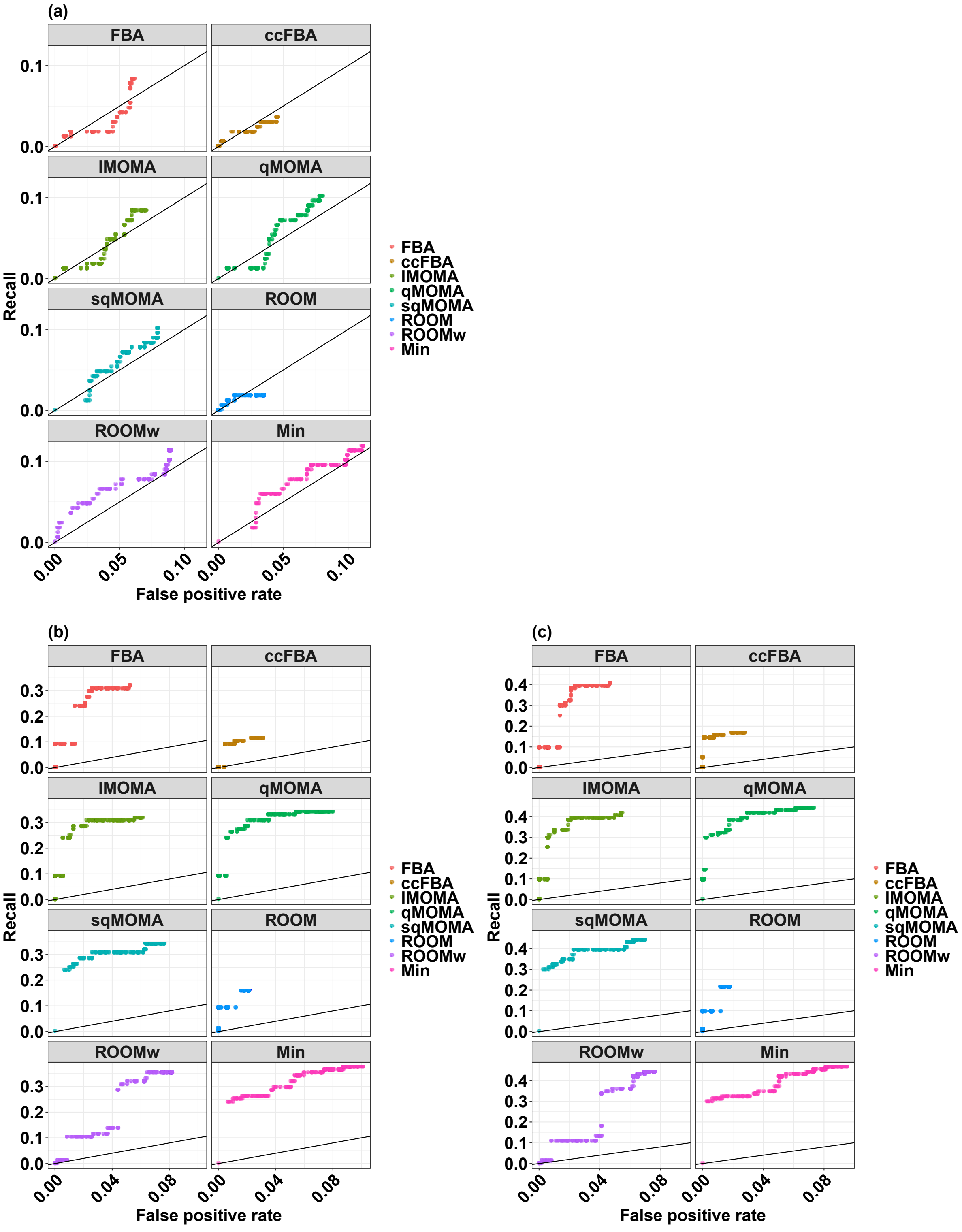


Figure S4. Individual panels for the *E. coli* ROC curves shown in **Figure 4**.
(a) Growth rate data from Fuhrer *et al.* 2014 (2017).
(b) Growth rate data from Takeuchi *et al.* 2014 (2014).
(c) Yield data from Takeuchi *et al.* 2014 (2014).

