Probabilistic Reconstruction and Comparative Systems Biology of Microbial Metabolism

Germán A. Plata

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY
2013

ABSTRACT

Probabilistic Reconstruction and Comparative Systems Biology of Microbial Metabolism

Germán Plata

With the number of sequenced microbial species soon to be in the tens of thousands, we are in a unique position to investigate microbial function, ecology, and evolution on a large scale. In this dissertation I first describe the use of hundreds of *in silico* models of bacterial metabolic networks to study the long-term the evolution of growth and gene-essentiality phenotypes. The results show that, over billions of years of evolution, the conservation of bacterial phenotypic properties drops by a similar fraction per unit time following an exponential decay. The analysis provides a framework to generate and test hypotheses related to the phenotypic evolution of different microbial groups and for comparative analyses based on phenotypic properties of species. Mapping of genome sequences to phenotypic predictions –such as used in the analysis just described– critically relies on accurate functional annotations. In this context, I next describe GLOBUS, a probabilistic method for genome-wide biochemical annotations. GLOBUS uses Gibbs sampling to calculate probabilities for each possible assignment of genes to metabolic functions based on sequence information and both local and global genomic context data. Several important functional predictions made by GLOBUS were experimentally validated in *Bacillus subtilis* and hundreds more were obtained across other species. Complementary to the automated annotation method, I also describe the manual reconstruction and constraints-based analysis of the metabolic network of the malaria parasite *Plasmodium falciparum*. After careful reconciliation of the model with available biochemical and phenotypic data, the high-quality reconstruction allowed the prediction and *in vivo* validation of a novel potential antimalarial target. The model was also used to contextualize different types of genome-scale data such as gene expression and metabolomics measurements. Finally, I present two projects related to population genetics aspects of sequence and genome evolution. The first project addresses the question of why highly expressed proteins evolve slowly, showing that, at least for *Escherichia coli*, this is more likely to be a consequence of selection for translational efficiency than selection to avoid misfolded protein toxicity. The second project investigates genetic robustness mediated by gene duplicates in the context of large natural microbial populations. The analysis shows that, under these conditions, the ability of duplicated yeast genes to effectively compensate for the loss of their paralogs is not a monotonic function of their sequence divergence.

TABLE OF CONTENTS

CHAPTER 1. THESIS OVERVIEW……………………………………………………001

CHAPTER 2. NEXT-GENERATION BIOCHEMICAL ANNOTATIONS FOR SYSTEMS MICROBIOLOGY…………………………………………..…….... 003

2.1. From genomes to networks to phenotypes……………………………………. 003

2.2. Towards a comprehensive sampling of the earth's microbiome……………… 005

2.3. The evolution of protein functional annotations……………………………… 007

2.4. Data integration for next-generation annotation strategies…………………… 010

2.5. Converting annotations to phenotypes across thousands of genomes………… 013

2.6. Comparing the phenotypic potential of microbial species……………………..017

2.7. The road ahead for functional annotations……………………………………. 020

CHAPTER 3. LONG TERM PHENOTYPIC EVOLUTION OF MICROBIAL SPECIES……………………………………………………………………...023

3.1. Introduction……………………………………………………………………023

3.2. Results…………………………………………………………………………025

3.2.1. *Species considered for* in silico *phenotyping*................................ 025

3.2.2. Long-term evolution of growth phenotypes…………………………. 026

3.2.3. Experimental validation of observed trends………………………… 029

3.2.4. Possible drivers of phenotypic divergence………………………….. 029

3.2.5. Phenotypic similarity in the context of bacterial taxonomy………… 031

3.2.6. Using metabolic phenotypes as classification markers…………….. 032

3.2.7. Evolution of gene essentiality and epistasis…………………………035

3.3. Discussion…………………………………………………………………….. 037

3.4 Methods………………………………………………………………... 039

3.4.1. Model selection…………………………………………………… 039

3.4.2. Flux balance analysis…………………………………………….. 039

3.4.3. Measuring phenotypic similarity…………………………………. 040

3.4.4. Validation with Biolog phenotypic arrays………………………... 041

3.4.5. Exponential decay fits…………………………………………….. 041

3.6. Supplementary figures and tables…………………………………... 042

CHAPTER 4. GLOBAL PROBABILISTIC ANNOTATION OF METABOLIC NETWORKS ENABLES ENZYME DISCOVERY……………………047

4.1. Introduction……………………………………………………………047

4.2. Results…………………………………………………………………050

4.2.1. Strategy of a global probabilistic reconstruction…………………… 050

4.2.2. *Optimization of the fitness function parameters*…………………….. 054

4.2.3. *GLOBUS precision-recall performance*…………………………….. 055

4.2.4. *Specific metabolic predictions and biochemical validation in* B. subtilis……………………………………………………………058

4.2.5. *GLOBUS automation and probabilistic predictions for multiple species*……………………………………………………………. 064

4.3. Discussion…………………………………………………………….. 066

4.4. Appendix:Using system-level properties as evidence for gene annotations….. 068

4.5 Methods………………………………………………………………... 072

4.5.1. *Construction of the generic EC network*……………………………..072

4.5.2. *Identification of potential metabolic genes and potential metabolicfunctions*……………………………………………………. 072

4.5.3. *The functional descriptors in the GLOBUS fitness function*…………073

4.5.3.1. *Sequence homology. F* $_{\text{homology}}$……………………………...073

4.5.3.2. *Orthology. F* $_{\text{orthology}}$………………………………………...074

4.5.3.3. *Gene-gene context correlations. F* $_{\text{context}}$…………………... 074

4.5.3.3.1. Phylogenetic correlation………………………… 075

4.5.3.3.2. Gene chromosomal clustering……………………076

4.5.3.3.3. Co-expression…………………………………… 076

4.5.3.4. *EC co-occurrences. F* $_{\text{Ecco-ocurrence}}$…………………………...077

4.5.4. *Calculating the marginal probability using Gibbs sampler*………… 078

4.5.5. *Computational requirements*……………………………………... 079

4.5.6. *Experimental validation of biochemical predictions*………………... 080

4.5.7. *MILP problem description*………………………………………...081

4.6. Supplementary figures and tables……………………………………... 083

CHAPTER 5. RECONSTRUCTION AND FLUX BALANCE ANALYSIS OF THE *Plasmodium falciparum* METABOLIC NETWORK…………………... 093

5.1. Introduction………………………………………………………………094

5.2. Results……………………………………………………………………096

5.2.1. *Scale of the reconstructed flux-balanced metabolic network*……….. 096

5.2.2. *Analysis of in silico single and double gene deletions*……………………100

5.2.3 *Resolution of disagreements between* in silico *predictions and experimental data*…………………………………………………..108

5.2.4. *Validation of the predicted drug target nicotinate nucleotide adenylyltransferase*………………………………………………. 111

5.2.5. *Prediction of metabolite concentration changes based on expression data*…………………………………………………………..114

5.3. Discussion………………………………………………………………... 116

5.4. Methods…………………………………………………………………….. 119

5.4.1. *Genome-scale metabolic reconstruction*…………………………...119

5.4.2. *Parasite culture, growth, and drug inhibition assays*………………..121

5.4.3. *Using expression-constrained network to predict shifts in external metabolic concentrations*…………………………………………... 121

5.5 Supplementary figures and tables…………………………………………… 124

CHAPTER 6. THE RATE OF THE MOLECULAR CLOCK AND THE COST OF GRATUITOUS PROTEIN SYNTHESIS……………………………… 131

6.1. Introduction…………………………………………………………………132

6.2. Results………………………………………………………………………134

6.2.1. *Destabilizing mutations in lacZ yield aggregated and partialy soluble proteins*………...…………………………………………… 134

6.2.2. *Misfolded proteins are no more toxic than WT proteins*……………. 137

6.2.3. *Nucleotide level selection, protein solubility, and stability in* E. coli..139

6.3. Discussion………………………………………………………………... 143

6.4. Conclusions…………………………………………………………………146

6.6. Materials and methods……………………………………………………... 147

6.6.1. Strains and mutant generation………………………………………147

6.6.2. Growth Curve Analysis……………………………………………… 147

6.6.3. SDS-PAGE and western blot……………………………………... 147

6.6.4. Structural analysis of E. coli proteins………………………………..148

6.6.5. Correlation of the synonymous (Ks) and non-synonymous (Ka) substitution rates with expression…………………………………... 149

CHAPTER 7. GENETIC ROBUSTNESS AND FUNCTIONAL EVOLUTION OF GENE DUPLICATES……………………………………………………151

7.1. Introduction……………………………………………………………151

7.2. Materials and methods………………………………………………… 153

7.3. Results…………………………………………………………………155

7.3.1. Compensation patterns based on quantitative fitness data…………. 155

7.3.2. Correlates of functional compensation………………………………158

7.4. Discussion……………………………………………………………... 167

7.5. Supplementary figures………………………………………………… 171

CHAPTER 9. CONCLUSIONS…………………………………………………..181

REFERENCES……………………………………………………………………183

LIST OF FIGURES

Figure 2.1. The genome to network to phenotype loop…………………………….... 005

Figure 2.2 Exploring the world's microbiome…………………………………..…007

Figure 2.3. Probabilistic metabolic annotations using multiple sources of evidence……... 011

Figure 2.4. Strategies for model auto-completion following gene annotation……………. 015

Figure 2.5. Applications of genome-scale metabolic reconstructions for a growing catalog of sequenced genomes…………………………………………... 019

Figure 3.1. Phylogenetic tree showing the families of 322 bacterial species used in this study…………………………………………………………... 025

Figure 3.2. Evolution of growth phenotypes across bacterial species…………………….. 027

Figure 3.3. Drivers of fast and slow phenotypic change……………………………….. 030

Figure 3.4. Phenotypic similarity distribution at different taxonomic ranks……………… 032

Figure 3.5. Phenotypic tree for 322 phylogenetically diverse bacterial species…………... 033

Figure 3.6. Similarity of metabolic gene essentiality and synthetic lethality across bacterial species……………………………………………………..…037

Supplementary figure 3.1. Evolution of phenotypic similarity based on the fraction of shared nitrogen sources that can be used for growth…………………………. 042

Supplementary figure 3.2. Evolution of phenotypic similarity for different sets of carbon sources……………………………………………………….... 042

Supplementary figure 3.3. Null expectation of phenotypic similarity…………………… 043

Supplementary figure 3.4. Phylogenetic tree for the 40 species used for experimental validation…………………………………………………………… 043

Supplementary figure 3.5. Correlation between predicted and measured phenotypic similarity……………………………………………………………… 044

Supplementary figure 3.6. Phenotypic variance explained by size differences as a function of genetic distance…………………….………………………….. 044

Figure 4.1. Overview of the GLOBUS method………………………………………… 053

Figure 4.2. GLOBUS precision-recall performance……………………………………057

Figure 4.3. In vitro biochemical assays for characterizing activities of SpsI and SpsJ using high-precision mass spectrometry……………...……………………. 062

Figure 4.4. In vitro biochemical assays for characterizing 6-phospho-gluconolactonase activity of YkgB………………………………………………………………063

Figure 4.5. GLOBUS database implementation…………………………………………... 066

Figure 4.6. Identification of necessary reactions based on GLOBUS and flux analysis….. 069

Figure 4.7. Improving GLOBUS performance through system-level information……….. 070

Supplementary figure 4.1. Conditional probabilities in the GLOBUS fitness function.… 083

Supplementary figure 4.2. Training GLOBUS parameters using different models…….. 084

Supplementary figure 4.3. Comparison of the precision-recall relationships obtained using homology information established by BLAST and PSI-BLAST………085

Supplementary figure 4.4. Distribution of sequence identities…………….……………085

Supplementary figure 4.5. Contribution of individual context correlations to the GLOBUS performance……………………………………………………….. 086

Supplementary figure 4.6. Potential utility of GLOBUS in refining manually curated metabolic models…………………………………………………………………087

Supplementary figure 4.7. GLOBUS predictions for *sps* genes in *B. subtilis*……….…... 088

Supplementary figure 4.8. Substrate consumption at different spsI concentrations…….. 089

Supplementary figure 4.9. Product accumulation at different YkgB concentrations….… 089

Supplementary figure 4.10. Illustration of Gibbs sampler used to derive marginal probabilities, *$P(g_i)$*, in GLOBUS………………………………………………… 090

Figure 5.1. Annotation of reactions in the genome-scale metabolic model of *P. falciparum*……………………………………………………………….. 099

Figure 5.2. Small-molecule inhibition of the parasite nicotinate mononucleotide adenylyltransferase (NMNAT)…………………………………………………...112

Figure 5.3. Comparison between the predicted and experimentally measured shifts in metabolite concentrations in infected red blood cells………………………… 116

Supplementary figure 5.1. Distribution of sequence identities to human transcripts for *Plasmodium* proteins predicted as essential for growth………. 124

Supplementary figure 5.2. Growth inhibition profile of compound *1_03* in *in vitro* cultures of *P. falciparum*…………………………………………….. 124

Supplementary figure 5.3. Accuracy of e metabolite exchange predictions made randomized expression data…………………………………………………... 125

Figure 6.1. Expression of destabilizing mutants and WT LacZ……………………………136

Figure 6.2. Growth rate comparisons between WT and misfolding-prone LacZ………… 138

Figure 6.3. Correlation of *E. coli* mRNA expression with Ka, protein solubility, and the fraction of charged residues…………………………………………………... 140

Figure 6.4. Relationship between protein stability and mRNA expression………………..143

Figure 7.1. Compensation patterns between yeast duplicates as a function of their evolutionary divergence……………………………………………………. 157

Figure 7.2. Growth phenotypes for individual and simultaneous deletion of duplicates…..160

Figure 7.3. Differences in the number of sensitive conditions between duplicates………..162

Figure 7.4. Diversification of duplicates function and regulation………………………… 165

Figure 7.5. Average number of duplicates retained in the genomes of yeast species……...167

Supplementary figure 7.1. Fraction of conditions with a significant growth decrease for yeast duplicate deletions as a function of their divergence…………………… 171

Supplementary figure 7.2. Fraction of conditions with a significant growth decrease for yeast duplicate deletions as a function of Ks………………………………172

Supplementary figure 7.3. Fraction of conditions with a significant growth decrease for yeast duplicate deletions……………………………………………………173

Supplementary figure 7.4. Robustness of observed compensation patterns……..………. 174

Supplementary figure 7.5. Fraction of conditions as a function of gene family size….….175

Supplementary figure 7.6. The fraction of conditions with a significant growth phenotype for duplicate deletions…………………………………………….. 176

Supplementary figure 7.7. The fraction of single duplicate deletions with a growth decrease below different thresholds………………………………………….. 177

Supplementary figure 7.8. The average number of known and predicted transcription factor binding motifs for yeast duplicates…………………………………….. 177

Supplementary figure 7.9. Fraction of shared motifs and cell locations….………………178

Supplementary figure 7.10. Duplicates retained in different yeast species……….……... 179

Supplementary figure 7.11. Compensation patterns for yeast duplicates described in research literature as redundant……………………………………………. 180

LIST OF TABLES

Supplementary table 3.1. Frequency of metabolite usage as a carbon source across 322 species…………………………………………………………………. 045

Supplementary table 3.2. Overrepresentation of functional categories for reaction differences between species at close genetic distances………………………. 045

Table 4.1. Prediction of gene function in *B. subtilis*………………………………………060

Table 4.2. Prediction of gene function in *S. aureus*………………………………………..061

Supplementary table 4.1. Highly connected metabolites that were not used in establishing connections between metabolic activities……………………….. 091

Table 5.1. Characteristics of the reconstructed metabolic network of *P. falciparum*……...097

Table 5.2. Total number of single and non-trivial double deletion phenotypes………… 101

Table 5.3. Literature support for essentiality predictions………………………………. 102

Table 5.4. Predicted essential genes with no homologs in the human genome…………… 106

Supplementary table 5.1. Biomass components and cofactors which can be synthesized or imported by the reconstructed metabolic network of *P. falciparum*………... 126

Supplementary table 5.2. Single deletions displaying a phenotype in the unconstrained metabolic network of *P. falciparum*…………………………..… 127

Suppelementary table 5.3. Single deletions of orphan metabolic activities predicted to be lethal in the unconstrained metabolic network of *P. falciparum*……… 129

Supplementary table 5.4. Double deletions predicted to be lethal in the unconstrained metabolic network of *P. falciparum*…………………………………………… 129

Supplementary table 5.5. Essentiality in *S. cerevisiae* of the orthologs for essential and non-essential *Plasmodium* metabolic genes…………………………………… 130

Table 6.1. Characteristics of destabilizing mutations engineered into *E. coli* β-galactosidase……………………………………………………………….135

ACKNOWLEDGEMENTS

First of all I am greatly thankful to my advisor, Dennis Vitkup, for his guidance, support, enthusiasm, and genuine interest in teaching me the ways of science.

I am grateful to the members of my dissertation committee –Harmen Bussemaker, Larry Shapiro, Max Gottesman and Saeed Tavazoie. It was always a pleasure to discuss these projects with them; their feedback is much appreciated. Special thanks go to Max and members of his lab for their help and guidance in the molecular clock project. I am particularly grateful to Christal Vitiello for her coaching in the wet lab.

I am also thankful to our various collaborators. Uwe Sauer and Tobias Fuhrer at ETH Zurich validated several enzymatic activities predicted by GLOBUS and continue to work with us in several projects. Manuel Llinás and Kellen Olszewski at Princeton University contributed to the reconstruction of the *P. falciparum* metabolic network and did the experimental work to validate our prediction of a novel drug target. Christopher Henry at Argonne National Laboratory and Barry Bochner at BiOLOG contributed the models and experimental data for the project on bacterial phenotypic evolution. Their contributions and disposition are much appreciated.

I am grateful to the members of the Vitkup lab, Tzu-Lin Hsiao, Sarah Gilman, Eugenia Lyashenko, Jie Hu, Mariam Konate and Johnathan Chang. It was a pleasure to collaborate, learn about their research and share our time at Columbia. I am especially thankful to Tzu-Lin, with whom I worked on the GLOBUS and malaria projects, and who, among other things, taught me the principles of constraints based analysis.

On a personal note, I want to thank my parents, my brother and sisters, and my closest friends for keeping me accompanied, happy and motivated throughout these years.

Chapter 1.

THESIS OVERVIEW

The document contains five main chapters (3 to 7) –each corresponding to a different research project– presented in approximately chronological order from most to least recent work. These projects are about two general topics: evolution at multiple levels of biological organization, and genotype to phenotype mapping in the context of microbial metabolic networks. Despite topic commonalities, the projects can be read independently of each other. In most cases I have kept the format with which they were presented to the corresponding scientific journals.

The introduction (**chapter 2**) is a review on the current state of biochemical annotation methods in the context of the rapidly expanding catalog of fully sequenced microbial species. I discuss the importance of probabilistic integrative strategies for assigning functions to metabolic genes and emphasize the role of genome-scale metabolic reconstructions in allowing phenotypic predictions across thousands of newly sequenced genomes.

In **chapter 3**, using metabolic network reconstructions for hundreds of species, I study the long-term evolution of bacterial phenotypes. The analysis shows that phenotypic similarity follows a clock-like exponential decay over billions of years of microbial evolution. Our observations provide a framework to analyze possible mechanisms governing the evolution of phenotypic properties and its variation across bacterial lineages and lifestyles.

In **chapter 4** I present work on a global probabilistic approach for biochemical annotations: GLOBUS [1]. The method uses Gibbs sampling to simultaneously assign probabilities to every possible gene-reaction assignment in a given genome. GLOBUS can integrate multiple sources of functional evidence and allows the identification of alternative functions of enzymes. I use GLOBUS to integrate global metabolic flux information with sequence and genome-context data in order to improve functional predictions.

In **chapter 5** I describe the reconstruction and flux analysis of the genome-scale metabolic model of *Plasmodium falciparum,* the causative agent of malaria. The article, published in Molecular Systems Biology [2], illustrates several applications of metabolic network models including the identification and experimental validation of a novel potential drug target.

In **chapter 6** I present computational and experimental work that explores the question of why highly expressed proteins evolve slowly. The article, published in Genome Biology [3], shows that at least for bacteria like *E. coli* the costs of producing gratuitous protein are a more important evolutionary constraint than those of misfolded protein toxicity.

In **chapter 7** I present work that explores the functional compensation between duplicated yeast genes. I focus on the fitness consequences of duplicate gene deletions that are relevant for wild yeast populations. Surprisingly, the results show that close duplicates are, on average, no more likely than random singletons to provide significant backup for their paralogs.

Finally, in **chapter 8**, I briefly summarize the conclusions of the research presented.

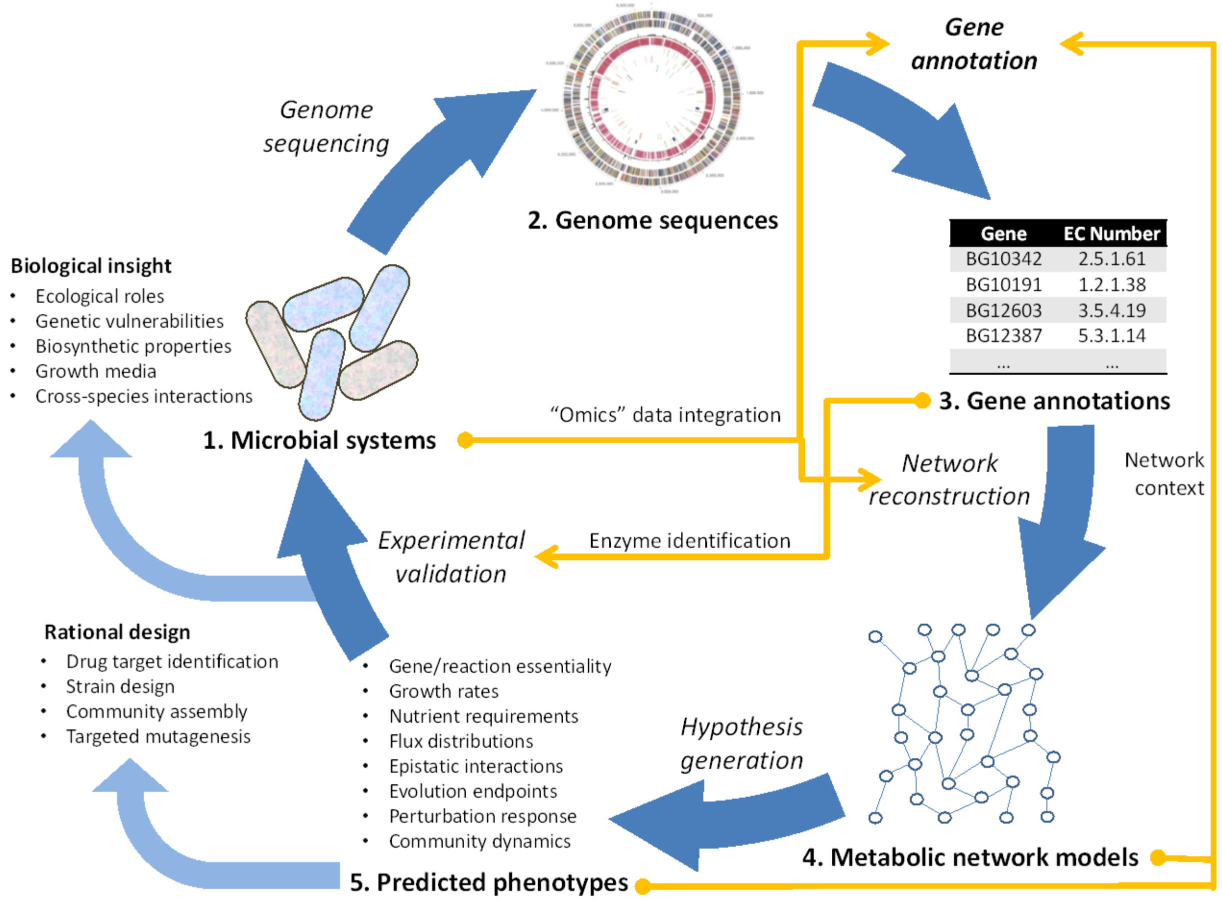
Chapter 2.

NEXT GENERATION BIOCHEMICAL ANNOTATIONS FOR SYSTEMS MICROBIOLOGY

As thousands of microbial genomes are sequenced every year, accurate annotation methods provide a path towards understanding the function, ecology, and evolution of microbes. Historically, gene functional assignments have relied on just a few sources of evidence and did not provide a quantitative measure of annotation confidence. This resulted in high rates of misannotations and contradictory results amongst databases. The explosive growth of genome-scale data has given way to a new generation of annotation tools. Methods are developing that seek to integrate sequence, structure, phylogenetic, and expression data, among several others, into probabilistic functional annotations. Using a metabolic network context, accurate gene-function assignments not only result in improved phenotypic predictions, but also allow the consideration of cell-level properties, measured by phenomics, metabolomics, or fluxomics data, as additional cues in the annotation process.

2.1. From genomes to networks to phenotypes

The vast majority of microbes in the planet cannot be cultured using standard techniques and their phenotypic properties are still unexplored. While it remains a challenge to estimate the total number of microbial species [4], the diversity of closely related strains within each taxonomic group –the pan-genome– may easily account for millions or even billions of phenotypically distinct prokaryotes [5-7]. Given the above, it is likely that most bacteria will never be studied in detail in the laboratory, and that most of the microbes that are indeed studied will be so mainly through their genome sequences or other types of high-throughput data. In order to understand the roles played by these species, and to identify those with potential for biotechnology, genome annotations should be as precise as possible. This allows the identification of the genetic elements responsible for specific functions and, through the reconstruction of molecular network models, facilitates the prediction of a wide range of phenotypic traits. As illustrated in **Figure 2.1**, the process from genome sequences to predicted phenotypes is not necessarily linear. I show here how annotation strategies are being developed that integrate information at the gene and whole-cell levels, further improving the accuracy of metabolic models and their corresponding predictions. This review focusses on biochemical annotations as a central element of the genome-network-phenotype loop. I first describe the ongoing avalanche of genomic data and the concurrent evolution of annotation methods. Then, I review the role of annotations in the reconstruction of genome-scale metabolic models, emphasizing the possibility of obtaining phenotypic predictions across thousands of microbial species. Finally, I take look at the future of biochemical annotation methods and highlight several applications in the context of comparative systems biology.



Gene	EC Number
BG10342	2.5.1.61
BG10191	1.2.1.38
BG12603	3.5.4.19
BG12387	5.3.1.14
...	...

Figure 2.1. The genome to network to phenotype loop. Most species on earth live in complex communities and cannot currently be cultured under standard laboratory conditions. Genome annotations, coupled with reconstruction and modeling of molecular networks (e.g. metabolic networks), allow the prediction of phenotypes that can aid in our understanding of interactions and functional properties that are not observable by field measurements. Additionally, the network framework provides a context for using multiple types of evidence in the annotation process.

2.2. Towards a comprehensive sampling of the earth's microbiome

With rapidly falling sequencing costs and doubling times of the number of bacterial genomes of around 20 months [8], the number of species with a sequenced genome is growing fast (**Fig. 2.2a**). Currently there are at least 3,767 bacterial genome projects either completed or as permanent drafts and 14,657 prokaryote sequencing projects in progress (GOLD, February

2013 [9]); it is likely that up to one to several hundred thousand genomes are sequenced within the next decade (**Fig. 2.2a**). Interestingly, while the first sequenced genomes were primarily selected because of prior knowledge of their medical, economic, or scientific value, attention is now turning towards poorly sampled phylogenetic groups and environments [10] (**Fig. 2.2b**). Sequencing of rare bacteria has become an important tool for the discovery of new cellular functions at specific ecological niches; for example, a combination of metagenomics and single cell sequencing revealed multiple genes and pathways for hydrocarbon degradation after a recent oil spill in the Gulf of Mexico [11]. Systematic sequencing of species across phylogenetic space can create taxonomic anchors for the classification of genes and species in metagenomic samples [12]. There is also a growing number of studies that sequence and compare the genomes of multiple strains rather than representative clones for a given microbial species (e. g. [13]). These diversity-driven studies provide tools to understand the evolution and functional diversification of microbes, as well as specific genetic features shared at different levels of phenotypic similarity. There are at least 857 metagenomic projects deposited in Genbank (June 2013 [14]), many of which include tens to hundreds of different samples. These data, coupled with a growing array of techniques for the molecular characterization of metagenomic samples (e.g. metatranscriptomics, metaproteomics, metametabolomics) are set to rapidly transform our view on microbial diversity and function in a community setting [15]. Making sense of this avalanche of data is one of the main challenges of modern biology.

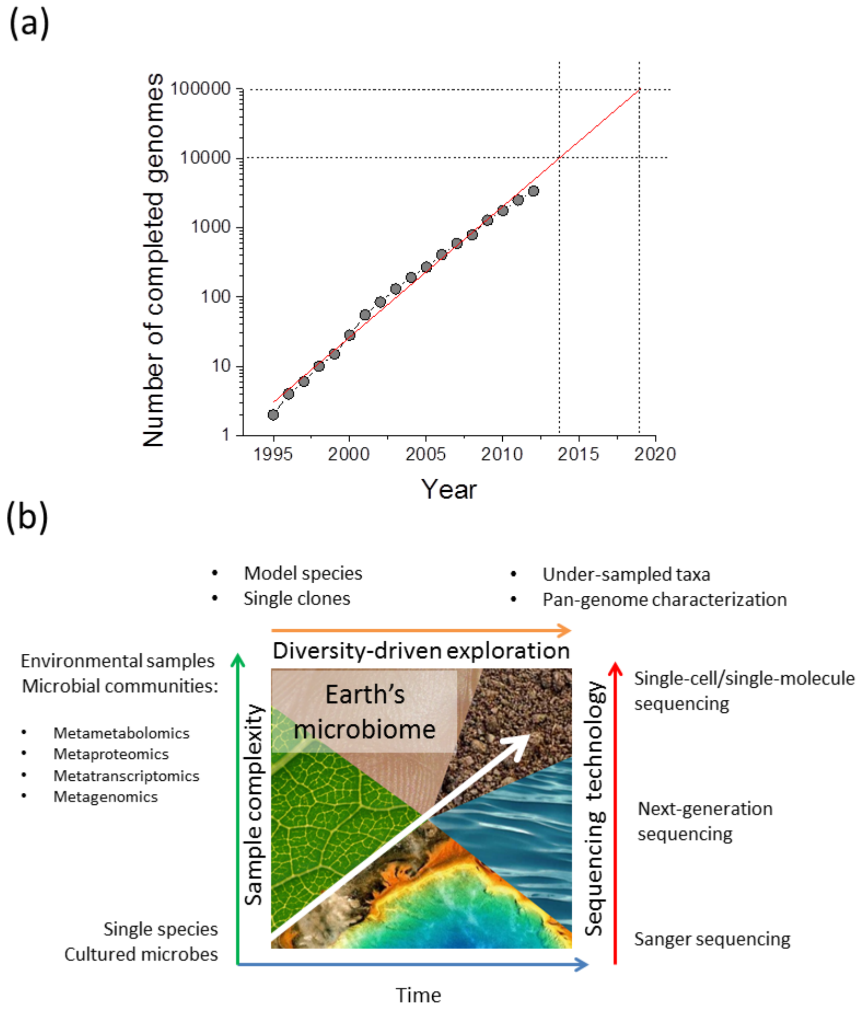


Figure 2.2 Exploring the world's microbiome. (a) Cumulative number of completed bacterial genome projects over time (Source: GOLD [9]), the red line indicates a linear fit of the log-transformed values. Dashed lines indicate projected times for ten thousand and one hundred thousand completed bacterial genomes. **(b)** The landscape of emerging technologies and exploration approaches to tackle the overwhelming diversity of existing microbes.

2.3. The evolution of protein functional annotations

Knowing the functions encoded in a particular genome or set of genomes can give clues as to the roles of genes, pathways, and whole molecular networks; these are essential steps towards understanding the mechanistic principles underlying phenotypes of interest. Obtaining such functional annotations, however, is not a trivial process. The functional characterization of genes relies on experimental evidence that is generally only available for a limited group of proteins and species. CharProtDB [16], a database of experimentally determined protein functions, contained a little over sixteen thousand proteins from 1588 organisms when it was published in 2012. UniProt KB/Swiss-Prot [17], which also includes human-reviewed computational predictions of protein functions, had about half a million sequences from thirteen thousand organisms as of May 2013. By contrast, UniProt KB/TrEMBL [17], which is not reviewed, had over 35 million sequences from about half a million organisms, 73% of which were prokaryotes. There is a huge gap between the number of sequences that have to be algorithmically annotated, and those from which annotations will be transferred. This can lead in many cases to genome annotations with poor coverage, annotations that rely on weak evidence, or annotations to functions that are too general for a detailed analysis; for example, a pyruvate dehydrogenase (EC: 1.2.4.1) may be only annotated as an oxidoreductase (EC: 1.-.-.-).

Sequence homology to proteins of known function, detected by programs like BLAST [18] or FASTA [19], constitutes a first and still popular approach to gene functional annotations. Transferring of functional assignments on the basis of sequence homology, however, is known to produce unreliable predictions when sequence identity is low (below 70%) [20]. This is an important problem as more of the species getting sequenced are not closely related to well-characterized organisms, and they are unlikely to have high sequence similarity to known proteins [21]. An additional problem is that even when sequence identity is high, the prevalence of incorrect annotations across databases can lead to the spread of annotation errors [22, 23]. Although such error rate is relatively low for manually curated repositories such as Swiss-Prot, it was shown to be increasing with time, and as high as 80% misannotation levels have been estimated for certain enzyme families across various automatic annotation repositories [24].

The above shortcomings led to a second generation of functional annotation methods that use various computational tools and sources of evidence to improve annotation coverage and accuracy. For example, hidden Markov models (HMMs) and other sequence-based methods have been used for the identification of functional protein domains or motifs in uncharacterized genes [25]. Structural homology is another strategy that allows inferring functional properties when sequence homology is too low to reliably infer function [26]. Other strategies rest on the idea that genes already assigned a particular function can help to predict the function of other genes of the same species. For example, if most members of a biochemical pathway are identified in a given genome, there is a high likelihood that other genes will encode the remaining reactions [27]. Expert definition of pathways has been a popular approach in the characterization of novel genome sequences. BioCyc [28] and KEGG [29] generate organism-specific pathway databases under controlled chemical vocabularies which have been widely used in metabolic reconstructions. The RAST annotation system [30] further complements information from biological subsystems with the simultaneous assignment of gene functions across multiple genomes, taking advantage of sequence and pathway conservation across species. So-called genomic context correlations have also been widely used to improve annotation accuracy and to find genes responsible for orphan activities [31-33]. For example, the patterns of gene conservation across species have been used to calculate correlations between phylogenetic profiles in order to assign functions to genes that co-evolve with genes of known function [33]. Similar principles have been applied to other functional correlations such as chromosomal gene clustering [34] or gene fusions [35], which can be readily inferred from multiple genome sequences and do not require the prior definition of metabolic pathways [1]. While each of these strategies can lead to improved functional predictions in their own context, integration of multiple data types under a unified framework has the potential to advance functional annotations beyond any single source of evidence.

2.4. Data integration for next-generation annotation strategies

Several partially independent evidence types should converge upon the real function of a gene. As illustrated in **Figure 2.3a**, while high sequence identity between proteins can often be interpreted as evidence of equivalent molecular functions, complementary information can either support or contradict a particular assignment. This concept was recently used to predict annotation errors based on a combination of sequence identity and the strength of genomic context correlations between metabolic neighbors. The method, developed by Hsiao *et al.* [36], allowed the re-annotation of the leucine degradation pathway in the model gram positive bacterium *Bacillus subtilis*. Although this "policing" of biochemical annotations is useful to keep annotation databases in check, it does not necessarily point to the correct function of the misannotated genes.

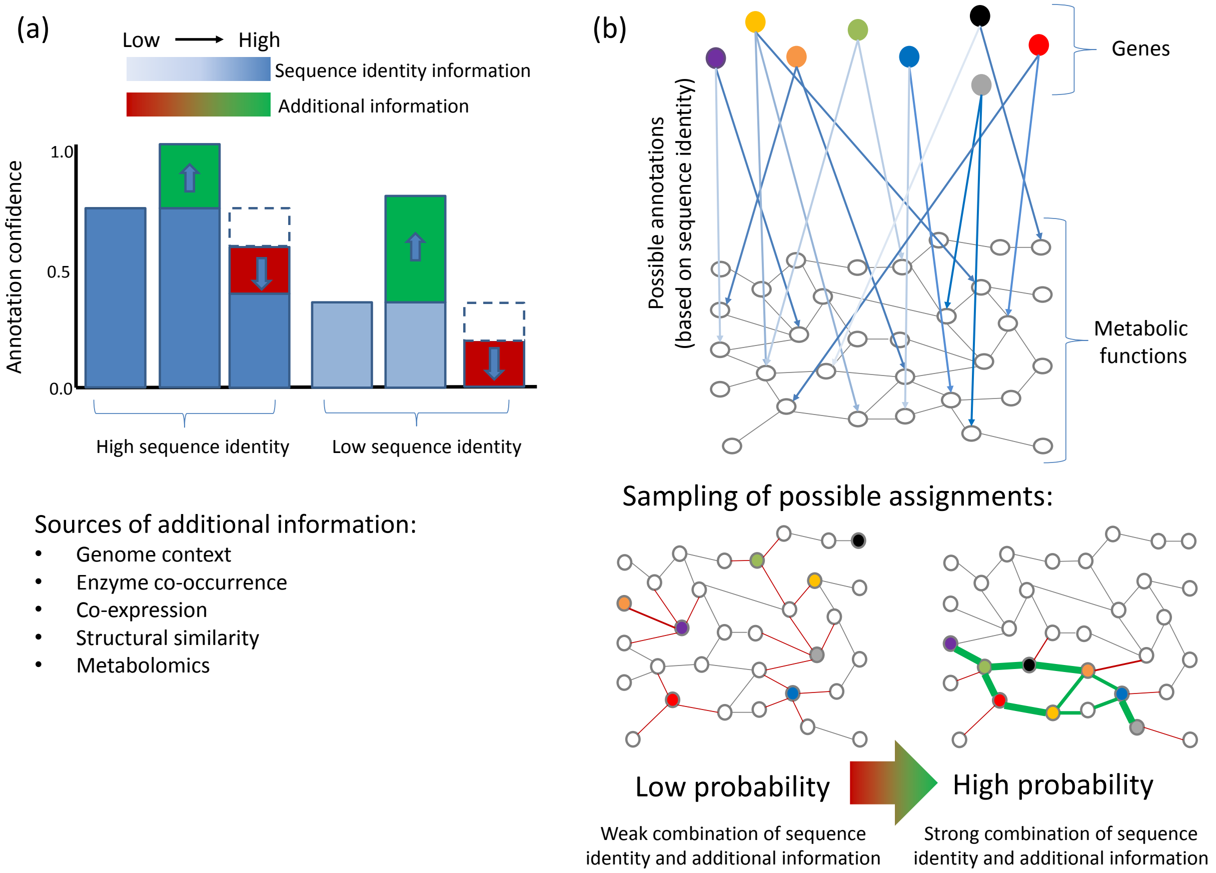


Figure 2.3. Probabilistic metabolic annotations using multiple sources of evidence. (a) While sequence identity to previously annotated proteins is typically used to transfer functional annotations, additional information such as genome context correlations, structural similarity, or expression measurements can either increase (green) or decrease (red) our belief that an annotation is correct. When sequence identity is high these additional data can be used to spot potential misannotations; when it is low and it supports sequence-based predictions, it facilitates enzyme discovery. **(b)** In a genome-scale context, when multiple potential functions are found for each candidate gene, probabilistic annotations can be obtained through sampling from only those gene-to-reaction assignments that display high likelihood configurations of sequence identity and complementary context information [1].

Multiple algorithms have been used to combine evidence sources for functional assignments of metabolic as well as non-metabolic genes; these include Bayesian networks, support vector machines, decision trees, and functional linkage networks, among others (see [37]). Metabolic annotations are special compared to other gene functions (e.g. in regulatory or signaling networks) in that catalytic activities can be mapped to a well-defined network that is common across all species. Each enzyme is associated to a reaction or set of reactions such that links between metabolic activities can be precisely defined on the basis of shared metabolites [38]. The structure of the metabolic network provides a context to consider functional correlations between genes, and gene annotations can be optimized to maximize the consistency of such correlations in the network. For example, assignments can be made such that co-expressed genes or genes that are clustered in bacterial chromosomes are responsible for consecutive metabolic reactions. Based on this idea, we recently used Gibbs sampling in GLOBUS [1] (**Chapter 4**) to simultaneously annotate all candidate metabolic genes of a species; the method assigns probabilities to each possible gene-reaction assignment based on the hypothesis that genes close in the metabolic network have strong functional associations with each other and have detectable sequence homology to known enzymes with the target functions (**Fig. 2.3b**). Importantly, several different types of context-based and homology-based information can be combined to derive such probabilistic predictions [1]. In the context of metagenomics data, sampling procedures were recently used to derive the probabilities for the presence of specific reactions in a bacterial community based on the enzymes detected in a metagenomics sample [39]. An important property of using a global network instead of pathway definitions to map gene-gene associations is generality; whereas not all pathways are conserved across organisms [40] and their boundaries are not always well defined, a global metabolic network represents all currently known biochemistry and is therefore able to represent every possible pathway architecture. The structure of the global network was recently combined with sequence homology to improve organism-specific annotations reported in pathway databases [41].

Binary annotations, which are common among several annotation repositories, can be misleading as not all functional assignments are equally good [42]. On the other hand, probabilistic approaches such as GLOBUS, not only provide a direct measure of error rates, but may result in higher accuracies compared to methods using a simple scoring cutoff. For example, a simple Bayesian approach that used sequence identity to predict the probability that a gene belongs to one of several functional classes was shown to outperform predictions based on direct functional transfer from the top BLAST hit [43]. In summary, a next generation of biochemical annotation methods is developing that is able to accommodate diverse types of data in a network context, and also provides confidence scores that differentiate reliable from unreliable predictions.

2.5. Converting annotations to phenotypes across thousands of genomes

While gene annotations, by themselves, give rise to important functional hypotheses about microbial systems, they are also the foundation for the reconstruction of network models used for system-level phenotypic predictions (**Fig. 2.1**). Although several methods exist to simulate metabolic network behavior (reviewed in [44]), constraints-based analysis has become one of the main tools to perform this step. This is mainly because this type of analysis does not require detailed knowledge of enzyme kinetics or metabolite concentrations, which are typically not available on a genome-wide scale even for the best studied model microorganisms. Constraints-based methods like Flux Balance Analysis (FBA) [45], rely on identifying metabolic fluxes that optimize one or several objectives (e.g. synthesis of biomass precursors) while keeping metabolite concentrations at steady-state given the metabolic networks' stoichiometric matrix [45]. Additional constraints can be applied to limit flux capacity and reaction directionality. This diverse set of techniques has been applied to dozens of different microbes to predict gene essentiality [46], viable media for growth [47], genetic interventions for metabolite overproduction [48], and potential cross-species interactions [49], among several other phenotypes. A thorough review of constraints based analysis methods and their applications can be found in [50].

Enzyme annotations are the main source of information for the reconstruction of genome-scale metabolic models used in FBA, however, the corresponding lists of reactions are often not enough to obtain a (gapless) model amenable for constraints-based analysis. For instance, spontaneous reactions or missing enzyme annotations may be needed to allow the synthesis of every molecule in the network. Historically, filling of those missing reactions has been done through the assembly of pathways and individual gene annotations coupled with manual curation supported by organism-specific literature [51]. Because many of the microbial species now sequenced and those to come are unlikely to have a vast body of published biochemical knowledge, network models must be produced based on high confidence annotations, universal sets of reactions and pathways, and experimental data, if available. Given the exponential growth of sequence databases, attention is now rising towards the problem of automating metabolic model generation (**Fig. 2.4**).

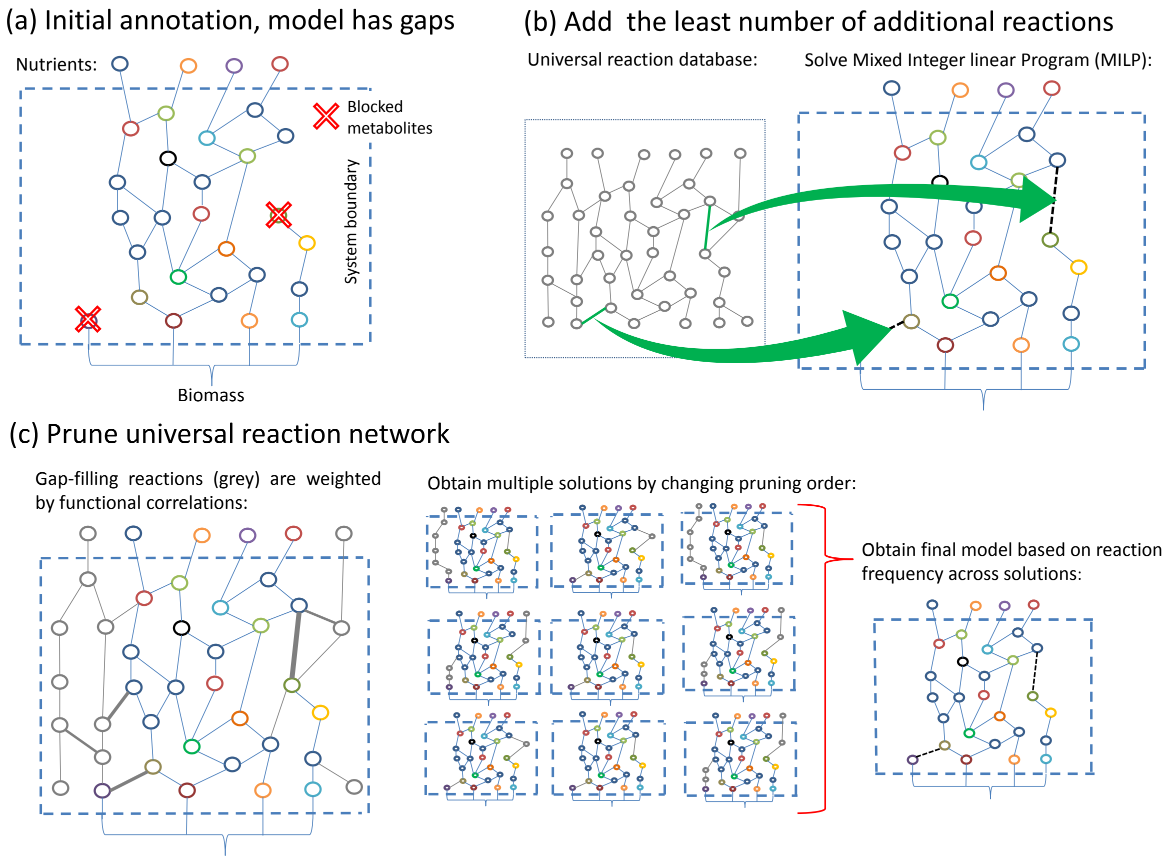

(a) Initial annotation, model has gaps
Nutrients:
Blocked metabolites
System boundary
Biomass
(b) Add the least number of additional reactions
Universal reaction database:
Solve Mixed Integer linear Program (MILP):
(c) Prune universal reaction network
Gap-filling reactions (grey) are weighted by functional correlations:
Obtain multiple solutions by changing pruning order:
Obtain final model based on reaction frequency across solutions:

Figure 2.4. Strategies for model auto-completion following gene annotation. (a) The initial assembly of reactions catalyzed by annotated enzymes may contain gaps that prevent production or consumption of specific metabolites (blocked metabolites, red crosses). **(b)** The GapFill algorithm [52] uses mixed integer linear programming (MILP) to find the least number of reactions from a reaction repository that allow flux to proceed through blocked metabolites. **(c)** An alternative strategy, starts with a flux-balanced model of all known reactions, and iteratively removes reactions that are not needed to maintain a gapless network or have weak correlations with already annotated genes; the reactions that are most frequently used across iterations are used to reconstruct the final model [53].

In 2010, Henry *et al*. [54] built into a single pipeline (Model SEED) the steps for annotation, assembly, and refinement of microbial metabolic models. The study demonstrated that a draft metabolic reconstruction can be produced for any sequenced species with currently available computational tools, and that at least several phenotypes predicted by such models do not display much lower accuracy relative to manually curated reconstructions. As access to additional sources of experimental data and novel annotation strategies are developed, the range and correctness of predictable phenotypes by these models is likely to increase. **Figure 2.4** describes two different strategies used for model auto-completion following genome annotation. Model SEED [54] and MetaFlux [55], for example, use predicted reactions in individual metabolic subsystems and mixed integer linear programming (MILP) to find the least number of additional reactions from a universal biochemical database needed to produce a gapless model [52], i.e. one in which all metabolites in the optimized objective can be synthesized at steady-state (**Fig. 2.4b**). A different strategy, implemented in MIRAGE [53], starts with a stoichiometric model that includes all reactions from a universal reaction database. The method then uses a pruning procedure that keeps high-confidence annotations and reactions likely to be present based on functional genomic correlations while maintaining a consistent flux-balanced model (**Fig. 2.4c**).