Supplemental Tables

Table S1. Medium definition for the simulations for *S. cerevisiae* Dataset 1

Metabolite	Reaction	Uptake rate
4-aminobenzoate [extracellular]	4-aminobenzoate exchange	-2,00E-06
adenine [extracellular]	adenine exchange	-3.01
L-alanine [extracellular]	L-alanine exchange	-0.36
L-asparagine [extracellular]	L-asparagine exchange	-0.36
L-aspartate [extracellular]	L-aspartate exchange	-0.36
biotin [extracellular]	biotin exchange	-1.42e-06
L-cysteine [extracellular]	L-cysteine exchange	-0.36
iron [extracellular]	iron(2+) exchange	-1000
D-glucose [extracellular]	D-glucose exchange	-22.6
L-glutamine [extracellular]	L-glutamine exchange	-0.36
L-glutamate [extracellular]	L-glutamate exchange	-3.6
L-glycine [extracellular]	glycine exchange	-0.36
L-isoleucine [extracellular]	L-isoleucine exchange	-0.36
myo-inositol [extracellular]	myo-inositol exchange	-0.11
potassium [extracellular]	potassium exchange	-4.44
L-leucine [extracellular]	L-leucine exchange	-1.8
L-methionine [extracellular]	L-methionine exchange	-0.36
pyruvate [extracellular]	pyruvate exchange	-0.75
nicotinate [extracellular]	nicotinate exchange	-2,00E-06
oxygen [extracellular]	oxygen exchange	-6.3
L-phenylalanine [extracellular]	L-phenylalanine exchange	-0.36
phosphate [extracellular]	phosphate exchange	-0.89
(R)-pantothenate [extracellular]	(R)-pantothenate exchange	-2,00E-04
L-proline [extracellular]	L-proline exchange	-0.36
riboflavin [extracellular]	riboflavin exchange	-0.00092
L-serine [extracellular]	L-serine exchange	-0.36
sulphate [extracellular]	sulphate exchange	-100
thiamine [extracellular]	thiamine(1+) exchange	-0.0032
L-threonine [extracellular]	L-threonine exchange	-0.36
L-tryptophan [extracellular]	L-tryptophan exchange	-0.36
L-tyrosine [extracellular]	L-tyrosine exchange	-0.36
uracil [extracellular]	uracil exchange	-3.36
L-valine [extracellular]	L-valine exchange	-0.36

Table S2. *M*edium definition for the simulations for *S. cerevisiae* Dataset 2, YPD

Metabolite	Reaction	Uptake rate
4-aminobenzoate [extracellular]	4-aminobenzoate exchange	-2,00E-06
adenine [extracellular]	adenine exchange	-3.01
L-alanine [extracellular]	L-alanine exchange	-0.36
L-asparagine [extracellular]	L-asparagine exchange	-0.36
L-aspartate [extracellular]	L-aspartate exchange	-0.36
biotin [extracellular]	biotin exchange	-1.42e-06
L-cysteine [extracellular]	L-cysteine exchange	-0.36
iron [extracellular]	iron(2+) exchange	-1000
D-glucose [extracellular]	D-glucose exchange	-22.6
L-glutamine [extracellular]	L-glutamine exchange	-0.36
L-glutamate [extracellular]	L-glutamate exchange	-3.6
L-glycine [extracellular]	glycine exchange	-0.36
L-isoleucine [extracellular]	L-isoleucine exchange	-0.36
myo-inositol [extracellular]	myo-inositol exchange	-0.11
potassium [extracellular]	potassium exchange	-4.44
L-leucine [extracellular]	L-leucine exchange	-1.8
L-methionine [extracellular]	L-methionine exchange	-0.36
pyruvate [extracellular]	pyruvate exchange	-0.75
nicotinate [extracellular]	nicotinate exchange	-2,00E-06
oxygen [extracellular]	oxygen exchange	-6.3
L-phenylalanine [extracellular]	L-phenylalanine exchange	-0.36
phosphate [extracellular]	phosphate exchange	-0.89
(R)-pantothenate [extracellular]	(R)-pantothenate exchange	-2,00E-04
L-proline [extracellular]	L-proline exchange	-0.36
riboflavin [extracellular]	riboflavin exchange	-0.00092
L-serine [extracellular]	L-serine exchange	-0.36
sulphate [extracellular]	sulphate exchange	-100
thiamine [extracellular]	thiamine(1+) exchange	-0.0032
L-threonine [extracellular]	L-threonine exchange	-0.36
L-tryptophan [extracellular]	L-tryptophan exchange	-0.36
L-tyrosine [extracellular]	L-tyrosine exchange	-0.36
uracil [extracellular]	uracil exchange	-3.36
L-valine [extracellular]	L-valine exchange	-0.36
ammonium [extracellular]	ammonium exchange	-1000
sodium [extracellular]	sodium exchange	-1000
choline [extracellular]	choline exchange	-1000
inosine [extracellular]	inosine exchange	-1000
pyridoxine [extracellular]	pyridoxine exchange	-1000
L-histidine [extracellular]	L-histidine exchange	-0.36
L-arginine [extracellular]	L-arginine exchange	-0.36
"L-lysine [extracellular]	L-lysine exchange	-0.36