Although network reconstructions generated by computational pipelines have yet to achieve the accuracy attained by manually curated networks, they significantly reduce the time and effort required to obtain a working model, making constraints-based analysis easily available for almost any sequenced species. These tools also facilitate cross-species comparisons by using standard rules to name reactions and metabolites. Because the same effort goes into building each model, this leads to predictions that can often be directly compared without extensive model reconciliation [56]. The Model SEED pipeline was recently applied to the reconstruction of metabolic models for 37 species of Actinomycetes allowing the analysis of gene and reaction conservation and essentiality across this important group of bacteria [57]. Despite their shortcomings, automated network reconstructions will play an important role in characterizing many of the possible metabolic phenotypes in the biosphere; the success of this approach will depend on both the quality of the annotations and the methods that transform them into predictive models.

2.6. Comparing the phenotypic potential of microbial species

In the context of an expanding catalog of sequenced genomes, it is perhaps the ability to do comparative biology that will benefit the most from novel annotation and model reconstruction strategies. It has been argued that genome-scale metabolic reconstructions can be used in prokaryotic systematics through the prediction of metabolic features (phenotypes) that could serve as evolutionary markers [58](**Fig. 2.5a**). As an example, the comparison of metabolic networks across *Pseudomonas* species served to reveal evolutionary events and strain-specific metabolic features associated with pathogenic and non-pathogenic lifestyles [59]. In the same vein, automated annotation and reconstruction methods can provide relatively uniform models for comparative drug target discovery. The ability to predict the essentiality of genes and reactions has long been one of the main applications of genome-scale metabolic reconstructions [60], as shown in **Chapter 5**, such efforts can readily suggest potential drug targets against pathogenic microbes. As more genomes become available, the comparative analysis of species genetic vulnerabilities can be used to detect novel narrow-spectrum drug targets that may help reduce the rate of resistance transfer among species [61] (**Fig. 2.5b**). Notably, as I show in **Chapter 3**, metabolic gene essentiality is typically conserved for about 50 to 70% of essential genes depending on the evolutionary distance between species, which provides ample ground for the comparative drug-target identification approach. A third example of comparative systems analysis is related to microbial cell factories. Several different methods have been proposed to facilitate rational strain design for metabolite overproduction [44]; among these efforts, attention has recently been paid to *in silico* screening for the most suitable strains/species for the synthesis of specific byproducts (**Fig. 2.5c**). For instance, Zakrzewski *et al.* [62] used 38 genome-scale actinobacterial metabolic models to predict the efficiency of each species in producing 15 different biotechnologically relevant compounds; the study suggested that a large number of secondary metabolite biosynthesis genes does not necessarily correlate with higher production efficiencies. Such strategies will continue to become more relevant as microbial species from diverse environments and their biosynthetic properties continue to be characterized.

Finally, because individual species do not live in isolation, as the number of high quality genomic and metagenomic annotations increases it should be possible to predict cross-species metabolic interactions and their joint phenotypes (**Fig. 2.5d**). Metabolic models were used in the past to study the association between sulfate-reducing bacteria and methanogens, producing accurate predictions for growth in co-culture [63]. Constraints-based models were also used to study general principles facilitating cooperative interactions between species [49]. Modeling frameworks that consider individual and community-level fitness criteria to simulate community behavior have been developed [64]; these tools, along with deeper or targeted sequencing of microbial communities [11] will provide means for the characterization of ecological features at higher levels of microbial organization.

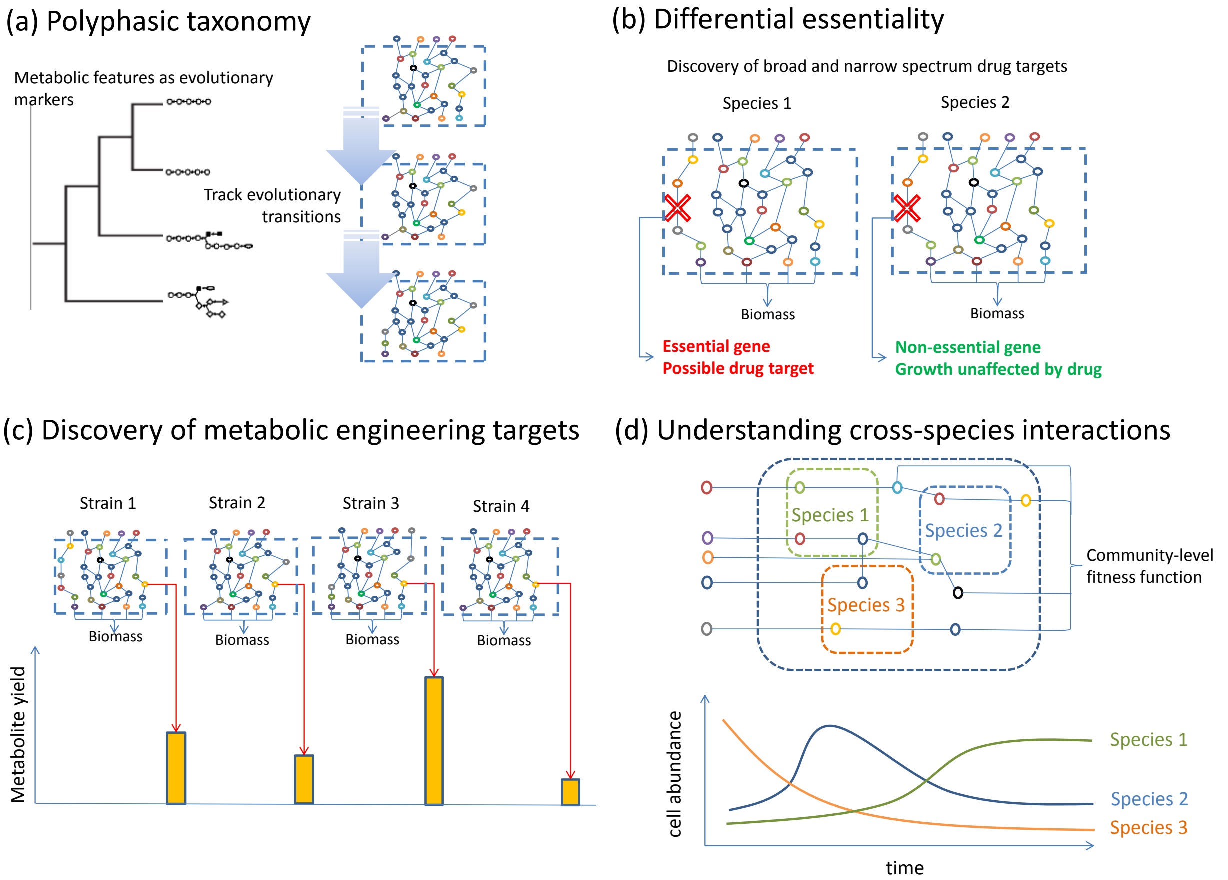

(a) Polyphasic taxonomy
Metabolic features as evolutionary markers
Track evolutionary transitions
(b) Differential essentiality
Discovery of broad and narrow spectrum drug targets
Species 1
Species 2
Biomass
Biomass
Essential gene
Possible drug target
Non-essential gene
Growth unaffected by drug
(c) Discovery of metabolic engineering targets
Strain 1
Strain 2
Strain 3
Strain 4
Biomass
Biomass
Biomass
Biomass
Metabolite yield
(d) Understanding cross-species interactions
Species 1
Species 2
Species 3
Community-level fitness function
cell abundance
time
Species 1
Species 2
Species 3

Figure 2.5. Applications of genome-scale metabolic reconstructions for a growing catalog of sequenced genomes. (a) Accurate metabolic annotations allow the utilization of the presence and absence of metabolic pathways and phenotypes to support taxonomic classifications and to track evolutionary events leading to phenotypic differences among bacteria [58]. **(b)** Comparative analysis can also be used to reveal differences in gene and reaction essentiality between species, which can be used to identify organism-specific drug-targets for narrow-spectrum antibiotic development [61]. **(c)** Different strains may respond differently to genetic perturbations aimed at metabolite overproduction; systems analysis allows *in silico* screening to identify the most suitable strains for specific bioengineering tasks [62]. **(d)** As metagenomics, metatranscriptomics, and targeted single-cell sequencing become more prominent, systems analysis can be used to model and understand cross-species interactions and community function through the combination of individual and multi-species fitness criteria [64].

2.7. The road ahead for functional annotations

Currently, our ability to understand microbial diversity is largely determined by the information that can be extracted from genome sequences and other types of high-throughput data. Biochemical annotations can uncover microbes' metabolic potential and allow tracing species properties to the activity of particular genes. Functional annotations are also the fundamental blocks for the reconstruction of molecular networks used for system-level phenotypic predictions and comparative studies. As metagenomics and single-cell studies extend genomic coverage into unexplored portions of the microbial world [10], annotations will more prominently rely on weak sequence homology to proteins of known function. This situation, together with a substantial prevalence of annotation errors in public databases [24, 36], and poor agreement between published annotations even for such well-studied species as *E. coli* [41], calls for annotation methods that do not rely on sequence identity alone and are able to deal with annotation uncertainty. I have argued that a new generation of methods for biochemical annotation should possess several important characteristics: **i)** allow diverse types of functional cues to be included in the annotation process and produce valid predictions even if certain kinds of data are missing. **ii)** Annotations should be probabilistic or at least provide a quantitative measure of annotation confidence that tells apart reliable and unreliable predictions. **iii)** Once made, functional predictions should be traceable to the evidence sources supporting a particular assignment; for instance by keeping track of the context-based and homology-based scores that were combined by the annotation method. **iv)** The next generation of biochemical annotation methods should be network oriented; this not only provides a common context for comparing diverse types of data (e.g. metabolomics or co-expression), but also provides a framework in which to consider biochemical functions that is complementary to subsystems or pathways that are not necessarily well conserved.

Considering biochemical annotations in the context of the genome-network-phenotype loop illustrated in **Figure 2.1** suggests that the processes of data acquisition, gene annotation, and model reconstruction and simulation are tightly connected and should not be considered in isolation. This opens the possibility of using high-level information such as flux distributions, or phenotypic and metabolomics measurements for the accurate identification of gene functions. In **Section 4.4**, using GLOBUS, I demonstrate how these high-order considerations have a significant effect on annotation coverage and accuracy. As exploration of the earth's microbiome continues to move forward, I envision flexible computational tools that weight-in virtually any available data to provide a genome or metagenome-wide probabilistic landscape of predicted metabolic functions and phenotypes. Unbiased whole-network functional annotations and model reconstructions will be essential for comparative studies within species and across phylogenetic space, which shall guide further efforts to understand and exploit microbial diversity.

Chapter 3.

LONG TERM PHENOTYPIC EVOLUTION OF MICROBIAL SPECIES

For many decades the comparative analysis of protein sequences and structures has been used to understand fundamental principles of molecular evolution [65-67]. In contrast, relatively little is known about the long-term evolution of phenotypic and genetic properties. This represents an important gap in our understanding of evolution, as exactly these proprieties play key roles in natural selection and adaptation to diverse environments. Here we present a comparative analysis of hundreds of genome-scale metabolic models to investigate the long-term phenotypic evolution of bacteria. The analysis reveals that, in spite of lineage-specific differences, the average long-term phenotypic divergence of bacterial species can be described by a clock-like exponential decay spanning billions of years. The observed trend, with approximately similar fractions of phenotypic properties changing per unit time, was validated through experimental profiling of 40 bacterial species across 62 growth conditions. Although fast phenotypic evolution is frequently observed between strains in the same species, a transition from high to low phenotypic similarity happens primarily at the genus level. A complementary analysis of gene knockout phenotypes suggests that gene essentiality diverges substantially slower than the ability to use different nutrients, with a relatively high conservation across lineages. Synthetic lethality, on the other hand, diverges much faster.

3.1. Introduction

Prokaryotes are the main agents of biogeochemical transformations and play essential roles in the life and health of all multicellular organisms [68]. Despite their recognized importance, we still know little about where bacteria live and what roles they play in their natural environments [69]. Understanding how bacterial species evolve and adapt to different niches is central to characterizing microbial diversity. For more than half a century, the comparative analysis of protein sequences and structures has been an essential resource to elucidate the mechanisms, constraints, and rates of protein evolution [70, 71]. More recently, the explosive growth of genome sequencing (see **Section 2.2**) made it possible to extend comparative analysis to complete genomes and the structures of several molecular networks[72-75]. In stark contrast, comparative studies of microbial phenotypic properties remain sparse; these analyses are necessary to address several fundamental questions in microbial ecology and evolution including: What is the typical variation of phenotypic properties in bacteria? How fast or slow do different phenotypic properties change? What processes or evolutionary models can explain observed trends of phenotypic evolution? And how does phenotypic variation relate to bacterial taxonomy and genetic distance between species?

Although a large-scale comparative analysis of microbial phenotypes –such as the ability to grow on different nutrients or withstand genetic perturbations– is currently challenging due to a relative paucity of experimental data, we rationalized that thoroughly validated computational methods can be used to investigate the phenotypic evolution of diverse bacterial species. Constraints-based analyses of metabolic networks have been successful in predicting various phenotypic properties of microbial species. Flux balance analysis (FBA), in particular, has been used to accurately predict gene and nutrient essentiality [2, 76, 77], microbial growth rates [78], metabolic flux distributions [79, 80], and evolutionary adaptations to environmental and genetic perturbations [81-83]. An important advantage of FBA methods is that they generally do not require kinetic parameters and rely on stoichiometric models that can be inferred from complete genome sequences. The accuracy of FBA methods has been independently demonstrated for many dozens of species encompassing diverse phylogenetic distributions and growth environments [84]. Recently, sophisticated methods have been developed for the automated reconstruction, optimization, and gap filling of flux-balanced stoichiometric models [1, 52, 54, 55, 85, 86]. Here we take advantage of these methods and use flux balance analysis to investigate the long-term evolution of microbial growth and gene-knockout phenotypic properties.

3.2. Results

3.2.1. *Species considered for* in silico *phenotyping*

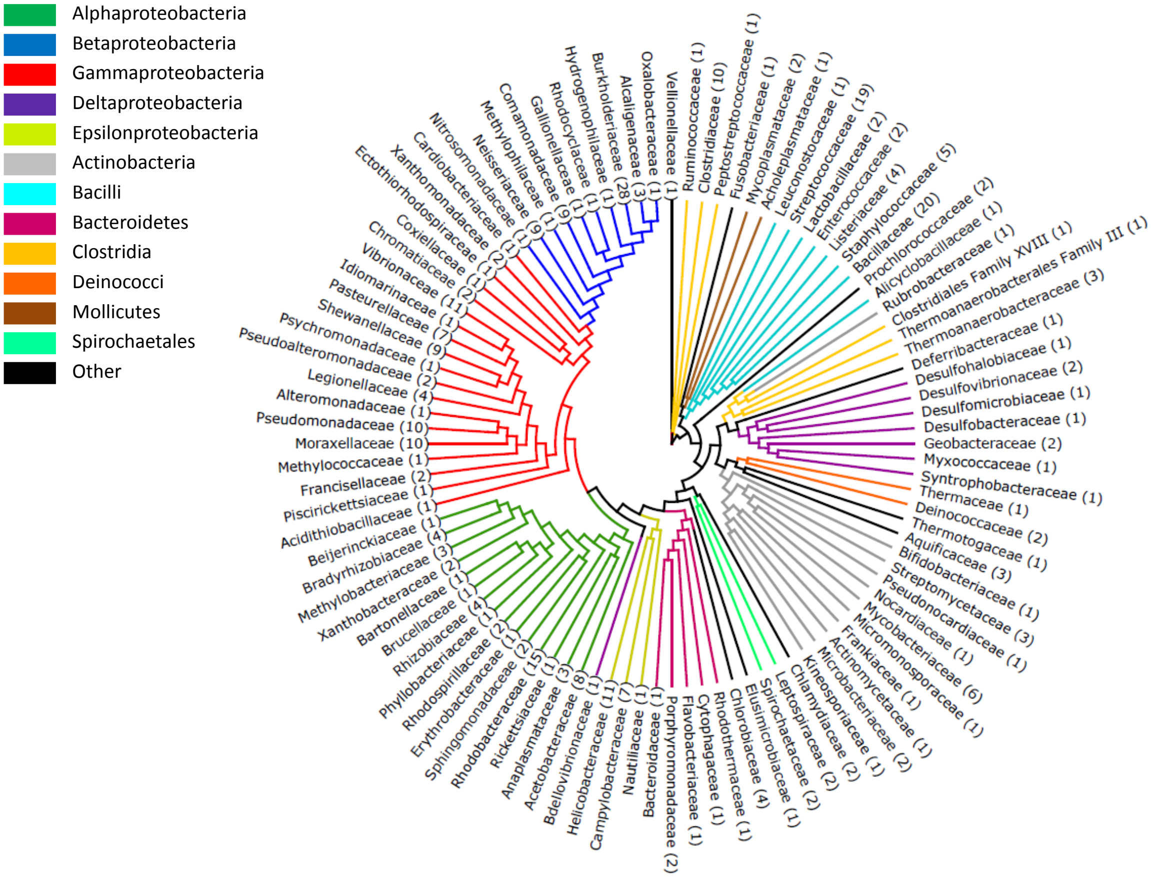


Figure 3.1. Phylogenetic tree showing 100 families of 322 bacterial species used in this study. The numbers of selected species per family are shown in parentheses. Different colors correspond to different classes of bacteria. The tree was built using the neighbor joining algorithm [87] based on the 16S rRNA distance between species.

We selected for our analysis 322 phylogenetically diverse bacteria (**Fig. 3.1**) for which genome-scale metabolic models were produced using the approach developed by Henry *et al.* [54] (See Methods). To quantify the evolutionary distance between bacterial species we used the divergence between their 16S rRNA sequences; 1% 16S RNA distance approximately corresponds to 50 million years of divergence since a common ancestor [88, 89].

3.2.2. Long-term evolution of growth phenotypes

To investigate the long-term evolution of growth phenotypes we considered 62 carbon sources that are commonly used by microbes for growth and energy production [90]. Specifically, for each of the considered species we used FBA to determine a subset of the compounds that could be used for biomass synthesis or generation of ATP –two of the essential metabolic objectives for bacterial growth [91]. This analysis resulted in binary phenotypic vectors that describe the ability of each microbial species to utilize each of the considered compounds (see Methods). The evolution of these phenotypic vectors, measured as the change in phenotypic similarity as a function of species divergence, is shown **Figure 3.2A,B**; the density plots in the figures were calculated based on pairwise comparisons of all considered species. Notably, this analysis demonstrates that the long-term evolution of growth phenotypes, averaged across species, can be approximated well by an exponential decay: the red lines in the figures show a running average of the density plots and the black lines show the exponential fits to the data (see Methods). A similar trend was also observed for phenotypic evolution with respect to compounds that can be utilized as a nitrogen source (see **Supplementary fig. 3.1**), and when the analysis was repeated testing a larger set of compounds as possible carbon sources (**Supplementary fig. 3.2**).

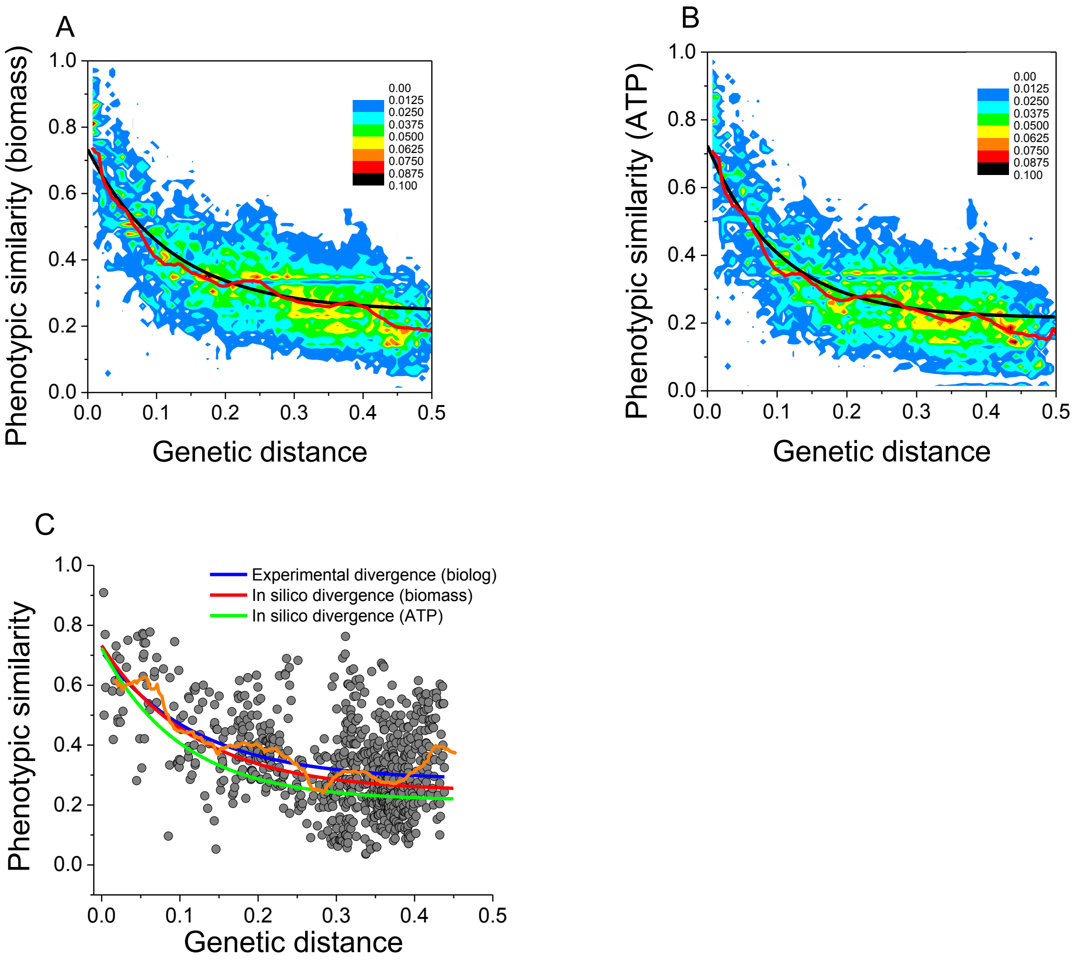


Figure 3.2. Evolution of growth phenotypes across bacterial species. The figures show the point density for all pairwise comparisons between 322 metabolic models at a given genetic distance. **A.** Similarity of carbon sources that can be used for biomass synthesis as a function of the genetic distance between species. The black line indicates an exponential fit of the data; the red line shows a 5% genetic distance moving average of the data. **B.** Similarity of carbon sources that can be used for ATP synthesis as a function of the genetic distance between species. **C.** Experimental validation of patterns of phenotypic divergence between bacteria. 40 species were tested for their ability to use 62 different carbon sources for growth using phenotypic microarrays. Each point in the figure indicates a pair of species. An exponential fit of the experimental data is shown in blue, the fits from A and B are shown in red and green, respectively. The orange line indicates a 5% genetic distance window moving average of the experimental data.

The observed exponential trends suggest that over long evolutionary distances approximately similar fractions of phenotypic properties are changing per unit time, as microbial species diverge and adapt to different environmental niches. Although the observed phenotypic evolution has been continuing for billions of years, for species separated by more than one billion years of evolution (~0.2 divergence in **Figure 3.2**) the divergence of growth phenotypes is approaching saturation. The average phenotypic similarity for biomass production (**Fig. 3.2A**) at these long evolutionary distances is ~25%, which is close to the value expected by chance given the relative frequency with which each of the considered compounds is used across models (**Supplementary fig. 3.3**). About half of the carbon sources shared at long evolutionary distances corresponds to metabolites that are used by more than 90% of the species (**Supplementary table 3.1)**. Frequently used compounds include sugars such as D-glucose, D-fructose, D-mannose, and D-maltose, which are typical substrates of glycolysis, and organic acids such as L-lactate, L-glutamate and L-malate.

Notably, before the evolution of growth phenotypes settles into the aforementioned average trend (distances <0.01 in **Fig. 3.2**), we observe an initial burst, .i.e. significantly higher rate, of phenotypic evolution. This results in an average phenotypic similarity at close genetic distances of about 75%. Such a rapid phenotypic evolution is a common hallmark of speciation and other important evolutionary transitions representing diversification to adapt to new habitats or host species [92-94]; it also reflects the underlying genomic plasticity of bacterial pan-genomes [6], which is a result of homologous recombination, mutation and horizontal gene transfer (HGT) acting on microbial populations [93, 95, 96].

3.2.3. Experimental validation of observed trends

To experimentally validate the predicted patterns of long-term phenotypic evolution, we selected 40 diverse microbial species (**Supplementary fig. 3.4**) for phenotyping using Biolog Phenotype MicroArrays (PMs) [90]. PM technology is based on the reduction of a tetrazolium dye which allows quantifying the metabolic activity of microbes across different growth conditions [97]. PM-based phenotyping has been previously used for bacterial strain identification, characterization of metabolic properties of diverse microbes, and the curation of genome-scale metabolic reconstructions, among several other applications [98, 99]. We used PMs to determine the ability of each of the selected bacteria to utilize the 62 different carbon sources used in the FBA simulations. As shown in **Figure 3.2C** we found a very good agreement between the average trends of phenotypic change as a function of genetic distance based on the experimental data and the computational predictions. The correspondence was further supported by a strong correlation (Pearson's r=0.75, p-value=3x10^{-9}) between the experimental and predicted similarity values for the 10 species present in both the computational and experimental analyses (**Supplementary fig. 3.5**).

3.2.4. Possible drivers of phenotypic divergence

There are several known examples of rapid bacterial evolution for species undergoing massive gene loss [100, 101]. Therefore, we investigated the extent to which phenotypic variance across bacteria can be explained by the size difference between species' metabolic networks (**Fig. 3.3A**). At close genetic distances the fraction of phenotypic variance explained by size differences for all microbial species is about 20% (**Fig. 3.3A**, gray bars). This fraction decreases to 5%-10% for diverged species. Notably, when only species associated with a host [9] are considered (**Fig. 3.3A**, red bars), the fraction of explained phenotypic variance increases to about 30% at close distances. Host-associated bacteria that acquire a symbiotic lifestyle frequently loose multiple biosynthetic pathways while adjusting their metabolism to the environments and nutrients available in their hosts [102-104].

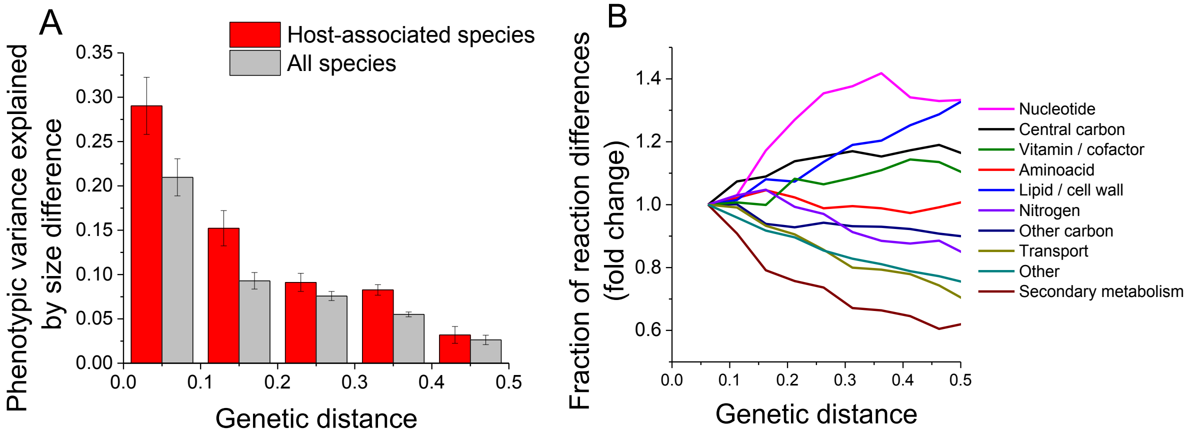


Figure 3.3. Drivers of fast and slow phenotypic change. A. Fraction of phenotypic variance explained by the relative size difference of metabolic networks at different genetic distances. Size differences are based on the number of reactions in each species' metabolic model; absolute differences were divided by the average number of reactions in each species pair. Error bars represent the R2 standard error. Values correspond to the square of the Pearson correlation coefficient between size difference and phenotypic similarity (biomass) at each bin. **Supplementary figure 3.6** shows the results for the similarity of carbons sources used for ATP production. **B.** Fold-change in the fraction of reaction differences for various metabolic subsets as a function of the genetic distance between species. Fold-change is calculated relative to the fraction of reaction differences between metabolic networks for the first bin. The lines correspond to a moving average across genetic distances (window=0.2, step=0.05).

Other than massive gene losses, phenotypic change can also be explained by the types of reactions that are changed between metabolic networks. In agreement with previous studies [40, 105], we found that at close genetic distances (<0.05, **Supplementary table 3.2**) there is a significant overrepresentation of reaction differences related to secondary metabolism and transport, while differences in central metabolic processes, such as nucleotide, and central carbon metabolism are underrepresented. As the genetic distance between species increases (**Fig. 3.3B**) the fraction of differences in central metabolic processes goes up while the fraction in peripheral pathways becomes less prominent.

3.2.5. Phenotypic similarity in the context of bacterial taxonomy

We next investigated the diversity of metabolic phenotypes within different levels of conventional taxonomic classification (**Fig. 3.4**). The figure shows the distribution of phenotypic similarities –based on carbon source usage for growth– for pairs of bacteria related at different taxonomic ranks. Bacteria from the same species show mostly similar phenotypic properties; nevertheless, as apparent from the long left tail of the distribution, some organisms can show substantial phenotypic differences even at this taxonomic level. The relatively high average phenotypic similarity at the species level indicates that currently used criteria to define microbial species [106], usually correctly delineate clusters of phenotypically related organisms. Notably, at the genus level the distribution is very broad and multimodal, with some bacteria displaying high levels of phenotypic similarity, and some others showing differences that are more typical of the higher ranks. Beyond the genus, for species related at the levels of family, order, class, and phylum there is in general little change in average phenotypic similarity, which remains low at about 30-40%. Comparison with experimental data at each taxonomic rank shows good agreement with the predicted distributions (stars in **Fig. 3.4**).

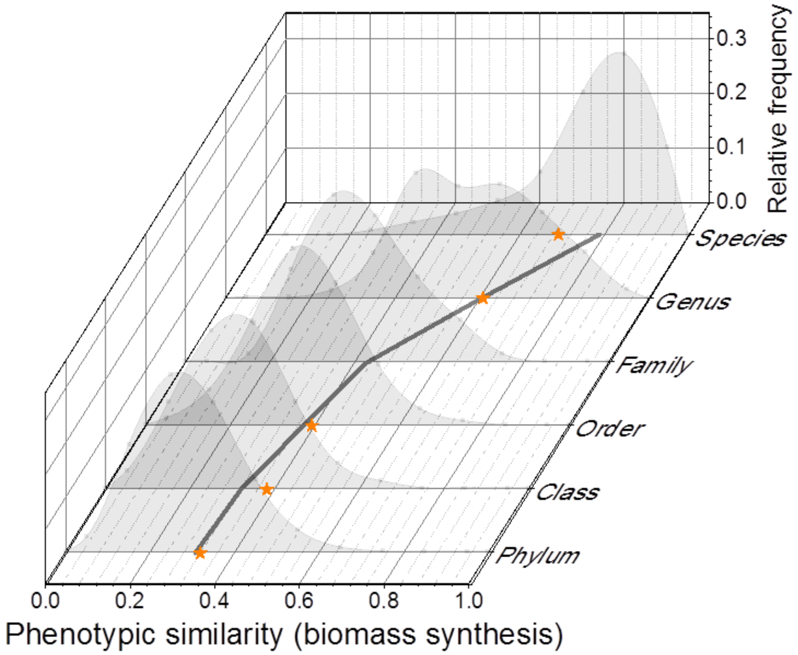


Figure 3.4. Phenotypic similarity distribution at different taxonomic ranks. The figure shows the frequency distribution of carbon source similarities for pairs of species related at a given taxonomic level. The stars indicate the average values for available experimental data; the dark gray line connects the mean values at each taxonomic level.

3.2.6. Using metabolic phenotypes as classification markers

Although it is generally accepted that bacterial species should be soundly predictive of the phenotypic potential of a strain [107], common species definition criteria (e.g. based on DNA-DNA hybridization or 16S rRNA gene sequence identity [106]) have been criticized for resulting in too much phenotypic variation within a named species [107]. Indeed, despite being more phenotypically similar than strains related at broader taxonomic ranks (**Fig. 3.4**), strains in the same species still differ by an average 20-25% of the carbon sources they can use for growth. The frequency distributions in **Figure 3.4** are very wide (range ~80%), and all of them overlap with each other to some extent; this suggests that multiple pairs of strains could be more or less ecologically and phenotypically related than expected from their taxonomic classification.

Although several methods have been proposed to build phylogenetic trees based on genomic data [108], the flux analysis approach can provide complementary information by directly considering a variety of phenotypic properties to determine associations between bacteria.

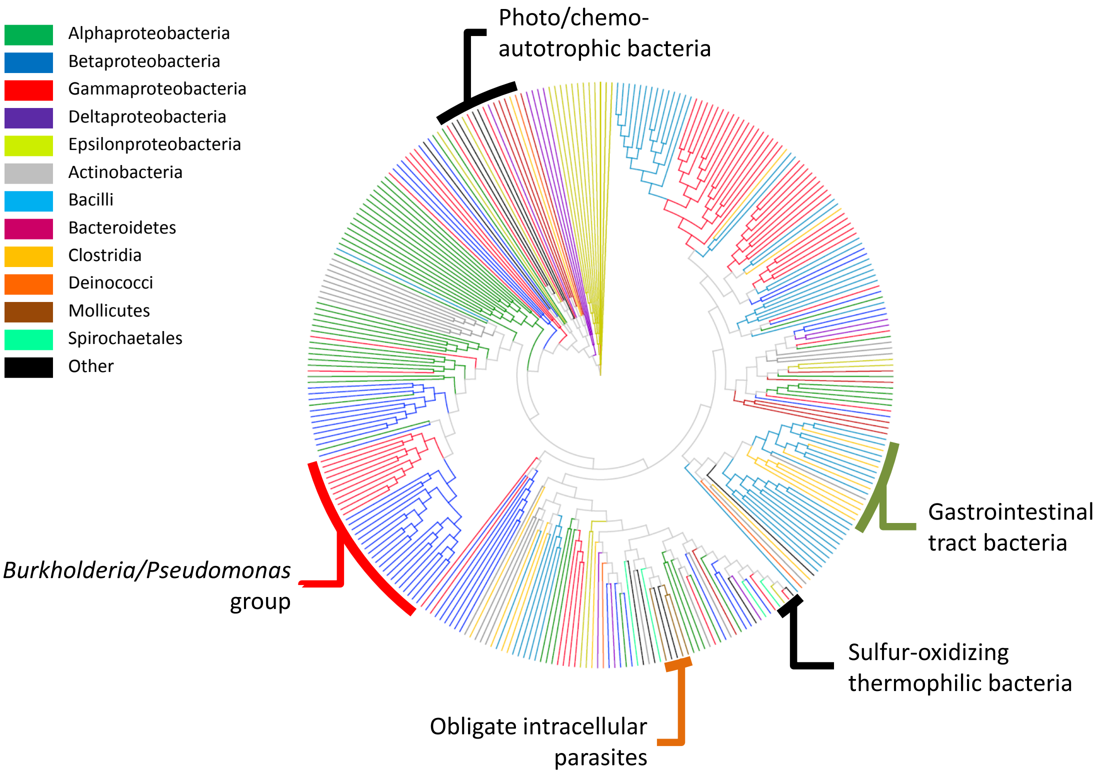


Figure 3.5. Phenotypic tree of 322 phylogenetically diverse bacterial species. The tree is based on the Jaccard distances of binary vectors representing the ability of species to use different carbon sources for growth. 443 carbon containing compounds were tested as possible carbon sources.

In **Figure 3.5**, using the neighbor joining algorithm [87], we show a 'phenotypic tree' for the 322 bacteria considered in this study. The tree is based on the Jaccard distance (1 minus phenotypic similarity) between carbon source utilization vectors; each node in the tree corresponds to one of the 322 strains and colors correspond to different classes of bacteria (similar to **Fig. 3.2**). Notably, while phenotypic proximity generally groups bacteria belonging to the same taxonomic class close in the tree, there are multiple regions where bacteria from different groups are closer to each other than they are to more phylogenetically related strains. A closer inspection of these cases revealed multiple ecological similarities between widely diverged bacteria. For example, we identified a tight cluster of phylogenetically distant species – *Aquifex aeolicus* (Aquificae), *Thiomicrospira crunogena* (Gammaproteobacteria) and *Sulfurimonas denitrificans* (Epsilonproteobacteria)– which are all sulfur-oxidizing, thermophilic bacteria. We also found a higher than expected phenotypic similarity between bacteria in the genera *Mycoplasma* (Mollicutes) and *Chlamydia/Chlamydiophila* (Chlamydiia), all displaying and obligate intracellular lifestyle. On the same vein, *Mycobacterium tuberculosis* (Actinobacteria) and *Legionella pneumophilla* (Gammaproteobacteria), both of which are known to infect human macrophages, were found together in the phenotypic tree. Another interesting association was that of *Clostridium difficile* and *C. botulinum* (Clostridia) with species that also find their habitat in the human gastrointestinal tract such as *Enterococcus faecalis* and *Lactobacillus plantarum* (Bacilli).

Our analysis identified multiple species of *Burkholderia* (Betaproteobacteria) and *Pseudomonas* (Gammaproteobacteria) which despite being related only at the phylum level show a relatively high level of phenotypic similarity (~60%). Interestingly, *Burkholderia* species were initially classified as part of the genus *Pseudomonas*, but they were later re-classified based on several genetic and biochemical characteristics [109]. Importantly, these two groups of species share several interesting biological properties such as their occurrence as human and plant pathogens [110], and their reported co-existence in specific microbial consortia in environments such as the rizosphere and waste water sludge [111, 112]. These observations suggest that *in silico* profiling of bacterial metabolic properties could, in some cases, support a more phenotype-oriented bacterial taxonomy [107].

The above results suggest that ‘phenotypic trees’ can be used as a tool to generate and evaluate hypotheses about microbial ecology. First, unrelated species that are close in phenotypic space potentially share common lifestyles and biochemical characteristics. Experimentally measured phenotypes, e.g. through phenotypic microarrays, and careful curation of the metabolic models can further validate these commonalities. Second, flux analyses of bacterial strains sequenced from a single location or microbial consortia could reveal the extent to which different environments select for bacteria with specific characteristics. In particular, the presented approach could be used to measure phenotypic distances between bacteria within a particular environment, and contrast these distances with the expected similarity based on their taxonomic/genetic proximity. As the quality of automated reconstructions improves, and sequencing of single cells directly from the environment [11] continues to expand, *in silico* phenotyping will play an important role in microbial ecological theory.

3.2.7. Evolution of gene essentiality and epistasis

To complement our analysis of metabolic growth phenotypes, we used FBA to investigate the long-term evolution of genetic phenotypes. Specifically, we considered the evolution of metabolic gene essentiality and gene-pair synthetic lethality (i.e. the evolution of genetic interactions) between bacterial species. To test these genetic phenotypes we determined the ability of each species to synthesize biomass on rich media, i.e. when all nutrients were available to the models, after *in silico* removal of model reactions corresponding to the metabolic genes considered (see Methods). The analysis demonstrated that the long-term evolution of gene essentiality can also be well described by an exponential decay (**Fig. 3.5A**, red line – moving average, black line –exponential fit). Nevertheless, the average evolution of reaction essentiality was significantly slower and reached a saturation level faster (around genetic distance 0.1) compared to the evolution of metabolic growth phenotypes (**Fig. 3.2**). The results suggest that even at long evolutionary distances, for an average pair of microbial species, about half of the conserved essential genes in one species remain essential in the other. Notably, the computationally predicted trend is consistent with available experimental data (**Fig. 3.5A,** black dots) for microbial species with high-quality genome-wide gene deletion screens.

In stark contrast to gene essentiality, our analysis revealed a very fast divergence of genetic interactions between metabolic genes (**Fig. 3.5B**). Even at close evolutionary distances synthetic lethality is conserved only for about 30% of metabolic gene pairs. As bacterial species diverge, synthetic lethality quickly drops further to about 5%. This demonstrates that synthetic lethality similarity is much more sensitive to changes of microbial genotypes than gene essentiality and growth-related phenotypic similarities. Only several comprehensive attempts have been made to experimentally assess the conservation of genetic interactions. Comparison of fitness data from budding and fission yeast revealed a similarity of ~26% [113]. On the other hand, only 5% of the orthologs of epistatic gene pairs in yeast were also found to be epistatic in *Caenorhabditis elegans* [114]. Although these data were obtained from eukaryotes, the experimental results (**Fig. 3.5B,** black dots) are generally consistent by the bacterial simulations, with a minor fraction of epistatic interactions shared at most genetic distances.

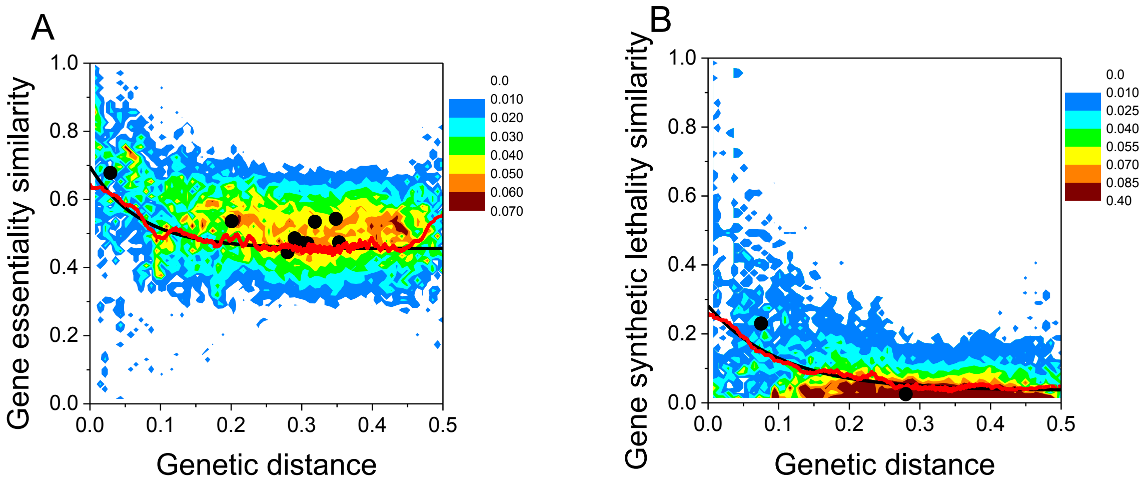


Figure 3.6. Similarity of metabolic gene essentiality and synthetic lethality across bacterial species. A. Density plot for the similarity of gene essentiality as a function of the genetic distance between species. Black lines indicate exponential fits of the data, red lines represent a 5% genetic distance moving average. Black dots indicate experimentally derived results from metabolic gene knockout experiments based on all pairwise comparisons between *Escherichia coli* [115], *Bacillus subtilis* [116], *Streptococcus sanguinis* [117], *Salmonella enterica* [118] and *Caulobacter crescentus* [119]. **B.** Density plot for the similarity of gene-pair synthetic lethality as a function of species' genetic distance. Black and red lines are as in A. The black dots represent experimentally determined values of synthetic lethality conservation between budding and fission yeast [113], and between budding yeast and *C. elegans* [114].

3.3. Discussion

Analysis of phenotypic evolution, such as the morphological variation of beaks in Darwin's finches [120], provided the original impetus and context for understanding natural selection. Understanding of phenotypic diversity within and between species continues to be an important area of biological research [50, 121-123]. Because the evolutionary significance and physiological role of different phenotypic traits change over time, it is often difficult to establish a clear mapping between genotypes and phenotypes for metazoans, in particular across long evolutionary distances. For microbial species, on the other hand, the ability to metabolize different nutrient sources, although clearly not the only important phenotype, always remains an essential determinant of bacterial fitness and lifestyle. Flux analyses of metabolic networks provide a relatively straightforward connection between genotype and phenotype. Using these tools and properties of microbial metabolism we were able to trace phenotypic change across four billions years of bacterial evolution. Notably, the observed clock-like behavior of phenotypic divergence is reminiscent of the molecular clock in protein evolution [65, 71]. Similar to protein evolution, it is likely that the long-term trends for phenotypic divergence are due both to bacterial adaptation to diverse environmental niches and neutral changes [93, 124-126]. The relative contribution of adaptive and neutral changes is likely to be different in each particular linage and evolutionary context. Our analysis shows that growth phenotypes, gene essentiality, and synthetic lethality diverge with different rates and respond differently to changes in bacterial genotypes. It is likely that other phenotypic properties, such as the ability to produce different compounds or withstand specific environmental perturbations, will also show distinct evolutionary patterns. Similar analyses performed on defined subsets of bacteria can be used to predict ecological relationships between species and provide clues about the evolutionary forces acting on different lineages. As genome sequencing is currently expanding at an unprecedented rate [8] and metabolic reconstruction methods are continuously improving (**Chapter 2**), it will be possible to construct in the near future a detailed map of phenotypic evolution across all sequenced microbial species.

3.4 Methods

3.4.1. Model selection

We obtained 322 genome-scale metabolic network models using the method of Henry *et al*. [54], including 123 models discussed in their original publication. The remaining 199 models were selected such that the reconstruction strategy used less than 20% reactions not supported by gene annotations to produce flux-balanced models of the corresponding species. We decided to ignore all genomes in the order *Enterobacteriales* for two reasons; first, there were significantly more genomes for this group than other sets of bacteria, which would significantly bias the all-against-all results. Second, gene annotation for these species displays a higher coverage than most other bacteria due to the close phylogenetic proximity to the well-studied model bacteria *E. coli*, which would produce numerous metabolic differences more often related to the completeness of the models than the underlying biology [56].

3.4.2. Flux balance analysis

Flux balance analysis finds feasible values of metabolic reaction fluxes subject to reaction stoichiometry constraints and the assumption of steady state of metabolites. These values are typically selected to maximize a specific objective such as biomass or ATP production. Additional constraints can be used to assign upper and lower bounds to fluxes going through specific reactions or reaction combinations [45]. In order to simulate the ability of species to use different carbon sources for growth or energy (ATP) production, we first simulated a carbon limited growth condition. To do this, a constraint was added such that the total number of carbon atoms going through the full set of metabolite uptake reactions was below a given constant. We then determined the maximum biomass or ATP produced given this constraint. For each compound tested, we ran FBA using the same conditions, but allowing a larger flux through the uptake reaction of the corresponding compound. If the resulting biomass or ATP produced was higher than the original value, we concluded that such compound could be used by the bacterium. The 62 compounds that we tested correspond to carbon sources typically assayed in Biolog [90] experiments that could be mapped to compounds in the metabolic reconstructions. An analogous procedure was used to test for nitrogen source phenotypic similarity.

To test for gene essentiality, based on the gene-reaction associations encoded in each model, we set the maximum flux through the reactions that depended on each gene to zero, and solved FBA for maximal biomass production. If biomass could not be produced, the corresponding genes were labeled as essential. A pair of non-essential genes was considered to be synthetic lethal if simultaneous deletion of the pair of genes resulted in zero biomass production as predicted by FBA. These simulations were made on an *in silico* rich medium, meaning that fluxes were allowed through every transport reaction present in the metabolic networks.

All FBA problems were solved using the COBRA toolbox [127].

3.4.3. Measuring phenotypic similarity

For a set of features, i.e. essential genes, epistatic gene pairs, or carbon sources that can support growth, similarity between two species was measured using Jaccard's similarity index [128], which is defined as the size of the intersection divided by the size of the union between two sets. For example, if A represents the set of carbon sources that can be used by species *a*, and B the set for species *b*, then carbon source similarity between *a* and *b* is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Importantly, to assess the similarity of gene essentiality and gene pair synthetic lethality we only considered genes shared between metabolic networks. Orthologous genes were identified as bi-directional BLASTP [18] hits with an e-value cutoff of (0.05).

3.4.4. Validation with Biolog phenotypic arrays

40 species spanning a wide range of phylogenetic distances (**Supplementary fig. 3.4**) were tested against a panel of 62 carbon sources using Biolog phenotypic microarrays [90]. The Biolog system uses a colorimetric technique to provide a quantitative measure of microbial respiration in the presence of specific compounds. Results for each species were normalized to a scale of 0 to 100; we considered any value above 5 to be positive for growth on a given carbon source and any value below to be negative. Similar results were obtained at other cutoff values.

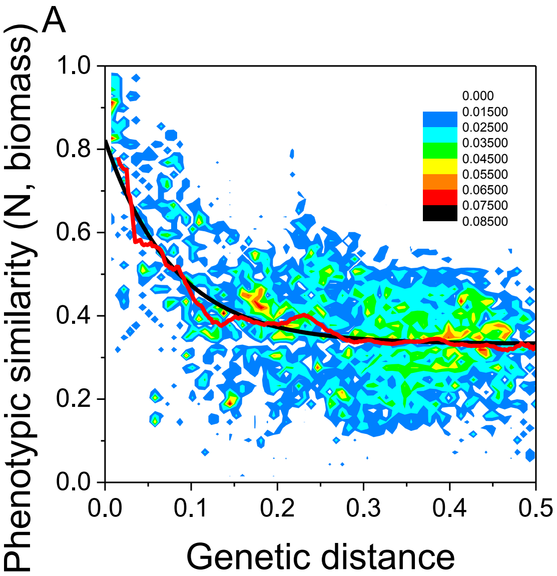
3.4.5. Exponential decay fits

Pairwise comparisons of phenotypic similarity (*y*) as a function of genetic distance (*x*) were fitted using the following formula:

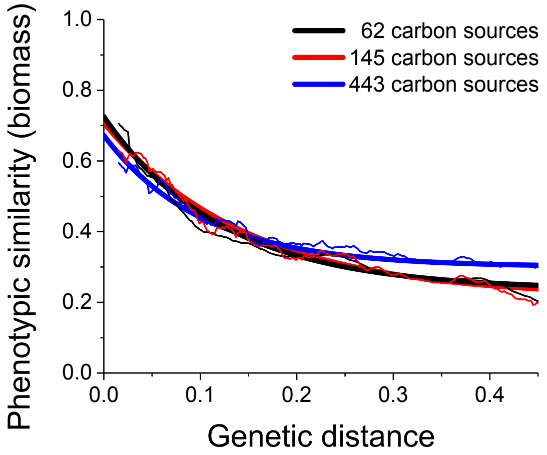
$$y = a + b \times e^{cx}$$

In the above equation, a represents the saturation level at long distances, *a*+*b* represent the phenotypic similarity at zero genetic distance, and *c* measures the decay rate, with lower (more negative) values of *c* associated with shorter decay times.

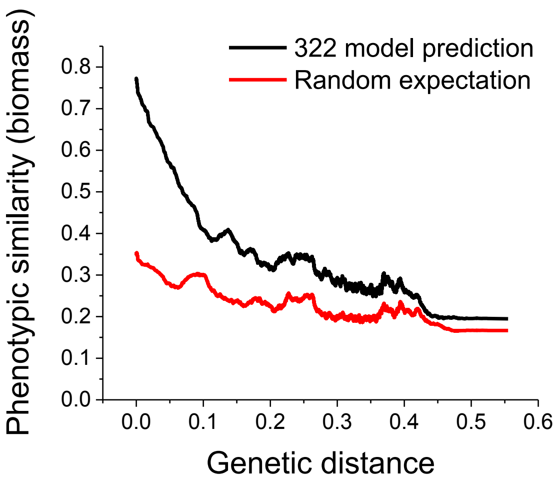
3.6. Supplementary figures and tables



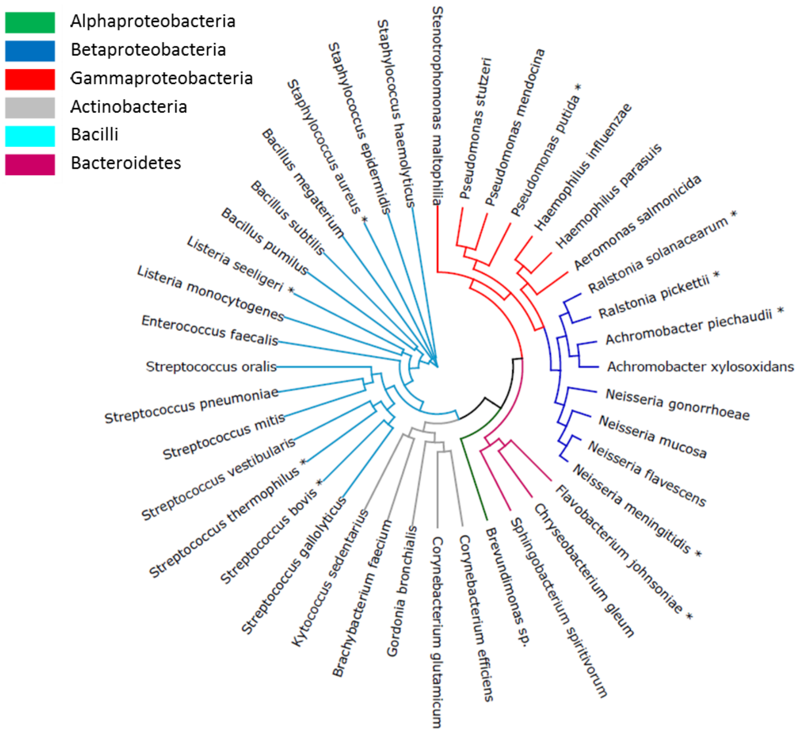
Supplementary figure 3.1. Evolution of phenotypic similarity based on the fraction of shared nitrogen sources that can be used for growth. The graph shows the point density at a given genetic distance for pairwise comparisons between 322 species. The red line is a 5% genetic distance moving average, and the black line is an exponential fit of the data. Results are based on the utilization of 74 nitrogen sources typically tested in phenotypic microarrays.



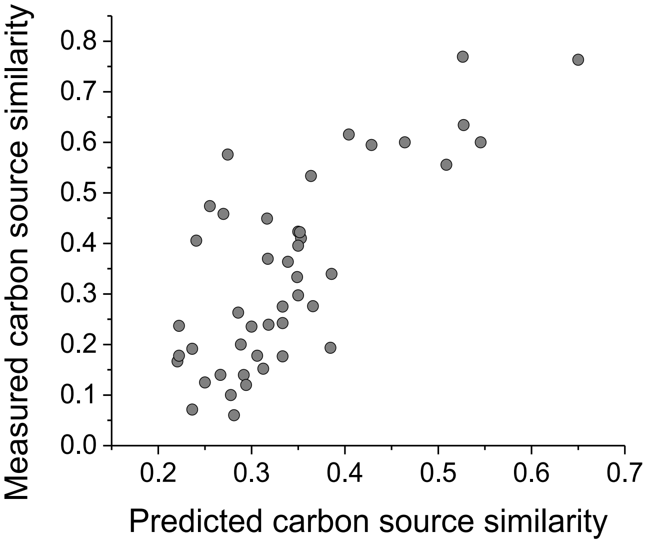
Supplementary figure 3.2. Evolution of phenotypic similarity for different sets of carbon sources. Thin lines are a moving average of the data (window size 1%, step size 0.2%) and thick lines are an exponential fit. The set of 145 carbon sources includes the 62 compounds used in our main analysis as well as 83 other carbon sources typically tested in phenotypic microarrays. The set of 443 carbon sources includes all possible compounds containing carbon that were present across the 322 models.



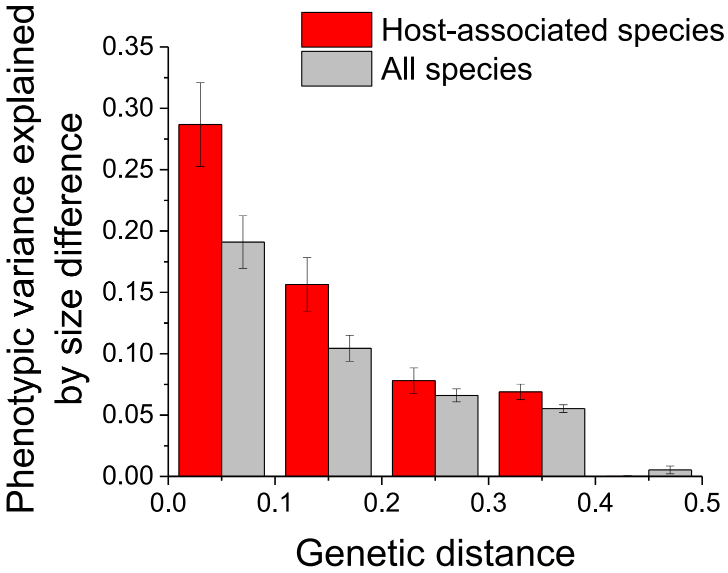
Supplementary figure 3.3. Null expectation of phenotypic similarity. The figure shows the average phenotypic similarity for all pairwise comparisons between 322 species as a function of genetic distance (black), and the average similarity when the same number of compounds used by each species is chosen at random according to the frequency with which those compounds are used across the 322 models (red).



Supplementary figure 3.4. Phylogenetic tree for the 40 species used for experimental validation. Different colors indicate different classes of bacteria. The tree is based on the 16S rRNA distance between species. Asterisks mark the species that were present in the set of 322 used for the computational predictions.



Suplementary figure 3.5. Correlation between predicted and measured phenotypic similarity. Data is for all pairwise comparisons between the 10 species in common between the set of 322 models used for computational predictions and the set of 40 used for validation. Pearson's r=0.75, p-value=3x10^{-9} Spearman's r: 0.66, p-value=7x10^{-7}.



Supplementary figure 3.6. Phenotypic variance explained by size differences as a function of genetic distance. Values correspond to the square of the Pearson correlation coefficient between size difference and phenotypic similarity at each bin. Phenotypic similarity is for the fraction of shared carbon sources that can be used for ATP production. Error bars represent the R^2 standard error.

Supplementary table 3.1. Frequency of metabolite usage as a carbon source across 322 species (ranked by frequency) as predicted by FBA.

Metabolite name	Number of models	Metabolite name	Number of models
L-Glutamic Acid	222	D-Sorbitol	63
a-D-Glucose	210	D-Gluconic Acid	58
D-Fructose	198	D-Galacturonic Acid	58
L-Malic Acid	173	Mucic Acid	55
L-Lactic Acid	171	D-Saccharic Acid	55
Maltose	158	D-Galactose	48
Glycerol	142	D-Raffinose	30
L-Serine	133	Dextrin	29
L-Arginine	132	N-Acetyl-b-D-Mannosamine	28
L-Aspartic Acid	127	g-Amino-Butyric Acid	27
D-Mannose	122	L-Rhamnose	27
Citric Acid	118	Salicin	26
N-Acetyl-D-Glucosamine	114	D-Glucose-6-Phosphate	24
D-Trehalose	104	Propionic Acid	16
Sucrose	94	D-Melibiose	16
L-Histidine	93	D-Aspartic Acid	16
Inosine	92	Quinic Acid	15
a-Keto-Glutaric Acid	89	L-Fucose	13
L-Alanine	81	m-Inositol	13
Formic Acid	80	Stachyose	9
D-Glucuronic Acid	79	N-Acetyl-Neuraminic Acid	8
a-D-Lactose	75	D-Arabitol	7
D-Serine	74	Acetic Acid	7
Acetoacetic Acid	74	N-Acetyl-D-Galactosamine	6
D-Cellobiose	71	D-Fructose-6-Phosphate	5
D-Mannitol	68	a-Keto-Butyric Acid	4
D-Malic Acid	64		

Supplementary table 3.2. Overrepresentation of functional categories for reaction differences between species at close genetic distances (genetic distance < 0.05), relative to random species pairs.

Functional category	Sense	Fraction of differences: Close species/Random pairs	P-value*
Nucleotide	Underrepresented	0.69	5E-34
Central carbon	Underrepresented	0.73	1E-26
Vitamin / cofactor	Underrepresented	0.86	2E-06
Aminoacid	Underrepresented	0.92	2E-04
Lipid / cell wall	Underrepresented	0.94	6E-04
Other carbon	--	1.04	3E+00
Other nitrogen	--	1.05	8E+00
Other	Overrepresented	1.25	3E-40
Transport	Overrepresented	1.32	7E-24
Secondary metabolism	Overrepresented	1.53	7E-51

* P-values correspond to the Bonferroni corrected chi-square test.

Chapter 4.

GLOBAL PROBABILISTIC ANNOTATION OF METABOLIC NETWORKS ENABLES ENZYME DISCOVERY

This chapter is based on "Plata, G., Fuhrer, T., Hsiao, T., Sauer, U., Vitkup, D. **Global Probabilistic Annotation of Metabolic Networks Enables Enzyme Discovery**. Nat. Chem. Biol. (10):848-54, (2012)"

Annotation of organism-specific metabolic networks is one of the main challenges of systems biology. Importantly, owing to inherent uncertainty of computational annotations, predictions of biochemical function need to be treated probabilistically. We present a global probabilistic approach to annotate genome-scale metabolic networks that integrates sequence homology and context-based correlations under a single principled framework. The developed method for global biochemical reconstruction using sampling (GLOBUS) not only provides annotation probabilities for each functional assignment but also suggests likely alternative functions. GLOBUS is based on statistical Gibbs sampling of probable metabolic annotations and is able to make accurate functional assignments even in cases of remote sequence identity to known enzymes. We apply GLOBUS to genomes of *Bacillus subtilis* and *Staphylococcus aureus* and validate the method predictions by experimentally demonstrating the 6-phosphogluconolactonase activity of YkgB and the role of the Sps pathway for rhamnose biosynthesis in *B. subtilis*.

4.1. Introduction

Advances in DNA sequencing technologies and high-throughput experiments provide a unique opportunity to study cellular function at the systems level. The systems biology perspective seeks to understand how the interaction between multiple genomic components determines cellular physiology. Genome-scale metabolic networks serve as an important platform for such systems analyses and have been very successful in predicting various emergent properties of biological systems. They also have great potential for guiding metabolic engineering [129] and aiding drug target discovery [130]. Unfortunately, accurate manual annotations of organism-specific metabolic networks are laborious and can take up to a year for a typical microbial genome. Efforts have been made to automate the reconstruction process, particularly the initial steps of genome annotation and network assembly [85, 86, 131, 132].

The annotation process usually relies on sequence homology methods, in which the function of a metabolic gene is assigned based on sequence similarity to known enzymes [133]. Although homology methods have been successful overall, annotations established based solely on weak sequence identity are often unreliable due to frequent functional divergence between distant homologues. It was demonstrated that a sequence identity above 60% is usually required to accurately transfer a precise enzyme function, i.e. all four digits of an Enzyme Commission (EC) number [20]. Consequently, homology-based methods fail to assign functions to a substantial fraction of genes in completely sequenced genomes [134] and have been known to produce multiple imprecise or incorrect annotations [24, 36, 135].

The metabolic network reconstruction for a given genome is usually performed based on a functional annotation of all metabolic genes. Functional databases such as BRENDA [136], GeneCards [137], KEGG [131], MetaCyc [138] or Swiss-Prot [139] are useful resources for establishing initial associations between metabolic genes and corresponding biochemical reactions. Draft metabolic models are typically reconstructed by assembling annotated biochemical reactions into a network. One disadvantage of this two-step approach is that genes are annotated individually rather than being considered together in a proper network context. Therefore, some successful computational approaches utilize pre-defined or manually curated metabolic pathways [28] and subsystems [30] to annotate network reactions. Naturally, the accuracy of such methods depends both on the quality of the initial annotation and the evolutionary conservation of reference pathways.

Context based methods such as phylogenetic profiles [140], protein fusions [35], gene co-expression [141], and chromosomal gene neighborhood [34, 142] capture conserved functional relationships and often provide information complementary to sequence homology [143]. The effectiveness of these methods has been shown by determining members of protein complexes, functional modules, and molecular pathways [144, 145]. Multiple studies have also demonstrated that context associations combined with local network structure can be used to identify genes responsible for orphan metabolic activities and to improve existing annotations of metabolic genes [31, 33, 146]. Therefore, it is natural to combine sequence homology and context functional descriptors using a unified probabilistic framework.

Although powerful probabilistic approaches, such as Bayesian and Boolean networks [147], have been applied to reconstruction of regulatory and signaling networks based on high-throughput data [148], global probabilistic methods to annotate metabolic networks have not been developed. Here, we present such a global probabilistic approach that integrates sequence homology and context associations to annotate genome-scale metabolic networks. The method for Global Biochemical reconstruction Using Sampling (GLOBUS) not only provides annotation probabilities for each gene and each metabolic activity, but also suggests possible alternative functions. We apply GLOBUS to the genomes of *Bacillus subtilis* and *Staphylococcus aureus*, evaluate the accuracy of the reconstructed networks, and experimentally validate three *B. subtilis* predictions that have important functional consequences.

4.2. Results

4.2.1. Strategy of a global probabilistic reconstruction

The conceptual outline of GLOBUS is shown in **Figure 4.1**. First, we built a generic metabolic network containing all possible metabolic activities characterized in the Enzyme Commission (EC) system (http://www.chem.qmul.ac.uk/iubmb/enzyme/). Nodes of this EC network represent known enzymatic activities (**Fig. 4.1a**), and network edges are established by metabolites shared between the activities either as substrates or products [38]. The usage of the global EC network allowed us to consider gene function in a proper network context without predefining metabolic pathways. With the EC network as a scaffold, the global metabolic reconstruction for a given organism is equivalent to assigning metabolic genes to their correct network locations (**Fig. 4.1b**). In this way, organism-specific networks will occupy a subset of all possible locations (activities) in the global EC network.

A gene assigned to its correct network location usually has at least remote sequence identity to enzymes known to catalyze the corresponding activity. In addition, a correctly assigned gene often has good context correlations with its network neighbors. As we demonstrated previously, the genes with high mutual context correlations tend to be located closer in metabolic networks [33]. For example, in **Supplementary figure 4.1** we show that the higher a context correlation between a pair of *Saccharomyces cerevisiae* genes, the more likely that the genes are direct network neighbors.

In GLOBUS we used sequence homology and context correlations to evaluate a given global assignment of multiple metabolic genes into a set of network locations using a Markov-like fitness function. The contribution of each gene to the fitness function depends on the sequence identity to the assigned location and the context correlations with the genes assigned to neighboring network positions. The overall GLOBUS fitness function $E(g_1, g_2, ..., g_n)$ (see Methods), which is calculated based on a given assignment of metabolic genes $(g_1, g_2, ..., g_n)$, consists of the following terms:

$$E(g_1, g_2, \dots, g_n) = -b_{homology} f_{homology} - b_{orthology} f_{orthology} - b_{context} f_{context} - b_{ECco-occurrence} f_{ECco-occurrence}$$

where f_s are various homology-based and context-based functional descriptors, and b_s are corresponding positive coefficients representing weights of each descriptor in the fitness function. For homology descriptors we used two separate terms: 1.) the highest sequence identity to a Swiss-Prot [149] protein annotated to catalyze the corresponding activity in other species (annotations marked as based exclusively on computational methods were excluded), and 2.) a binary (0 or 1) descriptor indicating if a protein ortholog in another species is annotated to catalyze the activity (see Methods). For context-based descriptors we used three types of gene-gene correlations: phylogenetic profiles (which quantify the co-occurrence of gene orthologs across species, see Methods), chromosomal gene clustering across sequenced genomes, and mRNA co-expression. For each context descriptor, we considered the maximum correlation Z-score (see Methods) between the gene under consideration and genes assigned to neighboring network locations. In addition, we also considered a context term describing the co-occurrence across sequenced genomes of various metabolic activities according to annotations available in the KEGG database.

Using the described fitness function the global probability for a particular assignment of multiple genes into their network locations is given by *P(g_1, g_2, …., g_n)* based on the relationship used in statistical physics and Markov Random Fields (MRF)[150]

$$P(g_1, g_2, ..., g_n) = \frac{1}{Z} \times e^{-E(g_1, g_2, ..., g_n)}$$

, where *E(g_1, g_2, …., g_n)* is the aforementioned fitness function, and *Z* is a normalizing partition function, which is necessary to insure that probabilities of all possible metabolic assignments sum to one. Using the defined probabilities we sampled from all possible assignments proportionally to their likelihood using Gibbs sampling [151]. Gibbs sampling is a version of Markov Chain Monte Carlo (MCMC) [152] and has been successfully used in many computational biology applications, such as finding transcription factor binding sites in a set of DNA sequences [153]. The efficiency of the Gibbs sampling in GLOBUS is due to the fact that although there is a combinatorially large number of possible metabolic assignments, the vast majority of them have very low probabilities. Gibbs sampling allows to efficiently sample the most relevant global assignments according to their probabilities.

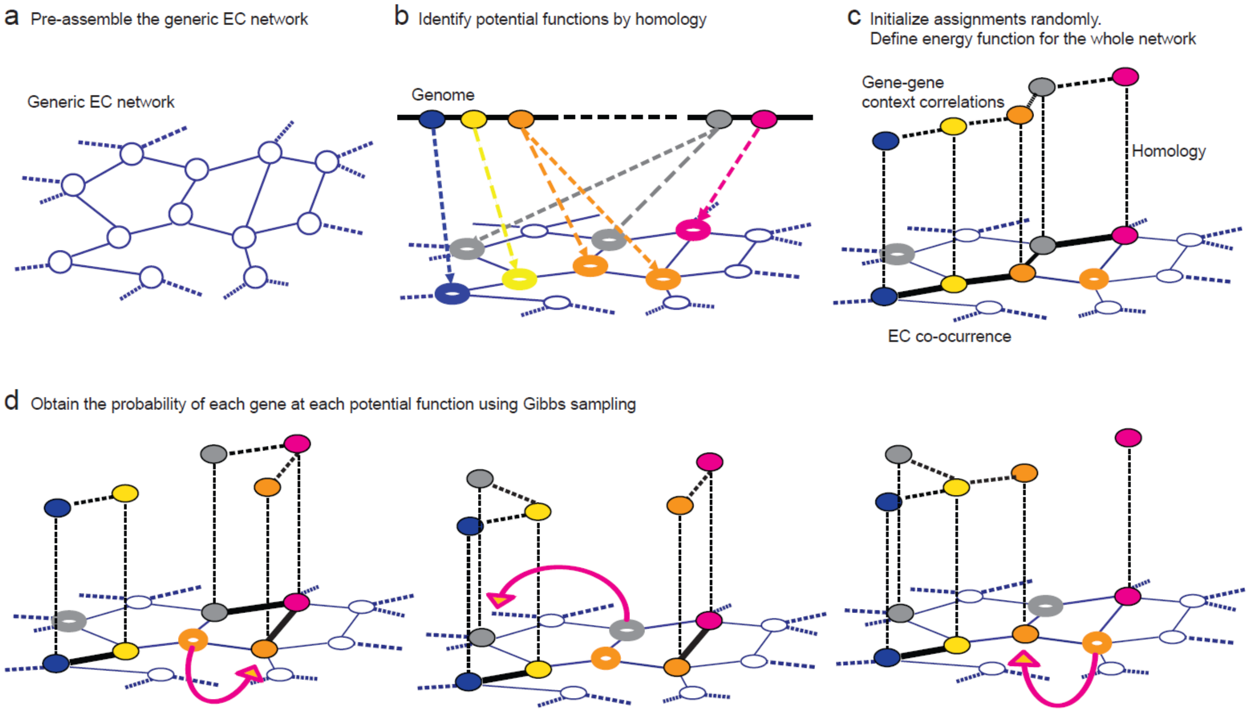

a Pre-assemble the generic EC network
Generic EC network
b Identify potential functions by homology
Genome
c Initialize assignments randomly.
Define energy function for the whole network
Gene-gene
context correlations
Homology
EC co-ocurrence
d Obtain the probability of each gene at each potential function using Gibbs sampling

Figure 4.1. Overview of the GLOBUS method. (a) A generic Enzyme Commission (EC) network is defined where nodes represent all known biochemical activities and edges represent connections between the activities established by shared metabolites. **(b)** For a genome of interest, the potential network locations of each gene are assigned based on sequence homology to known enzymes. **(c)** Each gene is initially assigned randomly to one of its possible locations. A fitness function is defined such that assignments to locations with high sequence identity and good context correlations with neighboring genes correspond to higher values of the fitness function (higher probability). **(d)** Gibbs sampling is used to sample all possible assignments of genes to their candidate network locations. At each step of a Gibbs chain a random gene is selected and re-assigned to one of the possible locations (arrows). The marginal probabilities for assigning every gene to each candidate network location are derived from converged Gibbs chains.

A step in a Gibbs chain was simulated by: 1.) selecting a random gene assigned to a particular network location, 2.) determining the probabilities for all possible locations of the selected gene, including the present location, and 3.) re-assigning the gene to a location according to the calculated probabilities (**Fig. 4.1c-f**). In the sampling we only considered the locations with at least remote sequence identity to the corresponding gene. In addition to possible locations in the network, a special out-of-the-network node was created, and in all Gibbs steps the move to the out-of-the-network node was also considered. The energy contribution to the fitness function for all genes located in the out-of-the-network node was the same. The energy in the out-of-the-network node is a parameter of the simulation (see below), it ensures that genes with little sequence identity or context correlation to any network location have a low probability of being assigned to an EC number. Importantly, we empirically established the absence of ergodicity problems in Gibbs sampling of microbial genomes. In other words, the annotation probabilities converged to essentially the same values for chains started from different random assignments; after about 20,000 iterations the maximum probability difference across all genes was < 1%. Based on the convergent Gibbs chains we obtained the marginal probabilities for each metabolic assignment, consistent with the global fitness function.

4.2.2. *Optimization of the fitness function parameters*

The GLOBUS fitness function contains several important adjustable parameters $b_{s,}$ that represent relative weights of several sequence and context correlations. The values of these parameters significantly affect the sampling and the resulting gene annotation probabilities. To learn the parameters we applied a maximal likelihood approach using a well-annotated metabolic model of *S. cerevisiae* (iLL672 [154]). Specifically, following the approach commonly used in MRF [150], we optimized the fitness function parameters to maximally increase the product of the probabilities for correct gene assignments in the yeast network. Multiple simulated annealing [155] runs were used to the search the parameter space for maximal likelihood values. Importantly, in searching for the parameters over-fitting was not an issue as many hundreds of known metabolic annotations (485 yeast genes with EC numbers in the iLL672 model) dominate the number of optimized parameters (7 parameters in total). As a result of the maximum likelihood optimization, the yeast genes in their correct network locations had a geometric mean probability of 0.617, and an overall prediction accuracy of 80.5%, i.e. the overlap with the iLL672 model when genes were assigned to their most probable locations. Using more recent metabolic models of *S. cerevisiae* (iMM904 [156]) or *B. subtilis* (iBsu1103 [157]) for optimization resulted in similar parameter values and similar GLOBUS probabilities (**Supplementary fig. 4.2**). Thus, we used the parameters optimized with the iLL672 model for GLOBUS metabolic annotations in other species.

4.2.3. *GLOBUS precision-recall performance*

To understand the utility of GLOBUS for metabolic network annotations we applied it to the genomes of a gram-positive model bacterium, *B. subtilis*, and a medically important bacterium, *S. aureus*. The genomes of these bacteria contain 1244 (*B. subtilis*) and 854 (*S. aureus*) genes with at least remote sequence identity to known enzymes in other species. Several curated metabolic models are also available for these species: iYO844 [158] and iBsu1103 [157] for *B. subtilis* and iSB619 [159] for *S. aureus.* The parameters optimized using the yeast model (see above) were used in Gibbs sampling of all possible metabolic assignments in the two bacteria. The GLOBUS annotation probabilities were generated and precision-recall curves calculated (**Fig. 4.2a**) based on comparison with the corresponding curated models. For comparison we also show in the figure the precision-recall curves calculated based only on sequence identity to enzymes in other species; similar results were obtained using either BLAST or PSI-BLAST [18] (**Supplementary fig. 4.3**). The precision-recall calculations demonstrate that GLOBUS significantly outperforms homology in the areas of high recall and high precision.

Further analysis (**Fig. 4.2b,c**) demonstrates that the main source of the superior GLOBUS performance lies in more accurate annotations of genes with low sequence identity to known enzymes. In **Figure 4.2b** we show the recall (at 70% precision) for gene annotations in *B. subtilis* and *S. aureus* as a function of sequence identity to known enzymes. GLOBUS recovers significantly more correct assignments compared to homology (10, $P < 4x10-4$ for *B. subtilis* and 14%, $P < 5x10-5$ for *S. aureus* using χ^2 test), especially for cases with less than 40% sequence identity. In Figure 2c we show that at the same level of recall (90%) GLOBUS achieves significantly higher precision (9-11% more). The difference in precision is again highest for genes with low sequence identity to known enzymes (**Supplementary fig. 4.4**).

To investigate the contribution of individual context correlations to the GLOBUS performance, we optimized the coefficients of the fitness function without each context descriptor. We then compared the precision and recall values for predictions using all context correlations and predictions obtained without individual correlations (see **Supplementary fig. 4.5**). This analysis showed that all correlations contribute to the method accuracy and that – similar to the complete fitness function - the effects of the individual context correlations are most apparent for cases with lower sequence identity.

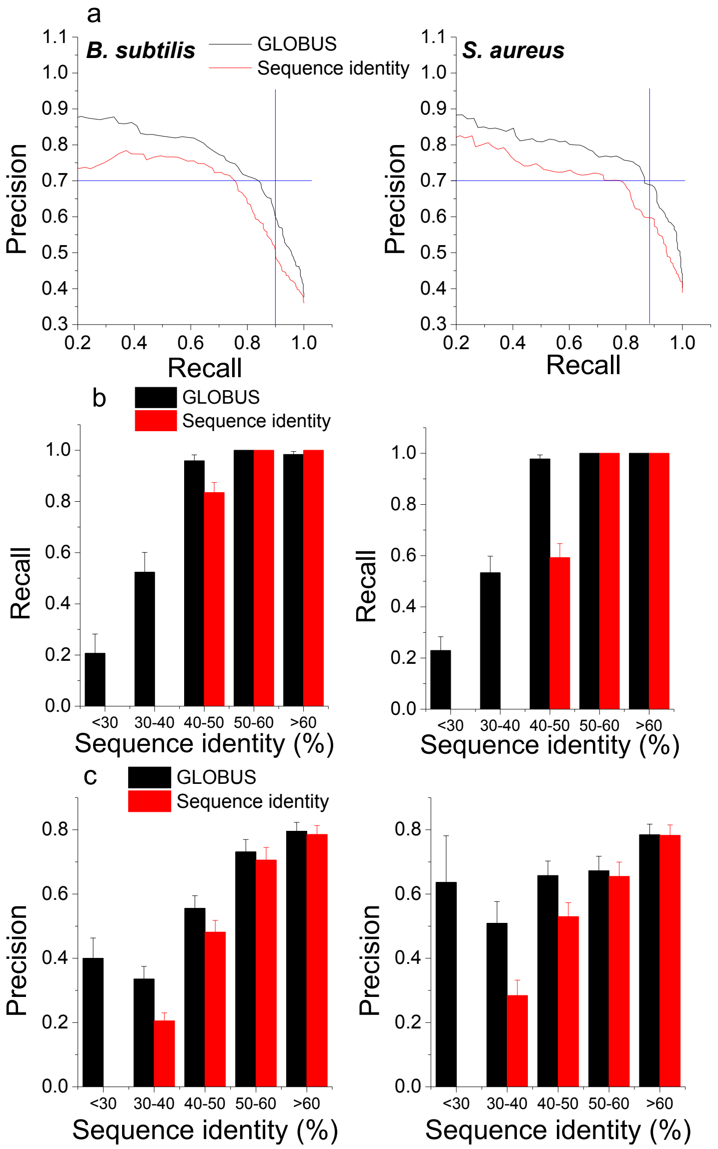


Figure 4.2. GLOBUS precision-recall performance. Using available metabolic models (iBsu1103 [157] for *B. subtilis* and iSB619 [159] for *S. aureus*) we compared predictions by GLOBUS to predictions made using sequence homology; predictions for *B. subtilis* are on the left, and predictions for *S. aureus* are on the right. **(a)** Precision–recall curves for GLOBUS (black lines) were calculated by ranking genes using assignment probabilities. Precision-recall curves for homology (red) were calculated by ranking genes using sequence identity. **(b)** Recall of known metabolic genes (at 70% precision) as a function of sequence identity to the closest enzymes from other species with the annotated functions. GLOBUS recovers significantly more enzymes with remote sequence identity (<40%). **(c)** Prediction precision (at 90% recall) for known metabolic genes as a function of sequence identity to the closest enzyme from other species with the annotated functions. In the figure error bars represent the SEM.

We investigated the potential utility of GLOBUS for refining existing metabolic reconstructions by comparing two curated models of *B. subtilis* [157, 158] (older iYO844, newer iBsu1103) and two models of *S. cerevisiae* [154, 156] (older iLL672, newer iMM904). Specifically, we considered all annotations with non-zero GLOBUS probabilities that were not included in the older metabolic models. We then subdivided these non-zero GLOBUS annotations into those that were included in the newer models for each species and those that were not included in the newer models. This analysis showed (see **Supplementary fig. 4.6**) that for both species, and across different sequence identity bins, higher GLOBUS probabilities corresponded to higher likelihoods of being included in the newer metabolic models.

4.2.4. *Specific metabolic predictions and biochemical validation in* B. subtilis

GLOBUS results indicate that in many cases context correlations provide crucial functional evidence determining correct annotations, especially when sequence identity is small. One example is the *B. subtilis* gene *hemD*, known to be responsible for the uroporphyrinogen-III synthase activity [160] (EC 4.2.1.75). The sequence identity of *hemD* to the closest Swiss-Prot sequence performing its correct function is only ~24%; however, GLOBUS assigned a high probability (P=0.86) to the correct EC number because of the excellent context associations with its neighboring enzymes at this location: the gene clustering Z-score (defined as the number of standard deviations from the mean based on all gene-gene context scores, see Methods) is 21.2, the co-expression Z-score is 5.64. Context correlations are also helpful in selecting between potential functions with comparable sequence identity. For instance, the *B. subtilis* 8-amino-7-oxononanoate synthase *bioF* [161] has ~39% sequence identity to both its correct function (EC 2.3.1.47) and to glycine C-acetyltransferase (EC 2.3.1.29). GLOBUS selected the correct assignment (P=0.64 vs. 0.02) despite the equivalent sequence identity due to high clustering and co-expression Z-scores (16.6 and 4.3, respectively) in the correct location compared to the alternative location (1.1 and 2.4).

In **Table 4.1** (*B. subtilis*) and **Table 4.2** (*S. aureus*) we list GLOBUS predictions without experimental validation that have high annotation probabilities despite low sequence identity to enzymes responsible for corresponding functions in other species. The annotations in the tables are ordered by averaging the prediction ranks sorted by decreasing annotation probability and the prediction ranks sorted by decreasing sequence identity distance to known enzymes. For each prediction in the table we also show the average Z-score for the three context correlations in the corresponding network location.

From the predictions listed in **Table 1** we selected the genes *spsI*, *spsJ*, and *ykgB* for experimental validation. The first two genes were selected because they were predicted to catalyze the first two steps in a rhamnose biosynthesis pathway (**Supplementary fig. 4.7**); the other two genes from the pathway (*spsK* and *spsL*, in **Table 1**) were also predicted by GLOBUS. Rhamnose is a main sugar component of the *B. subtilis* exosporium [162]. The *sps* genes are transcribed from a σ^K-controlled promoter [163] at late stages of *B. subtilis* sporulation when the outer components of the spore coat are being assembled. The gene *ykgB* was selected because GLOBUS predicted (with probability P=0.51) that this gene catalyzes the long elusive 6-phosphogluconolactonase activity of the *B. subtilis* pentose phosphate (PP) pathway. Importantly, despite a central role of PP pathway in the *B. subtilis* metabolism, this enzymatic activity remains without available experimental validation.

Table 4.1. Prediction of gene function in *B. subtilis*. In the table we show predictions without experimental validation that have GLOBUS-assigned probabilities above 0.5 and protein sequence identity to known enzymes below 50%. Shaded rows represent activities validated in this study. The annotations in the table are ordered by averaging the prediction ranks sorted by decreasing annotation probability and the prediction ranks sorted by decreasing sequence identity distance to known enzymes. The last column shows the average Z-score of phylogenetic correlations, gene clustering and gene co-expression when all sequences are assigned to their most probable locations. The Z-score for each type of data was calculated using the maximum context correlation between a gene and its immediate network neighbors (see Methods).

Gene	EC number	Enzyme name	Probability	Identity (%)	Average Context Z-score
murF	6.3.2.10	UDP-N-acetylmuramoyl-tripeptide-D-alanyl-D-alanine	0.98	32.8	9.0
spsL	5.1.3.13	dTDP-4-dehydrorhamnose-3,5-epimerase	0.95	33.1	8.4
ycgM	1.5.99.8	proline dehydrogenase	0.76	25.6	3.6
yfnG	4.2.1.45	CDP-glucose-4,6-dehydratase	0.76	27.5	11.0
birA	6.3.4.15	biotin-[acetyl-CoA-carboxylase] ligase	0.77	31.7	2.3
gcvPB	1.4.4.2	glycine dehydrogenase (decarboxylating)	0.97	41.5	12.3
yloI	4.1.1.36	phosphopantothenoylcysteine decarboxylase	0.99	44.5	2.6
fruK	2.7.1.56	1-phosphofructokinase	0.88	40.4	10.9
spsK	1.1.1.133	dTDP-4-dehydrorhamnose reductase	0.87	39.6	8.4
murB	1.1.1.158	UDP-N-acetylmuramate dehydrogenase	0.97	43	5.2
folK	2.7.6.3	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine	0.99	45.3	8.0
sul	2.5.1.15	dihydropteroate synthase	0.99	47	8.2
yitJ	2.1.1.13	methionine synthase	0.54	30.6	2.1
ybbF	2.7.1.69	protein-Npi-phosphohistidine-sugar phosphotransferase	0.85	40.5	11.3
yloI	6.3.2.5	phosphopantothenate-cysteine ligase	0.97	44.5	2.9
ykgB	3.1.1.31	6-phosphogluconolactonase	0.51	30.4	2.6
pheA	4.2.1.51	prephenate dehydratase	0.69	36.1	6.7
purK	4.1.1.21	phosphoribosylaminoimidazole carboxylase	0.89	43.5	13.3
spsI	2.7.7.24	glucose-1-phosphate thymidylyltransferase	0.93	44.4	11.6
ysnA	3.6.1.15	nucleoside-triphosphatase	0.56	33.3	7.7
ywbC	4.4.1.5	lactoylglutathione lyase	0.6	35.2	3.6
pucE	1.2.3.14	abscisic-aldehyde oxidase	0.62	35.8	1.0
ydhR	2.7.1.4	fructokinase	0.77	41.5	5.3
yfnH	2.7.7.33	glucose-1-phosphate cytidylyltransferase	0.88	43.2	11.0
ybbD	3.2.1.52	beta-N-acetylhexosaminidase	0.52	33.1	3.1
yngE	6.4.1.4	methylcrotonoyl-CoA carboxylase	0.64	36.2	8.6
kbl	2.3.1.29	glycine C-acetyltransferase	0.97	49	9.4
spsJ	4.2.1.46	dTDP-glucose-4,6-dehydratase	0.97	48	12.0
tenI	2.5.1.3	thiamine-phosphate diphosphorylase	0.7	40.6	6.6
pabB	4.1.3.27	anthranilate synthase	0.74	42.8	8.6

Table 4.2 Prediction of gene function in *S. aureus*. In the table we show predictions without experimental validation that have GLOBUS-assigned probabilities above 0.5 and protein sequence identity to known enzymes below 50%. The annotations in the table are ordered by averaging the prediction ranks sorted by decreasing annotation probability and the prediction ranks sorted by decreasing identity distance to known enzymes. The last column shows the average Z-score of phylogenetic correlations as described in Table 4.1.

Gene	EC number	Enzyme name	Probability	Identity (%)	Average Context Z-score
bioD	6.3.3.3	dethiobiotin synthase	0.99	31.2	7.9
hisG	2.4.2.17	ATP phosphoribosyltransferase	0.99	39.6	6.3
murE	6.3.2.13	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate-2,6-	0.96	39	7.0
thrB	2.7.1.39	homoserine kinase	1.00	42.4	7.0
mvaA	1.1.1.34	hydroxymethylglutaryl-CoA reductase (NADPH)	0.95	40.1	5.9
hemD	4.2.1.75	uroporphyrinogen-III synthase	0.76	27.4	6.4
SA2374	1.3.3.1	dihydroorotate oxidase	0.84	37.8	2.2
murF	6.3.2.10	UDP-N-acetylmuramoyl-tripeptide-D-alanyl-D-alanine	1.00	46	7.2
mvaK1	2.7.1.36	mevalonate kinase	0.73	35.1	6.7
ribB	2.5.1.9	riboflavin synthase	0.91	43.3	8.3
ribC	2.7.1.26	riboflavin kinase	0.96	45.5	2.3
lysC	2.7.2.4	aspartate kinase	0.86	41.6	5.4
scrB	3.2.1.26	beta-fructofuranosidase	0.82	40.5	5.9
folA	1.5.1.3	dihydrofolate reductase	0.89	42.8	9.6
SA1288	6.3.4.15	biotin-[acetyl-CoA-carboxylase] ligase	0.56	33.1	2.0
aroK	2.7.1.71	shikimate kinase	0.66	34.9	2.0
coaW	2.7.1.33	pantothenate kinase	0.71	36.6	2.1
nagA	3.5.1.25	N-acetylglucosamine-6-phosphate deacetylase	0.95	45.5	8.1
ansA	3.5.1.1	asparaginase	0.66	36	2.2
SA2317	4.3.1.17	L-Serine ammonia-lyase	0.90	43.9	2.8
bioA	2.6.1.62	adenosylmethionine-8-amino-7-oxononanoate	0.97	48.2	9.3
asd	1.2.1.11	aspartate-semialdehyde dehydrogenase	0.98	48.9	5.8
gcvPB	1.4.4.2	glycine dehydrogenase (decarboxylating)	0.81	42.3	9.2
thiD	2.7.4.7	phosphomethylpyrimidine kinase	0.89	44	8.9
hemY	1.3.3.4	protoporphyrinogen oxidase	0.94	47	4.1
trpG	4.1.3.27	anthranilate synthase	0.68	40.6	6.6
SA2006	4.1.1.5	acetolactate decarboxylase	0.90	46.6	4.9
alr1	5.1.1.1	alanine racemase	0.82	43.6	3.3
SA0511	1.1.1.103	L-threonine 3-dehydrogenase	0.78	43.5	2.2
SA2318	4.3.1.17	L-Serine ammonia-lyase	0.83	45.6	9.3

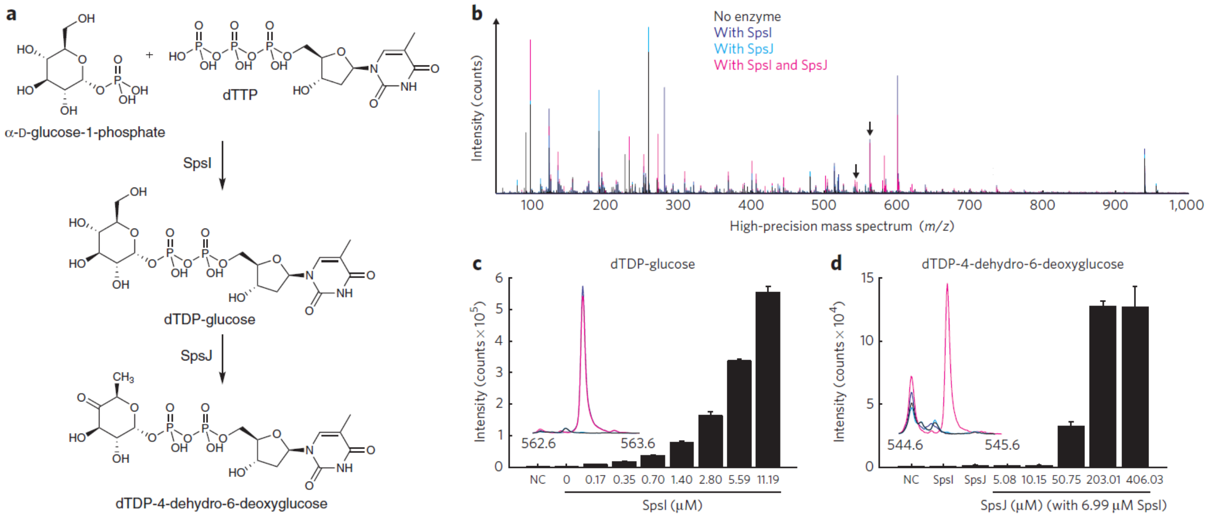


Figure 4.3. In vitro biochemical assays for characterizing activities of SpsI and SpsJ using high-precision mass spectrometry. (a) Reaction diagram. **(b)** Peak plots show intensities with masses corresponding to the products dTDP-glucose and dTDP-4-dehydro-6-deoxy-glucose of the reactions catalyzed by SpsI and SpsJ **(**black arrows, detailed in **c).** Observed masses deviate by less than 0.001 atomic mass units (amu) from the corresponding reference masses. Spectra were recorded from two independent assays. **(c,d)** Bar plots show dependency of dTDP-glucose **(c)** and dTDP-4-dehydro-6-deoxy-glucose **(d)** accumulation on protein concentration of SpsI and SpsJ, respectively. As negative control (NC), the protein free filtrate of 6.99 µM spsI or 203.01 µM SpsJ solution was used. Error bars represent standard deviations from two independent assays.

The three proteins selected for experimental validation were over-expressed in *E. coli* and purified by His-Tag affinity and anion exchange chromatography. The correct identity of the purified proteins was confirmed by in-gel tryptic digestion and subsequent peptide analysis using mass spectrometry. In vitro enzymatic assays for SpsI and SpsJ were performed using a published method [164]. Predicted SpsI substrates (dTTP and α-D-glucose-1-phosphate; **Fig. 3.1a**) were observed in negative ionization mode high-precision mass-spectra profiles at 259.022 m/z and 480.981 m/z (M-H$^+$) respectively. Intensities of both dTTP and α-D-glucose-1-phosphate decreased only when SpsI was present in the assays, indicating that the enzyme uses these compounds as substrates (**Supplementary fig. 4.8**). In addition, the predicted reaction product (dTDP-glucose) accumulated at 563.068 m/z (M-H$^+$) only in the presence of SpsI (**Fig. 4.3b,c**). The product of SpsJ (dTDP-4-dehydro-6-deoxy-glucose) was observed at 545.058 m/z (M-H$^+$) only in the presence of both SpsI and SpsJ (**Fig. 4.3b,d**), suggesting that SpsJ indeed converts dTDP-glucose into dTDP-4-dehydro-6-deoxy-glucose (**Fig. 4.3a**). Product accumulation as well as substrate consumption showed a clear dependence on the protein concentrations within a wide range around the estimated *in vivo* concentration of glucose-1-phosphate thymidylyltransferase (~1 µM for RfbA in *Escherichia coli* [165]).