Table S3. Medium definition for the simulations for *E. coli* Dataset 1

Metabolite	**Reaction**	**Uptake rate**
glc_DASH_D[e]	EX_glc(e)	-10
k[e]	EX_k(e)	-1000
thm[e]	EX_thm(e)	-1000
na1[e]	EX_na1(e)	-1000
ca2[e]	EX_ca2(e)	-1000
nh4[e]	EX_nh4(e)	-1000
mg2[e]	EX_mg2(e)	-1000
fe3[e]	EX_fe3(e)	-1000
zn2[e]	EX_zn2(e)	-1000
cu2[e]	EX_cu2(e)	-1000
mn2[e]	EX_mn2(e)	-1000
cobalt2[e]	EX_cobalt2(e)	-1000
cl[e]	EX_cl(e)	-1000
co2[e]	EX_co2(e)	-1000
h[e]	EX_h(e)	-1000
h2o[e]	EX_h2o(e)	-1000
so4[e]	EX_so4(e)	-1000
pi[e]	EX_pi(e)	-1000
o2[e]	EX_o2(e)	-1000
mobd[e]	EX_mobd(e)	-1000
ni2[e]	EX_ni2(e)	-1000
ala_DASH_L[e]	EX_ala_L(e)	-0.5
arg_DASH_L[e]	EX_arg_L(e)	-0.5
asn_DASH_L[e]	EX_asn_L(e)	-0.5
asp_DASH_L[e]	EX_asp_L(e)	-0.5
cys_DASH_L[e]	EX_cys_L(e)	-0.5
gln_DASH_L[e]	EX_gln_L(e)	-0.5
glu_DASH_L[e]	EX_glu_L(e)	-0.5
gly[e]	EX_gly(e)	-0.5
his_DASH_L[e]	EX_his_L(e)	-0.5
ile_DASH_L[e]	EX_ile_L(e)	-0.5
leu_DASH_L[e]	EX_leu_L(e)	-0.5
lys_DASH_L[e]	EX_lys_L(e)	-0.5
met_DASH_L[e]	EX_met_L(e)	-0.5
phe_DASH_L[e]	EX_phe_L(e)	-0.5
pro_DASH_L[e]	EX_pro_L(e)	-0.5
ser_DASH_L[e]	EX_ser_L(e)	-0.5
thr_DASH_L[e]	EX_thr_L(e)	-0.5
trp_DASH_L[e]	EX_trp_L(e)	-0.5
tyr_DASH_L[e]	EX_tyr_L(e)	-0.5
val_DASH_L[e]	EX_val_L(e)	-0.5