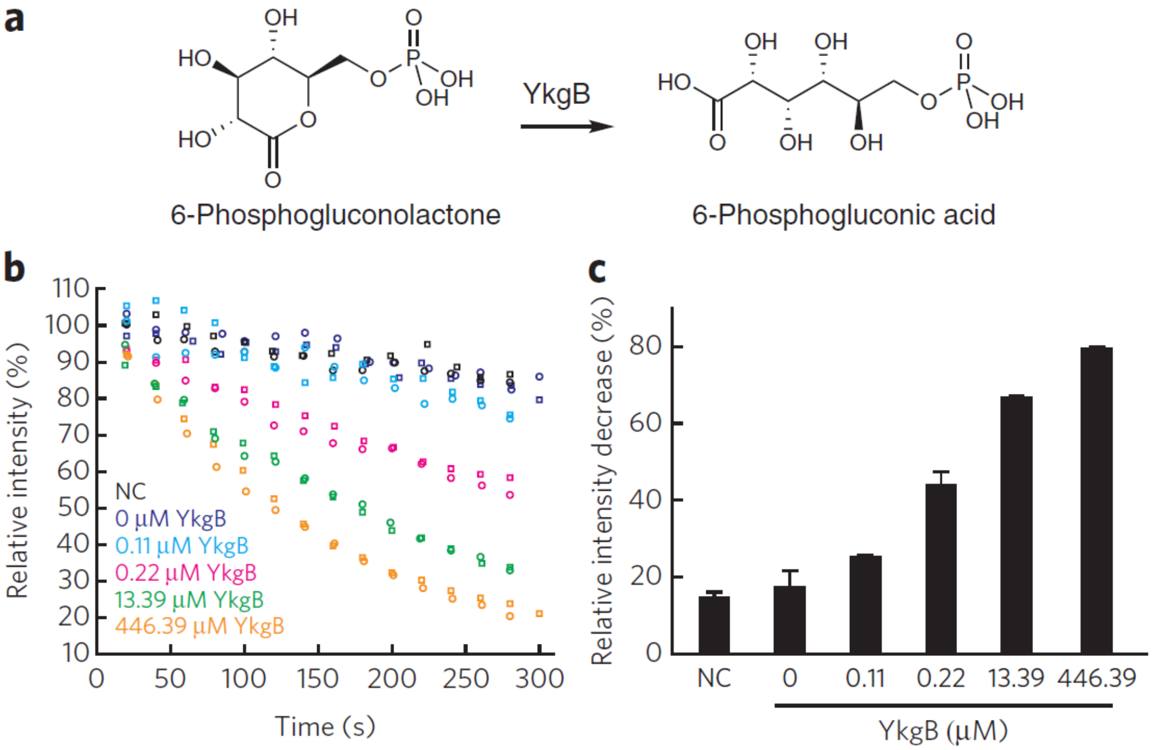


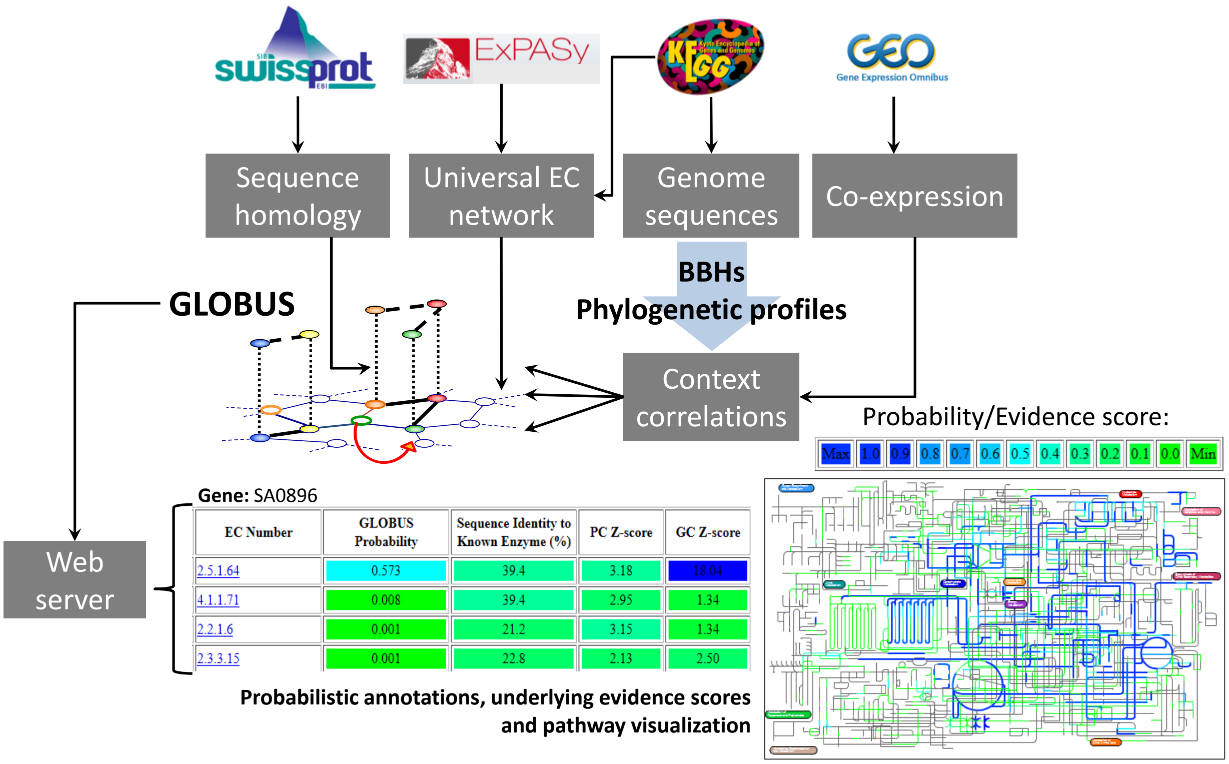
Figure 4.4 In vitro biochemical assays for characterizing 6-phospho-gluconolactonase activity of YkgB. (**a**) Reaction diagram for 6-phosphogluconolactonase. (**b**) Time courses of lactone degradation at different YkgB concentrations were recorded online by direct flow injection analysis. Different symbols represent replicate assays. **(c)** Bar plot shows relative intensity increase comparing final and initial intensities as a function of YkgB concentration. As negative control (NC), the protein free filtrate of 223.2 µM YkgB solution was used. Error bars represent standard deviations from two independent assays.

Similarly to SpsI and SpsJ, the YkgB activity (**Fig. 4.4a**) was followed by observing the 6-phospho-gluconolactone degradation with online flow injection into a high-precision mass-spectrometer operating in the negative ionization mode. The intensity at the mass of 257.007 m/z (M-H^+), corresponding to 6-phospho-gluconolactone, decreased with rates faster than the rate of spontaneous background hydrolysis only when YkgB was present in the assays (**Fig. 4.4b**). The 6-phospho-gluconolactone degradation rate also exhibited a clear dependence on the protein concentration (**Fig. 4c**) within a wide range around the estimated *in vivo* 6-phosphogluconolactonase concentration (~1.5 μM for YbhE in *Escherichia coli* [165]). Similarly, the production rate of 6-phosphogluconic acid was significantly higher than the background when YkgB was present in the assays (**Supplementary fig. 4.9**). Notably, available expression and proteomic data show that the *ykgB* gene is transcribed during several environmental conditions [166, 167], such as heat and phenol stress. This suggests that YkgB - similar to lactonases in other species [168] - is likely to play a role in removing toxic byproducts of the PP pathway.

4.2.5. *GLOBUS automation and probabilistic predictions for multiple species*

Our probabilistic annotation method can be applied to any sequenced microbial genome. Other than gene co-expression, which may be hard to come by for species with no or too few expression studies, phylogenetic correlations and gene chromosomal clustering, as well as the remaining functional descriptors can be easily calculated based on pre-defined open reading frames (see Methods). Using high-quality annotations from Swiss-Prot [17] and enzyme definitions from ExPASy [169] and KEGG [29] it is possible to automatically update the generic EC network and sequence database used to detect homology. Furthermore, data on gene order present in KEGG, as well as simple protocols of sequence alignment can be applied on a large scale to produce the information required to solve the energy function equation during Gibbs sampling (**Fig. 4.5**). To demonstrate the scalability of GLOBUS, we have applied it to 45 diverse prokaryotic species producing thousands of probabilistic gene annotations within a couple of days. Based on these tests, using GLOBUS to annotate all (~4,000) currently sequenced genomes could be achieved within 3-6 months.

We have set up a prototype database at http://vitkuplab.c2b2.columbia.edu/globus/ with probabilistic predictions for 10 species of medical importance. As illustrated in **Figure 4.5**, these predictions are based on our automated annotation pipeline and are linked to the underlying functional descriptor scores that resulted in the observed probabilities. These scores are important for manually reviewing annotations, for example, during the reconstruction of genome-scale metabolic models. Additionally, all predictions can be downloaded in bulk format to be used in downstream computational methods (see for example **Section 4.4**), and links to external resources such as BRENDA [136] are provided. The quantitative results are also linked to visualization tools [170], allowing the inspection/comparison of the annotated enzymatic activities and their probabilities in the context of KEGG metabolic pathways.



Gene: SA0896

EC Number	GLOBUS Probability	Sequence Identity to Known Enzyme (%)	PC Z-score	GC Z-score
2.5.1.64	0.573	39.4	3.18	18.04
4.1.1.71	0.008	39.4	2.95	1.34
2.2.1.6	0.001	21.2	3.15	1.34
2.3.3.15	0.001	22.8	2.13	2.50

Figure 4.5. GLOBUS database implementation. Probabilistic annotations can be obtained for any sequenced microbial genome based on public and subscription based (optional) data repositories. A prototype database (http://vitkuplab.c2b2.columbia.edu/globus/) was built to display GLOBUS results and the underlying functional descriptor scores. Links to BRENDA [136] and iPATH [170] for visualization are also provided.

4.3. Discussion

Owing to the inherent uncertainty of computational annotations, predictions of biochemical function need to be treated probabilistically. Currently, most publicly available biochemical databases do not provide quantitative probabilities or confidence measures for existing annotations. This makes it hard for the users of these valuable resources to distinguish between confident assignments and mere guesses. As the application and impact of genome-scale metabolic networks rapidly expands [76], a probabilistic treatment of annotations is essential. The GLOBUS approach, which is based on statistical sampling of possible biochemical assignments, provides a principled framework for such global probabilistic annotations. The method assigns annotation probabilities to each gene, as well as suggests likely alternative functions.

We demonstrate that context correlations can significantly improve the accuracy of biochemical predictions, especially when annotations are based on distant sequence identity. Over half of potential metabolic genes, even in such well-studied model organisms as *S. cerevisiae* and *B. subtilis*, have remote sequence identity (<40%) to known enzymes (**Supplementary fig. 4.4**). Application of GLOBUS to less-studied organisms should be straightforward, as context-based correlations, with the exception of gene co-expression, can be calculated directly from genomic sequences; the decrease in the overall accuracy without gene co-expression is relatively small (<1%). The accuracy of other context correlations should only improve with the rapid growth of fully sequenced genomes.

Probabilistic predictions generated by GLOBUS can be directly used to annotate sequences and genomes. GLOBUS annotations can be also used by various gap identification and gap filling approaches [31, 33, 52, 54, 146] to produce simulation-ready flux balanced networks [76]. In addition, recent advances in metabolomics, proteomics, and fluxomics offer complementary opportunities to expand and refine biochemical annotations and network reconstructions [171, 172]. The flexibility of the GLOBUS framework makes it easy to integrate metabolomics and proteomics data. For example, as genes are moved through the network to sample possible assignments, available data for corresponding proteins and metabolites can be included in the global fitness function. Additional functional descriptors, for example based on protein structure and information about protein localization, can be also considered in the framework. Such probabilistic integration of diverse biochemical data will be crucial for exploiting the ongoing avalanche of genomic sequencing.

4.4. Appendix: Using system-level properties as evidence for gene annotations

As described in **Chapter 2**, one of the main applications of biochemical annotations is the reconstruction of stoichiometric models suited for constraints-based analysis. The reconstruction process involves the mapping of annotated enzymatic activities to mass-balanced reactions that, represented as a stoichiometric matrix, enable the search for steady state flux distributions [51]. Typically, a pre-defined set of biomass precursors plays an important role in the reconstruction process: based on literature reports or computational predictions, reactions are included in the models such that all biomass precursors can be synthesized at steady state [2]. Given that all species possess the ability to obtain basic bio-molecules such as proteins, nucleic acids, lipids, and co-factors, we rationalized that prior knowledge of the reactions needed for these biosynthetic processes could be considered in GLOBUS to further improve functional assignments (**Fig. 4.6**). In particular, if a reaction is likely to be present in a given metabolic network, a global assignment of genes to EC numbers that has genes assigned to said reaction is expected to be more likely than an assignment with no genes associated with it.

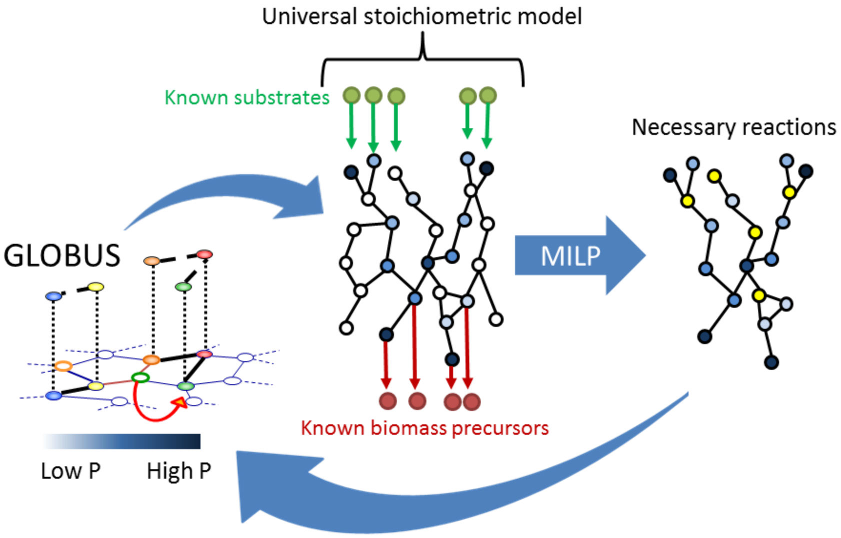


Figure 4.6. Identification of necessary reactions based on GLOBUS and flux analysis. System-level properties of metabolic networks, such as the ability to synthesize biomass precursors, can be used to find reactions that are likely to be in a metabolic network based on an initial annotation. The identified reactions can then be used to improve GLOBUS predictions by narrowing the set of possible locations with a high-probability. These reactions can be obtained from a universal stoichiometric network using an optimization approach such as mixed integer linear programming (MILP).

In order to test this approach, a new term representing the presence of a reaction in the metabolic network was added to the GLOBUS energy function. Similar to previous sections, a weight was learned for this new variable using simulated annealing and the yeast iLL672 model. Every node in the EC network was labeled according to its presence (1) or absence (0) in the iLL672 model, and the energy function was defined in such way that a higher probability (lower energy) was obtained when genes were assigned to reactions marked as present. Using the parameters obtained after training on the yeast model, GLOBUS was ran on the *B. subtilis* genome using the EC numbers present in the iBsu1003 model [157] as input. As shown in **Figure 4.7a**, there was a significant boost in precision and recall when prior knowledge of reactions was considered. Moreover, a substantial improvement was also observed when as many as 30% known EC numbers were ignored or 30% random functions were considered (**Fig. 4.7a**, orange and green lines).

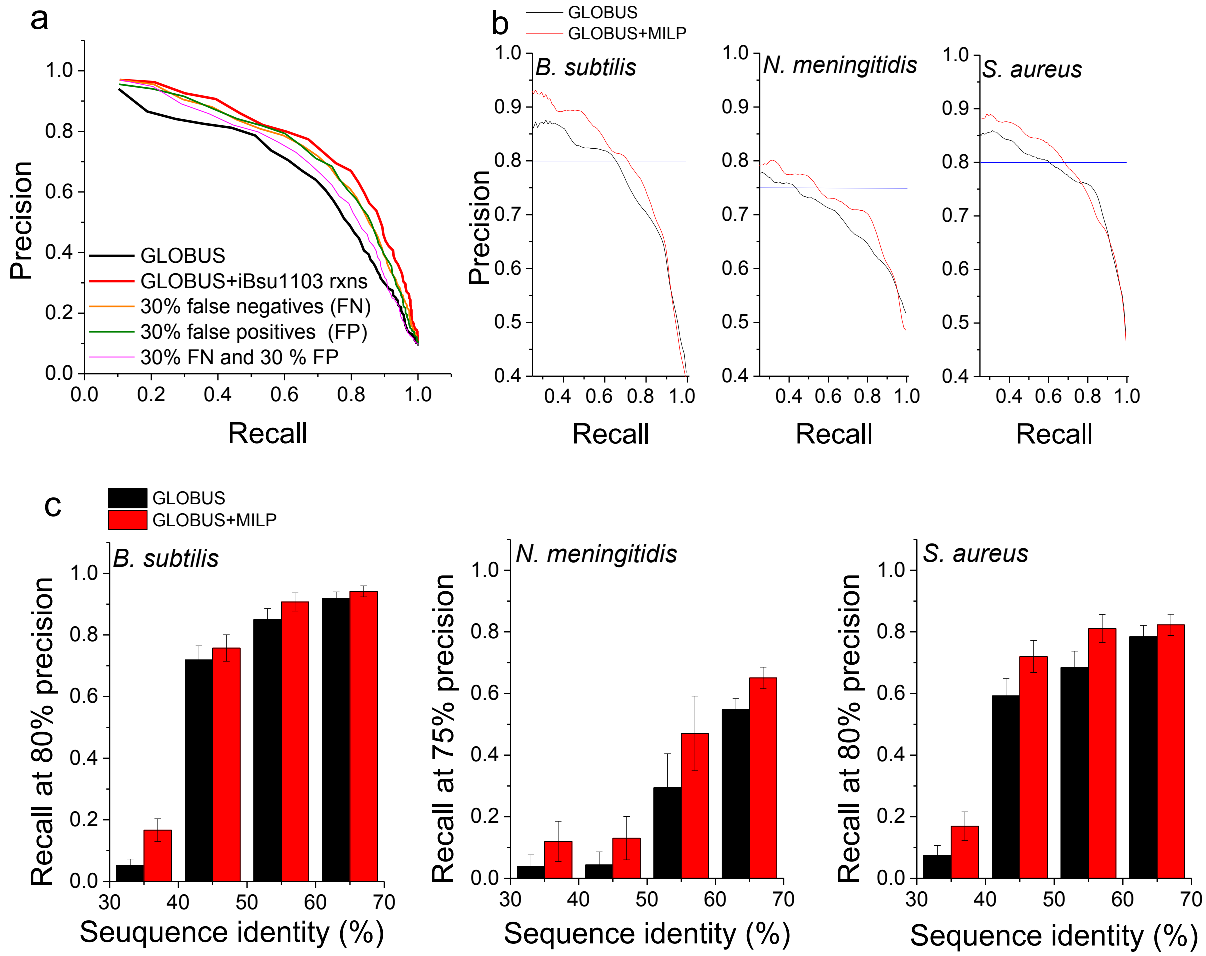


Figure 4.7. Improving GLOBUS performance through system-level information. (**a**) Precision-recall curves for GLOBUS (red), GLOBUS considering known reactions in the *B. subtilis* metabolic network (black), and GLOBUS considering known reactions with different levels of false positives and false negatives. (**b**) Precision-recall curves for GLOBUS (black) and GLOBUS using reactions required for the synthesis of biomass precursors (red) identified using MILP in three different genomes (*B. subtilis*, *N. meningitidis* and *S. aureus*). (**c**) Recall at a given level of precision (blue lines in panel b) for different bins of sequence identity to known enzymes across three different species. Error bars represent the SEM.

Because it is unlikely that one would know all reactions present in a given species beforehand, a previously described optimization procedure [52] was used to identify reactions likely to be present in a query metabolic network. For this, 10,625 mass-balanced reactions from the ModelSEED were used to build a universal stoichiometric model, and EC numbers in the generic EC network were mapped to those reactions. For a given species, GLOBUS probabilities were transferred to reactions in this stoichiometric network and mixed integer linear programming (MILP) was used to find the least number of low or zero-probability reactions that would allow the synthesis of all biomass components defined for such species (see Methods, **Section 4.5.7**). These reactions were then used as an additional input for GLOBUS using the newly defined energy function and the parameters obtained after training with known yeast reactions (**Fig. 4.6**). As shown in **Figure 4.7b** for three different species (*B. subtilis*, *Neisseria meningitidis* and *S. aureus*), considering the reactions identified this way produces a substantial (~10%) improvement in precision and recall of metabolic annotations. Interestingly, looking at recall values for genes with different levels of sequence identity to known enzymes (**Fig. 4.7c**), there is a similar improvement at every sequence identity bin using the new method. This suggests that GLOBUS can effectively combine local context information, which is most useful in cases of remote sequence identity (**Fig. 4.2**), and the new global context term related to the ability of genome-scale metabolic networks to synthesize essential and organism-specific biomass components.

While biomass synthesis (growth) is a phenotypic characteristic shared by all organisms, the described approach could be used to incorporate more specific phenotypic properties such as the ability to utilize various nutrients (measured by phenotypic microarrays [90]), or the ability to produce specific metabolites detected by untargeted mass-spectrometry. Both of these tasks can be achieved through a reformulation of the optimization problem described. On the other hand, considering GLOBUS probabilities in the context of a universal mass-balanced stoichiometric network provides useful information for deciding –through manual inspection, or optimization procedures– what reactions to include or leave out in the reconstruction of genome-scale metabolic models (see **Supplementary fig. 4.6**). In summary, integrative probabilistic methods such as GLOBUS can be used to integrate into a single process the annotation and *in silico* systems analysis of sequenced genomes (**Fig 2.1**).

4.5. Methods

4.5.1. *Construction of the generic EC network*

In the construction of the EC (Enzyme Commission) network we considered 3284 EC numbers (http://www.chem.qmul.ac.uk/iubmb/enzyme/) responsible for biochemical activities involving small compounds as substrates and products; activities such as “RNA polymerase” or “protein kinase” were excluded. In the global EC network, nodes represent EC numbers and edges represent connections established by metabolites shared by reactions. Following a common procedure [38], linkages through the top 30 most highly connected metabolites and cofactors were not considered (**Supplementary table 4.1**).

4.5.2. *Identification of potential metabolic genes and potential metabolic functions*

The program BLAST [18] (with E-value cutoff of $5*10^{-2}$) was used for homology searches against enzymes in Swiss-Prot [149], excluding sequences that were: 1.) from genomes of closely related species (species in the same taxonomic genus) or 2.) likely annotated based exclusively on computational methods, i.e., annotations with keywords *probable*, *like*, *by similarity*, *hypothetical*, or *putative*. Although many remaining annotations in Swiss-Prot are also derived using computational methods, they are usually curated, ensuring that the misannotation rate in this database is relatively low [24, 36].

To account for multi-functional enzymes, when non-overlapping regions of a query gene could be mapped to different enzymatic functions - indicating domains responsible for distinct metabolic activities - the mapped regions of the query gene were allowed to be assigned independently to different network locations.

4.5.3. *The functional descriptors considered in the GLOBUS fitness function*

The fitness (energy) function over all metabolic genes, $E(g_1, g_2, \ldots., g_n)$, was defined to reflect the hypotheses that a particular global assignment of genes into their network locations will be more probable if genes have significant homologies to the assigned locations, and also exhibit strong context correlations with their network neighbors. Accordingly, in GLOBUS calculations we used the following fitness function for genes included in the network:

$$E(g_1, g_2, \ldots, g_n) = -b_{homology} f_{homology} - b_{orthology} f_{orthology} - b_{context} f_{context} - b_{ECco-occurrence} f_{ECco-occurrence}$$

where, *b*(s) are positive coefficients representing weights of each functional feature, and *f*(s) are various functional features described below.

4.5.3.1. *Sequence homology.* f_{homology}

The term, f_{homology}, represents the descriptor of sequence homology. The higher the sequence identity between a protein and enzymes in other species known to catalyze the assigned activity, the more likely is the assignment to be correct [20, 173]. As the sequence homology descriptor we used the logarithm of the conditional probability that the gene performs the assigned function, given the highest sequence identity to a Swiss-Prot [149] protein annotated to catalyze the target activity:

$$f_{homology} = \sum_{i=1}^{n} \log P(\text{gene performs function}|\text{highest sequence identity to annotated Swiss-Prot protein})$$

The conditional probabilities were estimated using the well-curated yeast iLL672 metabolic model [154] (**Supplementary fig. 4.1d**).

4.5.3.2. *Orthology.* $f_{\mathrm{orthology}}$

An additional binary descriptor related to sequence homology was the possible gene orthology to a gene from another species annotated with the target activity. The orthology descriptor was based on bi-directional best hits by SSEARCH [174]; in these calculations we used the bi-directional best hits in the KEGG SSDB database [131] (http://www.genome.jp/kegg/ssdb/). For each gene, the orthology term was either 1, if at least one possible ortholog was annotated in Swiss-Prot to perform the target activity, or 0, if no orthologs with the target activity could be identified. Again we excluded annotations based exclusively on computational methods, and treated separately non-overlapping regions with homology to different activities (see above).

4.5.3.3. *Gene-gene context correlations.* f_{context}

Gene pairs that share similar biological functions tend to be either present or absent together in genomes of sequenced species (phylogenetic correlation), tend to be co-localized on chromosomes across multiple genomes (gene chromosomal clustering), and tend to be co-regulated. These context-based correlations were initially developed to infer gene functions and provide complementary information to sequence homology data [143, 175]. Multiple studies have also demonstrated that genes located close to each other in a metabolic network tend to have significantly stronger context associations [38, 176]. Previously, we and others used context associations in combination with local structure of the metabolic network to identify genes responsible for orphan metabolic activities [31, 33, 146, 177].

In GLOBUS we used the context correlations by first transforming them into Z-scores [178] using the distribution of correlations between all pairs of candidate metabolic genes, and then estimating the conditional probability that two genes are direct network neighbors given their context association Z-score. The conditional probabilities were derived based on the iLL672 yeast metabolic model (**Supplementary fig. 1a-c**). In the GLOBUS fitness function for each assigned gene we considered the maximum log probability among all network neighbors of the gene:

$$f_{context} = \sum_{i=1}^{n} max(\log P(\text{two genes are network neighbors}|\text{context correlation Z-score between the genes}))$$

4.5.3.3.1. Phylogenetic correlation

Phylogenetic correlation [140, 179] measures the co-occurrence (co-presence) of homologues for a pair of genes across genomes. Phylogenetic profiles were constructed using protein BLAST searches against a collection of 70 diverged genomes [33]. We used the binary phylogenetic profiles, i.e. 70-dimensional binary vectors representing the presence or absence of homologues in the target genomes. Pearson's correlation between the profile vectors was calculated using the following equation:

$$r = \frac{Nz - xy}{\sqrt{(Nx - x^2)(Ny - y^2)}}$$

, where *N* is the total number of target genomes. For genes X and Y, *x* is the number of genomes in which any homologue of X is present, *y* is the number of genomes in which any homologue of Y is present, and *z* in the number of genomes in which homologues of both X and Y are present.

4.5.3.3.2. Gene chromosomal clustering

For a pair of genes, chromosomal gene clustering [34, 142, 180] measures the degree of co-localization of their orthologues across a set of genomes. We considered gene order statistics instead of the exact nucleotide positions of genes, i.e. we defined a gene order distance d(X,Y) as the minimum number of genes separating genes X and Y. Under the null hypothesis that genes are distributed randomly within a genome, $P(d_\gamma(X, Y))$ is the probability of observing gene order distance equal or less than $d_\gamma(X, Y)$ between a pair of genes X and Y in a genome γ. $P(d_\gamma(X, Y))$ can be calculated directly as the fraction of gene pairs in genome γ that are separated by gene order distance $d_\gamma(X, Y)$ or smaller. Assuming gene order distances are independent across a set of 108 evolutionary divergent organisms Γ, and given that X and Y are orthologues of genes A and B from the target genome, we calculated the clustering of genes A and B:

$$C_\Gamma(A, B) = -log \prod_{\gamma \in \Gamma} \left(P\left(d_\gamma(X, Y)_\gamma \right) \right)$$

For a given set of genomes, this clustering measure can be biased by the variable phylogenetic proximity between different organisms. Therefore we deliberately filtered the genome set to eliminate species closely related to the target genome using a mutual information threshold of 0.9 for ortholog occurrences [31]. Orthology mapping required for the chromosomal clustering calculations was established using best bi-directional hits in the KEGG SSDB dataset [131] (http://www.genome.jp/kegg/ssdb/).

4.5.3.3.3. Co-expression

Numerous studies [141, 181] demonstrated that genes with similar mRNA expression profiles usually have related biological functions. Descriptors of mRNA co-expression used in

GLOBUS were calculated as Spearman's rank correlation between expression profiles obtained from the Rosetta "compendium" dataset [182] for *S. cerevisiae* and the GEO database [183] for *B. subtilis and S. aureus*. In all calculations Log_{10} intensity ratio values were used.

4.5.3.4. *EC co-occurrences.* $f_{ECco\text{-}ocurrence}$

In addition to gene phylogenetic profiles, we used in GLOBUS a functional descriptor based on likely co-occurrence between different metabolic activities (EC numbers) across species. This descriptor measures the correlation between the occurrences of different metabolic activities across species without considering specific genes assigned to the activities. To calculate the correlation between different metabolic activities (EC numbers) we used a 70-dimentional binary vector for each EC number representing its presence or absence in a set of 70 genomes (see **section 4.5.3.3.1**) according to the KEGG database [131] (http://www.genome.jp/kegg). For every pair of EC numbers the Pearson's correlation between their profile vectors was calculated (see **section 4.5.3.3.1**).

In the GLOBUS fitness function for each assigned gene we considered the EC co-occurrence descriptor equal to the average correlation between the EC activity of the assigned gene and the EC activities for all its network neighbors. The most relevant information about homology usually comes from annotated enzymes with the highest sequence identity to a protein under consideration. On the other hand, the EC co-occurrence reflects common presence and absence of metabolic activities across multiple KEGG genomes. Thus, this term quantifies tendencies of closely related activities to be filled together.

4.5.4. *Calculating the marginal probability using Gibbs sampler*

The marginal probability, *P(g$_i$)*, represents the probability that a gene is responsible for a metabolic activity (EC number) consistent with all possible assignments of other genes into the network. Formally, given all parameters of the GLOBUS fitness function, b_{homology}, $b_{\text{orthology}}$, b_{context}, and $b_{\text{EC co-occurrence}}$, *P(g$_i$)* can be calculated by summation:

$$P(g_i) = \sum_{g_1} \cdots \sum_{g_{i-1}} \sum_{g_{i+1}} \cdots \sum_{g_n} \frac{1}{Z} EXP\{-E(g_1, \ldots, g_{i-1}, g_i, g_{i+1}, \ldots, g_n)\}$$

, where Z is a normalizing partition function. Suppose that there are *n* metabolic genes in the genome and each metabolic gene has *m* potential network assignments, obtaining *P(g$_i$)* then requires summing over m^n possible terms. Because a typical genome contains many hundreds to thousands of metabolic genes, this summation is computationally intractable. Nevertheless, the success of the GLOBUS approach is due to the fact that the vast majority of all possible gene assignments have very low probabilities. Consequently, it is possible to recover correct marginal probabilities for each gene using an efficient sampling of high probability configurations (assignments).

To sample probable gene assignment we applied a widely used algorithm, the Gibbs sampler [151, 184, 185]. The Gibbs sampler is a special case of Markov Chain Monte Carlo (MCMC) and the Metropolis-Hasting algorithm [152, 186]. The Gibbs sampler allows obtaining marginal probabilities using sampling based on conditional probabilities. Starting with a random initial assignment of *n* metabolic genes to a network, a Gibbs chain of *t* steps: G^1,G^2,…,G^t is obtained iteratively by selecting a random gene *i* and re-assigning to a location g_i according to the following conditional probability:

$$G_i^{k+1} \sim P(g_i \mid g_1 = G_1^k, \ldots, g_{i-1} = G_{i-1}^k, g_{i+1} = G_{i+1}^k, g_n = G_n^k)$$

Where G_i^k represents the location of gene *i* at step *k* of the Gibbs chain *G.* If at each iteration the location of every gene was recorded; it can be proven that the distribution of G_i converges to $P(g_i)$ as the number of iterations t → ∞.

The conditional probability used in the iterative sampling, $P(g_i| g_1,\ldots,g_{i-1}, g_{i+1},\ldots,g_n)$, is:

$$P(g_i \mid g_1,\ldots,g_{i-1},g_{i+1},\ldots,g_n) = \frac{P(g_1,\ldots,g_{i-1},g_i,g_{i+1},\ldots g_n)}{P(g_1,\ldots,g_{i-1},g_{i+1},\ldots,g_n)}$$

Since in each iteration the denominator of the above equation and Z (the partition function) are constant, the conditional probability can be derived from the fitness function, $E(g_1, g_2, \ldots., g_n)$:

$$\begin{aligned} P(g_i \mid g_1,\ldots,g_{i-1},g_{i+1},\ldots,g_n) &\propto P(g_1,\ldots,g_{i-1},g_i,g_{i+1},\ldots,g_n) \\ &\propto EXP\{-E(g_1,\ldots,g_{i-1},g_i,g_{i+1},\ldots,g_n)\} \end{aligned}$$

A schematic illustration of a Gibbs sampler chain generating iterative gene assignments is shown in **Supplementary figure 4.10**.

4.5.5. *Computational requirements and statistical analysis*

The calculations were performed using the 3GHz Intel Xeon processor with 256MB of RAM. GLOBUS run times depended both on the number of iterations and the number of genes for a given species. For the *S. cerevisiae*, *S. aureus*, and *B. subtilis* genomes, 10,000 iterations over all genes took about 10 minutes to complete. The run time increased linearly with the number of iterations and the number of genes. Importantly, 20,000-50,000 iterations (20-50 minutes) were required to achieve 1% convergence of annotation probabilities, i.e. a convergence where not a single annotation probability varied by >1% between independent sampling runs.

P values used to compare precision-recall performances for GLOBUS and sequence identity were calculated using the one tailed χ^2 test, n=332 to 717 annotations.

4.5.6. *Experimental validation of biochemical predictions*

Different amounts of purified SpsI or SpsJ were incubated at 37 °C in 1 mL of 10 mM potassium phosphate buffer pH 7.4, 2.5 mM $MgCl_2$, 1 mM glucose-1-phosphate, 1 mM dTTP and 1 U pyrophosphatase[164]. The enzyme reaction samples were assayed after 4 hours by flow-injection into a time of flight mass spectrometer (6520 Series QTOF, Agilent Technologies) operated in the negative ionization mode. High-precision mass spectra were recorded from 50-1000 m/z and analyzed as described previously [187]. Acquired masses were deviating less than 0.001 atomic mass units (amu) from the reference masses 259.022, 480.982, 545.058, and 563.068 for α-D-glucose-1-phosphate, dTTP, dTDP-glucose, and dTDP-4-dehydro-6-deoxy-glucose, respectively.

Purified YkgB was assayed in 1 mL 5 mM potassium phosphate buffer pH 7, 2.5 mM $MgCl_2$, and freshly prepared 6-phospho-gluconolactone. The lactone was prepared freshly from 6-phospho-gluconic acid by lyophilization, and its degradation due to the YkgB activity was followed by direct online flow-injection into a time of flight mass spectrometer as described above. Acquired masses were deviating less than 0.001 atomic mass units (amu) from the reference masses 257.007 and 275.017 for 6-phospho-gluconolactone and 6-phosphogluconic

4.5.7. *MILP problem description*

A universal stoichiometric reaction network was built from the set of reactions in the ModelSeed database (http://seed-viewer.theseed.org/). Unbalanced reactions and reactions of compounds with unspecified numbers of monomers were discarded; transport and exchange reactions were added based on the set of bacterial genome-scale metabolic models described in **Chapter 3**. For a particular species a template biomass composition was selected as described by Henry *et al*. [54], making sure that the universal network was able to produce all biomass components; i.e. that flux balance analysis (FBA [45]) predicted a positive flux through a biomass reaction with all considered precursors as reactants. Having defined a biomass reaction, flux variability analysis (FVA [188]) was used in order to remove reactions that could not carry flux under any condition while still producing biomass.

For a given genome GLOBUS probabilities were obtained as described (**Section 4.2.1**). The probability of each EC number in the generic EC network was estimated as the maximum probability among all genes assigned to the EC number. EC probabilities were assigned to the corresponding reactions based on the ModelSeed reaction definitions. When more than one EC number was associated to a particular reaction, the highest EC probability was used as the reaction probability. Reactions that were not associated with an EC number, reactions associated with EC numbers with zero probability, and reactions with a probability below 0.05 were all assigned a default probability of 5%.

In order to identify a set of likely reactions, a subset of reactions from the universal stoichiometric network was selected at random with likelihood proportional to the reaction probabilities. Starting with these reactions and using the previously defined biomass composition we solved the following optimization problem using mixed integer linear programming (MILP):

$$minimize \sum_{i=1}^{N} \left(1 + (1 - Probability_i)\right) z_i$$

$$subject\ to:$$

$$S \times v = 0,$$

$$0 \leq v_i \leq ub_i,$$

$$v_{biomass} > k$$

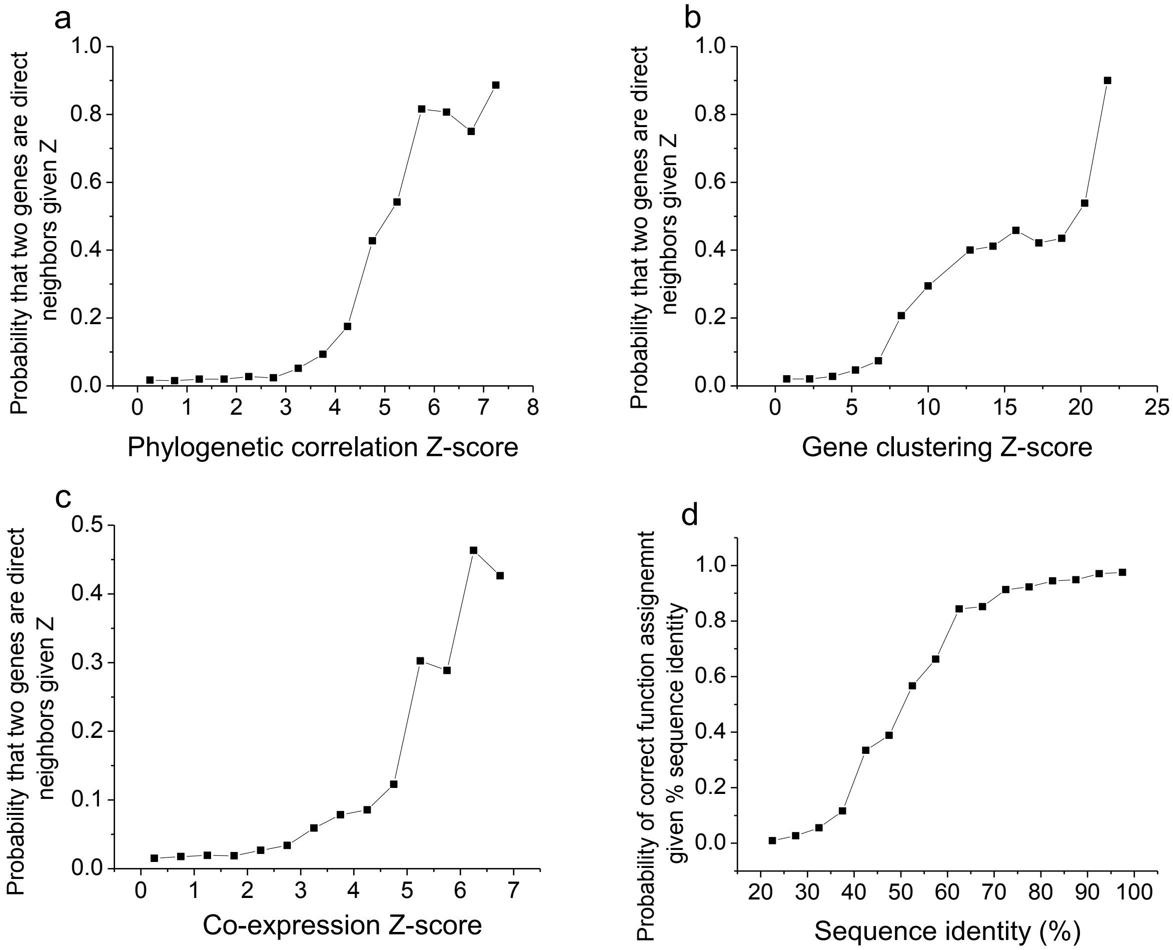
The solution identifies the least number of low probability reactions *i* not included in the initial subset that allow flux through the biomass reaction ($v_{biomass}$) larger than a small constant *k*. Z_i is a binary variable indicating the presence or absence of every reaction not present in the initial reaction set in the solution. The first condition ensures that the solution is at steady state; *S* represents the network's stoichiometric matrix and *v* is a vector containing all metabolic fluxes. The second condition sets upper bounds (ub_i) to fluxes in *v* when reversible reactions are represented as pairs of forward and backward reactions.

A binary vector of the reactions carrying flux in the MILP solution was used in GLOBUS with a modified objective function:

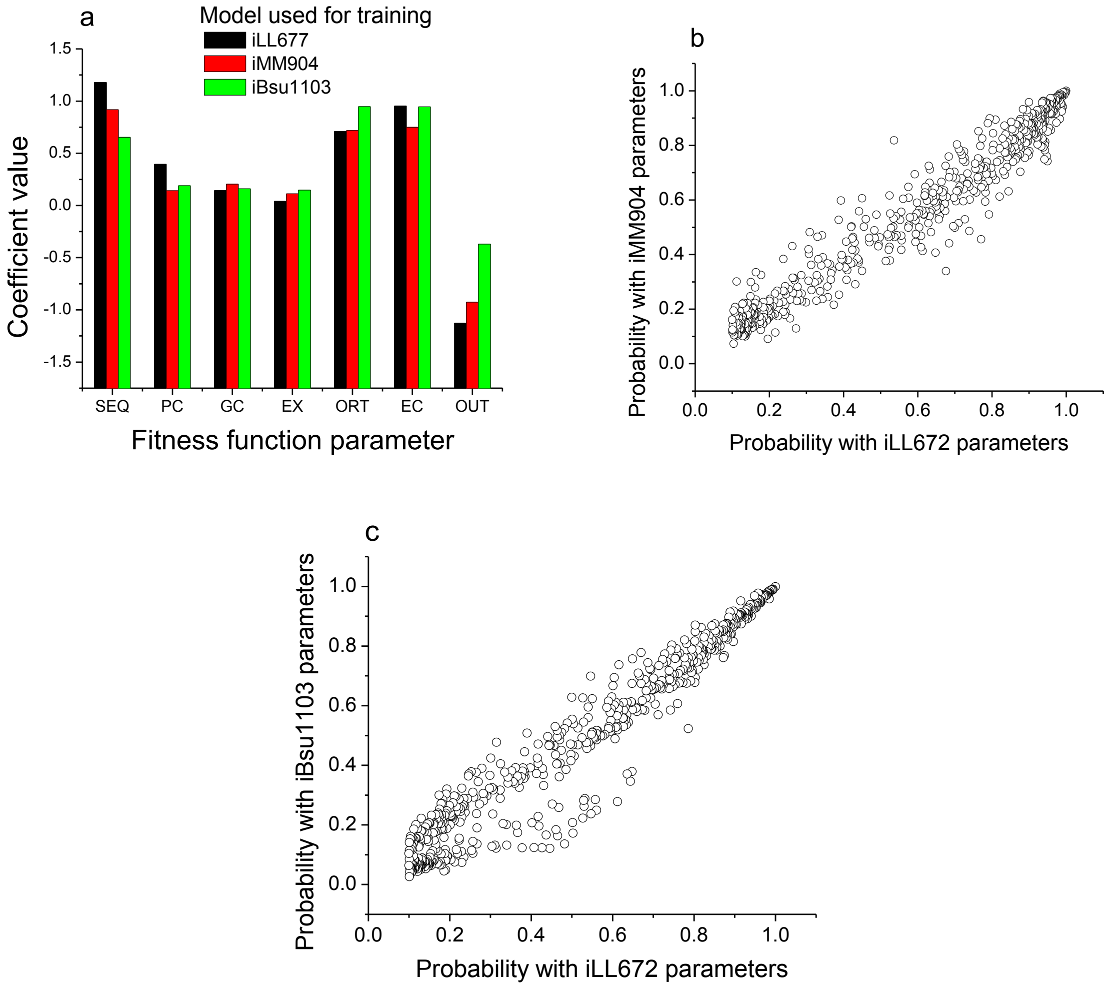
$$E(g_1, g_2, \dots, g_n) = -b_{homology} f_{homology} - b_{orthology} f_{orthology} - b_{context} f_{context} - b_{ECco-occurrence} f_{ECco-occurrence} \\ - b_{likely-rxn} f_{likely-rxn}$$

, where the additional term $f_{likely\text{-}rxn}$ had a value of 1 when genes were assigned to EC numbers associated to a reaction with non-zero flux and 0 otherwise.

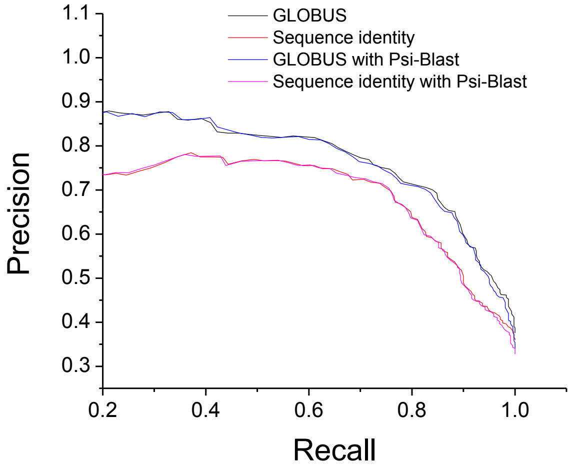
4.6. Supplementary figures and tables



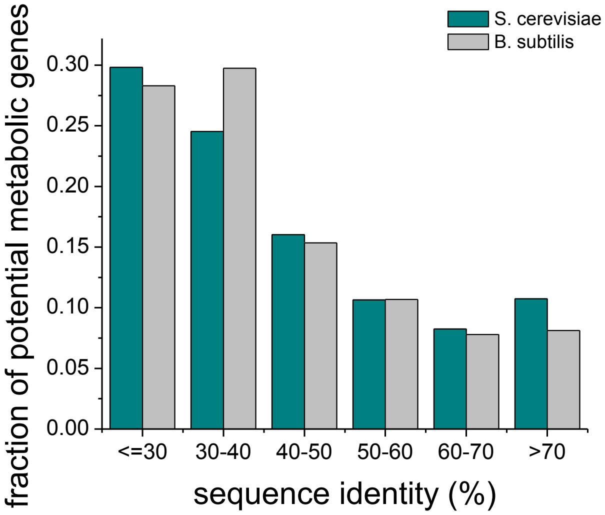
Supplementary figure 4.1. Conditional probabilities used in the GLOBUS fitness function. The context correlations used in GLOBUS (a-c) were first transformed into Z-scores[178] using the distribution of correlations for all pairs of candidate metabolic genes. Then we estimated the conditional probability that two genes are direct network neighbors given their context association Z-score. The greater the context correlation Z-score, the more likely the two genes are network neighbors. The conditional probabilities were estimated based on the iLL672 yeast metabolic model[154]. **(a)** The conditional probabilities for phylogenetic profiles, **(b)** The conditional probabilities for chromosomal gene clustering, **(c)** The conditional probabilities for mRNA co-expression. **(d)** As a sequence homology term in GLOBUS we used the conditional probability that a gene performs the assigned function, given the highest sequence identity to a Swiss-Prot[149] protein annotated to catalyze the target activity. The conditional probabilities for sequence homology were estimated using the well-curated yeast iLL672 metabolic model[154].



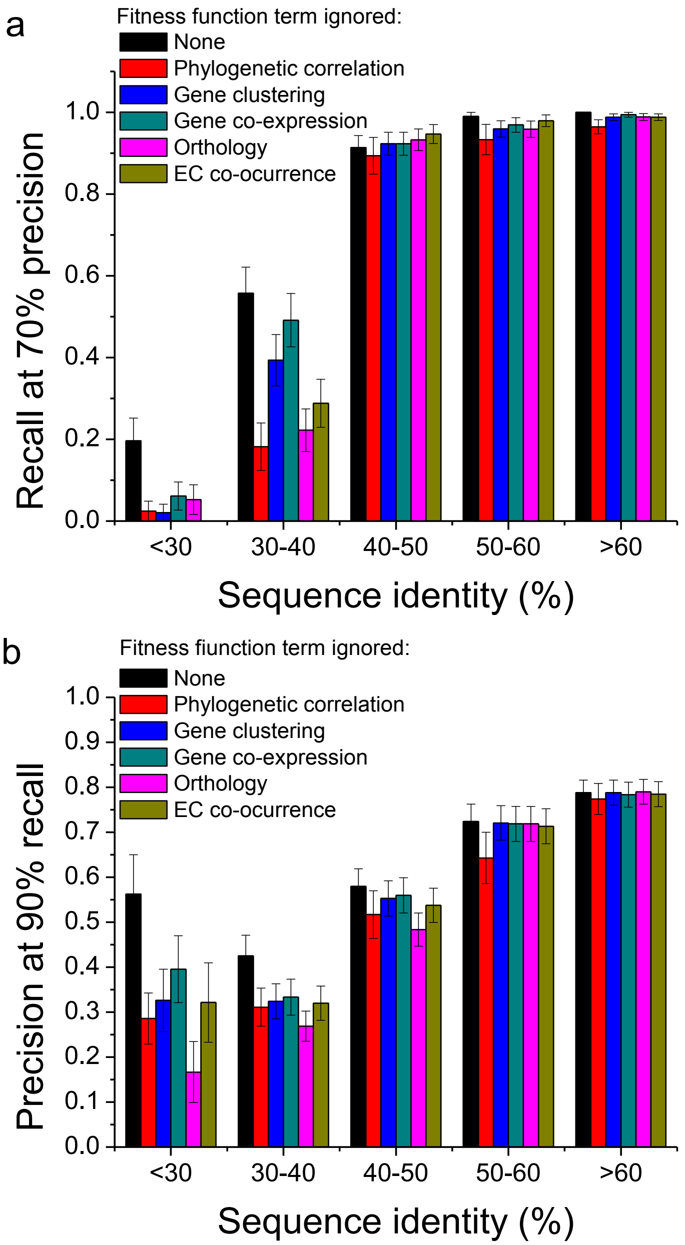
Supplementary figure 4.2. Training GLOBUS parameters using different models. **(a)** Maximum likelihood values of the context weight coefficients derived using the iLL672[154] and iMM904[156] *S. cerevisiae* models and the iBsu1103 [157] model for *B. subtilis*. SEQ: sequence identity; PC: phylogenetic correlation; GC: gene clustering; EX: co-expression; ORT: Orthology; EC: EC co-occurrence; OUT: not in the network **(b)** The correlation of probabilities with values higher than 0.1 in *S. aureus* based on GLOBUS parameters obtained by training with the two different yeast models (Pearson's r=0.94, median probability difference = 0.04, maximum probability difference = 0.33). **(c)** The correlation of probabilities with values higher than 0.1 in *S. aureus* based on GLOBUS parameters obtained by training with the yeast iLL672 and the iBsu1103 metabolic models (Pearson's r=0.96, median probability difference = 0.05, maximum probability difference = 0.35).



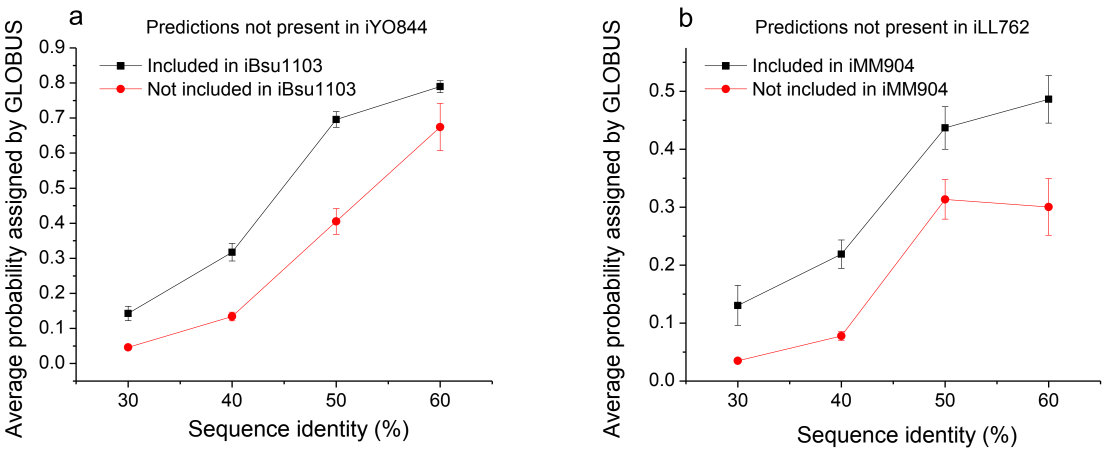
Supplementary figure 4.3. Comparison of the precision-recall relationships obtained using homology information established by BLAST and PSI-BLAST. Using PSI-BLAST, instead of regular BLAST, does not improve the performance significantly because additional sequences with low identity (detected by PSI-BLAST) only rarely have the target function.



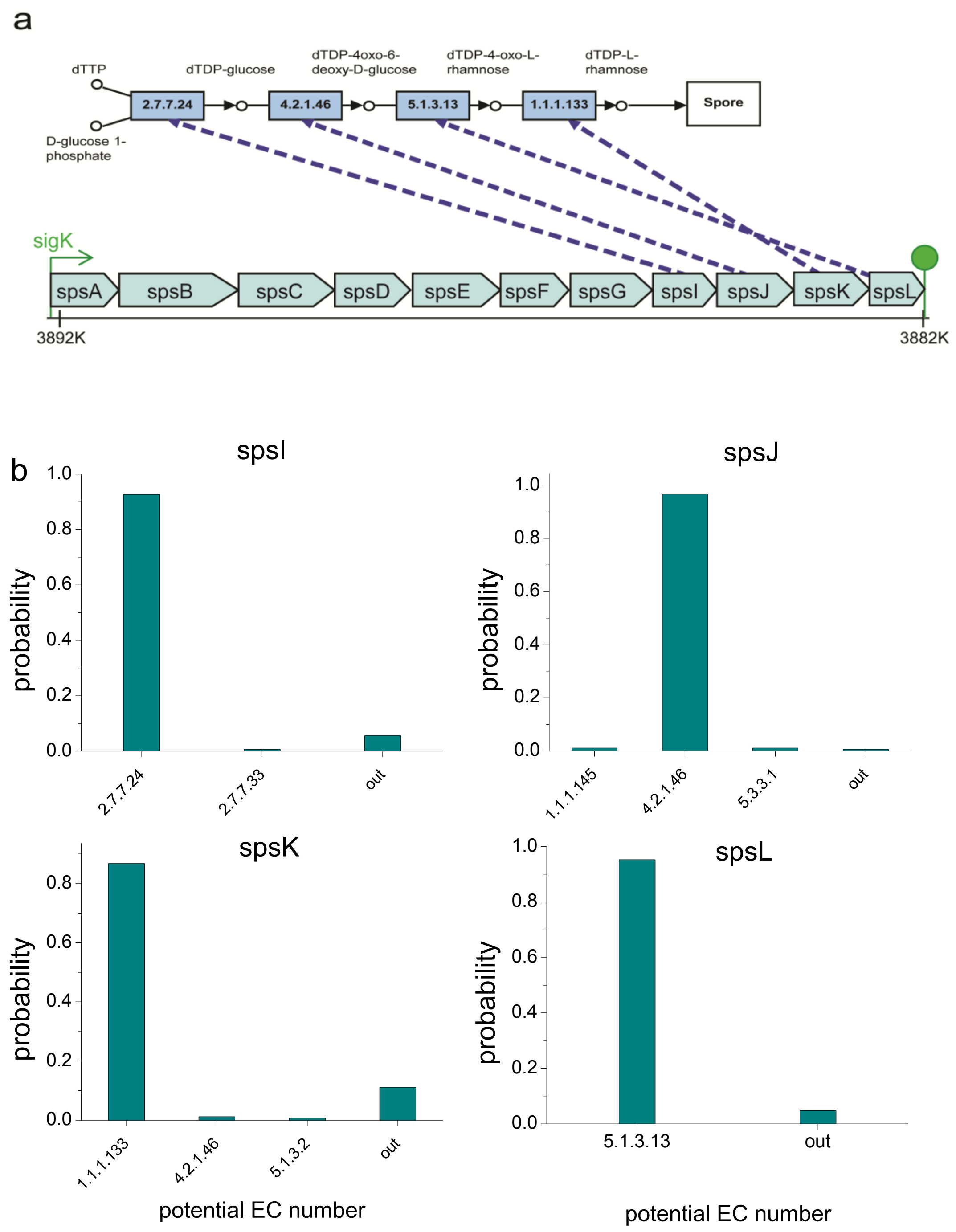
Supplementary figure 4.4. Fractions of potential metabolic genes in *S. cerevisiae* (green) and *B. subtilis* (grey) are shown as a function of sequence identity to annotated enzymes in other species. Over half of metabolic genes have relatively small sequence identity (<40%) to known enzymes in both model organisms.



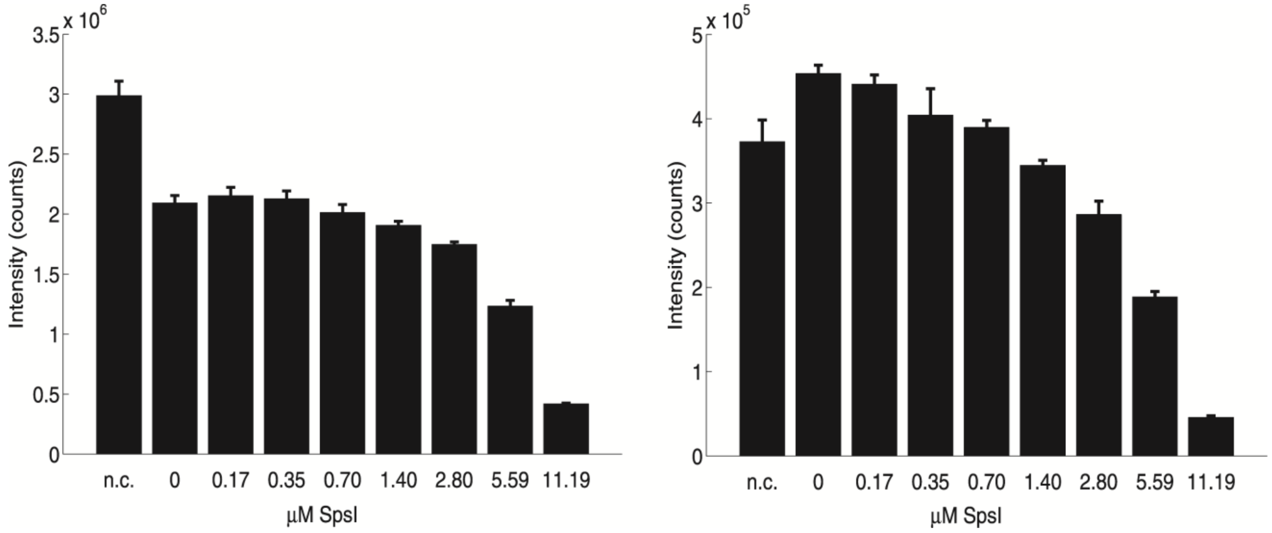
Supplementary figure 4.5. Contribution of individual context correlations to the GLOBUS performance. Different columns in the figure represent precision/recall values - across sequence identity bins - achieved by GLOBUS without using individual context correlations. The corresponding GLOBUS parameters were determined by simulated annealing optimizations performed without using each of the context correlations. The results show that at the same level of precision (70%) **(a)** and recall (90%) **(b)**, there is a marked reduction in performance; such reduction is most apparent for cases with low sequence identity.



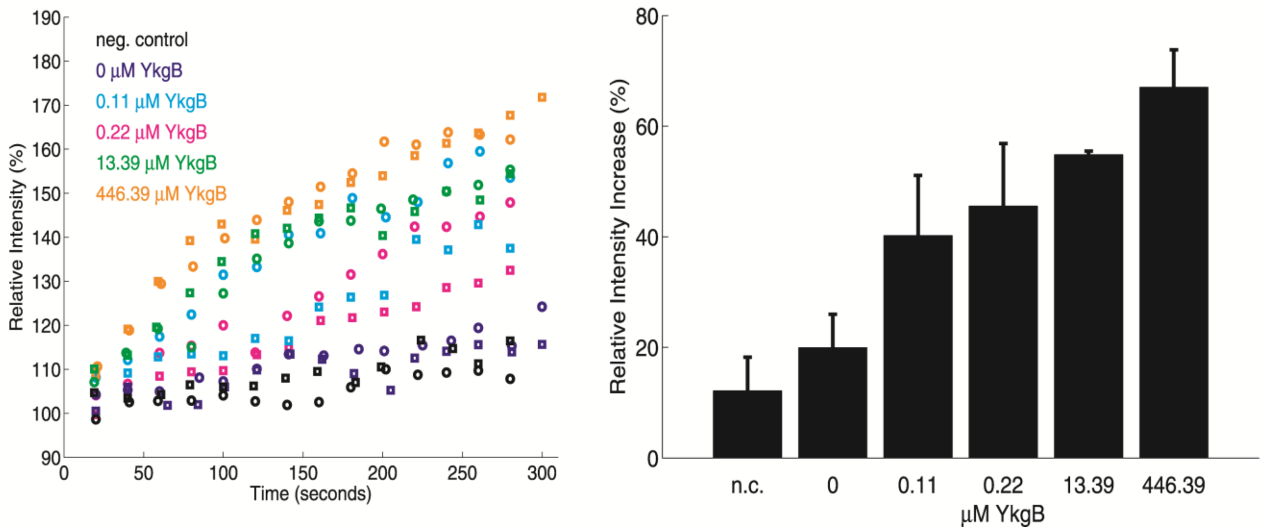
Supplementary figure 4.6. Potential utility of GLOBUS in refining manually curated metabolic models. **(a)** Annotations with non-zero GLOBUS probabilities that were not included in the older iYO844[158] *B. subtilis* model were subdivided into those included (black) and those that were not included (red) in the newer iBsu1103 model. Results show that, for different bins of sequence identity, higher GLOBUS probabilities correspond to higher likelihoods that annotations were included in the newer model. **(b)** Similar result is observed for yeast. Annotations with non-zero GLOBUS probabilities that were not included in the older iLL672[154] *S. cerevisiae* model were subdivided into those included (black) or not included (red) in the updated iMM904[156] model.



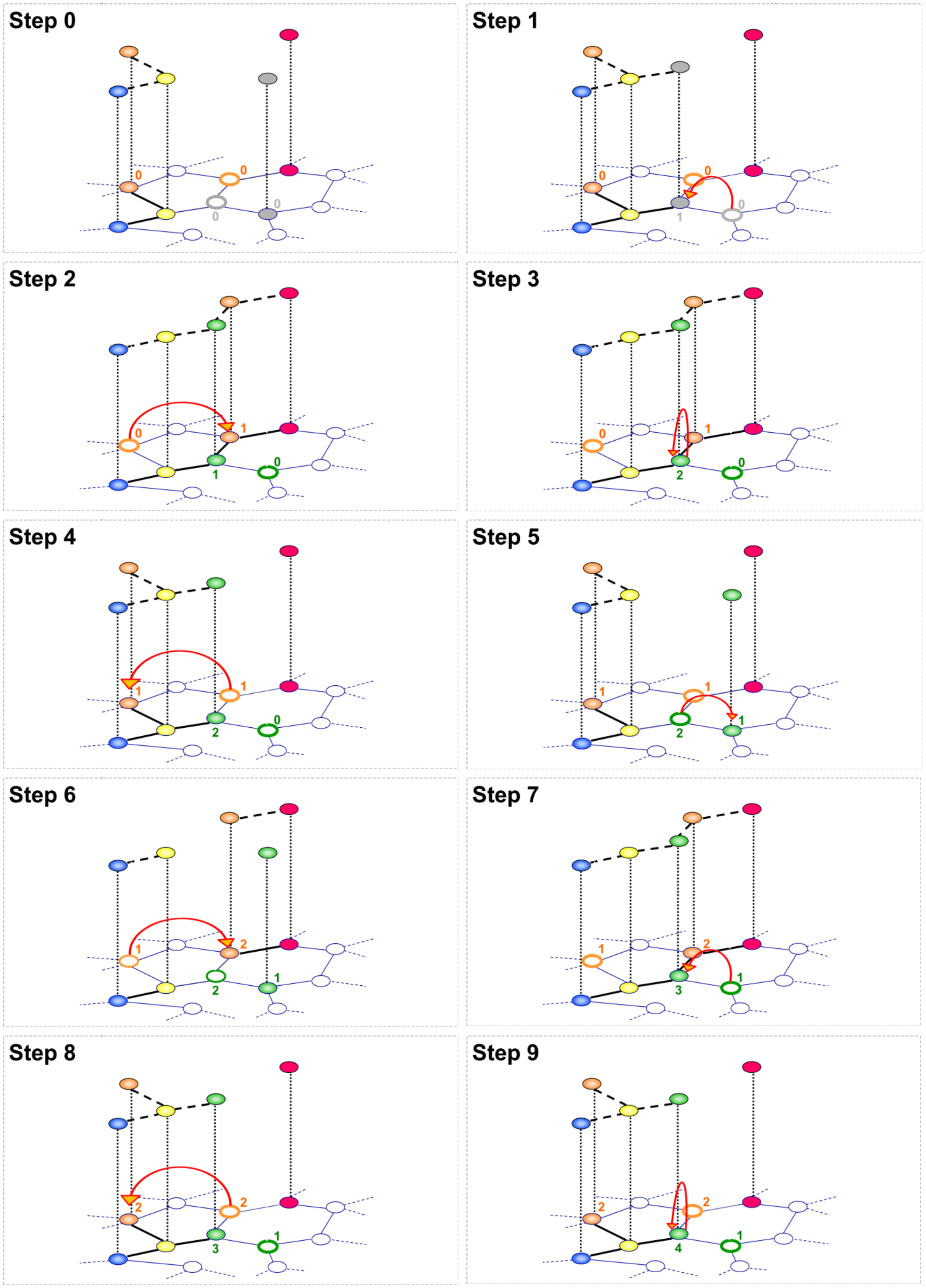
Supplementary figure 4.7. GLOBUS predictions for *sps* genes in *B. subtilis*. (a) Genomic positions of the *sps* genes, as well as gene mapping (dashed arrows) to the dTDP-L-rhamnose biosynthesis pathway. The expression of *sps* genes is controlled by the σ^K transcription factor[163]. **(b)** GLOBUS-derived probabilities for potential functions of *spsI*, *spsJ*, *spsK*, and *spsL*.



Supplementary figure 4.8. Substrate consumption at different spsI concentrations. Mass spectrometry intensities of α-D-glucose-1-phosphate (left panel) and dTTP (right panel) are shown as a function of the SpsI concentration. As negative control (n.c.), the protein free filtrate of 6.99 μM spsI solution was used. Error bars represent standard deviations from two independent assays.



Supplementary figure 4.9. Product accumulation at different YkgB concentrations. Mass spectrometry intensities of 6-Phosphogluconic acid (left panel) and its relative intensity increase (right panel) comparing final and initial intensities are shown as a function of the YkgB concentration. As negative control (n.c.), the protein free filtrate of 232 μM YkgB solution was used. Error bars represent standard deviations from two independent assays.


Step 0
Step 1
Step 2
Step 3
Step 4
Step 5
Step 6
Step 7
Step 8
Step 9

Supplementary figure 4.10. Illustration of Gibbs sampler used to derive marginal probabilities, *P(g$_i$)*, in GLOBUS. A marginal probability for a gene assignment reflects the likelihood of the gene to catalyze the corresponding activity consistent with, i.e. integrated or summed over, all possible assignment of other genes in the network. Probabilities were calculated directly from the Gibbs chain counts.

Supplementary table 4.1. Highly connected metabolites that were not used in establishing connections between metabolic activities (EC numbers).

Metabolite	Number of connected EC numbers	Metabolite	Number of connected EC numbers
H_2O	1224	Reduced acceptor	115
H+	703	AMP	111
NADP+	435	Pyruvate	109
NADPH	433	S-Adenosyl-L-homocysteine	107
NAD+	422	Acetyl-CoA	103
NADH	412	H_2O_2	102
Oxygen	379	L-Glutamate	100
ATP	375	2-Oxoglutarate	96
Orthophosphate	306	UDP-glucose	76
ADP	294	Acetate	73
CO_2	254	D-Glucose	56
CoA	230	Carboxylate	48
Pyrophosphate	185	Succinate	43
NH_3	183	Oxaloacetate	41
UDP	150	Glycine	41
S-Adenosyl-L-methionine	115		

Chapter 5.

RECONSTRUCTION AND FLUX BALANCE ANALYSIS OF THE *Plasmodium falciparum* METABOLIC NETWORK

This chapter is based on "Plata, G., Hsiao, T., Olszewski, K., Llinás, M., Vitkup, D. **Reconstruction and Flux-balance Analysis of the Plasmodium falciparum Metabolic Network**. Mol. Syst. Biol. 6:408 (2010)"

Genome-scale metabolic reconstructions can serve as important tools for hypothesis generation and high-throughput data integration. Here we present a metabolic network reconstruction and flux-balance analysis of *Plasmodium falciparum*, the primary agent of malaria. The compartmentalized metabolic network accounts for 1001 reactions and 616 metabolites. Enzyme-gene associations were established for 366 genes and 75% of all enzymatic reactions. Compared with other microbes, the *P. falciparum* metabolic network contains a relatively high number of essential genes, suggesting little redundancy of the parasite metabolism. The model was able to reproduce phenotypes of experimental gene knockout and drug inhibition assays with up to 90% accuracy. Moreover, using constraints based on gene expression data, the model predicted the direction of concentration changes for external metabolites with 70% accuracy. Using flux-balance analysis, we identified 40 enzymatic drug targets (i.e. *in silico* essential genes) with no or very low sequence identity to human proteins. To demonstrate that the model can be used to make clinically relevant predictions, we experimentally tested one of the identified drug targets, nicotinate mononucleotide adenylyltransferase, using a recently discovered small molecule inhibitor.

5.1. Introduction

Malaria is an ancient disease which can be dated back to 2800 B.C. [189] and remains one of the most severe public health challenges worldwide. Currently, about half of the Earth's population is at risk from this infectious disease according to the World Health Organization [190]. Malaria inflicts acute illness on hundreds of millions of people worldwide and leads to at least one million deaths annually [190, 191]. It ranks as a leading cause of death and disease in many developing countries, where the most affected groups are young children and pregnant women [190]. The disease is transmitted to humans by the female *Anopheles* mosquito and is caused by at least five species of *Plasmodium* parasites. The lifecycle of the parasite is highly complex and includes various hosts and tissue types. During a blood meal, sporozoites are transmitted from the mosquito to humans and initiate infection in the liver where they reproduce prolifically but are asymptomatic. In the next stage of infection the parasites are released from the liver cyst into the bloodstream in the form of merozoites, where they invade red blood cells and reproduce asexually [192]. The destruction of red blood cells coupled with the significant load imposed on the host metabolism is ultimately responsible for the major clinical symptoms of malaria, which are often fatal [193].

Although several anti-malarial drugs are currently available, most of them are losing efficacy due to acquired drug resistance in the most lethal causative agent, *Plasmodium falciparum* [194, 195]. The loss of drug efficiency in resistant strains poses a great threat to malaria control and has been linked to increases in worldwide malaria mortality [196]. There is an urgent need for new anti-malarial drugs coupled with better administration strategies. Understanding the molecular mechanisms and interactions of the parasite's cellular components is essential for identification of new drug targets, especially given the difficulties associated with *in vivo* drug testing [197].

Various systems biology approaches have been applied to improve our understanding of *P. falciparum* physiology and to facilitate drug development [198]. The sequencing of the *P. falciparum* genome has provided researchers with a complete collection of parasite proteins and likely regulatory interactions [199, 200]. Several large scale transcriptome [201-204], proteome [205-207] and metabolome [208] analyses have been conducted in order to dissect functional interactions and define essential biological pathways. In addition to experimental studies, several databases have been developed to integrate functional knowledge of the parasite and its metabolism. For example, PlasmoCyc is an integrated database that links genomic data, protein annotation, enzymatic reactions, and pathway information [209]; the Malaria Parasite Metabolic Pathway Database, on the other hand, is a manually curated resource which assembles annotated enzymes into likely metabolic pathways [210].

A stoichiometric representation of metabolism can be effectively used to study functional properties of biochemical networks using a growing number of computational methods [211]. For example, flux balance analysis (FBA) considers steady state distributions of metabolic fluxes satisfying a set of biophysical constrains, such as bounds and mass balance of fluxes [45]. Given the applied constraints, a likely distribution of fluxes in the network can be obtained by maximizing an appropriate objective function (e.g. biomass production) [78] or applying minimal perturbation principles [79, 80]. The analysis of flux-balanced genome-scale metabolic networks is useful not only for the discovery of essential genes and potential drug targets, but also as a tool to better understand species-specific biology [76, 171]. For example, among other applications, these models have been used to identify minimal media requirements for growth

[212], explore metabolic weaknesses in bacterial pathogens [213], integrate gene expression and other types of data [214], and investigate objective functions important under different growth conditions [215]. Given the complex life cycle of *Plasmodium*, a flux-balanced model of this organism is of direct relevance to the ongoing search to identify new therapeutic drug targets.

In this study we reconstructed the genome-scale flux balanced metabolic network of *P. falciparum* and used it to perform a systems level analysis of the parasite's metabolism. Based on *in silico* gene deletions we identified potential new anti-malarial drug targets with low sequence identity to human proteins. One of these targets, nicotinate mononucleotide adenylyltransferase, was experimentally tested in a growth inhibition assay using a recently discovered small molecule inhibitor. We also illustrate, using a previously published methodology, how the reconstructed metabolic model can be used to integrate flux analysis with expression data to more accurately simulate the physiology of this complex eukaryotic pathogen.

5.2. Results

5.2.1. *Scale of the reconstructed flux-balanced metabolic network*

The reconstructed flux-balanced model is based on gene-reaction associations reported in public domain databases as well as on a careful literature analysis. We used well-curated microbial metabolic models and enzyme databases to determine the stoichiometry of most reactions. To produce a functional reconstruction we also searched the literature for missing steps necessary for the model to produce a set of required biomass components (see Methods). The model accounts for 366 genes, corresponding to 7% of all genes identified in *P. falciparum*. Compared to 61 metabolic models of various organisms compiled by Feist *et al.* [216], our model ranks tenth in terms of the smallest number of genes. Not surprisingly, the other metabolic models with small gene numbers include many parasitic/symbiotic species, such as *Mycoplasma genitalium, Buchnera aphidicola, Haemophilus influenzae and Helicobacter pylori*. The *P. falciparum* network also includes 616 metabolites and 1001 reactions, 657 of which are metabolic transformations (**Table 5.1**). In addition, there are 231 reactions corresponding to transport between different cellular compartments and 111 input–output exchange reactions that allow extracellular metabolites to enter and end products to be excreted from the network.

Table 5.1. Characteristics of the reconstructed metabolic network of *P. falciparum*.

Reactions	**1001**	**Metabolites**	**616**
Cytosolic reactions	503	Cytosolic metabolites	537
Mitochondrial reactions	47	Mitochondrial metabolites	83
Apicoplast reactions	105	Apicoplast metabolites	135
Transport reactions	231	Extracellular metabolites	159
Cytosolic transport reactions	130	**Genes**	**366**
Mitochondrial transport reactions	50		
Apicoplast transport reactions	48		
Exchange reactions	111		

The metabolic reconstruction includes four distinct compartments: parasite cytosol, mitochondria, apicoplast (a non-photosynthetic plastid), and the extracellular space (representing the host cell cytosol and host serum). The majority of all reactions (50%) occur in the cytosol. The apicoplast accounts for 10% of all reactions, such as the synthesis of isopentenyl diphosphate, fatty acids, and heme [217]. A special reaction is included in the model to account for the biomass components and essential metabolites needed for growth (Table S1). Supplementary file 1 provides a complete list of all network reactions and metabolite abbreviations.

Excluding metabolite-exchange reactions, 74% of the reactions in the model are directly associated with *P. falciparum* genes, which compares well to other models of eukaryotes such as the iND750 yeast model (70%) [218] and the iAC560 model for *Leishmania major* (63%) [212]. The remaining reactions include spontaneous transformations that can proceed without enzymatic catalysis and reactions required for the proper functioning of the metabolic model. Intracellular and inter-compartmental transport reactions, most of which are not currently associated with any gene, account for about 6% and 15% of all reactions in the model, respectively (**Figure 5.1A**). Most of the transporter proteins in *Plasmodium spp.* are currently uncharacterized. However, it is well-established that the parasite significantly modifies the permeability of the host cell membrane [219, 220] and several metabolic processes occur across different organelles. For instance, such metabolic pathways as heme biosynthesis and antioxidant defense have been shown to involve both host and parasite enzymes localized to multiple intra-cellular compartments [221, 222]. Given the importance of metabolite exchange, many transport reactions were included in the model, although the identities of the corresponding genes remain unknown (**Figure 5.1A**).

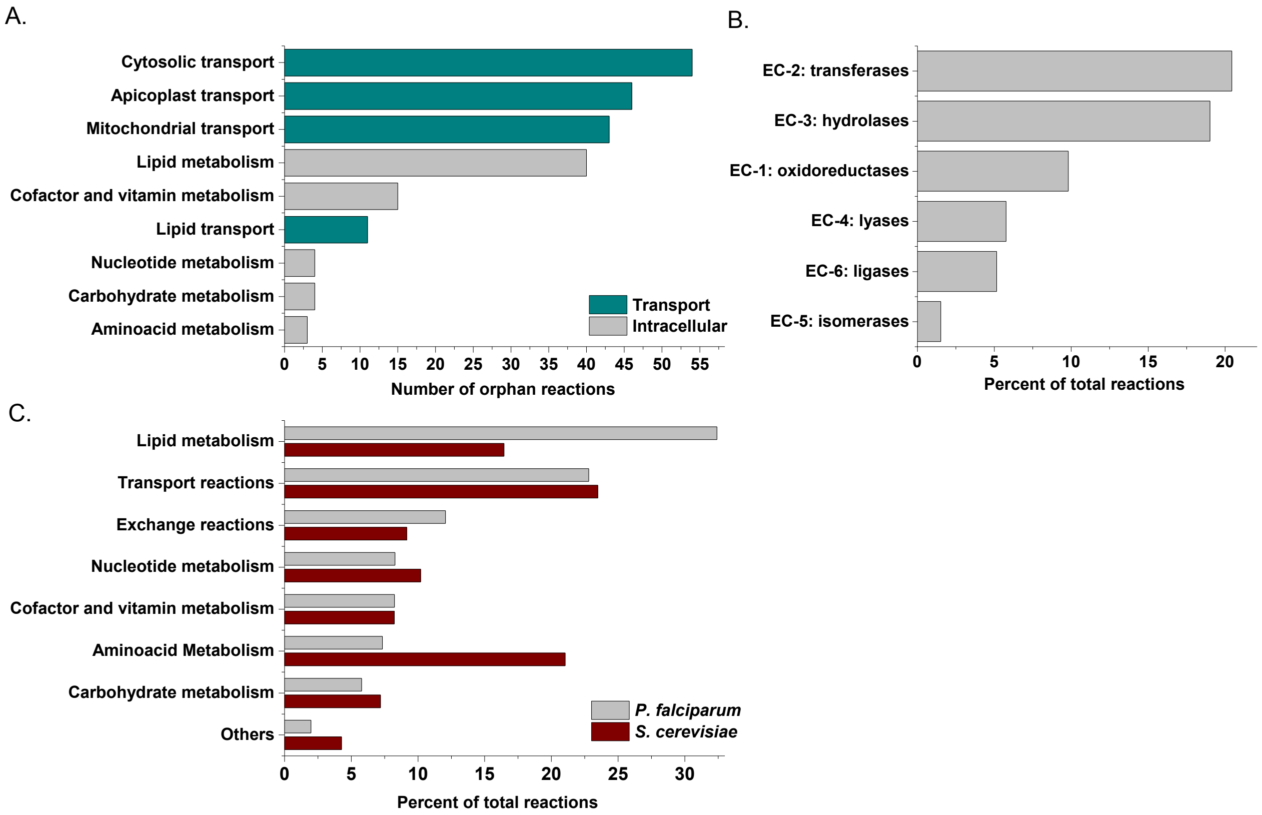


Figure 5.1. Annotation of reactions in the genome-scale metabolic model of *P. falciparum*. A. Number of orphan (non-gene associated) reactions in *P. falciparum* grouped by metabolic processes. B. Reactions grouped by Enzyme Commission (EC) classifications. C. Reactions grouped by metabolic processes in *P. falciparum* and *S. cerevisiae* [218].

Transferases and hydrolases comprise the largest fraction of enzymatic reactions in the network (**Fig. 5.1B**). In terms of specific metabolic processes (**Fig. 5.1C**), most reactions in the network are related to lipid metabolism, followed by transport and exchange reactions. In comparison with *S. cerevisiae*, a free living eukaryote of similar genome size, the most significant metabolic difference is the fraction of reactions involved in amino acid metabolism (**Fig. 5.1C**). About 20% of the reactions in the iND750 yeast metabolic network [218] are responsible for amino acid pathways; in contrast, this fraction is only 7% in *P. falciparum*. Amino acid biosynthesis pathways are absent in *P. falciparum* metabolism due to the unique ability of the parasite to catabolize the erythrocyte hemoglobin [223] and to scavenge free amino acids from the host serum (human stages) or hemolymph (mosquito stages).

5.2.2. *Analysis of in silico single and double gene deletions*

We simulated gene deletions using flux balance analysis (FBA) of the reconstructed metabolic network. Even though sugars other than glucose do not support *P. falciparum* growth in culture [224], in performing the *in silico* deletions we initially allowed all exchange reactions to carry non-zero metabolic fluxes, thereby permitting the potential import and utilization of other hexoses. Purines, such as inosine and adenosine, which are not normally included for *in vitro* culture but can be imported *in vivo* [225]*,* were also made available to the network. We note that genes identified as essential when all exchange reactions are allowed will also be essential under more specific (constrained) conditions. The phenotypic effects of *in silico* deletions were classified into four groups: lethal (L), growth reducing (GR, growth between 0% and 95% of the wild type network), slight growth reducing (SGR, growth between 95% and 100%), and with no effect (NE). About 15% of all single gene deletions (**Table 5.2**) were lethal, approximately 1% were growth reducing, and 3.5% were slightly growth reducing.

Table 5.2.Total number of single and non-trivial double deletion phenotypes.

Predicted phenotype[a]	Single deletion (# genes)	Single deletion (%)	Double deletion (# non-trivial pairs)	Double deletions (%)
No Effect (NE)	295	80.60	43160 + 4974 (trivial GR,SGR)	99.85
Lethal (L)	55	15.02	16	0.03
Growth Reducing (GR)	3	0.82	48	0.10
Slight Growth Reducing (SGR)	13	3.55	7	0.01
TOTAL	366	100.0	48205	100.0

[a] Single and double gene deletion predictions were classified into lethal, growth reducing (growth between 0% and 95% of wild type), slight growth reducing (growth between 95% and 100%), and no effect.

Out of 366 genes in the *P. falciparum* metabolic network, 55 genes were predicted to be essential for growth (**Supplementary table 5.2**). To assess the accuracy of these predictions we compiled a list of experimentally validated gene-knockouts and phenotypes resulting from targeted inhibitions of enzymatic activities with drugs (**Table 5.3**). In the computational analysis we assumed that the drug treatments resulted in a complete inhibition of targeted enzymatic activities. In this way, the available drug phenotypes were simulated with computational deletions of the corresponding genes. In total, 14 metabolic gene knockouts and 25 drug inhibition phenotypes for genes were retrieved from the literature for *P. falciparum* and *P. berghei*, a murine malaria parasite commonly used in experimental studies [226, 227].

The FBA analysis was able to achieve 100% accuracy for predictions of both essential and non-essential gene knockouts (14 cases in total). In contrast, about 70% accuracy was achieved for phenotypes resulting from drug inhibitions of metabolic enzymes. Interestingly, all mispredicted drug phenotypes (8 cases) involved genes for which the computational analysis predicted a non-zero growth phenotype, while corresponding drugs were lethal to the parasite in experimental studies. In three cases inconsistencies between the FBA predictions and experimental drug inhibitions can be explained by considering functions that are not explicitly represented in our model. These included the degradation of spontaneously forming toxic metabolites (e.g. methylglyoxal), and the synthesis of metabolites that are involved in the progression between the intraerythrocytic stages (e.g. sphingolipid-ceramide) [228]. The remaining discrepancies (5 cases) can be resolved by taking into account specific literature-based evidence (**Table 5.3**, green rows), i.e. by considering nutrient availabilities and directionality of exchange reactions. Interestingly, in one case the source of discrepancy between the model and experiments was clearly related to off-target drug effects. Specifically, the inhibitor of enoyl-Acyl carrier reductase (FabI), triclosan, has been shown to kill *P. falciparum in vitro* and *in vivo* (Surolia and Surolia 2001) despite the fact that its presumed target, PfFabI, can be deleted with no apparent blood-stage phenotype [229-231]; this deletion phenotype is correctly predicted by our model.

Table 5.3. Literature support for essentiality predictions[a].

Gene	Reaction	EC number	Sp	Exp	Pred	Biological Process	Reference
PFL2510w	Chitinase	3.2.1.14	*Pber*	NL[b]	NL	Aminosugars metabolism	[232]
PFI0320w	Arginase	3.5.3.1	*Pber*	NL[c]	NL	Arginine and proline metabolism	[208]
PF14_0200*	Pantothenate kinase	2.7.1.33	*Pfal*	L	L	CoA biosynthesis	[233]
PF14_0354*							
PF13_0128	Beta-hydroxyacyl-ACP dehydratase	4.2.1.58	*Pber*	NL[d]	NL	Fatty acid synthesis	[231]
		4.2.1.60					
		4.2.1.61					
PFF0730c	Enoyl-acyl carrier reductase (FABI)	1.3.1.9	*Pber*	NL[d]	NL	Fatty acid synthesis	[229, 231]
PFF1275c	3-oxoacyl-acyl-carrier protein synthase I/II (FABB/F)	2.3.1.41	*Pber*	NL[d]	NL	Fatty acid synthesis	[231]
PF08_0095*	Dihydropteroate synthetase	2.5.1.15	*Pfal*	L	L	Folate biosynthesis	[234]
PFD0830w*	Dihydrofolate reductase. Thymidylate synthase	1.5.1.3	*Pfal*	L	L	Folate biosynthesis Pyrimidine metabolism	[235]
		2.1.1.45					
PF13_0269	Glycerol kinase	2.7.1.30	*Pfal*	NL	NL	Glycolysis	[236]
PF14_0425*	Fructose-bisphosphate aldolase	4.1.2.13	*Pfal*	NL[e,q]	NL[c]	Glycolysis	[237]
PF13_0141*	Lactate dehydrogenase	1.1.1.27	*Pfal*	L	L[m]	Glycolysis	[238]

PF13_0144*							
PF14_0641*	1-deoxy-D-xylulose-5-phosphate reductoisomerase	1.1.1.267	*Pfal*	L	L	Isoprenoids metabolism	[239]
PF14_0788	Adenylyl cyclase	4.6.1.1	*Pber*	NL	NL	Isoprenoids metabolism	[240]
PF10_0322*	Ornithine decarboxylase	4.1.1.17	*Pfal*	L[f]	L[f]	Methionine and polyamine metabolism	[241-243]
	S-adenosylmethionine decarboxylase	4.1.1.50					
PF11_0301*	Spermidine synthase	2.5.1.16	*Pfal*	L[g]	L[g]	Methionine and polyamine metabolism	[244]
PF10_0275*	Protoporphyrinogen oxidase	1.3.3.4	*Pfal*	L	L	Porphyrin metabolism	[245]
PF14_0381*	δ-Amino-levulinic acid dehydratase	4.2.1.24	*Pfal*	L	L	Porphyrin metabolism	[245]
PF10_0121*	Hypoxanthine phosphoribosyl transferase	2.4.2.8	*Pfal*	L	L	Purine metabolism	[246]
PF13_0287*	Adenylosuccinate synthase	6.3.4.4	*Pfal*	L	L	Purine metabolism, Asparagine and aspartate metabolism	[247]
PFB0295w*	Adenylosuccinate lyase	4.3.2.2	*Pfal*	L	L	Purine metabolism	[248]
MAL13P1.301	Guanylyl cyclase	4.6.1.2	*Pber*	NL[h]	NL	Purine metabolism Porphyrin metabolism	[249]
PFF0160c*	Dihydroorotate dihydrogenase	2.4.2.8	*Pfal*	L	L	Purine metabolism	[250]
PF10_0289*	Adenosine deaminase	3.5.4.4	*Pfal*	L[i]	L[i]	Purine metabolism Methionine and polyamine metabolism	[251]
PFE0660c*	Purine nucleoside phosphorylase	2.4.2.1	*Pfal*	L[j]	L[j]	Purine metabolism Methionine and polyamine metabolism	[252]
PF10_0154* PF14_0053* PF14_0352*	Ribonucleoside reductase	1.17.4.1	*Pfal*	L	L	Pyrimidine metabolism Purine metabolism Redox metabolism	[253]
PF10_0225	Orotidine-monophosphate decarboxylase	4.1.1.23	*Pber*	L	L	Pyrimidine metabolism	Leiden Malaria Research Group (unpublished)
PF11_0410*	Carbonic anhydrase	4.2.1.1	*Pfal*	L	L	Pyrimidine metabolism Fatty acid synthesis Pyruvate metabolism	[254]
PF13_0044*	Carbamoyl-phosphate synthase	6.3.5.5	*Pfal*	L	L	Pyrimidine metabolism Glutamate metabolism	[255]
PF11_0282*	Deoxyuridine 5’-triphosphate nucleotidohydrolase	3.6.1.23	*Pfal*	L	NL[n]	Pyrimidine metabolism	[256]
PF11_0145* PFF0230c*	Lactoylglutathione lyase	4.4.1.5	*Pfal*	L	NL[o]	Pyruvate metabolism	[257]
PF14_0368	Thioredoxin peroxidase	1.11.1.15	*Pber*	NL[k]	NL	Redox metabolism	[258, 259]
PFI0925w	Gamma-glutamylcysteine synthase	6.3.2.2	*Pber*	NL[b,c]	NL	Redox metabolism Glutamate metabolism	[260]
PFI1170c	Thioredoxin reductase (NADPH)	1.8.1.9	*Pfal*	L	L	Redox metabolism Pyrimidine metabolism Purine metabolism	[261]

PFB0280w*	3-phosphoshikimate 1-carboxyvinyl transferase	2.5.1.19	*Pfal*	L	L	Shikimate biosynthesis	[262]
PFF1105c*	Chorismate synthase	4.2.3.5	*Pfal*	L	L	Shikimate biosynthesis	[263]
PFL1870c*	Sphingomyelinase	3.1.4.12	*Pfal*	L	NL[p]	Sphingomyelin and ceramide metabolism	[228]
PF11_0295*	Farnesyl diphosphate synthase	2.5.1.10 2.5.1.1	*Pfal*	L	L	Terpenoid metabolism	[264]
PF11_0338	Aquaglyceroporin	--	*Pber*	NL[c]	NL	Transport	[265]
PFF1420w	Phosphatidylcholine-sterol acyltransferase	2.3.1.43	*Pber*	NL[l]	NL	Utilization of phospholipids	[266]

[a] Genes are grouped and sorted by biological process. Yellow rows indicate cases for which the model is unable to reproduce the experimental phenotype. Green rows highlight cases for which predictions coincide with experiments after specific experimental conditions are included in the simulation. *Drug targets with experimental evidence, mostly as compiled by [209]. L: Lethal, NL: Non-Lethal. *Pber: P. berghei, Pfal: P. falciparum.*
[b] reduced mosquito stage viability.
[c] slightly reduced growth.
[d] liver stage not viable.
[e] reduced parasitemia.
[f] lethal in the absence of putrescine.
[g] lethal in the absence of spermidine.
[h] mosquito stage not viable.
[i] lethal in the absence of inosine and hypoxanthine.
[j] lethal in the absence of hypoxanthine.
[k] lower gametocyte production.
[l] reduced liver stage viability.
[m] lethal when pyruvate export is not allowed in the model.
[n] this activity is required to maintain a low dUTP/dTTP ratio to prevent DNA damage. This is not reflected in the biomass function and cannot be predicted by FBA.
[o] this activity is required for detoxification of methylglyoxal, which is forms spontaneously and must eventually converted to lactate and excreted. This cannot be predicted by FBA.
[p] slight growth reducing, sphingomyelinase is required for progression from the trophozoite to schizont stage, this is not captured by our objective function.
[q] incomplete inhibition

For 15 metabolic genes identified in our analysis as essential, knockout experiments or drug inhibition assays in *P. falciparum*/*P. berghei* are already available in the literature (see Tables III, S2). The remaining predictions include 24 genes coding for proteins with relatively low sequence identity (20%-40%) to human transcripts (**Supplementary fig. 5.1**, see Methods), and 16 genes with no significant sequence identity to any human protein (BLAST E-value > 10^{-2}). This last group comprises six genes associated with isoprenoid metabolism, three genes involved in nucleotide metabolism, and genes related to CoA, shikimate, and folate biosynthesis

(**Table 5.4**). Nine of the genes with no homology to human proteins are homologous to plant proteins; this is consistent with the essential functions of apicoplast-associated genes in the Apicomplexa [267]. These forty enzymes are of immediate interest as potential drug targets, as low homology to human proteins suggests that side effects for drugs targeting these enzymes may be minimized or avoided. Interestingly, five of the 16 enzymes with no detectable homology correspond to enzymatic activities (EC numbers) that are unlikely to be present in human metabolism according to the KEGG [131], HumanCyc [268] and UniProt [17] databases. Among the predicted essential genes with low but significant sequence identity to human transcripts, only aminodeoxychorismate lyase/synthetase (2.6.1.85, 4.1.3.38, PFI1100w) is associated with EC numbers not reported in human metabolism. In addition to genes predicted as essential, 9 internal metabolic reactions with no associated network genes (orphan reactions) were also predicted to be essential for growth. Four of these reactions are associated with the shikimate biosynthetic pathway; three with ubiquinone metabolism, one with nicotinamide, and one with porphyrin metabolism (**Supplementary table 5.3**).

One metabolic pathway of significant interest in the parasite is the mitochondrial tricarboxylic acid (TCA) cycle. In most free-living microbes this pathway fully oxidizes available carbon sources to carbon dioxide, in the process generating high-energy phosphate bonds (ATP/GTP). Within the *Plasmodium* species, however, the nature and function of the TCA cycle remains unclear [269]. In the malaria parasites the sole pyruvate dehydrogenase complex, which normally provides the key link between glycolysis and TCA metabolism, localizes not to the mitochondrion but to the apicoplast. In that compartment it is likely to be used primarily to generate acetyl-CoA for lipogenesis [270]. Incorporating this fact into our model, we find that the only TCA cycle enzyme predicted to be essential is the 2-oxoglutarate dehydrogenase complex. This enzyme converts 2-oxoglutarate into succinyl-CoA, which is required for heme biosynthesis. Intriguingly, metabolic labeling experiments indicate that TCA cycle of the parasite also reduces 2-oxoglutarate to malate, generating acetyl-CoA from citrate cleavage (Olszewski and Llinas, unpublished). A non-zero flux through this reaction is observed in our model when additional constraints are applied to mitochondrial transport reactions.

Table 5.4. Predicted essential genes with no homologs in the human genome[a].

Gene name	Enzyme name	EC	Biological Process
MAL8P1_81	Phosphopantothenoylcysteine decarboxylase	4.1.1.36	CoA biosynthesis
PF07_0018	Pantetheine-phosphate adenylyltransferase	2.7.7.3	CoA biosynthesis
PFF1490w	Methenyltetrahydrofolate cyclohydrolase Methylenetetrahydrofolate dehydrogenase (NADP+)	3.5.4.9 1.5.1.5	Folate biosynthesis
MAL13p1_186	1-deoxy-D-xylulose-5-phosphate synthase	2.2.1.7	Isoprenoid metabolism
PF10_0221	(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase	1.17.7.1[b]	Isoprenoid metabolism
PFA0225w	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	1.17.1.2[b]	Isoprenoid metabolism
PFA0340w	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	2.7.7.60[b]	Isoprenoid metabolism
PFB0420w	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	4.6.1.12[b]	Isoprenoid metabolism
PFE0150c	4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase	2.7.1.148[b]	Isoprenoid metabolism
PF13_0159	**Nicotinate-nucleotide adenylyltransferase**	**2.7.7.18**	**Nicotinate and nicotinamide metabolism**
PF14_0697	Dihydroorotase	3.5.2.3	Pyrimidine metabolism
PFE0630c	Orotate phosphoribosyltransferase	2.4.2.10	Pyrimidine metabolism
MAL13P1_221	Aspartate carbamoyltransferase	2.1.3.2	Pyrimidine metabolism, asparagine and aspartate metabolism
MAL13P1_292	Riboflavin kinase	2.7.1.26	Riboflavin metabolism
PF11_0059	Pantothenate transporter	--	Transport
PF11_0169	Pyridoxine/pyridoxal 5-phosphate biosynthesis enzyme	--	Vitamin B6 metabolism

[a] In bold: nicotinate-nucleotide adenylyltransferase, which was selected for experimental validation.
[b] Enzymatic activities not annotated in the human databases.

We also extended the computational analysis of essential *Plasmodium* genes to pairs of genes that are not essential on their own, but are lethal if deleted simultaneously, i.e. synthetic lethal enzyme pairs with non-trivial genetic interactions [271] (see **Table 5.2**). In total, deletion of 16 gene pairs gave rise to such synthetic lethality in the unconstrained model. The enzymes that were essential upon double deletions participate in glycolysis, metabolism of nucleotides, lipids, porphyrin, the pentose phosphate cycle, and transport of NO_2 and phosphate (**Supplementary table 5.4**).