Table S4. Medium definition for the simulations for *E. coli* Dataset 2

Metabolite	Reaction	Uptake rate
cbl1[e]	EX_cbl1(e)	-0.01
ser_DASH_D[e]	EX_ser_D(e)	-0.50
ade[e]	EX_ade(e)	-0.50
ala_DASH_D[e]	EX_ala_D(e)	-0.50
ala_DASH_L[e]	EX_ala_L(e)	-0.50
arg_DASH_L[e]	EX_arg_L(e)	-0.50
asn_DASH_L[e]	EX_asn_L(e)	-0.50
asp_DASH_L[e]	EX_asp_L(e)	-0.50
btn[e]	EX_btn(e)	-0.50
csn[e]	EX_csn(e)	-0.50
cys_DASH_D[e]	EX_cys_D(e)	-0.50
cys_DASH_L[e]	EX_cys_L(e)	-0.50
gln_DASH_L[e]	EX_gln_L(e)	-0.50
glu_DASH_L[e]	EX_glu_L(e)	-0.50
gly[e]	EX_gly(e)	-0.50
gua[e]	EX_gua(e)	-0.50
his_DASH_L[e]	EX_his_L(e)	-0.50
hom_DASH_L[e]	EX_hom_L(e)	-0.50
ile_DASH_L[e]	EX_ile_L(e)	-0.50
leu_DASH_L[e]	EX_leu_L(e)	-0.50
lys_DASH_L[e]	EX_lys_L(e)	-0.50
met_DASH_L[e]	EX_met_L(e)	-0.50
nmn[e]	EX_nmn(e)	-0.50
phe_DASH_L[e]	EX_phe_L(e)	-0.50
pro_DASH_L[e]	EX_pro_L(e)	-0.50
pydxn[e]	EX_pydxn(e)	-0.50
ser_DASH_L[e]	EX_ser_L(e)	-0.50
thm[e]	EX_thm(e)	-0.50
thr_DASH_L[e]	EX_thr_L(e)	-0.50
thym[e]	EX_thym(e)	-0.50
trp_DASH_L[e]	EX_trp_L(e)	-0.50
tyr_DASH_L[e]	EX_tyr_L(e)	-0.50
ura[e]	EX_ura(e)	-0.50
val_DASH_L[e]	EX_val_L(e)	-0.50
glc_DASH_D[e]	EX_glc(e)	-10
h[e]	EX_h(e)	-1000
h2o[e]	EX_h2o(e)	-1000
ca2[e]	EX_ca2(e)	-1000
cl[e]	EX_cl(e)	-1000
co2[e]	EX_co2(e)	-1000
cobalt2[e]	EX_cobalt2(e)	-1000
cu2[e]	EX_cu2(e)	-1000
fe2[e]	EX_fe2(e)	-1000
fe3[e]	EX_fe3(e)	-1000
k[e]	EX_k(e)	-1000
mg2[e]	EX_mg2(e)	-1000
mn2[e]	EX_mn2(e)	-1000
mobd[e]	EX_mobd(e)	-1000
na1[e]	EX_na1(e)	-1000
nh4[e]	EX_nh4(e)	-1000
ni2[e]	EX_ni2(e)	-1000
o2[e]	EX_o2(e)	-1000
pi[e]	EX_pi(e)	-1000
sel[e]	EX_sel(e)	-1000
slnt[e]	EX_slnt(e)	-1000
so4[e]	EX_so4(e)	-1000
tungs[e]	EX_tungs(e)	-1000
zn2[e]	EX_zn2(e)	-1000

Table S5. Residual sum of squares for fitness predictions calculated across all non-essential genes in *S. cerevisiae.* Colony size data from Szappanos *et al.* (2011).

Method	**Residual sum of squares (RSS)[1]**	**Coefficient of determination**	**Spearman rank correlation coefficient *ρ***	***P***	**Deleterious knockout fitness predictions**	**Distinct predicted fitness values for deleterious knockouts**
FBA	5.435	-0.029	0.278	6.47E-11	42	18
lMOMA	7.319	-0.385	0.269	2.74E-10	50	23
qMOMA	13.256	-1.509	0.258	1.57E-09	58	28
sqMOMA	12.956	-1.452	0.264	6.56E-10	60	34
ccFBA	5.214	0.013	0.261	1.05E-09	40	15
ROOM	10.207	-0.932	0.237	3.07E-08	22	10
ROOMw	17.593	-2.330	0.244	1.23E-08	58	34

[1] Observed “fitness” values >0.9985 were set to 1.0 for all methods except ROOM; for ROOM, values >0.955 were set to 1.0 (see “Numerical calculations” in Methods).

Table S6. Residual sum of squares for fitness predictions calculated across all non-essential genes in *S. cerevisiae.* Parallel fitness profiling data from Deutschbauer *et al.* (2005) on YPD.

Method	**Residual sum of squares (RSS)[1]**	**Coefficient of determination**	**Spearman rank correlation coefficient *ρ***	***P***	**Deleterious knockout fitness predictions**	**Distinct predicted fitness values for deleterious knockouts**
FBA	64.497	-0.086	0.197	4.43E-08	14	7
lMOMA	64.404	-0.085	0.294	1.25E-16	59	22
qMOMA	69.190	-0.165	0.296	6.70E-17	62	18
sqMOMA	68.204	-0.149	0.263	1.68E-13	68	28
ccFBA	67.539	-0.138	0.193	8.06E-08	43	16
ROOM	65.684	-0.106	0.171	2E-06	14	7
ROOMw	78.468	-0.322	0.199	2.9E-8	53	31

[1] Observed “fitness” values >0.9985 were set to 1.0 for all methods except ROOM; for ROOM, values >0.955 were set to 1.0 (see “Numerical calculations” in Methods).

Table S7. Residual sum of squares for fitness predictions calculated across all non-essential genes in *S. cerevisiae.* Parallel fitness profiling data from Deutschbauer *et al.* (2005) on minimal medium.

Method	**Residual sum of squares (RSS)[1]**	**Coefficient of determination**	**Spearman rank correlation coefficient *ρ***	***P***	**Deleterious knockout fitness predictions**	**Distinct predicted fitness values for deleterious knockouts**
FBA	68.876	-0.347	0.217	7.85E-09	56	19
lMOMA	75.900	-0.484	0.219	6.36E-09	64	25
qMOMA	77.451	-0.514	0.227	1.72E-09	73	32
sqMOMA	81.421	-0.592	0.220	5.28E-09	74	36
ccFBA	61.365	-0.200	0.179	2.19E-06	64	23
ROOM	73.001	-0.427	0.164	1.56E-05	21	5
ROOMw	81.631	-0.596	0.169	7.65E-06	55	35

[1] Observed “fitness” values >0.9985 were set to 1.0 for all methods except ROOM; for ROOM, values >0.955 were set to 1.0 (see “Numerical calculations” in Methods).

Table S8. Residual sum of squares for fitness predictions calculated across all non-essential genes in *S. cerevisiae.* Competitive fitness data from Breslow *et al.* (2008) on minimal medium.

Method	**Residual sum of squares (RSS)[1]**	**Coefficient of determination**	**Spearman rank correlation coefficient *ρ***	***P***	**Deleterious knockout fitness predictions**	**Distinct predicted fitness values for deleterious knockouts**
FBA	7.368	-3.175	0.131	0.0021	32	16
lMOMA	11.510	-5.521	0.138	0.0012	40	19
qMOMA	14.073	-6.974	0.135	0.0015	45	26
sqMOMA	17.550	-8.943	0.128	0.0027	46	29
ccFBA	3.201	-0.814	0.119	0.0052	40	17
ROOM	5.907	-2.347	0.061	0.15	5	3
ROOMw	26.807	-14.188	0.104	0.014	40	32

[1] Observed “fitness” values >0.9985 were set to 1.0 for all methods except ROOM; for ROOM, values >0.955 were set to 1.0 (see “Numerical calculations” in Methods).

Table S9. Residual sum of squares for fitness predictions with the NAD-corrected iMM904 model, calculated across all non-essential genes in *S. cerevisiae*. Colony size data from Szappanos *et al.* (2011).

Method	**Residual sum of squares (RSS)**[1]	**Coefficient of determination**	**Spearman rank correlation coefficient *ρ***	***P***	**Deleterious knockout fitness predictions**	**Distinct predicted fitness values for deleterious knockouts**
FBA	4.989	0.013	0.275	3.38E-11	37	13
lMOMA	5.964	-0.180	0.259	4.16E-10	39	17
qMOMA	10.657	-1.109	0.247	2.87E-09	48	23
sqMOMA	7.929	-0.569	0.251	1.63E-09	53	26
ccFBA	5.071	-0.004	0.224	7.48E-08	48	18
ROOM	5.963	-0.180	0.289	2.94E-12	24	13
ROOMw	22.064	-3.366	0.281	1.2E-11	53	34

[1] Observed “fitness” values >0.9985 were set to 1.0 for all methods except ROOM; for ROOM, values >0.955 were set to 1.0 (see “Numerical calculations” in Methods).