The analysis of essential genes in the *S. cerevisiae* metabolic network iND750 [218] can be used to put the results of the *P. falciparum in silico* deletions into an appropriate context. To make the proper comparison, we only considered deletions of genes carrying non-zero metabolic fluxes in wild type models of both networks; the focus on non-zero fluxes is necessary to prevent the difference in network sizes (*S. cerevisiae* 750 genes/1266 reactions, *P. falciparum* 366 genes/1001 reactions) from biasing the results. When both networks were allowed to simultaneously use all carbon sources the fraction of essential genes associated with non-zero fluxes in *P. falciparum* was 37%, while in *S. cerevisiae* it was 5% (Fisher's exact test *P*-value $<10^{-10}$). The fraction of essential genes was also significantly higher in the parasite when single carbon sources were used in both models. For example, when glucose was used as the sole source of carbon, 50% of genes were essential in the parasite versus 26% in yeast (*P*-value $=10^{-6}$). To understand whether the significantly higher fraction of essential genes in *P. falciparum* arises exclusively from a smaller number of isoenzymes in that network (111 in *P. falciparum* versus 276 in *S. cerevisiae*) or if it is also related to the inherent differences in the networks' architecture, we performed deletions of all isoenzymes associated with metabolic reactions, instead of individual gene deletions. As a result, when networks used all carbon sources, 39% of the reactions with non-zero fluxes were essential in *P. falciparum*, compared to only 6% in *S. cerevisiae* (*P*-value <10^{-10}). In a glucose minimal media, 62% of the reactions are essential in the parasite and 47% in yeast (*P*-value <10^{-10}). These results demonstrate that a significantly smaller genetic robustness of the parasite's network arises, at least in part, due to a paucity of alternative metabolic pathways [272]. The lower redundancy of the *P. falciparum* network is likely to be a consequence of the adaptation to the relatively homeostatic and nutrient-rich environments of the hosts in which it proliferates [200].

5.2.3 *Resolution of disagreements between* in silico *predictions and experimental data*

While the presented metabolic model achieves a high accuracy in predicting phenotypes of the experimental knockouts and drug inhibiton assays, it is the disagreement between the model and experiments that often leads to model improvement [51]. Thus, it is important to discuss the inconsistencies between modeling and experimental results which were corrected in the process of model construction. In four cases the disagreements between the predicted gene essentiality and experimental results reported in the literature were resolved through additional flux constraints. The adjustments included purine nucleoside phosphorylase (PFE0660c), adenosine deaminase (PF10_0289), ornithine decarboxylase (PF10_0322) and lactate dehydrogenase (PF13_0141/PF13_0144). The additional constraints applied to the network were based either on specific experimental conditions or known details of *Plasmodium spp*. physiology (Table III, Green rows). For example, lactate is believed to be the main byproduct of glucose metabolism in *P. falciparum* [273], and lactate dehydrogenase is an essential enzyme that regenerates NAD$^+$ from NADH [274]. In contrast, in our initial model pyruvate was exported as the primary glycolysis byproduct and NAD$^+$ was regenerated through the transformation of pyrroline-5-carboxylate to proline (EC: 1.5.1.2). Because there is no evidence of extensive pyruvate export in *Plasmodium*, a constraint was added to the corresponding pyruvate exchange reaction. As a result, we observed reduction of pyruvate to lactate, followed by lactate export. In the adjusted model lactate dehydrogenase carried a non-zero flux and was correctly predicted to be essential for growth. In the other three cases, constraints on exchange fluxes were added to reproduce the composition of the media used in drug inhibition experiments. Specifically, purine nucleoside phosphorylase has been shown to be essential if hypoxanthine is not available in the environment [252], while adenosine deaminase is essential in the absence of both inosine and hypoxanthine [251]. When the fluxes through the inosine and hypoxantine exchange reactions were set to zero, the experimentally observed knockout phenotypes were reproduced. Similarly, ornithine decarboxylase was correctly predicted to be essential without putrescine in the growth media [241].

In two cases, the inability of our initial model to reproduce experimental results was due to reactions involving metabolic dead-ends; i.e. metabolites that are either only produced or only consumed in the network [275]. In the first case the model was not able to synthesize spermidine. The synthesis of spermidine from putrescine by spermidine synthase was accompanied by the production of 5-methylthioadenosine (MTA) [244], which was a metabolic dead end in the initial model. Consequently, the spermidine synthesis caused MTA to accumulate, violating the steady state assumption of the constraint-based approach. However, while it is known that in *Plasmodium* spp. MTA is first converted to 5-methylthioinosine by adenosine deaminase and then recycled into methionine and hypoxanthine [276], not all enzymes involved in these reactions have been fully characterized. We addressed this problem by including in the model the PfADA and PfPNP activities, responsible for the hypoxanthine synthesis from MTA [276], and an additional hypothetical reaction which converts the resulting byproduct, 5-methylthioribose-1-

PO_4, to methionine in order to represent the methionine salvage pathway. The second case was related to the folate biosynthesis pathway. In this pathway, the reaction catalyzed by 6-pyruvoyltetrahydropterin synthase (4.2.3.12), in addition to the folate biosynthesis intermediate 6-hydroxymethyl-7,8-dihydropterin, is known to produce a small amount of 6-pyruvoyl-5,6,7,8-tetrahydropterin (6PTHP) [277]. Initially, both products were included in the model and, because 6PTHP represented another metabolic dead end, the cell was not able to synthesize folate. We resolved this problem by including in the model separate reactions for each of the two alternative products.

The total number of remaining dead end metabolites in the final model (266) is comparable to that of other recently published genome-scale metabolic networks; for example the iBsu1103 model for *Bacillus subtilis* (270 dead-end metabolites,[157]), the iAC560 model of *L. major* (261,[212]), or the iND750 *S. cerevisiae* model (194, [218]). The remaining dead-ends in the *Plasmodium* network include 109 metabolites that are consumed but are not currently produced or imported in the model, 80 metabolites that are produced but not consumed, and 79 metabolites, associated with reversible reactions, that can be either exclusively consumed or produced. Because protein synthesis is not explicitly included in the metabolic model, a significant number (42) of the remaining dead end compounds correspond to tRNAs; this compares to 68 dead end tRNAs in the iND750 yeast network. Among the other functional categories associated with a large number of the metabolic dead ends are lipid metabolism (45%), transport reactions (15%), the metabolism of carbohydrates (5%), amino acids (8%), and nucleotides (5%). More precise measurements of the *P. falciparum* biomass composition, for example a detailed lipid composition of the parasite membrane [278, 279], can be used in the future to significantly shrink the pool of the remaining dead end metabolites.

5.2.4. *Validation of the predicted drug target nicotinate nucleotide adenylyltransferase*

The most urgent motivation for flux-balance reconstruction of the pathogen metabolism is to facilitate drug development. To illustrate the potential of the model to make clinically relevant predictions, we experimentally tested a predicted target for which candidate drugs have been reported in other microbial species. The ideal drug target will be an essential enzyme with no homolog in the human genome and in a pathway not currently targeted by any pharmaceutical. Based on these criteria we selected for validation nicotinate mononucleotide adenylyltransferase (NMNAT, PlasmoDB ID PF13_0159, EC 2.7.7.18) (**Table 5.4**). This enzyme, a member of the plasmodial nicotinamide adenine dinucleotide (NAD) synthesis and recycling pathway, catalyzes the conversion of nicotinic acid mononucleotide to nicotinic acid adenine dinucleotide (**Fig. 5.2A**). NMNAT has recently been the focus of novel antimicrobial agent development due to structural and metabolic differences between the enzyme in microbial and human cells [280]. As NAD(P) is one of the most promiscuous redox molecules in metabolism, as well as a cofactor for important histone regulatory proteins such as sirtuins [281], inhibition of NAD(P) synthesis and recycling should have a profound impact on parasite metabolism. However, to the best of our knowledge, this pathway has not been previously targeted by pharmaceutical interventions in *P. falciparum.*

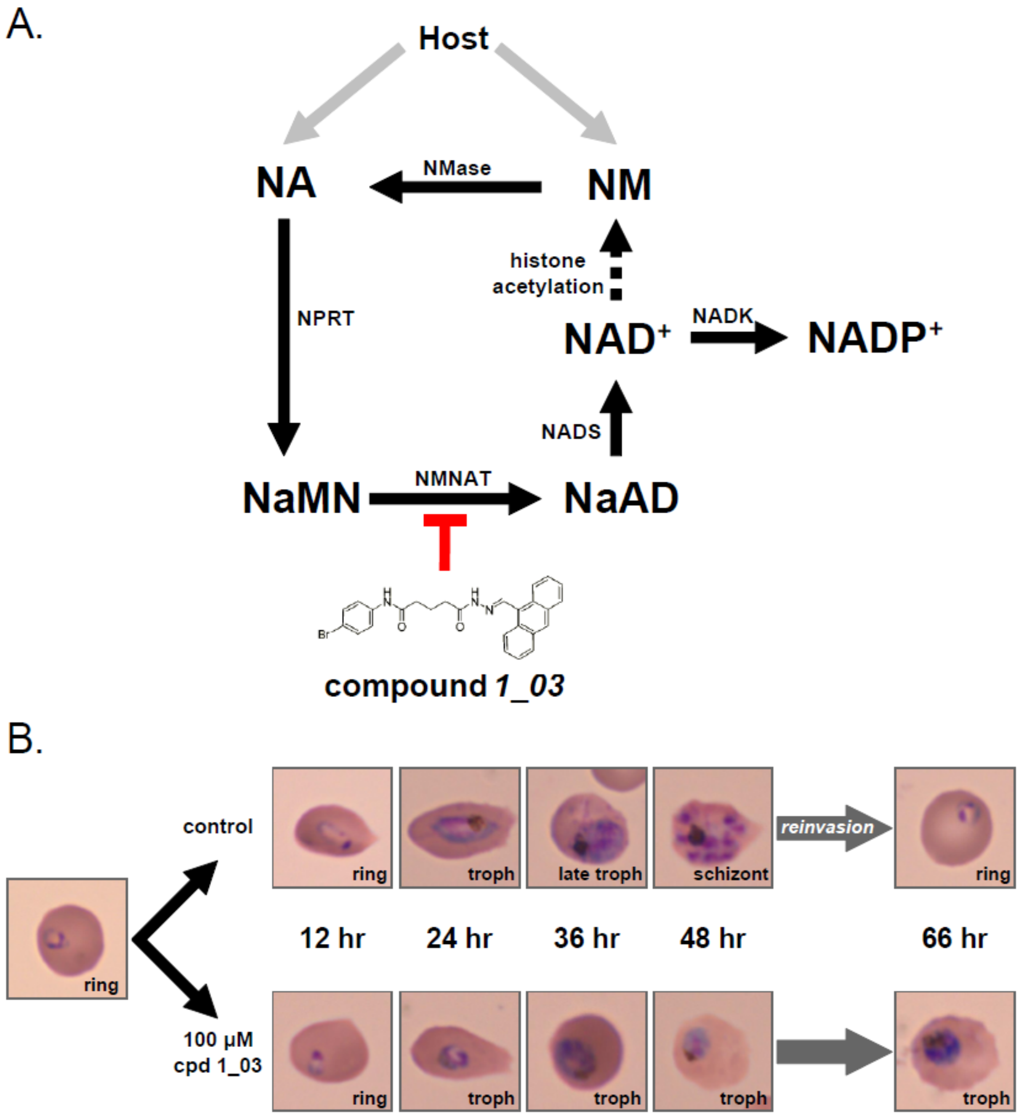


Figure 5.2. Small-molecule inhibition of the parasite nicotinate mononucleotide adenylyltransferase (NMNAT). A. Schematic of the *P. falciparum* NAD(P) synthesis and recycling pathway determined from the genome sequence. Nicotinamide (NM) and nicotinic acid (NA) can be scavenged from the host. Compound *1_03* is an inhibitor targeting NMNAT. B. Compound *1_03* causes growth arrest of intraerythrocytic *P. falciparum*. Cultures were resuspended in niacin-free medium containing 0 or 100 µM of compound *1_03* at early ring stage and observed for 66 hours (see Methods). Untreated parasites undergo normal development and reinvasion, while drug-treated parasites arrest at the trophozoite ("troph") stage and do not reinvade. Abbreviations: NM, nicotinamide; NA, nicotinic acid; NaMN, nicotinate mononucleotide; NaAD, nicotinate adenine dinucleotide; NAD(P)$^+$, nicotinamide adenine dinucleotide (phosphate), reduced; NMase, nicotinamidase; NPRT, nicotinate phosphoribosyltransferase; NMNAT, nicotinate mononucleotide adenylyltransferase; NADS, NAD synthase; NADK, NAD kinase.

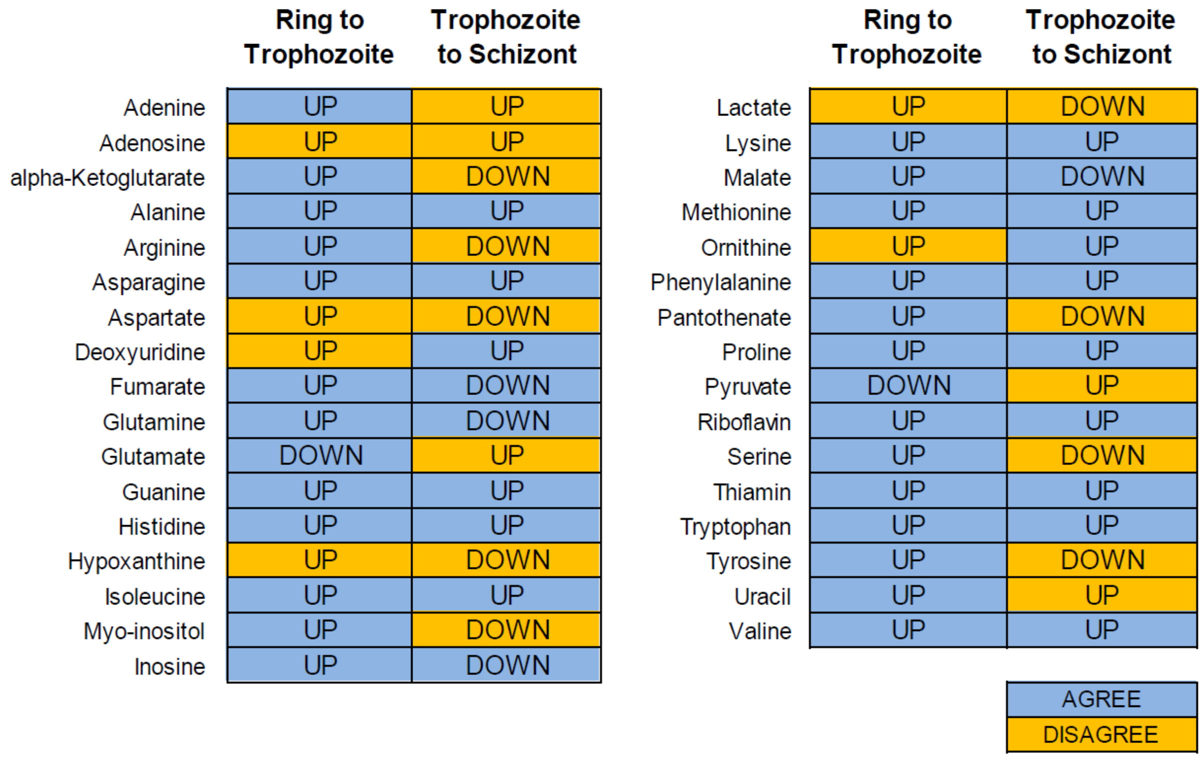
Recently, Sorci *et al.* used a combination of *in silico* structure modeling and enzyme inhibition assays to identify several classes of small molecules that inhibit bacterial (*E. coli* and *B. subtilis*) but not human NMNAT [282]. Several of these drug candidates were able to inhibit bacterial growth in culture. We tested two of the designed candidate compounds (*1_03* and *3_02*), representing two different chemotypes, for their ability to inhibit *P. falciparum* growth using both the SYBR Green I fluorescence assay [283] to measure DNA synthesis and microscopic examination of morphological effects. The two compounds were tested at a range of concentrations for growth inhibitory effects in nicotinamide-free culture medium so as not to rescue any metabolic blocks induced by the drugs. Nicotinamide removal has been previously shown not to affect normal growth in culture [284], which we confirmed before running our growth assay experiments (data not shown). While the compound *3_02* did not significantly affect parasite's growth at moderate concentrations (MIC_{50} > 100 μM), the compound *1_03* (N'-[(E)-anthracen-9-ylmethylideneamino]-N-(4-bromophenyl)pentanediamide) exerted an inhibitory effect in a growth assay (MIC_{50} = 50 μM, **Supplementary fig. 5.2**) comparable to that previously observed for bacteria (MIC_{50} > 80 μM for *E. coli*, MIC_{50} = 10 μM for *B. subtilis*). At 100 μM the compound ***1_03*** completely blocked host cell escape and reinvasion by arresting parasites in the trophozoite growth stage (**Fig. 5.2B**). Importantly, the human NMNAT isoforms are insensitive to the compound at least up to the concentrations used in our assay [282]. This suggests that the parasite NMNAT enzyme and, more generally, the NAD(P) synthesis pathway are indeed potentially effective and druggable targets. The experimental results also highlight the ability of our model to identify promising candidates for pharmaceutical intervention.

5.2.5. *Prediction of metabolite concentration changes based on expression data*

Genome-scale metabolic networks can be used not only to predict the effects of gene deletions, but also as a tool for integration of diverse genomic and physiological data [76, 171]. For example, information on nutrient availability, uptake rates, and maximal rates of internal reactions can be used to further constrain the space of feasible metabolic fluxes. To illustrate the ability of the model to combine genomic data we investigated whether available gene expression datasets can be used to predict shifts in concentrations of external metabolites caused by the *P. falciparum* exchange fluxes at different developmental stages. Investigation of the exchange fluxes is essential for understanding perturbations caused by parasitic infections in the metabolic state of their host tissues, and, consequently, main mechanisms of pathogenesis. Because the flux-balance analysis operates in the space of the fluxes and not in the space of metabolic concentrations, the model cannot be directly used to predict absolute concentration changes. Nevertheless, it is possible to use the model to investigate the direction of concentration changes for external metabolites, following the simple logic that an increase in the uptake rate or decrease in the excretion (output) rate should lead to a drop in the concentration of the corresponding external metabolite; similarly, an uptake rate decrease or excretion rate increase should increase the metabolite concentrations.

Although gene expression level does not perfectly correlate with the flux through the associated enzyme [285, 286], the recent study by Colin *et al.* [214] demonstrated that mRNA abundance data, if used as additional constraints on maximal reaction fluxes, can significantly improve stoichiometric model predictions. For instance, if the expression level of a particular enzyme is low, it is unlikely that the enzyme will be used by the metabolic network to carry a large flux. Consequently, it should be possible to use gene expression data to obtain a more accurate view of the *in vivo* metabolic state. To test this, we used DNA microarray results collected from synchronized cultures of the *P. falciparum* 3D7 strain during the red blood cell phase of the parasite's lifecycle [204, 208]. The expression data were collected at the ring, trophozoite, and schizont developmental stages (see Methods). Following Coljin *et al.*, the maximum flux allowed through enzymes was constrained proportionally to the relative expression level of the corresponding genes [214].

We compared the accuracy of our predictions to the experimentally measured metabolic changes in *Plasmodium*-infected red blood cells [208]. In **Figure 5.3** we show the predicted and experimentally measured changes, indicating either an increase or decrease in metabolic concentrations for the transition from the ring to trophozoite and from trophozoite to schizont stages. The predicted shifts in metabolic concentrations agree with the experimental results in 70% of the measurements (Binomial *P*-value $=9*10^{-4}$). In addition, we found a significant correlation between the magnitudes of the change in metabolite concentrations and the predicted flux values (Pearson's correlation: 0.34, *P*-value $=6*10^{-3}$, Spearman's correlation: 0.25, *P*-value = 0.04).



	Ring to Trophozoite	Trophozoite to Schizont
Adenine	🟦 UP	🟨 UP
Adenosine	🟨 UP	🟨 UP
alpha-Ketoglutarate	🟦 UP	🟨 DOWN
Alanine	🟦 UP	🟦 UP
Arginine	🟦 UP	🟨 DOWN
Asparagine	🟦 UP	🟦 UP
Aspartate	🟨 UP	🟨 DOWN
Deoxyuridine	🟨 UP	🟦 UP
Fumarate	🟦 UP	🟦 DOWN
Glutamine	🟦 UP	🟦 DOWN
Glutamate	🟦 DOWN	🟨 UP
Guanine	🟦 UP	🟦 UP
Histidine	🟦 UP	🟦 UP
Hypoxanthine	🟨 UP	🟨 DOWN
Isoleucine	🟦 UP	🟦 UP
Myo-inositol	🟦 UP	🟨 DOWN
Inosine	🟦 UP	🟦 DOWN

	Ring to Trophozoite	Trophozoite to Schizont
Lactate	🟨 UP	🟨 DOWN
Lysine	🟦 UP	🟦 UP
Malate	🟦 UP	🟦 DOWN
Methionine	🟦 UP	🟦 UP
Ornithine	🟨 UP	🟦 UP
Phenylalanine	🟦 UP	🟦 UP
Pantothenate	🟦 UP	🟨 DOWN
Proline	🟦 UP	🟦 UP
Pyruvate	🟦 DOWN	🟨 UP
Riboflavin	🟦 UP	🟦 UP
Serine	🟦 UP	🟨 DOWN
Thiamin	🟦 UP	🟦 UP
Tryptophan	🟦 UP	🟦 UP
Tyrosine	🟦 UP	🟨 DOWN
Uracil	🟦 UP	🟨 UP
Valine	🟦 UP	🟦 UP

🟦 AGREE
🟨 DISAGREE

Figure 5.3. Comparison between the predicted and experimentally measured shifts in metabolite concentrations in infected red blood cells. UP/DOWN indicates direction of experimentally measured changes in metabolic concentrations in infected versus uninfected cells. Blue color indicates agreement between experiment and predictions, while yellow indicates disagreement. In most cases (70%, *P*-value =9*10^{-4}) the shifts in metabolic concentrations from one stage to the next can be predicted based on changes in the *P. falciparum* metabolic exchange fluxes. The *in silico* predictions of exchange fluxes were made based on the expression-constrained flux-balance analysis [214]. Briefly, for genes with available mRNA expression data, the maximum flux through the associated metabolic reactions was constrained proportionally to their expression level; with the highest expression value corresponding to the maximum allowed flux.

5.3. Discussion

The presented flux-balanced model can serve as a valuable tool for quantitative predictions of *P. falciparum* metabolic states under various growth conditions and perturbations.

The results of *in silico* gene deletions demonstrate that the model achieves high accuracy in reproducing available experimental measurements. In addition, our analysis suggests several dozen essential metabolic targets for therapeutic intervention. Although several studies which assemble and analyze plasmodial metabolic pathways have been performed previously [209, 210, 287, 288], our contribution is important because the genome-scale model can be used to investigate and predict genetic perturbations from a network-level perspective.

Interestingly, our results suggest a limited degree of robustness in the *P. falciparum* network, which should lead to a relatively high success rate for inhibitors of metabolic genes. It is possible that the small robustness of the reconstructed model, for example in comparison with the yeast metabolic network, is due primarily to unannotated *P. falciparum* genes without significant homology to known enzymes in other organisms. To investigate this possibility further we used the available collection of single metabolic gene knockouts/inhibitions in *P. falciparum* or *P. berghei* (**Table 5.3**) and all metabolic gene knockouts in *S. cerevisiae* [289]. We calculated the fraction of orthologs for essential metabolic knockouts in the parasite which are also essential in yeast, and, *vice versa,* the fraction of orthologs for essential metabolic knockouts in yeast which are essential in the parasite (see **Supplementary table 5.5**). Interestingly, while the majority of orthologs for essential metabolic genes in *P. falciparum* are not essential in *S. cerevisiae*, about half of the essential metabolic genes in yeast are also essential in the parasite (**Supplementary table 5.5**, Fisher's exact test P=0.04). This result independently supports the conclusion of the flux balance simulations about the relatively small robustness of the *P. falciparum* network.

We anticipate several immediate extensions of our work. First, the presented network can be used for effective integration of multiple genomic data types. For example, known regulatory interactions can be incorporated into the model [290]. Accurate measurements of gene expression [202], key protein-DNA regulatory interactions [291], and post-translational modifications [292] in the parasite will be especially important for modeling the dynamic behavior of the network under varying environmental conditions. Second, it will be important to model exchanges and interactions between the metabolic networks of the parasite and its hosts. The analysis of the combined parasite-host metabolic network should significantly improve understanding of the *P. falciparum* vulnerabilities. For example, several host cell enzymes are actively used by the parasite during its life cycle [276, 293]. Although we did not consider these human enzymes in our analysis, they can be easily included in future applications of the model. The available global flux-balanced metabolic human network [294], metabolic network specifically active in the liver [295], and well-curated models of human red blood cell metabolism [296-300] make such combined analyses possible. Third, it will be interesting to reconstruct stoichiometric metabolic networks for other clinically relevant *Plasmodium* species (e.g. *P. vivax, P. malariae, P. ovale*, and *P. knowlesi*) as well as the important model species *P. berghei* and *P. yoelii*. The comparative analysis of these networks may reveal important physiological and evolutionary differences between *Plasmodium* spp., and also help in the identification of common metabolic drug targets.

Taking into account the global health burden of malaria, it is essential to develop and implement new effective pharmaceuticals as quickly as possible. Systems biology approaches can be used to significantly facilitate drug identification and development [301, 302]. To date, we have only begun to see the application of such integrative methods in the context of malaria research (reviewed in [198]). We believe that the presented network represents an important step in this direction. The experimental validation of a candidate drug, compound *1_03*, targeting the parasite nicotinate mononucleotide adenylyltransferase illustrates the ability of the model to speed up development of novel anti-malaria targets and pharmaceuticals. Importantly, although the identified compound is available, it has not been previously tested against *Plasmodium spp*. While *1_03* inhibited parasite growth only at relatively high concentrations (MIC_{50} = 50 μM), these were comparable to the inhibitory concentrations for the bacteria against which the drug was initially developed [282]. The incomplete growth inhibition at lower compound concentrations could be explained by the incomplete drug inhibition. Our model predicts linear decrease in the *P. falciparum* biomass production as the level of NMNAT inhibition increases; for example, 90% inhibition results in 90% growth decrease. Since Sorci *et al*. initially screened for compounds that could selectively inhibit pure NMNAT enzyme, these compounds have not been optimized for cellular permeability, accumulation or other pharmacokinetic parameters, and thus should primarily serve as a structural basis for further malaria drug development.

Future improvements to the reconstructed *P. falciparum* metabolic network, including adding experimental details for missing activities and precise metabolic measurements necessary to describe the growth-related objective function, will lead to a better understanding of parasite physiology. Ultimately, such models should significantly accelerate the identification of desperately needed new drug leads against this devastating disease.

5.4. Methods

5.4.1. *Genome-scale metabolic reconstruction*

The reconstruction process started with the identification of enzyme coding genes in the *P. falciparum* genome. We considered a variety of resources, including PlasmoDB [303], the Malaria Parasite Metabolic Pathway Database (MPMP) [287], PlasmoCyc [209], and the Kyoto

Encyclopedia of Genes and Genomes (KEGG) [131]. The identified enzymes were mapped to the corresponding metabolic reactions by consulting several well studied metabolic models, including the iAF1260 model for *E. coli* [304], the iND750 model for *S. cerevisiae* [218], the genome scale human metabolic network by Duarte *et al.* [294], and the KEGG database [131]. Based on this set of enzymatic activities and their stoichiometry, we used flux balance analysis to see if the network was able to produce a set of basic biomass components; e.g. aminoacids, lipids, nucleotides and cofactors. For each metabolite that the network was unable to synthesize we searched the literature for relevant publications concerning pathways and genes associated with the metabolite production or transport. Network enzymes were assigned to different cellular compartments based on experimental evidence, when available, and on computationally predicted localization information [217, 269, 305, 306]. Transport and exchange reactions reported in the literature or in databases such as PlasmoDB and MPMP were initially included in the model. We added additional transport reactions required for production of the biomass components. All metabolic and transport reactions were used to formulate a stoichiometric flux balance model [307]. The model was improved following an iterative procedure as previously described [51, 216].

The assembled network was manually inspected and compared to the Malaria Parasite Metabolic Pathway Database [210]. Metabolic network gaps [31] were identified and included in the assembled network model [308, 309]. The reactions for which no literature support is available and which are not essential for the biomass production were removed from the network. Additional adjustments related to reaction directionalities and metabolite availabilities were made following the computational analyses described in this work.

Because the *P. falciparum* biomass objective function cannot be completely established based on the available literature, in our calculations we used a modified version of the yeast objective function reported in the iND750 model [218]. The objective function modifications included the lipid composition, which was adjusted as reported for *Plasmodium* [279], and amino acid and nucleotide compositions adjusted based on the proteome and genome sequences weighted by available expression data [204]. In particular, the percent prevalence of each ribonucleotide and aminoacid across all open reading frames (ORFs) was calculated as the relative frequency of each monomer; the counts at each ORF were multiplied by the ORF's expression level (when available). The percent prevalence of the dNTPs was derived from the genome A+T content of 80.6%. These percentages were converted to mmol/gDW as described [212]. Systems Biology Research Tool [310] was used to perform flux balance analysis [307] of the network, including single and double *in silico* deletions of network enzymes. The reconstructed network was able to either synthesize or import all the biomass components presented in **Supplementary Table 5.1**. The assembled metabolic model is available in the Systems Biology Markup Language format from the BioModels database [311] with accession number MODEL1007060000.

5.4.2. *Parasite culture, growth, and drug inhibition assays*

The cultures of *P. falciparum* were maintained and synchronized by standard methods [312, 313]. Briefly, red blood cells (RBCs), infected by *P. falciparum* (3D7 strain), were grown in the RPMI 1640 culture medium supplemented with sodium carbonate (2 mg/mL), hypoxanthine (100 μM), Albumax II (0.25%) and gentamycin (50 μg/mL) in a humidified incubator at 5% CO_2, 6% O_2 and 37° C. The growth synchronizations were carried out by incubating parasite-infected RBCs in phosphate-buffered saline (PBS) containing 5% w/v sorbitol for 5 minutes at room temperature, washing once with sorbitol-free PBS, and resuspending in culture medium.

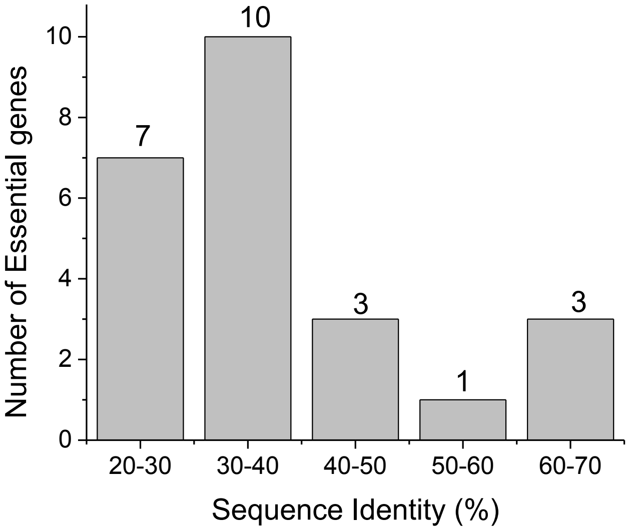
The compounds *1_03* and *3_02* were acquired from ChemDiv (http://chemdiv.emolecules.com; ChemDiv IDs 8003-6329 and 5350-0029, respectively) and resuspended at 100 mM in DMSO. Growth inhibition studies were carried out using the SYBR Green I fluorescence assay [283]. Briefly, synchronized parasite cultures (early ring stage, 1% parasitemia, 1% hematocrit, 100 μL total volume) were suspended in nicotinamide-free RPMI 1640 containing 0.1% DMSO and differing concentrations of drug in 96-well plates. After 72 hour incubation the plates were frozen at -80° C overnight, then thawed and mixed with 100 μL lysis buffer (20 mM Tris-HCl, pH 7.5; 5 mM EDTA; 0.008% w/v saponin; 0.08% v/v Triton X-100; 1 × SYBR Green I) per well, incubated 1 hour at room temperature and quantified using a BioTek SynergyMX plate reader (excitation 488 nm, emission 522 nm). The concentrations 0, 0.1, 1, 5, 10, 50, 100 and 250 μM were tested in triplicate in two independent growth assays.

5.4.3. *Using expression-constrained network to predict shifts in external metabolic concentrations*

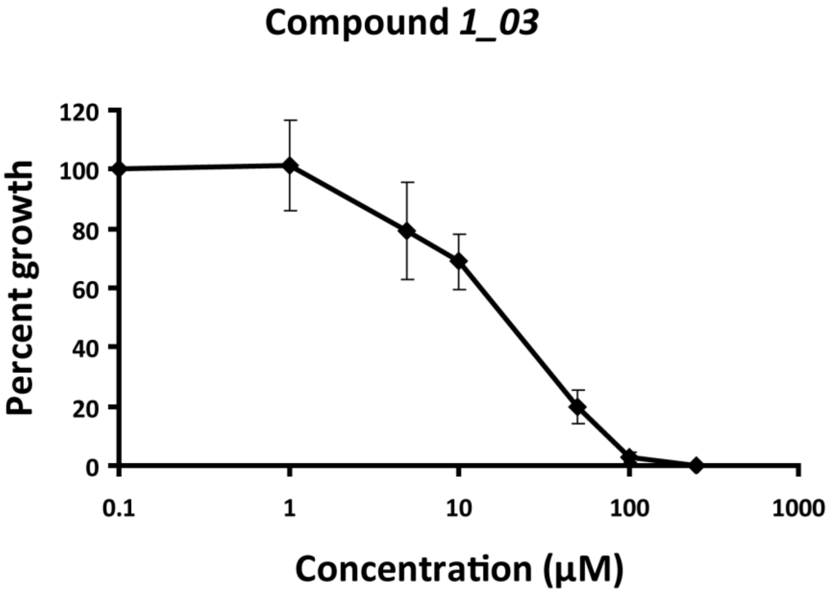
In the expression-constrained flux analysis we used *P. falciparum* expression data [208] as described previously in Coljin *et al.* [214]. Specifically, for genes with available mRNA expression data, the maximum flux through the associated metabolic reactions was constrained proportionally to their expression level; with the highest expression value corresponding to the maximum possible metabolic flux. In order to obtain absolute expression values, rather than the ratios between the microarray intensities at a given time point and those of a pooled sample, we multiplied each ratio with the average sum of the median intensity across the full intraerythrocytic developmental cycle [204].

The ring, throphozoite, and schizont developmental stages in cultures were defined for hours 1-18, 19-30, and 31-48 after synchronization with D-sorbitol, respectively. In the analysis we used intracellular RBC metabolite concentrations data obtained by Olszewski *et al.* [208]. We compared the changes in metabolite abundances between infected and uninfected red blood cells, from one development stage to the next. For each metabolite the predicted concentration changes were considered to agree with experimental data, if the metabolite consumption in the network increased (or the metabolite production decreased) when the experimentally measured metabolite concentration also decreased, or alternatively, if consumption of the metabolite decreased (or production increased) when the metabolite concentration increased.

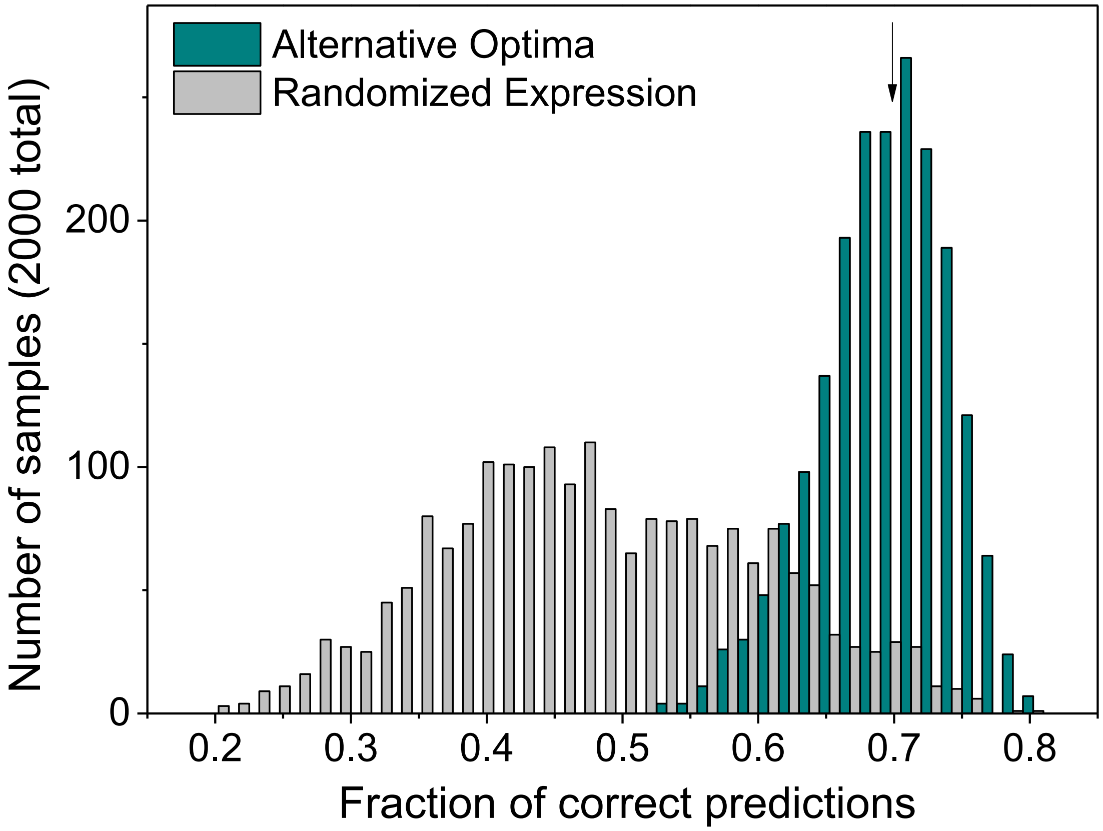
5.5 Supplementary figures and tables



Supplementary figure 5.1. Distribution of sequence identities to human transcripts for *Plasmodium* proteins predicted as essential for growth.



Supplementary figure 5.2. Growth inhibition profile of compound ***1_03*** in *in vitro* cultures of *P. falciparum*. See Methods for experimental details. Error bars represent the standard deviation of the mean based on triplicate experiments.



Supplementary figure 5.3. The accuracy of the metabolite exchange predictions made using 2000 trials with randomized expression data (the distribution in grey), and prediction based on sampling of 2000 alternative FBA optimal solutions (the distribution in cyan). The single prediction discussed in the paper (70% accuracy) is shown using a vertical black arrow. In the randomized trials expression values were randomly shuffled between parasite metabolic genes. Only in about 2% of the randomized trials the prediction accuracy was higher than the average accuracy obtained using the true (non-shuffled) expression values. Alternative optima were obtained using the artificial centering hit-and-run algorithm [314], implemented in the COBRA toolbox [127], with biomass fluxes fixed to their maximum value. The difference in the distributions is highly statistically significant (Mann-Whitney U *P*-value $<10^{-10}$).

Supplementary table 5.1. Biomass components and cofactors which can be synthesized or imported by the reconstructed metabolic network of *P. falciparum*. Coefficients indicate mmol/gDW

Amino acids	
Alanine	0.0885
Arginine	0.0876
Asparagine	0.4053
Aspartate	0.1943
Cysteine	0.0438
Glutamine	0.0879
Glutamate	0.2275
Glycine	0.1186
Histidine	0.0690
Isoleucine	0.2314
Leucine	0.2210
Lysine	0.3325
Methionine	0.0662
Phenylalanine	0.1183
Proline	0.0704
Serine	0.1982
Threonine	0.1327
Tryptophan	0.0136
Tyrosine	0.1440
Valine	0.1318
Ribonucleotides	
ATP	0.1475
CTP	0.0320
GTP	0.0456
UTP	0.1004
Deoxyribonucleotides	
dATP	2.12E-03
dCTP	4.97E-04
dGTP	4.97E-04
dTTP	2.12E-03
Cofactors	
10-Formyltetrahydrofolate	2.23E-04
2-Octaprenyl-6-hydroxyphenol	2.23E-04
Coenzyme A	5.76E-04
Flavin adenine dinucleotide	2.23E-04
5,10-Methylenetetrahydrofolate	2.23E-04
Nicotinamide adenine dinucleotide	1.83E-03
Nicotinamide adenine dinucleotide phosphate	4.47E-04
Protoheme	2.23E-04
Pyridoxal 5-phosphate	2.23E-04
Riboflavin	2.23E-04
5,6,7,8-Tetrahydrofolate	2.23E-04
Thiamine diphosphate	2.23E-04
Polyamines	
Putrescine	0.0350
Spermidine	0.0070
Lipids	
Phosphatidylethanolamine	2.88E-04
Phosphatidylcholine	5.43E-04
Cholesterol	1.55E-02
Others	
S-Adenosyl-L-methionine	2.23E-04
Fe^{2+}	7.11E-03
Fe^{3+}	7.11E-03
Ammonium	1.18E-02
SO_4	3.95E-03
Water	59.8100

Supplementary table 5.2. Single deletions displaying a phenotype in the unconstrained metabolic network of *P. falciparum.* L: Lethal, GR: Growth Reducing, SGR: Slight Growth Reducing

Gene	Enzyme name	EC Number	Pred.
MAL13P1_186	1-deoxy-D-xylulose-5-phosphate synthase	2.2.1.7	L
MAL13P1_221	1-deoxy-D-xylulose-5-phosphate synthase	2.1.3.2	L
MAL13P1_292	FAD synthetase, riboflavin kinase	2.7.7.2, 2.7.1.26	L
MAL13P1_326	Ferrochelatase	4.99.1.1	L
MAL8P1_81	Phosphopantothenoylcysteine decarboxylase	4.1.1.36	L
PF07_0018	Pantetheine-phosphate adenylyltransferase	2.7.7.3	L
PF08_0095	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase, dihydropteroate synthase	2.7.6.3, 2.5.1.15	L
PF10_0121	Hypoxanthine phosphoribosyltransferase	2.4.2.8	L
PF10_0155	Phosphopyruvate hydratase	4.2.1.11	L
PF10_0221	(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase	1.17.7.1	L
PF10_0225	Orotidine-5'-phosphate decarboxylase	4.1.1.23	L
PF10_0275	Protoporphyrinogen oxidase	1.3.3.4	L
PF10_0363	Pyruvate kinase	2.7.1.40	L
PF11_0059	Pantothenate transporter	--	L
PF11_0169	Pyridoxine/pyridoxal 5-phosphate biosynthesis enzyme	--	L
PF11_0295	Geranyltranstransferase, dimethylallyltranstransferase	2.5.1.10, 2.5.1.1	L
PF11_0410	Carbonate dehydratase	4.2.1.1	L
PF11_0436	Coproporphyrinogen oxidase	1.3.3.3	L
PF13_0044	Carbamoyl-phosphate synthase (ammonia), carbamoyl-phosphate synthase (glutamine-hydrolysing)	6.3.4.16, 6.3.5.5	L
PF13_0140	Dihydrofolate synthase, tetrahydrofolate synthase	6.3.2.12, 6.3.2.17	L
PF13_0159	Nicotinate-nucleotide adenylyltransferase	2.7.7.18	L
PF13_0287	Adenylosuccinate synthase	6.3.4.4	L
PF14_0415	Dephospho-CoA kinase	2.7.1.24	L
PF14_0598	Glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)	1.2.1.12	L
PF14_0641	1-deoxy-D-xylulose-5-phosphate reductoisomerase	1.1.1.267	L
PF14_0697	Dihydroorotase	3.5.2.3	L
PFA0225w	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	1.17.1.2	L
PFA0340w	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	2.7.7.60	L
PFB0200c	Aspartate transaminase, tyrosine transaminase, phenylalanine(histidine) transaminase	2.6.1.1, 2.6.1.5, 2.6.1.58	L
PFB0280w	Shikimate kinase, 3-phosphoshikimate 1-carboxyvinyltransferase	2.7.1.71, 2.5.1.19	L
PFB0295w	Adenylosuccinate lyase	4.3.2.2	L
PFB0420w	Adenylate cyclase, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	4.6.1.1, 4.6.1.12	L
PFC0831w	Triose-phosphate isomerase	5.3.1.1	L
PFD0830w	Thymidylate synthase, dihydrofolate reductase	2.1.1.45, 1.5.1.3	L
PFE0150c	4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase	2.7.1.148	L

PFE0410w	Dihydroxyacetone phosphate transporter (apicoplast)	--	L
PFE0630c	Orotate phosphoribosyltransferase	2.4.2.10	L
PFE1510c	Phosphoenolpyruvate transporter (apicoplast)	--	L
PFF0160c	Dihydroorotate oxidase	1.3.3.1	L
PFF0360w	Uroporphyrinogen decarboxylase	4.1.1.37	L
PFF0370w	3-octaprenyl-4-hydroxybenzoate carboxy-lyase	4.1.1.-	L
PFF0450c	Iron transporter	--	L
PFF0530w	Transketolase	2.2.1.1	L
PFF1105c	Chorismate synthase	4.2.3.5	L
PFF1410c	Nicotinate phosphoribosyltransferase	2.4.2.11	L
PFF1490w	Methylenetetrahydrofolate dehydrogenase (NADP+), methenyltetrahydrofolate cyclohydrolase	1.5.1.5, 3.5.4.9	L
PFI1090w	Methionine adenosyltransferase	2.5.1.6	L
PFI1100w	Aminodeoxychorismate synthase, aminodeoxychorismate lyase	2.6.1.85, 4.1.3.38	L
PFI1105w	Phosphoglycerate kinase	2.7.2.3	L
PFI1170c	Thioredoxin-disulfide reductase	1.8.1.9	L
PFI1195c	Thiamine diphosphokinase	2.7.6.2	L
PFI1310w	NAD+ synthase (glutamine-hydrolysing)	6.3.5.1	L
PFI1420w	Guanylate kinase	2.7.4.8	L
PFL2210w	5-aminolevulinate synthase	2.3.1.37	L
PFL2465c	dTMP kinase	2.7.4.9	L
PF14_0378	Triose-phosphate isomerase	5.3.1.1	GR
PF14_0425	Fructose-bisphosphate aldolase	4.1.2.13	GR
PFL0960w	Ribulose-phosphate 3-epimerase	5.1.3.1	GR
MAL13P1_206	Phosphate transporter (citoplasm)	--	SGR
MAL13P1_82	CDP-diacylglycerol-inositol 3-phosphatidyltransferase	2.7.8.11	SGR
MAL13P1_86	Choline-phosphate cytidylyltransferase	2.7.7.15	SGR
PF10_0122	Phosphoglucomutase	5.4.2.2	SGR
PF13_0259	dCTP deaminase	3.5.4.13	SGR
PF14_0097	phosphatidate cytidylyltransferase	2.7.7.41	SGR
PF14_0341	Glucose-6-phosphate isomerase	5.3.1.9	SGR
PF14_0511	6-phosphogluconolactonase, glucose-6-phosphate dehydrogenase	3.1.1.31, 1.1.1.49	SGR
PF14_0520	phosphogluconate dehydrogenase (decarboxylating)	1.1.1.44	SGR
PFE0660c	Purine-nucleoside phosphorylase	2.4.2.1	SGR
PFF1215w	Sphingomyelin synthase	2.7.8.27	SGR
PFF1375c	Ethanolaminephosphotransferase, diacylglycerol cholinephosphotransferase	2.7.8.1, 2.7.8.2	SGR
PFL1870c	Sphingomyelin phosphodiesterase	3.1.4.12	SGR

Suppelementary table 5.3. Single deletions of orphan metabolic activities predicted to be lethal in the unconstrained metabolic network of *P. falciparum*

Reaction Name	Enzyme	EC Number	Biological Process
R_NADK	NAD+ kinase	2.7.1.23	Nicotinate and nicotinamide metabolism
R_UPP3S_ap	Uroporphyrinogen-III synthase	4.2.1.75	Porphyrin metabolism
R_DDPA	3-deoxy-7-phosphoheptulonate synthase	2.5.1.54	Shikimate biosynthesis
R_DHQS	3-dehydroquinate synthase	4.2.3.4	Shikimate biosynthesis
R_DHQTi	3-dehydroquinate dehydratase	4.2.1.10	Shikimate biosynthesis
R_SHK3Dr	Shikimate dehydrogenase	1.1.1.25	Shikimate biosynthesis
R_CHRPL	Chorismate lyase	4.1.3.40	Ubiquinone metabolism
R_OPHBDC_mt	3-octaprenyl-4-hydroxybenzoate carboxy-lyase	4.1.1.-	Ubiquinone metabolism
R_OPHHX_mt	2-octaprenylphenol hydroxylase	1.14.13.-	Ubiquinone metabolism

Supplementary table 5.4. Double deletions predicted to be lethal in the unconstrained metabolic network of *P. falciparum.* *Isoenzyme pairs

Pair	Gene Name	Enzyme Name	EC Number	Biological Process
1*	PF11_0036	Phosphopantothenate-cysteine ligase	6.3.2.5	CoA biosynthesis
	PFD0610w	Phosphopantothenate-cysteine ligase	6.3.2.5	CoA biosynthesis
2*	PF11_0208	Phosphoglycerate mutase	5.4.2.1	Glycolysis
	PFD0660w	Phosphoglycerate mutase	5.4.2.1	Glycolysis
3	PF14_0378	Triose-phosphate isomerase	5.3.1.1	Glycolysis, isoprenoid metabolism
	PFL0960w	Ribulose-phosphate 3-epimerase	5.1.3.1	Pentose phosphate cycle
4	PF10_0122	Phosphoglucomutase	5.4.2.2	Glycolysis, pentose phosphate cycle
	PFE0730c	Ribose-5-phosphate isomerase	5.3.1.6	Pentose phosphate cycle
5*	PF13_0143	Ribose-phosphate diphosphokinase	2.7.6.1	Pentose phosphate cycle
	PF11_0157	Ribose-phosphate diphosphokinase	2.7.6.1	Pentose phosphate cycle
6	PFE0660c	Purine-nucleoside phosphorylase	2.4.2.1	Purine, methionine and polyamine metabolism
	PFE0730c	Ribose-5-phosphate isomerase	5.3.1.6	Pentose phosphate cycle
7*	PF10_0086	Adenylate kinase	2.7.4.3	Purine metabolism
	PFD0755c	Adenylate kinase	2.7.4.3	Purine metabolism
8*	PF14_0541	Inorganic diphosphatase	3.6.1.1	Purine, terpenoid metabolism
	PFC0710w	Inorganic diphosphatase	3.6.1.1	Purine, terpenoid metabolism
9	MAL13P1_206	Phosphate transporter	--	Transport

	PFL0305c	5'-nucleotidase	3.1.3.5	Purine metabolism
10*	PF13_0349	Nucleoside-diphosphate kinase	2.7.4.6	Pyrimidine, purine, dolichol metabolism
	PFF0275c	Nucleoside-diphosphate kinase	2.7.4.6	Pyrimidine, purine, dolichol metabolism
11	MAL13P1_82	CDP-diacylglycerol-inositol 3-phosphatidyltransferase	2.7.8.11	Inositol phosphate metabolism
	PF14_0100	CTP synthase	6.3.4.2	Pyrimidine metabolism
12	MAL13P1_82	CDP-diacylglycerol-inositol 3-phosphatidyltransferase	2.7.8.11	Inositol phosphate metabolism
	MAL13P1_206	Phosphate transporter	--	Transport
13	PF14_0097	Phosphatidate cytidylyltransferase	2.7.7.41	Phosphatidylethanolamine, phosphatidylserine metabolism
	PF14_0100	CTP synthase	6.3.4.2	Pyrimidine metabolism
14	MAL13P1_206	Phosphate transporter	--	Transport (Pi)
	PF14_0097	Phosphatidate cytidylyltransferase	2.7.7.41	Phosphatidylethanolamine, phosphatidylserine Metabolism
15	PFC0725c	NO_2 transporter	--	Transport
	PFI0735c	NADH dehydrogenase (ubiquinone)	1.6.5.3	Mitochondrial electron flow
16	PF10_0334	Succinate dehydrogenase (ubiquinone)	1.3.5.1	Mitochondrial electron flow
	PFC0725c	NO_2 transporter	--	Transport

Supplementary table 5.5. Essentiality in *S. cerevisiae* of the orthologs for essential and non-essential *Plasmodium* metabolic genes. While all of the orthologs for essential metabolic genes in yeast are also essential in *Plasmodium*, only about half of the orthologs for *Plasmodium* essential metabolic genes are essential in yeast. Fisher's exact test *P*-value =0.04.