Table S10. True and false predictions of a deleterious fitness effect in *S. cerevisiae* by the seven constraint-based methods based on the NAD-corrected iMM904 model. Colony size data for *S. cerevisiae* from Szappanos *et al.* (2011).

Method	Recall	Precision	Specificity	Accuracy	Matthews' correlation coefficient	**TP**	**FP**	**FN**	**TN**
FBA	0.184	0.676	0.972	0.782	0.269	25	12	111	415
lMOMA	0.184	0.641	0.967	0.778	0.255	25	14	111	413
qMOMA	0.206	0.583	0.953	0.773	0.244	28	20	108	407
sqMOMA	0.221	0.566	0.946	0.771	0.244	30	23	106	404
ccFBA	0.199	0.563	0.951	0.769	0.229	27	21	109	406
ROOM	0.140	0.792	0.988	0.783	0.271	19	5	117	422
ROOMw	0.243	0.623	0.953	0.782	0.287	33	20	103	407

Table S11. True and false predictions of a deleterious fitness effect by the seven constraint-based methods. Colony size data for *S. cerevisiae* from Szappanos *et al.* (2011).

Method	Recall	Precision	Specificity	Accuracy	Matthews' correlation coefficient	**TP**	**FP**	**FN**	**TN**
FBA	0.195	0.619	0.960	0.769	0.250	26	16	107	383
lMOMA	0.218	0.580	0.947	0.765	0.245	29	21	104	378
qMOMA	0.241	0.552	0.935	0.761	0.244	32	26	101	373
sqMOMA	0.248	0.550	0.932	0.761	0.247	33	27	100	372
ccFBA	0.180	0.600	0.960	0.765	0.230	24	16	109	383
ROOM	0.113	0.682	0.982	0.765	0.207	15	7	118	392
ROOMw	0.233	0.534	0.932	0.758	0.230	31	27	102	372

Table S12. True and false predictions of a deleterious fitness effect by the seven constraint-based methods. Parallel fitness profiling data for *S. cerevisiae* from Deutschbauer *et al.* (2005) on YPD.

Method	Recall	Precision	Specificity	Accuracy	Matthews' correlation coefficient	**TP**	**FP**	**FN**	**TN**
FBA	0.067	0.929	0.998	0.761	0.212	13	1	181	567
lMOMA	0.211	0.695	0.968	0.776	0.293	41	18	153	550
qMOMA	0.216	0.677	0.965	0.774	0.289	42	20	152	548
sqMOMA	0.216	0.618	0.954	0.766	0.261	42	26	152	542
ccFBA	0.134	0.605	0.970	0.757	0.197	26	17	168	551
ROOM	0.062	0.857	0.996	0.759	0.189	12	2	182	566
ROOMw	0.155	0.566	0.960	0.755	0.195	30	23	164	545

Table S13. True and false predictions of a deleterious fitness effect by the seven constraint-based methods. Parallel fitness profiling data from Data for *S. cerevisiae* from Deutschbauer *et al.* (2005) on minimal medium.

Method	Recall	Precision	Specificity	Accuracy	Matthews' correlation coefficient	**TP**	**FP**	**FN**	**TN**
FBA	0.180	0.714	0.966	0.713	0.250	40	16	182	452
lMOMA	0.198	0.688	0.957	0.713	0.250	44	20	178	448
qMOMA	0.221	0.671	0.949	0.714	0.257	49	24	173	444
sqMOMA	0.221	0.662	0.947	0.713	0.253	49	25	173	443
ccFBA	0.185	0.641	0.951	0.704	0.218	41	23	181	445
ROOM	0.086	0.905	0.996	0.703	0.221	19	2	203	466
ROOMw	0.162	0.655	0.959	0.703	0.210	36	19	186	449

Table S14. True and false predictions of a deleterious fitness effect by the seven constraint-based methods. Competitive fitness data for *S. cerevisiae* from Breslow *et al.* (2008) on minimal medium.

Method	Recall	Precision	Specificity	Accuracy	Matthews' correlation coefficient	**TP**	**FP**	**FN**	**TN**
FBA	0.115	0.469	0.960	0.760	0.136	15	17	115	403
lMOMA	0.146	0.475	0.950	0.760	0.157	19	21	111	399
qMOMA	0.154	0.444	0.940	0.755	0.146	20	25	110	395
sqMOMA	0.154	0.435	0.938	0.753	0.141	20	26	110	394
ccFBA	0.131	0.425	0.945	0.753	0.124	17	23	113	397
ROOM	0.031	0.800	0.998	0.769	0.127	4	1	126	419
ROOMw	0.146	0.475	0.950	0.760	0.157	19	21	111	399

Table S15. Residual sum of squares for fitness predictions calculated across all non-essential genes in *E. coli.* Growth rate data for *E. coli* from Fuhrer *et al.* (2017).

Method	**Residual sum of squares (RSS)**[1]	**Coefficient of determination**	**Spearman rank correlation coefficient *ρ***	***P***	**Deleterious knockout fitness predictions**	**Distinct predicted fitness values for deleterious knockouts**
FBA	14.834	-0.475	0.057	0.062	69	24
lMOMA	19.342	-0.923	0.045	0.15	77	32
qMOMA	26.388	-1.624	0.067	0.028	89	42
sqMOMA	41.563	-3.133	0.067	0.028	88	41
ccFBA	13.619	-0.354	0.052	0.092	47	15
ROOM	16.875	-0.678	-0.015	0.63	34	13
ROOMw	32.181	-2.200	0.079	0.0095	99	47

[1] Observed “fitness” values >0.9985 were set to 1.0 for all methods except ROOM; for ROOM, values >0.955 were set to 1.0 (see “Numerical calculations” in Methods).

Table S16. Residual sum of squares for fitness predictions calculated across all non-essential genes in *E. coli.* Growth rate data for *E. coli* from Takeuchi *et al.* (2014).

Method	**Residual sum of squares (RSS)**[1]	**Coefficient of determination**	**Spearman rank correlation coefficient *ρ***	***P***	**Deleterious knockout fitness predictions**	**Distinct predicted fitness values for deleterious knockouts**
FBA	7.072	0.105	0.218	4.13543E-14	85	28
lMOMA	8.071	-0.022	0.190	5.50213E-11	95	40
qMOMA	14.782	-0.871	0.183	2.41046E-10	117	48
sqMOMA	32.506	-3.115	0.173	2.376E-09	113	50
ccFBA	8.198	-0.038	0.067	0.021904913	44	15
ROOM	7.611	0.037	0.149	3.08291E-07	37	12
ROOMw	26.426	-2.345	0.154	1.04315E-07	120	58

[1] Observed “fitness” values >0.9985 were set to 1.0 for all methods except ROOM; for ROOM, values >0.955 were set to 1.0 (see “Numerical calculations” in Methods).

Table S17. Residual sum of squares for fitness predictions calculated across all non-essential genes in *E. coli.* Yield data for *E. coli* from Takeuchi *et al.* (2014).

Method	**Residual sum of squares (RSS)**[1]	**Coefficient of determination**	**Spearman rank correlation coefficient *ρ***	***P***	**Deleterious knockout fitness predictions**	**Distinct predicted fitness values for deleterious knockouts**
FBA	5.988	0.136	0.249	4.2E-18	85	28
lMOMA	7.347	-0.060	0.238	1.2E-16	95	40
qMOMA	14.396	-1.077	0.222	1.6E-14	117	48
sqMOMA	32.774	-3.728	0.229	2.1E-15	113	50
ccFBA	6.824	0.016	0.070	0.016	44	15
ROOM	6.498	0.063	0.178	7.6E-10	37	12
ROOMw	25.869	-2.732	0.207	7.5E-13	120	58

[1] Observed "fitness" values >0.9985 were set to 1.0 for all methods except ROOM; for ROOM, values >0.955 were set to 1.0 (see "Numerical calculations" in Methods).