	Yeast essential	Yeast non-essential	**Total**
Plasmodium essential	6	4	10
Plasmodium non-essential	0	5	5
Total	6	9	15

Chapter 6.

THE RATE OF THE MOLECULAR CLOCK AND THE COST OF GRATUITOUS PROTEIN SYNTHESIS

This chapter is based on “Plata, G., Gottesman, M., Vitkup, D. **The Rate of the Molecular Clock and the Cost of Gratuitous Protein Synthesis**. Genome Biol. (9):R98, (2010)”

The nature of the protein molecular clock, the protein-specific rate of amino acid substitutions, is among the central questions of molecular evolution. Protein expression level is the dominant determinant of the clock rate in a number of organisms. It has been suggested that highly expressed proteins evolve slowly in all species mainly to maintain robustness to translation errors that generate toxic misfolded proteins. Here we investigate this hypothesis experimentally by comparing the growth rate of *E. coli* expressing wild type and misfolding-prone variants of the LacZ protein. We show that the cost of toxic protein misfolding is small compared to other costs associated with protein synthesis. Complementary computational analyses demonstrate that there is also a relatively weaker, but statistically significant, selection for increasing solubility and polarity in highly expressed *E. coli* proteins. Although we cannot rule out the possibility that selection against misfolding toxicity significantly affects the protein clock in species other than *E. coli*, our results suggest that it is unlikely to be the dominant and universal factor determining the clock rate in all organisms. We find that in this bacterium other costs associated with protein synthesis are likely to play an important role. Interestingly, our experiments also suggest significant costs associated with volume effects, such as jamming of the cellular environment with unnecessary proteins.

6.1. Introduction

Once the first protein sequences became available, their comparison led to the conclusion that the number of accumulated substitutions between orthologs was mainly a function of the evolutionary time elapsed since the last common ancestor of corresponding species [65, 315]. Consequently, orthologous proteins accumulate substitutions at approximately constant rate over long evolutionary intervals. This observation suggests that one can use available protein sequences as a molecular clock to estimate divergence times between different species [316]. Further studies revealed that while the pace of the molecular clock is similar for orthologous proteins in different lineages, it varies by several orders of magnitude across non-orthologous proteins [317, 318].

For several decades the dominant hypothesis explaining the large variability of the molecular clock rate between non-orthologous proteins was based on the concept of functional protein density: the higher the fraction of protein residues directly involved in its function, the slower the protein molecular clock [319, 320]. It was not until high-throughput genomics data became widely available that multiple molecular and genetic variables were used to investigate the dominant factors influencing the molecular clock rates of different proteins. Surprisingly, such features as gene essentiality [321-324], the number of protein-protein interactions [72, 325], and specific functional roles [70, 326], have been shown to have, on average, either non-significant, or significant but relatively weak correlations with protein evolutionary rates. On the other hand, quantities directly related to gene expression, such as codon bias, mRNA expression, and protein abundance, showed the strongest correlation with the rate of protein evolution [73, 327]. For example, expression alone explains about a third of the variance in the substitution rates in several microbial species [73, 326, 328] and about a quarter of the variance in *C. elegans*

[329]. In these and many other organisms highly expressed genes accept significantly less synonymous and non-synonymous (amino acid changing) substitutions than genes with with low expression levels [330].

Considering the major role played by expression in setting the rate of amino acid substitutions, it is important to understand the main molecular mechanisms of this effect [71]. A popular theory by Drummond *et al.* [328, 331, 332] suggests that highly expressed proteins may evolve slowly in all organisms, from microbes to human [331], due to the selection against toxicity associated with protein misfolding. The logic behind this interesting hypothesis is that a significant fraction (>10%) of cellular proteins may contain translation errors [333, 334] that could cause cytotoxic protein misfolding. If misfolded proteins indeed incur substantial toxicity costs, greater pressure to avoid misfolding will affect highly expressed genes since they generate relatively more misfolded proteins [328]. Consequently, adaptive pressure will maintain sequences of highly expressed proteins robust to translation errors, which will in turn slow the amino acid substitution rate, i.e. the protein molecular clock. The misfolding toxicity hypothesis was supported by the results of computer simulations [331], but to the best of our knowledge, it has never been tested experimentaly.

In this study we specifically investigated whether the toxicity of misfolded proteins or other costs associated with protein synthesis make a dominant contribution to cellular fitness (growth rate), and consequently constrain the molecular clock in *E. coli*. To test this, we used wild type (WT) and misfolding-prone variants of the *E. coli* β-galactosidase gene, *lacZ*. We also computationally analyzed the contribution of other related factors, such as protein stability and solubility.

6.2. Results

The native biological function of the LacZ protein is to cleave lactose for use as a source of carbon and energy [335]; in the absence of lactose, β -galactosidase does not participate in the *E. coli* carbon metabolism. Therefore, we used *lacZ* expression in a lactose-free medium to measure the cost of gratuitous protein expression [336, 337]. To compare that expression cost to the cost of potentially toxic protein misfolding, we used site-directed mutagenesis to engineer several destabilizing single-residue substitutions into *lacZ*. Single amino acid substitutions should serve as a good model for translational errors because only rarely, in about 10% of the proteins that contain translation errors, two or more residues will be simultaneously mistranslated in the same protein. We expressed the misfolding-prone mutants at the same level as the wild type protein. Because the misfolded LacZ proteins are both potentially toxic and also devoid of biological function, the comparison of the growth rates of bacteria carrying the WT and each of the destabilized mutants allowed us to evaluate the additional fitness cost specifically arising from misfolding toxicity.

6.2.1. *Destabilizing mutations in lacZ yield aggregated and partialy soluble proteins*

Amino acid substitutions in protein cores are significantly more destabilizing than substitutions on protein surfaces [338, 339]. Therefore, we selected five buried residues encoding non-polar amino acids which could be mutated to polar residues with single nucleotide substitutions while maintaining a similar level of codon preference (**Table 6.1**). We used the DPX server [340] to identify buried residues of the LacZ protein based on its crystal structure (PDB code: 1dp0). We then applied the I-mutant_2.0 algorithm [341] to confirm that the selected substitutions would be indeed destabilizing. Using site-directed mutagenesis the five selected substitutions were introduced separately into plasmids containing *lacZ* under transcriptional control of the IPTG-inducible *lac* promoter [342]. We then used a β-galactosidase assay [343] to experimentally confirm reductions in the catalytic activity of LacZ in all of the generated mutants (see **Table 6.1**).

Table 6.1. Characteristics of destabilizing mutations engineered into *E. coli* β-galactosidase. In the table ΔΔG values represent destabilizing effects predicted by the I-mutant2.0 server [341]. The experimentally determined enzymatic activities of the mutants (in percentages) are shown in the table relative to WT.

Mutant	V567D	F758S	I141N	G353D	A880E
Predicted ΔΔG kcal/mol	-2.6	-2.9	-2.4	-1.6	-0.6
Relative protein activity (%)	31	4	17	2	61
Codon Substitution (WT/Mutant)	GTC/GAC	TTT/TCT	ATT/AAT	GGC/GAC	GCG/GAG
Codon preference % (WT/Mutant)	13.5/53.9	29.0/32.4	33.5/17.3	42.8/53.9	32.3/24.7
Found in inclusion bodies (see Figure 1A)	No	Yes	Yes	Yes	No

To determine whether the destabilized proteins tended to aggregate, we separated soluble proteins and proteins in inclusion bodies (see Methods) and analyzed them by SDS-PAGE (**Fig. 6.1A**). The three mutants with the lowest catalytic activity (F758S, I141N and G353D) were found in inclusion bodies (**Table 6.1**), the remaining two mutants (V567D and A880E) and WT proteins were found mainly in the soluble protein fraction. Next, by inspecting total cell extracts at different time points after IPTG induction, we confirmed that the total amount of protein synthesized in each mutant strain was similar to WT. As shown in **Figure 6.1B** similar amounts of LacZ are produced in WT and either soluble (V567D) or insoluble (F758S) mutants.

Quantitative analysis of the Coomasie stained bands also did not reveal any significant difference between the LacZ synthesis rates in WT and mutant strains (**Fig. 6.1C**). Finally, because expression of misfolded proteins is expected to generate a heat shock response [344, 345], we used western blots to monitor the amount of the GroEL heat shock protein in induced and un-induced cells carrying WT and mutant *lacZ* (**Fig. 6.1D**). In cells carrying WT *lacZ*, the concentration of GroEL increased when IPTG was added. However, in both the V567D and F758S mutants, the levels of GroEL in either induced or uninduced cells were equal or higher than that in induced WT cells.

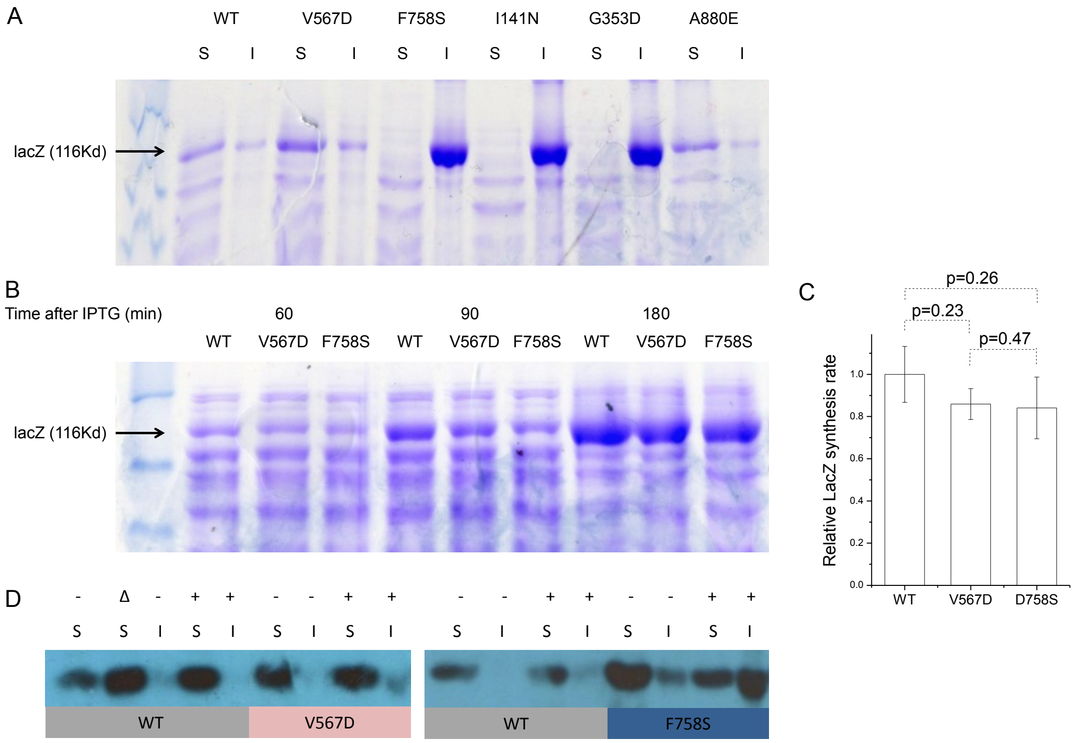


Figure 6.1. Expression of destabilizing mutants and WT LacZ. (A) SDS-PAGE of soluble and insoluble fractions of cells expressing WT LacZ and five destabilizing mutants induced with 10 μM IPTG. (B) Total β-galactosidase at different times after IPTG induction. The LacZ band is indicated by the black arrow. (C) Relative synthesis rate of β-galactosidase. P-values were obtained using a t-test of the linear regression slopes based on quantification of the gel images. (D) GroEL western blots in cells exprerssing WT and LacZ mutants. S: Soluble fraction, I: Insoluble fraction.-: No IPTG, +: 20 μM IPTG, Δ: Heat shock (1h shift from 37 to 42°C).

Overall, the results described in this section demonstrate that: 1) all engineered mutants have significantly reduced catalytic activities, 2) soluble and insoluble mutants are expressed at the same level as WT, and 3) the mutants induce a heat shock response, and in some cases aggregate in inclusion bodies.

6.2.2. *Misfolded proteins are no more toxic than WT proteins*

The synthesis of WT or mutant β-galactosidase was initially induced by adding 10 μM IPTG. Using WT LacZ activity as a reference [346], we estimated that about 30,000 molecules of β-galactosidase were present in each bacterial cell at this induction level. This approximately corresponds to half of the protein molecules expressed by a fully induced WT *lacZ* operon [343]. Cells expressing WT LacZ grew approximately 13% slower on glycerol as the sole carbon source compared to uninduced cells (**Fig. 6.2A**). If misfolded proteins indeed impose a significant extra cost on the bacterium, then similarly expressed mutant strains with destabilizing substitutions should lead to a more pronounced growth decrease compared to the one observed with WT LacZ. However, as shown in **Fig. 6.2A**, the mutant strains grew as well as cells expresing WT LacZ, and, despite inclusion body formation, two of the mutants even grew significantly faster (see Discussion).

To further explore the potential toxicity of the destabilized proteins we focused on two mutants (F758S and V567D). These mutants are examples of a completely aggregated and a soluble but destabilized LacZ protein, respectively. By varying the concentration of IPTG we monitored the growth of cells with different levels of expressed LacZ proteins (**Fig. 6.2B**). Importantly, no additional growth decrease was observed in the mutant strains compared to the WT at all IPTG induction levels. When no IPTG was added, resulting in a low expression level from the un-induced promoter, we also observed the same growth rate reduction in all constructs relative to cells carrying an empty pBR322 plasmid (**Fig. 6.2B**).

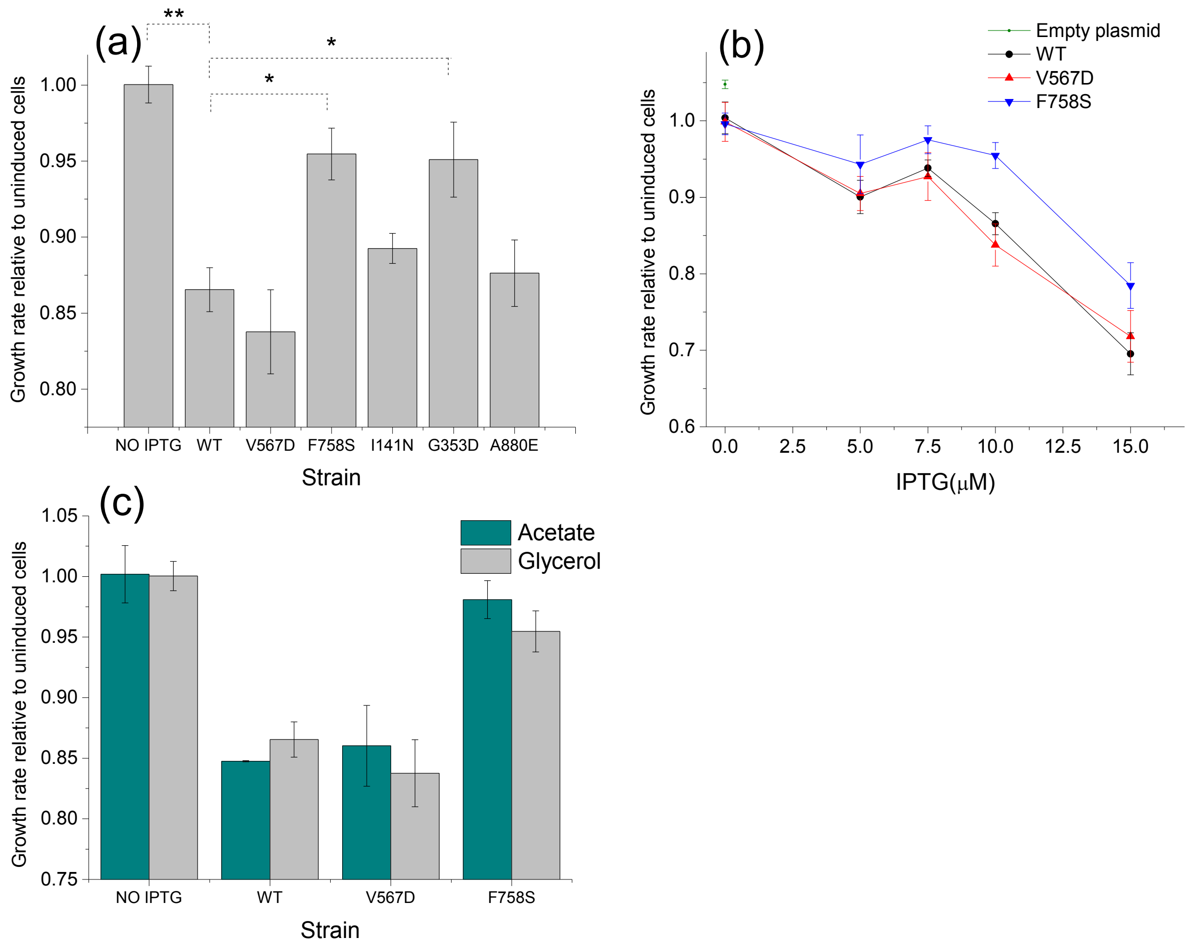


Figure 6.2. Growth rate comparisons between WT and misfolding-prone LacZ. (A) Growth rates of cells expressing WT LacZ relative to uninduced cells and cells expressing each of the five destabilizing mutants (10 μM IPTG). Mann-Whitney U P-value *:0.02,**:7.6*10^{-4}. (B) Growth rates of cells expressing WT LacZ and two mutants at different induction (IPTG) levels, the growth rate of cells carrying an empty plasmid is also shown for comparison. (C) Growth rates of cells expressing LacZ and two destabilizing mutants on acetate and glycerol as the main carbon source (expression in both cases was induced with 10 μM ITPG). Error bars represent the standard error of the mean (SEM) based on triplicate experiments.

We investigated the possibility that the toxicity of misfolded proteins was more pronounced on a relatively poor carbon source by measuring the growth of *E. coli* cells expressing V567D, F758S, and WT on acetate. Although the overall growth rate on acetate was only about 60% of that on glycerol, we again did not observe any additional fitness (growth) decrease due to the destabilizing mutations (**Fig. 6.2C**). This experiment confirmed that the observed results are not specific to a particular carbon source.

6.2.3. *Nucleotide level selection, protein solubility, and stability in* E. coli

Nucleotide sequences of highly expressed genes are significantly constrained by selection for amino acid codons corresponding to abundant tRNAs [347-349]. A recent experimental analysis by Kudla *et al.* [350] suggests that non-optimal codons can directly influence *E. coli* growth (fitness). Using 154 variants of Green Fluorescent Protein (GFP) with multiple random synonymous substitutions, these authors found a significant positive correlation between codon optimality and bacterial growth rate. An important role played by the nucleotide-level selection in evolution of *E. coli* proteins is also supported by a high correlation between the rates of non-synonymous (Ka) and synonymous (Ks) substitutions (**Fig. 6.3B**, Spearman's rank correlation $r=0.66$, P-value$<10^{-10}$). In addition, the partial correlation between Ka and mRNA expression, controlling for Ks, is small ($r=-0.14$, P-value$=7*10^{-9}$), whereas the partial correlation between Ks and expression, controlling for Ka, is significantly higher ($r=-0.38$, P-value$<10^{-10}$).

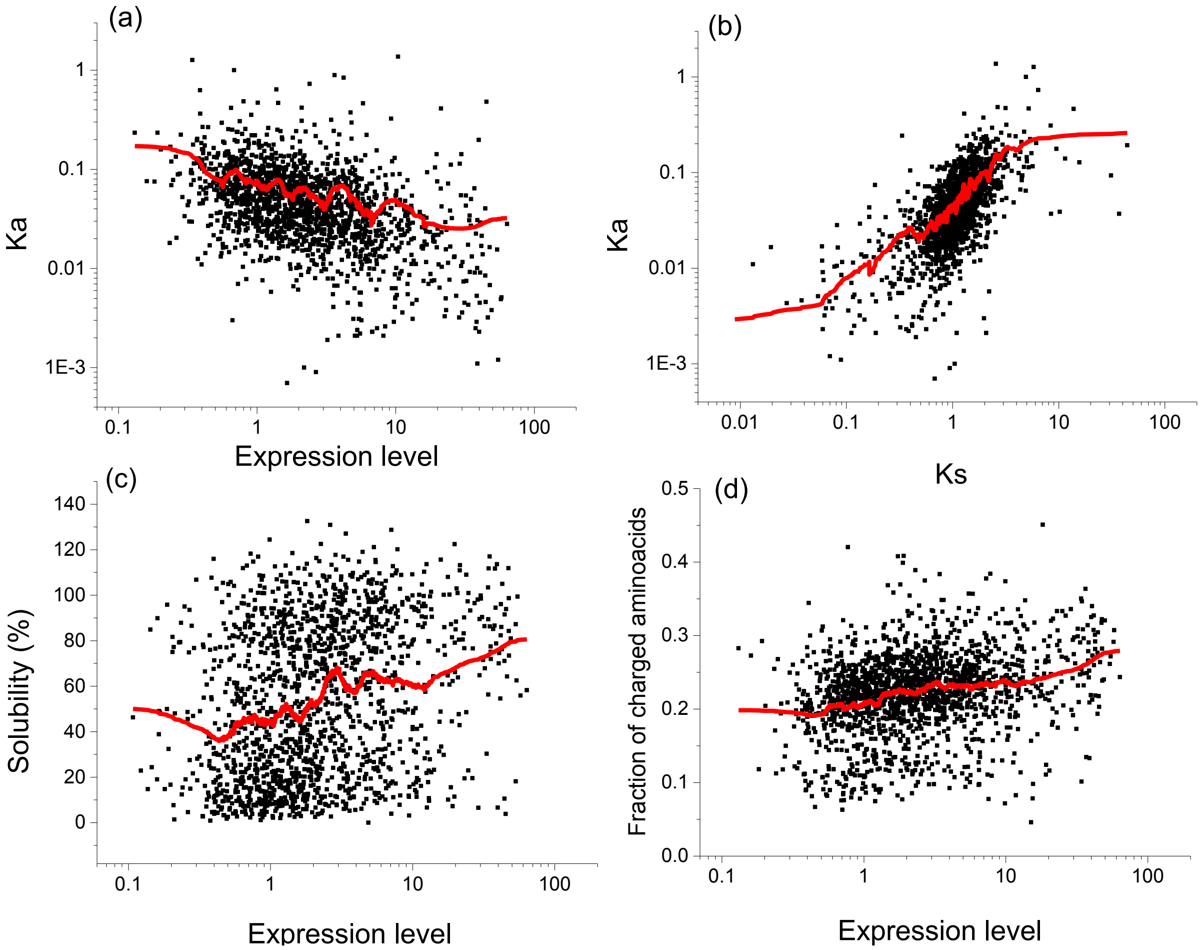


Figure 6.3. Correlation of *E. coli* mRNA expression with Ka, protein solubility, and the fraction of charged residues. (A) Correlation between expression and the rate of non-synonymous substitutions (Ka) (Spearman's r=-0.45, P-value<10^{-10}). (B) Correlation between Ka and the rate of synonymous substitutions (Ks) (r=0.66, P-value<10^{-10}). (C) Correlation between expression and protein solubility measured *in vitro* [351] (r=0.27, P-value< 10^{-10}). (D) Correlation between expression and the fraction of charged residues (r=0.28, P-value<10^{-10}). The red lines on each panel represent a 200-point moving average of the data.

Although selection for optimal codons at the nucleotide level should significantly affect the rates of both synonymous and non-synonymous substitutions [349], there are additional constraints specifically acting on non-synonymous sites (see [71, 352]). Many of these additional constraints affect the propensity of proteins to misfold and aggregate. For example, it has been reported that highly expressed *E. coli* proteins are more soluble than proteins with lower expression [353-355]. It is likely that the observed increase in solubility is necessary to avoid protein aggregation and non-functional binding [356] mediated by non-specific hydrophobic interactions. Using the genome-wide protein solubility data for *E. coli* proteins obtained by Niwa *et al* [351] we indeed observed a significant correlation between solubility and expression (**Fig. 6.3C**; Spearman's r= 0.27, P-value<10^{-10}). Importantly, the observed selection for solubility does not explain the correlation between the protein evolutionary rate and expression (**Fig. 6.3A**, r=-0.45, P-value<10^{-10}); the partial correlation between Ka and expression, controlling either for solubility or for the fraction of charged residues, is still significant, r=-0.42 and -0.41 respectively (P-value<10^{-10}).

The positive correlation between solubility and expression is in agreement with an increase in the fraction of charged residues (**Fig. 6.3D**, r=0.28, P-value<10^{-10}) and a simultaneous decrease in the fraction of hydrophobic residues (r=-0.16, P-value<10^{-10}) in highly expressed *E. coli* proteins. We observed similar results by analyzing *E. coli* protein duplicates (paralogs) with different expression levels. By directly comparing duplicates expressed at different levels, many confounding factors, such as differences in folding topology or protein secondary structure, are removed. The analysis of 370 *E. coli* paralogs (see Methods) demonstrated a decrease in the fraction of hydrophobic residues (paired Wilcoxon signed rank test, P-value=7*10^{-4}) and a simultaneous increase in the fraction of charged residues (P-value=7*10^{-6}) in the duplicates with higher expression levels.

The analysis of 602 *E. coli* protein structures currently available in PDB (see Methods,) confirmed a significant increase in the fraction of solvent-exposed charged residues in highly expressed proteins (r=0.18, P-value=6*10^{-6}). While such an increase may lead to higher protein stabilities [357], a proposed consequence of selection for translational robustness [331], we did not detect strong correlations between mRNA expression and other structural features usually associated with increased protein stability [328, 331]. For example, we did not observe a significant increase in the fraction of buried hydrophobic residues (r=0.06, P-value=0.13) [358-360] or an increase in the average number of contacts per residue (contact density) in highly expressed *E. coli* proteins (r=0.02, P-value=0.96). Neither did we find a decrease in the fraction of residues in loops or unstructured protein regions (r=0.07, P-value=0.06) [361]. Our analysis of experimentally determined *E. coli* protein stabilities assembled in the ProTherm database [362] also failed to reveal any significant correlation between protein stability, measured either by protein melting temperature (r=-0.14, P-value=0.46) or folding free energy (ΔG, r=-0.08 P-value=0.70), and mRNA expression level (**Fig. 6.4A, B**). We also did not detect significant changes in the contact order, a structural measure strongly associated with folding speed [363, 364], in highly expressed bacterial proteins (r=-0.01, P-value=0.8).

Overall, the computational analysis described above suggests that, at least based on the currently available datasets, an increase in folding speed and/or protein stability for highly expressed bacterial proteins are unlikely to play a major role in constraining the protein molecular clock in *E. coli*.

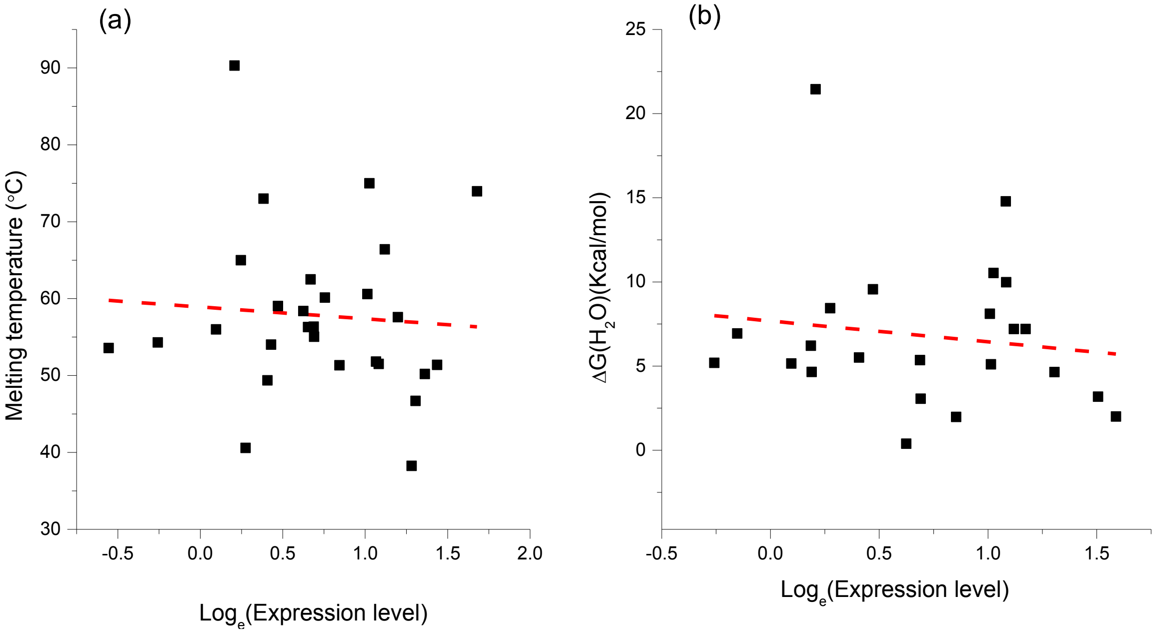


Figure 6.4. **Relationship between protein stability and mRNA expression**. The experimentally measured stability data were obtained from the ProTherm database [362], the expression data for *E. coli* was obtained from the study of Lu *et al.* [365] (A) Correlation between mRNA expression and melting temperature for 28 proteins (r=-0.14, P-value=0.45). (B) Correlation between mRNA expression and folding free energy for 23 proteins (r=-0.08, P-value=0.70). The dashed red line represents the linear regression line between each feature and the natural logarithm of the expression values.

6.3. Discussion

The results presented here demonstrate that, at least in *E. coli*, the cost associated with the gratuitous expression of a protein is significantly higher than the additional toxicity cost incurred by destabilization or misfolding of the same amount of protein; by "gratuitous" we imply here that the protein has no effect on fitness through its biological function. It is important to emphasize that our growth measurements are not sensitive enough to detect small fitness effects, for example decreases in the growth rate on the order of 1% or less, and consequently we cannot rule out additional costs specifically related to misfolding toxicity[366]. In fact, a detailed study by Lindner *et al.* [367] using time-lapse microscopy, showed that the presence of protein aggregates in *E. coli* has an effect on growth rate at the level of individual cells. Nevertheless, our experiments do show that the misfolding toxicity cost is significantly smaller than other costs associated with protein expression.

We believe that the main expression costs specifically in this bacterium are related to translational efficiency and jamming of the cell's cytoplasm with useless proteins. Importantly, expression costs associated with amino acid waste, or the energy required for gratuitous expression were recently shown by Stoebel *et al.* [368] to play a relatively minor role. On the other hand both gratuitous protein expression and suboptimal codons can significantly slow bacterial growth, for instance, by reducing the pool of free ribosomes in the cell [342, 350]. This effect will preferentially affect highly expressed genes bound by a relatively larger number of ribosomes. A gene with non-optimal codons will slow the rate of translation (speed of ribosomal motion) and thus titrate more ribosomes. A reduced pool of free ribosomes will necessarily slow expression of all bacterial genes and thus decrease the rate of biomass synthesis [369].

Interestingly, we observed that bacteria expressing two of the mutants (F758S and G353D) grew significantly faster than cells expressing native LacZ protein (**Figure 6.2A**), although still not as fast as uninduced *E. coli*. This intriguing result demonstrates that titration of ribosomes cannot be the only explanation for the costs associated with gratuitous protein synthesis. The F758S and G353D proteins had the lowest catalytic activities of all constructs (**Table 6.1**) and both mutants, as well as I141N, were found mostly in inclusion bodies. It is likely that the localization of the LacZ proteins to inclusion bodies prevents jamming of the cytoplasm and relieves effects associated with non-functional binding. It was previously shown that an asymmetric partition of inclusion bodies during cell division may result in a cell rejuvenation phenotype [367]. We would like to emphasize that this result does not support the misfolding toxicity hypothesis, as these mutants grew faster than the strain expressing WT LacZ. Based on the growth rates of mutants primarily localized to inclusion bodies (V567D, F758S, I141N; average growth decrease 6.7%) and the proteins remaining in the cytoplasm (WT, V567D, A880E; average growth decrease, 14%), one can conclude that effects of jamming and translational efficiency make approximately similar contributions to fitness.

An important separate question in the context of the mistranslation-induced misfolding hypothesis is whether phenotypic (transcriptional or translational) mutations can cause enough protein misfolding to be significantly cytotoxic. Although suboptimal codons are expected to substantially increase the translational error rate [348], no correlation was observed between codon optimization and the fraction of properly folded GFP by Kudla *et al.*[350]. Even if relatively rare, phenotypic mutations can be still significantly damaging if they occur in functionally and structurally important sties. This may explain a well-established correlation between codon optimization and evolutionary conservation of corresponding protein sites [370-372]. This correlation is not necessarily a consequence of selection against mistranslation induced toxicity, and again may be primarily related to the loss of functional proteins and the cost of additional protein synthesis necessary to compensate for the misfolding. In fact, it has been reported that essential bacterial proteins have lower aggregation propensities than those predicted for non-essential proteins [355].

While our study demonstrates that misfolding toxicity is unlikely to be a universally dominant factor connecting expression and the protein molecular clock in all species, we cannot rule out the possibility that toxicity may play an important role in other species. We note, however, that in higher organisms the correlations between mRNA expression and the protein molecular clock are generally much weaker than in some microbes. For example, Liao *et al*. [373] demonstrated that expression plays a relatively minor role in constraining the molecular clock in mammalian species. Also, by comparing evolutionary rate of separate and fused protein domains in human and *Arabidopsis,* Wolf *et al.* [374] found a comparable contribution from expression and structural-functional constraints.

A number of elegant experimental studies have demonstrated cytotoxic effect of several misfolded or marginally stable proteins in higher organisms [375, 376]. For instance, several hundred mutations in the SOD1 protein were shown to result in aggregates associated with amyotrophic lateral sclerosis (ALS) in humans [377]; also, non-natural peptides have been used to induce cytotoxic aggregates of GFP in *C. elegans* [378]. Although these studies directly demonstrate the importance of misfolding and aggregation for some specific proteins, the extent to which these effects dominate the molecular clock for *all* proteins in these and other species needs to be investigated and again compared to other contributing factors.

6.4. Conclusions

Our experimental results suggest that selection against toxic protein misfolding is unlikely to be the universal and dominant factor determining protein molecular clock in all species. We demonstrate that, at least in *E. coli,* other factors associated with gratuitous protein synthesis, such as translational efficiency and possibly jamming of the cytoplasm, are likely to be the primary constraints. Our computational analyses also suggest a relatively weaker, but statistically significant, selection for increasing solubility and polarity in highly expressed *E. coli* proteins.

6.6. Materials and methods

6.6.1. Strains and mutant generation

Escherichia coli K12 strain GP4 (W3102, XA 21Z, *lac*I^q) was used in all experiments. *lacZ* was expressed from the IPTG inducible Lac promoter in plasmid PIV18 [342]; PIV18 is a pBR322 derivative that carries a mutation in the Shine Dalgarno sequence of the *lacZ* transcript which increases translation efficiency. Site directed mutagenesis was carried out using Stratagene's QuikChange Lightning kit (Stratagene*,* Cedar Creek, TX). pBR322 was used as the empty plasmid control.

6.6.2. Growth Curve Analysis

For each construct, a sweep of colonies was grown overnight on LB liquid media supplemented with 100μg/mL ampicillin. Overnight cultures were diluted by a 1:100 factor and grown on M9 minimal media suplemented with 0.5% casaminoacids, 0.25μg/mL thiamine, 100 μg/mL ampicillin and either 0.4% glycerol or acetate as carbon sources. 300 μL of cells with an OD600 of 0.5 were transfered to flasks containing 5.5 mL of prewarmed media suplemented with the appropiate amount of IPTG. Two hours after induction, OD600 was measured every 45 minutes. Growth rate was determined as the regression line slope of time and the logarithm of OD600.

6.6.3. SDS-PAGE and western blot

The equivalent of 200 μL of cells at OD600 of 0.7 was collected by centrifugation and lysed using Novagen's BugBuster (primary amine-free) Protein Extraction Reagent (Novagen, Merck, Darmstadt, Germany). Soluble proteins were retrieved after centrifugation of the lysed cells and aggregated proteins were then harvested following instructions for inclusion body purification described in the BugBuster reagent manual. Both fractions were saved in a 50 μL volume including 10 μL 4X SDS loading buffer, boiled, and electrophoresed on a 10% SDS polyacrylamide gel. Gels were stained with coomassie blue and scanned for analysis. For the analysis of total protein, cells were lysed in BugBuster reagent containing rLysozyme and boiled after addition of 4X SDS loading buffer. Bands were quantified using the ImageJ program [379].

Protein samples separated by SDS-PAGE as described above were blotted overnight onto a nitrocellulose membrane and incubated with Anti-GroEL antibody produced in rabbit 1:10 000 (Sigma Aldrich, St. Louis, MO). Blots were blocked with 5 % non-fat dry milk, incubated with 1:3000 anti-rabbit horseradish peroxidase conjugate antibody and visualized with Amersham's ECL Plus Western Blotting Reagent (GE Healthcare, Munich, Germany).

6.6.4. Structural analysis of E. coli proteins

In the analysis we used 602 *E. coli* protein structures currently available in PDB [380]. To prevent sampling biases, we filtered available PDB entries so that no two protein structures used in the calculations, had sequence identity higher than 90%; similar results were obtained without filtering. We defined buried residues as those with a solvent accessible area smaller than 16% [381, 382]. Solvent accessibility was calculated by the DSSP [383] program. The fraction of protein residues in loops was also calculated using DSSP. Two non-adjacent protein residues were considered to be in contact if any two of their non-hydrogen atoms were closer than 4.5 Å [384]. The protein contact density was defined as the average number of non-adjacent contacts per residue. Contact order was calculated as $(L\cdot N)^{-1}\cdot\Sigma\Delta S_{ij}$; where N is the total number of contacts, L is the total number of residues in the protein and ΔS_{ij}, which is summed over all contacts, is the number of aminoacids separating contacting residues[364]. *In vitro* solubility data for *E. coli* proteins was obtained directly from the study of Niwa *et al*. [351]

6.6.5. Correlation of the synonymous (Ks) and non-synonymous (Ka) substitution rates with expression

Orthologous ORFs and protein sequences from *E. coli* and *Salmonella enterica* were used to calculate Ks and Ka values. The *E. coli - Salmonella* orthologs were determined as bi-directional best hits using protein BLAST [18]. Ka and Ks values were calculated using the Maximum Likelihood method implemented in the PAML package [385]. The mRNA expression data reported by Lu *et al*. [365] was used to calculate the correlations. For the analysis of duplicated genes we defined duplicates as pairs of *E. coli* proteins having more than 40% sequence identity that could be aligned for at least 80% of their total length using BLAST. In the analysis of duplicates we used expression data from 466 experiments in the Many Microbes Microarrays Database [386]. We selected for the analysis only the pairs for which one paralog had higher expression values in more than 80% of the reported experiments.

Chapter 7.

GENETIC ROBUSTNESS AND FUNCTIONAL EVOLUTION OF GENE DUPLICATES

Gene duplications are a major source of evolutionary innovations. Understanding the functional divergence of duplicates and their role in genetic robustness is an important challenge in biology. Previously, analyses of genetic robustness were primarily focused on duplicates essentiality and epistasis in several laboratory conditions. In this study we use several quantitative datasets to understand compensatory interactions between *S. cerevisiae* duplicates that are likely to be relevant in natural biological populations. Our study suggests that duplicates acquire new functional roles substantially faster than has been anticipated, possibly from the moment of duplication. Due to their high functional load, very close duplicates are unlikely to provide substantial backup in the context of long term evolution. Interestingly, for gene pairs that survive an initial period of high duplicate loss, the overall functional load is reduced. At intermediate divergence distances the quantitative decrease in fitness due to removal of one duplicate becomes smaller. At these distances, yeast duplicates display more balanced functional loads and their transcriptional control becomes significantly more complex. In the context of growth phenotypes relevant in natural populations, yeast duplicates diverged beyond 70% sequence identity are not likely to provide substantial compensation for their paralogs.

7.1 Introduction

Survival of biological systems crucially depends on robustness to harmful genetic mutations, i.e. genetic robustness, and to changes in environmental conditions [387-389]. Two distinct mechanisms of genetic robustness have been previously discussed. First, alternative signaling and metabolic pathways provide an important mechanism for re-routing in many molecular networks [390, 391]. Second, a major role in genetic robustness is attributed to gene duplicates [387, 392]. Gene duplications are frequent in evolution and range in size from small-scale (SSD) to whole-genome events (WGD) [393, 394]. While in about 90% of the cases one duplicate is eventually lost in evolution [392], duplicated genes can, at least partially, back-up each other's function. Importantly, functional compensation by duplicates plays a significant role in buffering deleterious human mutations [395].

Genetic robustness due to gene duplicates is inherently tied to their functional divergence. Duplicates that acquire distinct molecular functions are naturally unable to compensate for one another. In addition, even if molecular function is conserved, incomplete compensation between duplicates is possible due to different expression patterns or dosage effects. Gene duplications are the major source of new genes [396] and several conceptual models of duplicates' evolution have been proposed [397, 398]. In the neofunctionalization model one duplicate gains new functions, i.e. functions not associated with the ancestral gene, while the other duplicate retains the ancestral functions [396, 399, 400]. In contrast, in the subfunctionalization model both duplicates become indispensable and are retained in evolution by partitioning the ancestral gene functions [401, 402]. Both these models imply an eventual loss of the ability of duplicates to fully substitute for each other. It is also likely that a significant fraction of duplicates are fixed and retained in genomes due to selective advantages, such as dosage effects or condition-specific expression patterns, present from the moment of duplication [403, 404]. In cases of fixation due to a selective advantage, full compensation between duplicates is unlikely.

Even though full compensation between duplicates is not expected in the long term, the ability of duplicates to buffer deleterious mutations of their paralogs has been now demonstrated by several independent approaches. These include a lower than expected fraction of essential genes with close duplicates [387], a paucity of pairwise epistatic interactions involving duplicated genes [405], and an excess of aggravating genetic interactions between paralogs [406, 407]. The contribution of duplicates to robustness has been primarily considered in the context of qualitative growth phenotypes either in nutrient rich or in a small number of laboratory conditions [387, 408, 409]. Although popular in experiments, these conditions are unlikely to approximate well a natural *milieu* of living systems which are constantly bombarded by a diverse array of environmental stresses and stimuli. Perhaps more importantly, even if there is a strong compensatory interaction between a pair of duplicates, an evolutionary relevant decrease in fitness can still persist – due to an incomplete buffering – after a damaging mutation in one of the duplicates [366]. In the context of long-term evolution, there may not be much difference between mutations leading to the lethal phenotype and mutations associated with a fitness decrease substantially larger than the inverse of the effective population size [410, 411]. Given that typical population sizes of free-living microbial species are quite large (>10^6-10^8) [412], even a small fitness decrease can be effectively lethal for these organisms. Consequently, quantitative analyses of growth phenotypes, preferably in multiple environmental conditions, are necessary to understand the extent to which compensation between duplicates plays an important role in natural biological populations. Here we perform such an analysis and show that in the context of natural populations, genetic buffering mediated by duplicates is likely to be rare and, surprisingly, it is not a monotonic function of duplicates' divergence.

7.2. Materials and methods

Gene and protein sequences for *Saccharomyces cerevisiae, S. paradoxus*, *S. bayanus*, *S. castelli*, *S. mikatae*, *S. kudriavzevii* and *S. kluyveri* were obtained from the Saccharomyces

Genome Database (SGD; http://downloads.yeastgenome.org/) and the study by Kellis *et al.* [413]. Pairs of gene duplicates were identified by sequence homology between proteins within each genome using BLASTP [18]. Only duplicates that were bi-directional best hits and could be aligned by more than 80% of each ORF's sequence length were considered in our analysis [414]. Following previous studies [387], we excluded ribosomal genes from the analysis due to their high expression, dominant impact on growth, and strong codon adaptation bias. Evolutionary distances between duplicated genes were estimated using the method of Yang and Nielsen [415] implemented in the PAML package [385]; the use of other methods, such as maximum likelihood, to estimate Ka and Ks did not significantly change the observed patterns (**Supplementary fig. 7.1A**).

We used the data obtained by Hillenmeyer *et al*. [416] to measure the fitness contribution of duplicates across multiple environmental conditions and chemical perturbations. Using a P-value cutoff of 0.01, we obtained the number of experimental conditions for which a growth defect was observed for every single gene deletion mutant. We also analyzed quantitative growth measurements for double and single deletion yeast strains obtained from De Luna *et al.* [417] and Costanzo *et al*. [418]. Gene essentiality data was obtained from the study of Giaever *et al*. [289].