Table S18. True and false predictions of a deleterious fitness effect by the seven constraint-based methods. Growth rate data from Fuhrer *et al.* (2017).

Method	Recall	Precision	Specificity	Accuracy	Matthews' correlation coefficient	**TP**	**FP**	**FN**	**TN**
FBA	0.083	0.203	0.939	0.803	0.032	14	55	154	840
lMOMA	0.083	0.182	0.930	0.796	0.018	14	63	154	832
qMOMA	0.101	0.191	0.920	0.790	0.027	17	72	151	823
sqMOMA	0.101	0.193	0.921	0.791	0.029	17	71	151	824
ccFBA	0.036	0.128	0.954	0.809	-0.018	6	41	162	854
ROOM	0.018	0.088	0.965	0.816	-0.035	3	31	165	864
ROOMw	0.113	0.192	0.911	0.785	0.030	19	80	149	815

Table S19. True and false predictions of a deleterious fitness effect by the seven constraint-based methods. Growth rate data from Takeuchi *et al.* (2014).

Method	Recall	Precision	Specificity	Accuracy	Matthews' correlation coefficient	**TP**	**FP**	**FN**	**TN**
FBA	0.318	0.329	0.948	0.900	0.270	28	57	60	1030
lMOMA	0.318	0.295	0.938	0.892	0.248	28	67	60	1020
qMOMA	0.341	0.256	0.920	0.877	0.229	30	87	58	1000
sqMOMA	0.341	0.265	0.924	0.880	0.236	30	83	58	1004
ccFBA	0.114	0.227	0.969	0.905	0.114	10	34	78	1053
ROOM	0.159	0.378	0.979	0.917	0.208	14	23	74	1064
ROOMw	0.352	0.258	0.918	0.876	0.235	31	89	57	998

Table S20. True and false predictions of a deleterious fitness effect by the seven constraint-based methods. Yield data from Takeuchi *et al.* (2014).

Method	Recall	Precision	Specificity	Accuracy	Matthews' correlation coefficient	**TP**	**FP**	**FN**	**TN**
FBA	0.405	0.400	0.953	0.914	0.356	34	51	50	1040
lMOMA	0.417	0.368	0.945	0.907	0.342	35	60	49	1031
qMOMA	0.440	0.316	0.927	0.892	0.316	37	80	47	1011
sqMOMA	0.440	0.327	0.930	0.895	0.324	37	76	47	1015
ccFBA	0.167	0.318	0.973	0.915	0.189	14	30	70	1061
ROOM	0.214	0.486	0.983	0.928	0.290	18	19	66	1072
ROOMw	0.440	0.308	0.924	0.889	0.310	37	83	47	1008

Acknowledgments

First of all, I thank "almighty ALLAH” for the achievement of this work. I would like to express my deepest gratitude to my supervisor Professor Dr. Martin Lercher for his constructive supervision, valuable ideas, suggestions, feedbacks, and great support during my doctoral work.

I would also like to thank many of my fellow members of the Computational Cell Biology lab for their sincere support, insights, friendships, and constant encouragement; they truly helped me a lot. Special thanks are devoted to Dr. Abdelmoneim Desouki, Dr. Jonathan Fritzemeier, Uli Wittelsbuerger, and Tin Y Pang who helped me during my doctoral work.

I wish to express my special thanks and gratitude to Dr. Abdelmoneim Desouki for his sincere support, constant encouragement and counseling in overcoming various bottlenecks during the projects and also for his comments.

Above all I would like to seize this opportunity to express my dearest gratitude to wife Ala, whose unconditional love, support, motivation, and patience have made the completion of this work possible.

Besides giving me the opportunity to conduct my research in Germany, I gratefully acknowledge the financial support received from DAAD (Deutscher Akademischer Austauschdienst). Also, I would like to thank STIBET Completion of Studies Grant of Heinrich Heine University Düsseldorf for their funding that made my doctoral work possible.

Finally, I present my profound thanks to my family: my wife Ala, my daughters Rahaf and Tala, my mother, my father, sisters, and brothers whose support and endless love have always been with me in the course of my doctoral work. Thank you.

Düsseldorf, 2019
Deya Alzoubi