To functionally characterize duplicated genes, Gene Ontology [419] annotations were collected from SGD and EC annotations from the Comprehensive Yeast Genome Database (CYGD)[420]. Transcription factor binding motifs used in our work were compiled from Kafri *et al*. [421] and the high-confidence predictions in Kellis *et al.* [413]. We used protein localization data from Huh *et al*. [422], Codon Adaptation Index (CAI) calculations based on the dataset by Lu *et al*. [365], and the annotation of protein complexes in CYGD.

7.3. Results

Hillenmeyer *et al.* [416] quantified growth phenotypes of single-gene yeast deletion strains in a large collection of environmental conditions. The assembled dataset contains approximately 5.5 million phenotypes of heterozygous and homozygous mutants in about 400 conditions. The sampled conditions represent 27 different environmental stresses and hundreds of perturbations with diverse chemical compounds. Environmental stresses comprised different growth media, media lacking specific vitamins or amino acids, as well as different pH and temperature regimes. This comprehensive collection of phenotypes allowed us to investigate in detail the diversification of duplicates' functions and their contribution to genetic robustness in multiple conditions.

7.3.1. Compensation patterns based on quantitative fitness data

We first investigated how the average number of sensitive conditions, i.e. conditions with a significant growth decrease due to deletion of one duplicate, depends on sequence divergence (Ka) between the duplicated genes (**Fig. 7.1A, B**). We considered the fraction of different conditions with a growth phenotype as a quantitative measure of compensation capacity for duplicates at various divergence distances. For very close duplicates the average number of sensitive conditions is not significantly different from that of a random pair of yeast singletons (**Fig. 7.1B** red line). Importantly, this result does not suggest that random gene pairs and close duplicates are equivalent in terms of similarity of their molecular function. As we demonstrate below, the observed pattern is likely due to a higher overall functional load of close duplicates; here and throughout the paper we use the term *functional load* of a gene to characterize the average fitness decrease – across considered conditions – due to deletion of the gene.

Interestingly, the number of sensitive conditions initially drops as duplicates diverge, decreasing about 30% at the distances corresponding to Ka ≈ 0.1 (Ks ≈ 1, see **Supplementary fig. 7.2A, B**). As duplicates diverge further, the average number of sensitive conditions increases again, reaching the average for a random pair of yeast singletons at Ka ≈ 0.25. The trend shown in **Figure 7.1B** is not sensitive to the P-value cutoff used to determine the significance of the growth decrease observed in mutant strains (**Supplementary fig. 7.3**). A similar trend was also observed for the average growth decrease, measured either by log-ratios or Z-scores across all tested conditions (**Supplementary fig. 7.4A, B**). Bin-free analyses of the data (**Supplementary fig. 7.1B, C, 7.2B**) also revealed a smaller fitness cost due to the loss of duplicates at intermediate distances (Ka ≈ 0.1).

Because most actively growing wild-type yeast populations are diploid [423], we mainly focused our analysis on heterozygous mutant strains. The patterns of functional compensation for heterozygous and homozygous mutants are similar when multiple-drug resistance genes, as defined by Hillenmeyer *et al.* [416], are not considered (**Supplementary fig. 7.4C**). The trends also remain similar when only environmental perturbations are analyzed in the homozygous experiments (**Supplementary fig. 7.4D**). We also checked that the observed compensation patterns due to closest duplicates are not significantly influenced by additional, i.e. more diverged, paralogs (**Supplementary fig. 7.4E**). This lack of significant compensation by diverged duplicates results in an approximately linear relationship between the number of sensitive conditions per yeast protein family and the family size (**Supplementary fig. 7.5**). Finally, the observed compensation patterns were not affected by removal of gene pairs with a high Codon Adaptation Index (CAI, **Supplementary fig. 7.6A**), suggesting that the observed trend cannot be explained by expression-based constraints on the rate of duplicate sequence evolution (Ka) [73].

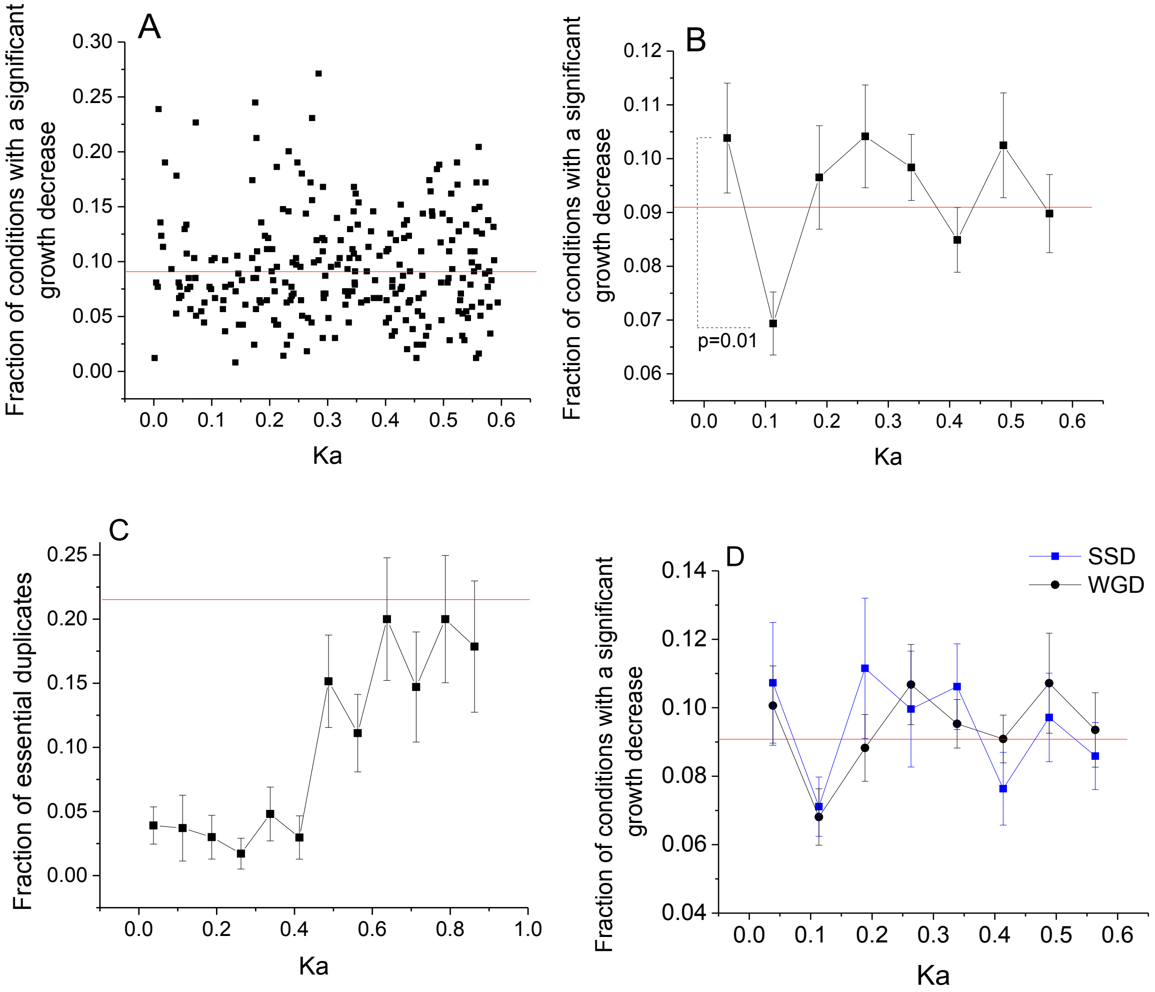


Figure 7.1. **Compensation patterns between yeast duplicates as a function of their evolutionary divergence**, Ka, the number of non-synonymous substitutions per site. **A.** Scatter plot of the fraction of sensitive conditions, i.e. conditions with detectable growth phenotypes resulting from deletion of one duplicate gene, versus Ka. Each dot in the figure represents a pair of yeast duplicates. **B.** The average fraction of sensitive conditions per duplicate pair. The P-value was calculated using the Mann Whitney U test. The red lines in A and B indicate the average fraction of sensitive conditions for a random pair of yeast singletons. **C.** The average fraction of essential duplicates, i.e. duplicates with a lethal phenotype upon deletion, as a function of Ka. The red line indicates the fraction of essential yeast singletons. Gene essentiality data were obtained from the *Saccharomyces* Genome Deletion Project [289]. **D.** Fraction of conditions with a significant growth decrease for deletion of yeast duplicates arising from small-scale (SSD, blue) and whole-genome duplications (WGD, black). The duplicates were classified

as SSD or WGD based on the study by Kellis *et al*. [394] The red line shows the average fraction of sensitive conditions for a random pair of yeast singletons. In the figures, error bars represent the standard error of the mean (SEM).

It is interesting to compare the ability of duplicates to buffer mutations leading to any detectable growth decrease beyond a given fitness threshold (**Fig. 7.1B**) and their role in protecting against the no-growth phenotype, i.e. the likelihood to observe essential genes in duplicate pairs. In **Figure 7.1C**, using data from the study by Giaever *et al.* [289], we show the fraction of essential duplicates as a function of their divergence. In agreement with previous studies [387, 408, 409] we found that the fraction of essential genes remains low and approximately constant for close duplicates, and increases substantially only at divergence distances corresponding to Ka > 0.4. Notably, this pattern is qualitatively different from the compensation for quantitative growth phenotypes (**Fig. 7.1B**), demonstrating the aforementioned impact of using quantitative phenotypes to assess the evolutionarily relevant consequences of mutations. Also in contrast to patterns obtained in studies based on essential genes [408], we observed similar compensation profiles for gene pairs originating from small-scale and genome-wide duplications (**Fig. 7.1D**). Because all WGD duplicates have the same age, this result suggests that the ability of duplicates to buffer each-other's function across multiple conditions depends more strongly on their sequence divergence than on the time since duplication.

7.3.2. Correlates of functional compensation

It is likely that the observed decrease in the number of sensitive conditions at intermediate divergence distances (Ka ≈ 0.1) is due to a decrease of the functional load carried at these distances by the union of duplicate genes. To explore this possibility we considered the quantitative fitness data from DeLuna *et al.* [417] and the synthetic genetic array (SGA) data from Costanzo *et al.* [418]. In these studies the authors performed quantitative growth measurements of yeast strains with individual and simultaneous deletions of duplicates. Using the single deletion phenotypes from the DeLuna *et al.* (**Fig. 7.2A**) and Costanzo *et al.* studies (**Fig. 7.2B**), we observed fitness profiles similar to the one obtained based on the data from Hillenmeyer *et al.* (**Fig. 7.1B**) as a function of Ka, with smaller phenotypic effects at intermediate distances. Interestingly, the overall functional importance of duplicate pairs, measured by the phenotype of double deletions, indeed substantially decreases with their divergence (**Fig. 7.2C, D**). This result suggests that while close duplicates are more likely to have similar functions, their higher functional load makes complete compensation less likely. Because the overall functional load of duplicates remains approximately constant for Ka > 0.1, the higher fraction of detectable growth phenotypes at these distances is likely due to a decreased ability for functional compensation as duplicates diverge. Indeed, compensation between duplicates quantified by the presence of aggravating interactions between duplicate pairs decreases as a function of sequence divergence (**Fig. 7.2E, F**) (see [407]).

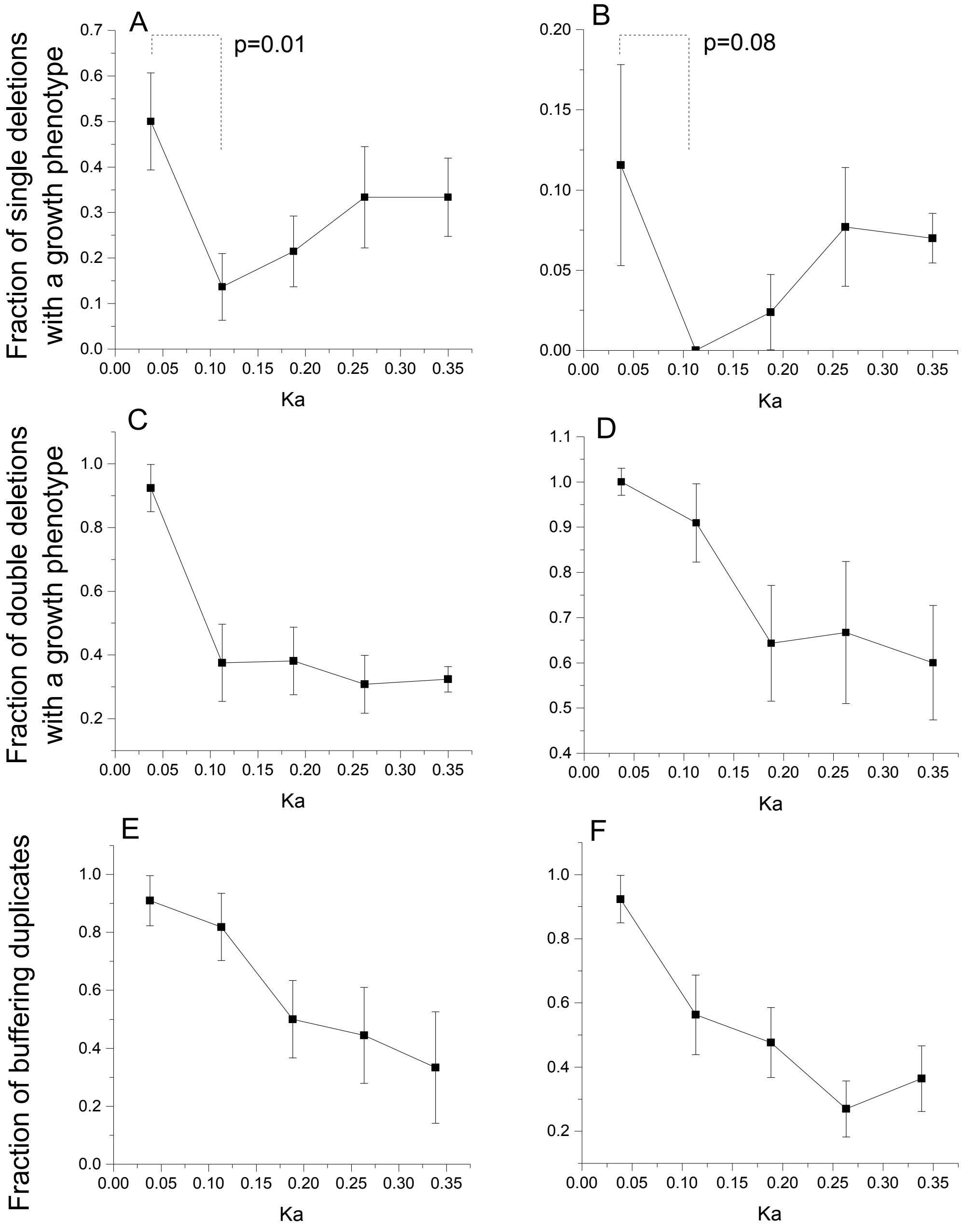


Figure 7.2. Growth phenotypes for individual and simultaneous deletion of duplicates as a function of their sequence divergence (Ka). The results in the first column (A, C, E) are based on the competition experiments by DeLuna *et al.* [417], and in the second column (B, D, F) on the synthetic genetic arrays (SGA) by Costanzo *et al.* [418]. **A, B.** Fractions of single duplicate deletions with a significant growth decrease. **C, D.** Fractions of simultaneous (double) duplicate deletions with a significant growth decrease. Due to different measurement sensitivities of the two studies, different cutoffs were used to determine a significant growth decrease: 1% for

DeLuna *et al.* (A, C) and 10% for Costanzo *et al.* (B, D); the presented results are not sensitive to the exact cutoff values (see **Supplementary fig.7.6**). P-values were obtained using Fisher's exact test. **E, F.** Fraction of paralogs with a significant negative epistatic interaction from the studies of DeLuna *et al.* and Costanzo *et al.*, respectively. In the figures error bars represent the SEM.

Besides a smaller overall functional load, it is possible that duplicates at intermediate distances have other properties that favor genetic robustness. To explore this possibility, for each duplicate pair we looked at the gene with the largest and the gene with the smallest number of sensitive conditions (**Fig. 7.3A**). Notably, while the duplicate with more conditions (**Fig. 7.3A**, black) follows the average trend for all duplicates (**Figure 7.1B**), the duplicate with fewer conditions (**Fig. 7.3A**, red) shows a steady gain in the number of conditions as a function of Ka. Consequently, the functional load of close duplicates, measured by the number of sensitive conditions, is very different, and this difference becomes significantly smaller as the genes diverge (Fig. 3B. Pearson's r=-0.64, P-value=7*10^{-4}, see also Supplementary **Fig. 72C**). Close duplicates with the larger number of sensitive conditions also show a higher evolutionary constraint, evaluated by the normalized ratio of non-synonymous to synonymous substitutions per nucleotide site, Ka/Ks (Wilcoxon Signed Rank test P-value=7*10^{-3}, **Fig. 7.2C**). This result agrees with previous reports of asymmetric evolution of duplicates in the context of co-expression, genetic interaction and protein-protein interaction networks [405, 424, 425]. The observed asymmetry in the functional load between close duplicates can make buffering difficult. For example, if the less sensitive duplicate is expressed only under very specific environmental conditions.

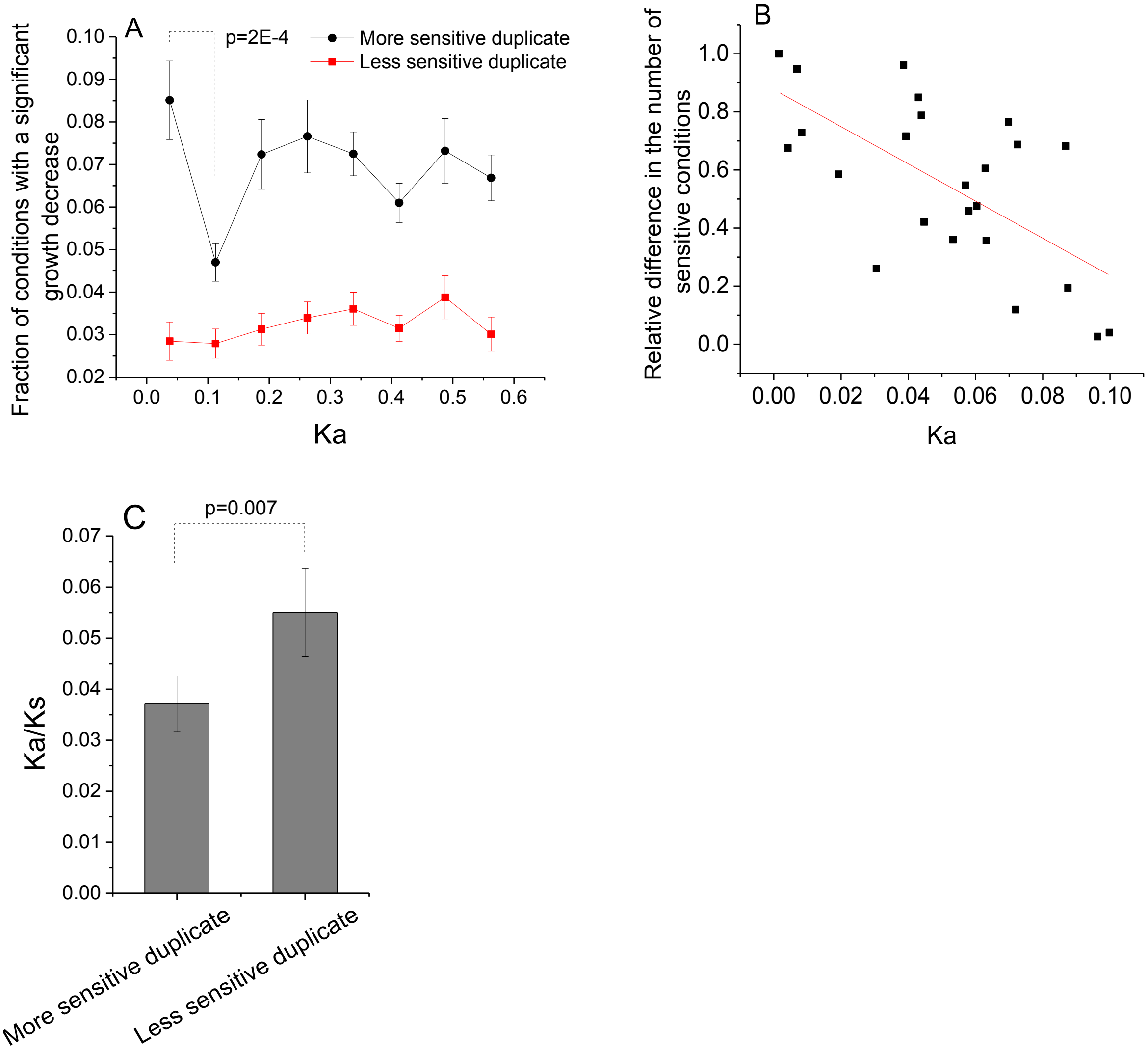


Figure 7.3. Differences in the number of sensitive conditions between duplicates. A. The average fraction of sensitive conditions for the duplicates with the higher and lower number of sensitive conditions in each pair; Ka values represent sequence divergence between duplicates. The P-value is for the Mann Whitney U test. **B.** The relative difference in the number of sensitive conditions between duplicates as a function of their initial divergence; Ka values represent sequence divergence between duplicates. The relative difference was calculated as the absolute difference in the number of sensitive conditions between duplicates normalized to the total number of sensitive conditions for the pair (Spearman's r=-0.60, P-value=$2*10^{-3}$; Pearson's r=-0.64, P-value=$7*10^{-4}$). **C.** The average Ka/Ks ratio for the paralogs with the largest (more sensitive) and smallest (less sensitive) number of conditions with a significant growth decrease. Ka/Ks ratios were calculated relative to orthologous sequences in *S. bayanus*. Only duplicates with Ka<0.15 to each other were considered. The P-value is for the Wilcoxon signed rank test.

To further explore the mechanism behind the observed back-up patterns, we analyzed the functional diversification of yeast duplicates as a function of their sequence divergence (Ka). First, for genes encoding metabolic enzymes we calculated the fraction of gene pairs with conserved Enzyme Commission (EC) numbers (**Fig. 7.4A**); the conservation of EC numbers indicates that corresponding proteins catalyze identical biochemical reactions. Second, we calculated the fraction of shared Gene Ontology (GO) terms describing protein molecular function (MF) for all duplicates (**Fig. 7.4B**). Both measures showed that the molecular function of yeast duplicates typically starts to substantially diverge only at about Ka > 0.4. The timing of this divergence approximately coincides with a significant increase in the fraction of essential duplicates (**Fig. 7.1C**). On the other hand, the significant changes in the number of quantitative growth phenotypes are observed when the molecular function of duplicates is usually still conserved.

A complementary analysis of transcription factor binding sites suggests that gene regulation plays an important role in establishing the observed compensation patterns. It was previously demonstrated that duplicated yeast genes have, on average, a higher number of cis-regulatory motifs than singleton genes [426]. Using a comprehensive dataset of about 150 known and predicted DNA binding motifs in yeast [413, 421], we found that the average number of different motifs regulating a duplicate pair increases significantly at Ka ≈ 0.1 (**Fig. 7.4D**, red, Mann-Whitney U test, P-value=0.06). At this divergence distance, the average number of different motifs per duplicate pair is more than twice the number of motifs for a pair of yeast singletons (**Fig. 7.4D**, red line). The number of regulatory motifs increases both for the duplicate with the highest and the duplicate with the smallest number of sensitive conditions (**Supplementary fig. 7.8A, B**). The increase in complexity of the duplicates regulation at Ka ≈

0.1 is also confirmed by a significant increase (Mann-Whitney U test, P-value=1*10^{-3}) at these distances of the number of transcription factor mutants [427] affecting duplicate gene expression (**Fig. 7.4D**, black).

While the total number of DNA motifs regulating duplicates initially increases with divergence, the fraction of shared motifs (**Supplementary fig. 7.9A**, se also [428]), the overlap in GO terms describing biological processes (**Fig. 7.4C**), and the overlap in cellular localization observed in fluorescence-tagging experiments [422] decrease (**Supplementary fig. 7.8B**). Such a pattern suggests that the increase in regulatory complexity allows duplicates to specialize for different biological processes while mostly preserving common molecular functions. The ability of duplicates with partially diverged regulatory regions to compensate for each other through expression changes of the intact gene was previously described by Kafri *et al* [421, 429]. Also, the recent study by DeLuna *et al*. [430] showed that upon deletion of one duplicate, expression changes of the remaining paralog are often need-based, i.e. they happen primarily when the corresponding function is required. Such regulatory back-up circuits should, at least in some cases, enable functional compensation between homologs with different expression patterns in wild type. Notably, based on the data from recent study by Springer *et al.* [431], who measured the expression changes of yeast genes when one of two genomic copies was deleted in diploid cells, we observed a significant dosage response only for genes forming recently duplicated pairs (Ka < 0.15, **Fig. 7.4E**). This suggests that genes with close duplicates are most responsive to dosage effects.

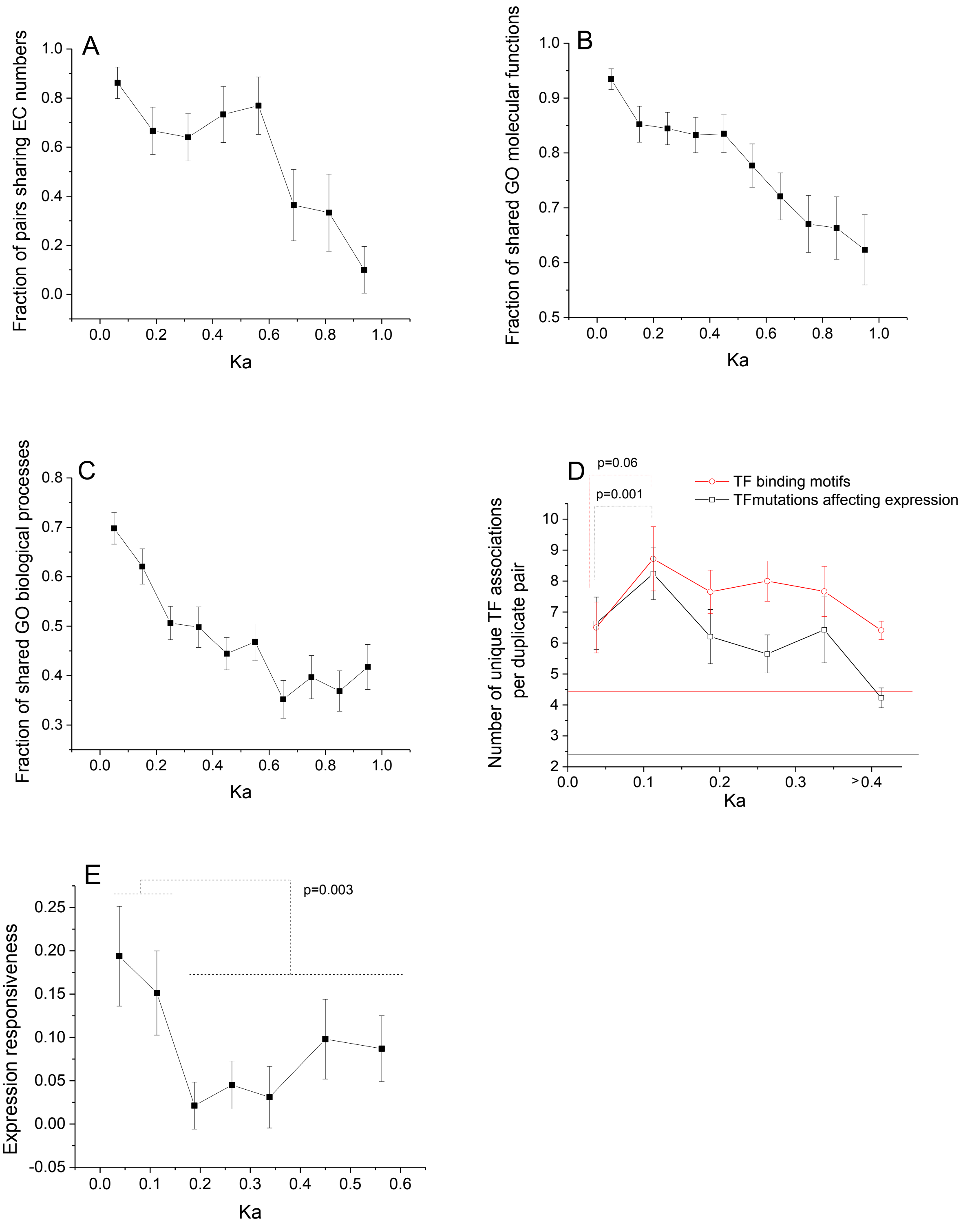


Figure 7.4. Diversification of duplicates function and regulation. A. Fraction of metabolic duplicates sharing the same Enzyme Commission (EC) numbers; conservation of EC numbers indicates catalysis of identical biochemical reactions. **B.** Fraction of Gene Ontology (GO) Molecular Function (MF) terms shared between duplicates. **C.** Fraction of GO Biological

Process (BP) terms shared between duplicates. In figures B and C we considered only GO terms with a distance of 3 or more to the corresponding GO root hierarchy term. **D.** In red, the average number of different transcription factor binding motifs per duplicate pair. Transcription factor (TF) binding motifs were compiled from the studies of Kafri *et al.* [421] and Kellis *et al.* [413]. In black, the average number of transcription factor deletions in *S. cerevisiae* that significantly affect the expression of duplicate genes. The data were obtained from the study by Hu et al. (2007). For comparison we also show the average number of motifs and TF mutants affecting expression for random pairs of yeast singletons (horizontal red and black lines); the P-values were calculated using the Mann Whitney U test. **E.** The average dosage compensation (responsiveness) of duplicates as a function of sequence divergence (Ka). The data for the average expression responsiveness was obtained from the work of Springer *et al*. [431]. In that study responsiveness was measured in diploid yeast strains as the Log_2 ratio (perturbed vs. normal) of expression changes for the remaining gene copy following deletion of the equivalent gene copy on a sister chromosome. The P-value was calculated using the Mann-Whitney U test. In all figures error bars represent the SEM.

Finally, the patterns of diversification and functional compensation described above should correlate with the process of duplicate loss in evolution. We investigated the retention of yeast duplicates using the complete genomic sequences of seven species: *S. cerevisiae, S. paradoxus*, *S. bayanus*, *S. castelli*, *S. mikatae*, *S. kudriavzevii* and *S. kluyveri*. We calculated the number of remaining duplicates as a function of their evolutionary divergence (**Fig. 7.5**, see also **Supplementary fig. 7.10A** for the corresponding relationships in individual yeast species). This analysis suggests that a relatively brief initial period of high duplicate loss [392] is followed by a long evolutionary period (Ka > 0.1) during which the average loss rate decreases more than tenfold (red in **Fig. 7.5**). Interestingly, the loss rate significantly decreases approximately at the divergence distance when duplicates become more similar in terms of their functional load (**Fig. 7.2B**) and when their regulatory complexity significantly increases (**Fig. 7.5D**). It is likely that the duplicates surviving the initial loss stage develop independent functionalities and are preserved for long times in the genomes of yeast species.

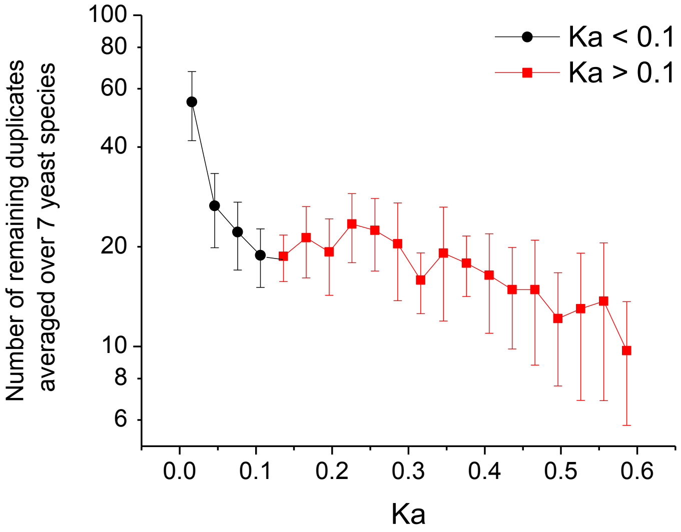


Figure 7.5. The average number of duplicates retained in the genomes of yeast species as a function of the duplicates divergence Ka, the number of non-synonymous substitutions per site. The number of remaining duplicates was averaged over the genomes of seven yeast species: *S. cerevisiae, S. paradoxus*, *S. bayanus*, *S. castelli*, *S. mikatae*, *S. kudriavzevii* and *S. kluyveri*. See also **Supplementary figure 7.10** for the number of remaining duplicates in the individual species, and for the number of remaining duplicates as a function of Ks. The rate of duplicate loss in evolution is more than 10 times lower for the distances shown in red (Ka>0.1) compared to the distances in black (Ka<0.1). In the figure error bars represent the SEM.

7.4. Discussion

In the present study we analyzed genetic robustness due to duplicates in the context of quantitative growth phenotypes and sensitivities to gene deletions in multiple environmental conditions. Such robustness is important for understanding the buffering of deleterious mutations in large natural biological populations. Our results demonstrate that, contrary to commonly held view, close gene duplicates are unlikely to provide a high level of back-up in the context of long term evolution. Consequently, it is unlikely that many duplicates are fixed in natural populations specifically due to selection for robustness.

Our analysis also suggests that duplicate redundancies described in genomics databases, and frequently observed in laboratory experiments, should be considered with caution, at least with respect to their functions in natural biological populations. To investigate this point further, we analyzed a set, compiled by Kafri *et al*. [432], of 112 yeast duplicates reported to be at least partially redundant in research publications. These duplicates have been described as redundant based on their functional overlap and compensatory interactions observed in small-scale experimental studies. Interestingly, based on the number of conditions with quantitative growth phenotypes from the study by Hillenmeyer *et al.* [416], and the quantitative growth measurements by Costanzo *et al.* [418], the duplicates annotated as redundant are not significantly different from all other yeast duplicates (Mann Whitney U P=0.13 and 0.35 respectively, **Supplementary fig. 7.11**). This demonstrates that, although many yeast duplicates indeed may show functional overlap in some laboratory conditions, their compensation properties will probably be significantly less important in long term evolution due to the ability of purifying selection to efficiently prune mutations causing even a small fitness decrease.

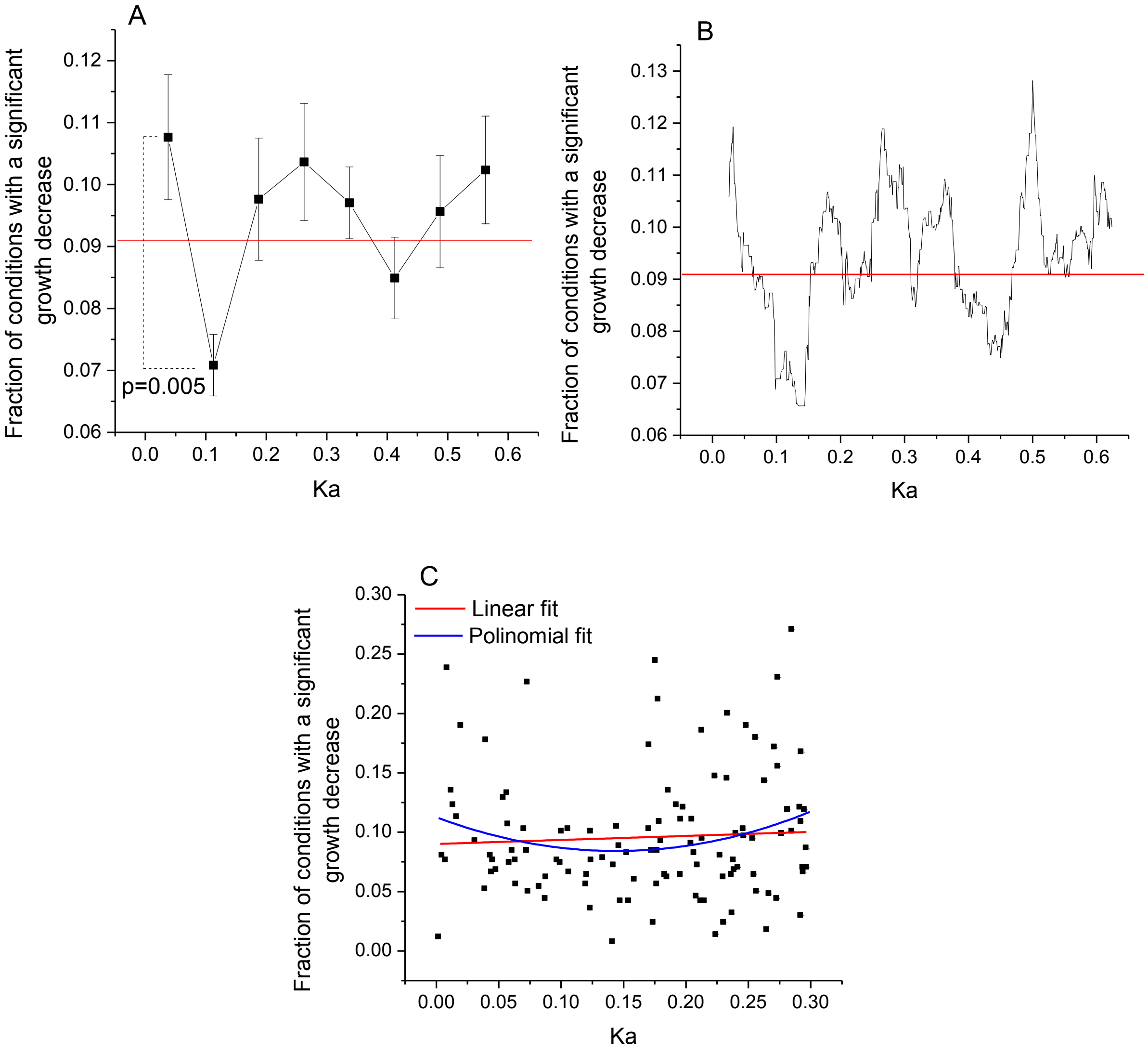
It is likely that several different factors contribute to the relative paucity of functional compensation between paralogs at small divergence distances. A significant fraction of duplications are likely to be fixed due to dosage effects [403], and functional compensation between such duplicates in the context of long term evolution is unlikely. For example, the lack of significant compensation between histone pairs, HTA1-HTA2 and HHT1-HHT2, is likely to be a consequence of their role in maintaining proper histone levels in yeast cells. Gene dosage may explain the inability of some duplicates to backup each other, but it is unlikely to be the only explanation. As we demonstrated, even when all duplicate pairs with a high CAI

(**Supplementary fig. 7.6A**) or forming known protein complexes (**Supplementary fig. 7.6B**) were removed from the analysis, the pattern of functional compensation remained similar.

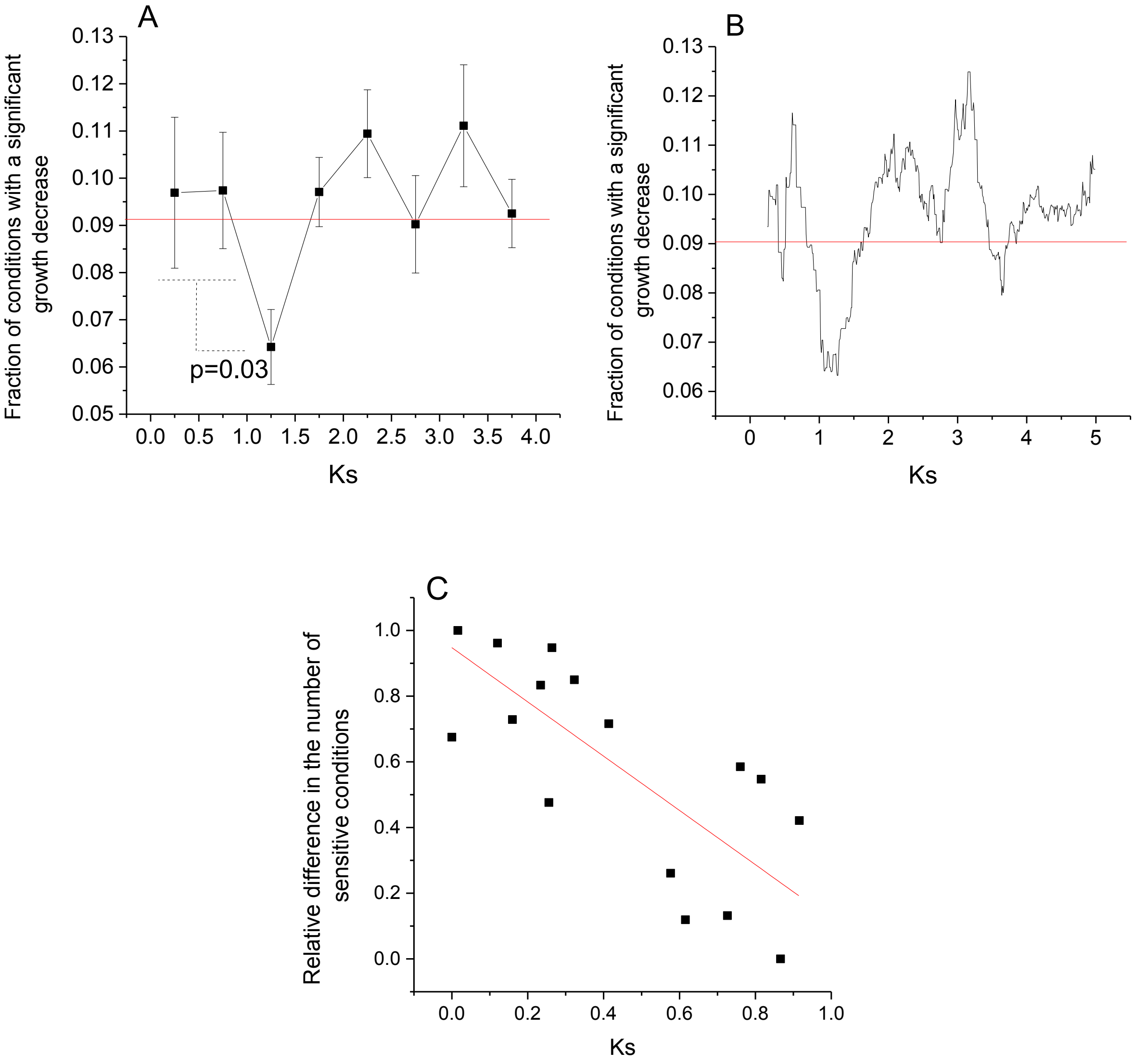
Close duplicates are also less likely to compensate for each other probably due to the aforementioned dichotomy in their functional loads (**Fig. 7.3A, B**). Indeed, many close duplicates can be classified, based on their activity and breadth of expression, into a major and a minor functional isoforms. For example, glyceraldehyde-3-phosphate dehydrogenase TDH1 is active under various stress conditions, while its isoenzyme TDH2 is used primarily during exponential growth [433]. Similarly, the ubiquitin conjugating enzyme UBC4 is expressed during exponential growth, while its duplicate UBC5 is active during stationary phase [434]. The difference in functional load for close yeast duplicates is also consistent with the asymmetric partition of functions, interactions, and gene expression, observed between close duplicates in other organisms, for example *Arabidopsis* and Human [425, 435, 436]. This suggests that duplicate-dependent compensation in the context of long term evolution may be limited in other species as well.

Our analysis suggests that a typical lifecycle of gene duplicates in yeast consists of several distinct evolutionary stages [397, 398]. In the first stage (at duplicate distances corresponding to $Ka < 0.05$), duplicates tend to have high overall functional loads and significant asymmetry in the number of sensitive conditions; both of these factors make complete compensation unlikely. The high functional load of close duplicates suggests that adaptive selection plays an important role in their fixation. In the second stage ($0.05 < Ka < 0.25$), as duplicates diverge further, their overall functional load usually decreases. This may happen, for example, due to relaxation of the environmental conditions which facilitated the original duplicate fixation. The vast majority of duplicates, likely the paralogs with relatively smaller functional loads (**Fig. 7.2C**), are lost at this stage (**Fig. 7.5**). Gene pairs which survive the period of high duplicate loss display more balanced functional loads and complex regulation; these gene pairs are usually retained for long evolutionary times in yeast genomes (**Fig. 7.5**). Surviving duplicates can provide at least partial compensation at intermediate divergence distances and also serve as an important source of new protein functions. In the third stage (Ka > 0.3 or ~70% sequence identity), the lifecycle of duplicates is completed when their functional roles diverge, and their quantitative compensation properties become indistinguishable from those of random pairs of yeast singletons.

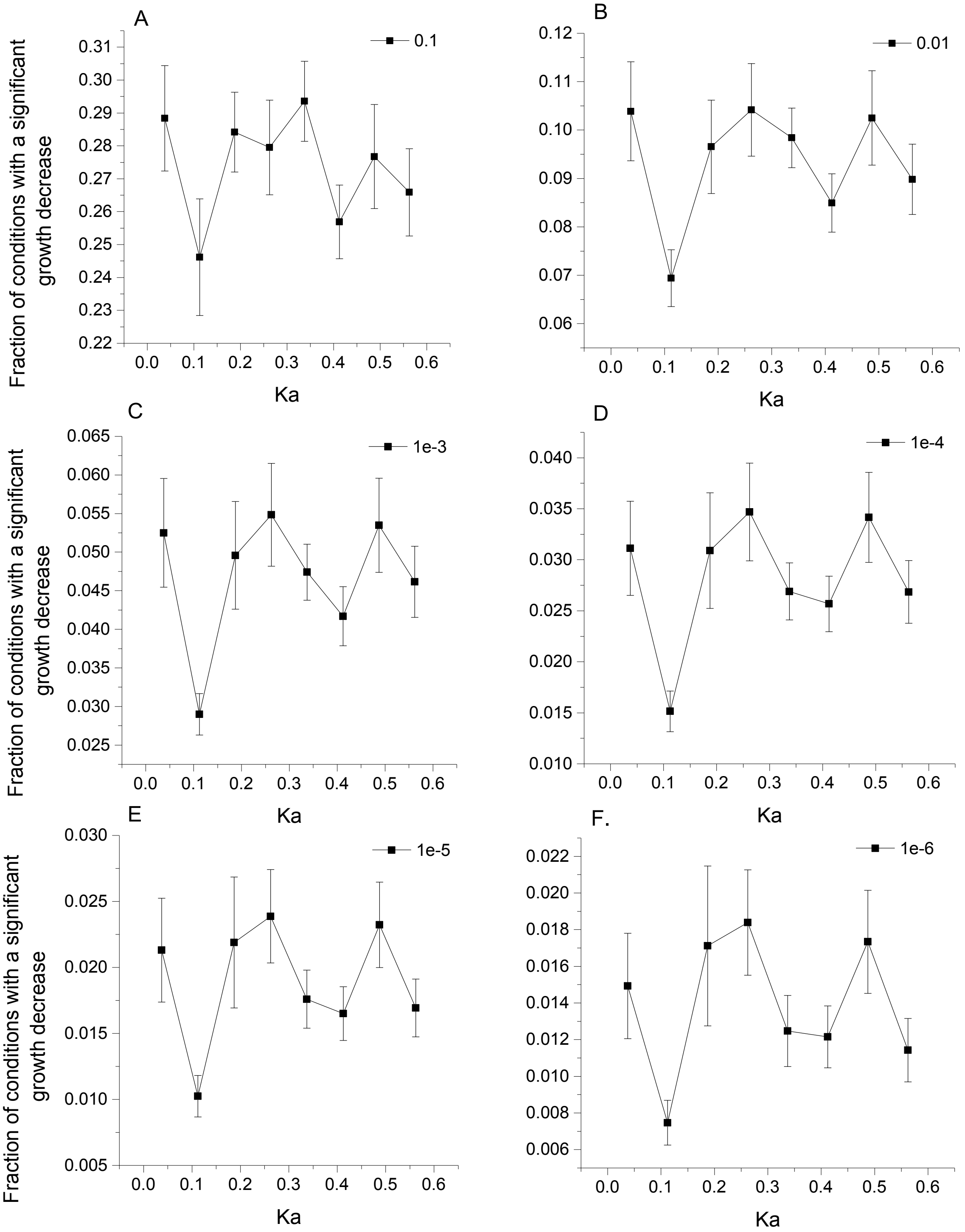
7.5 Supplementary figures



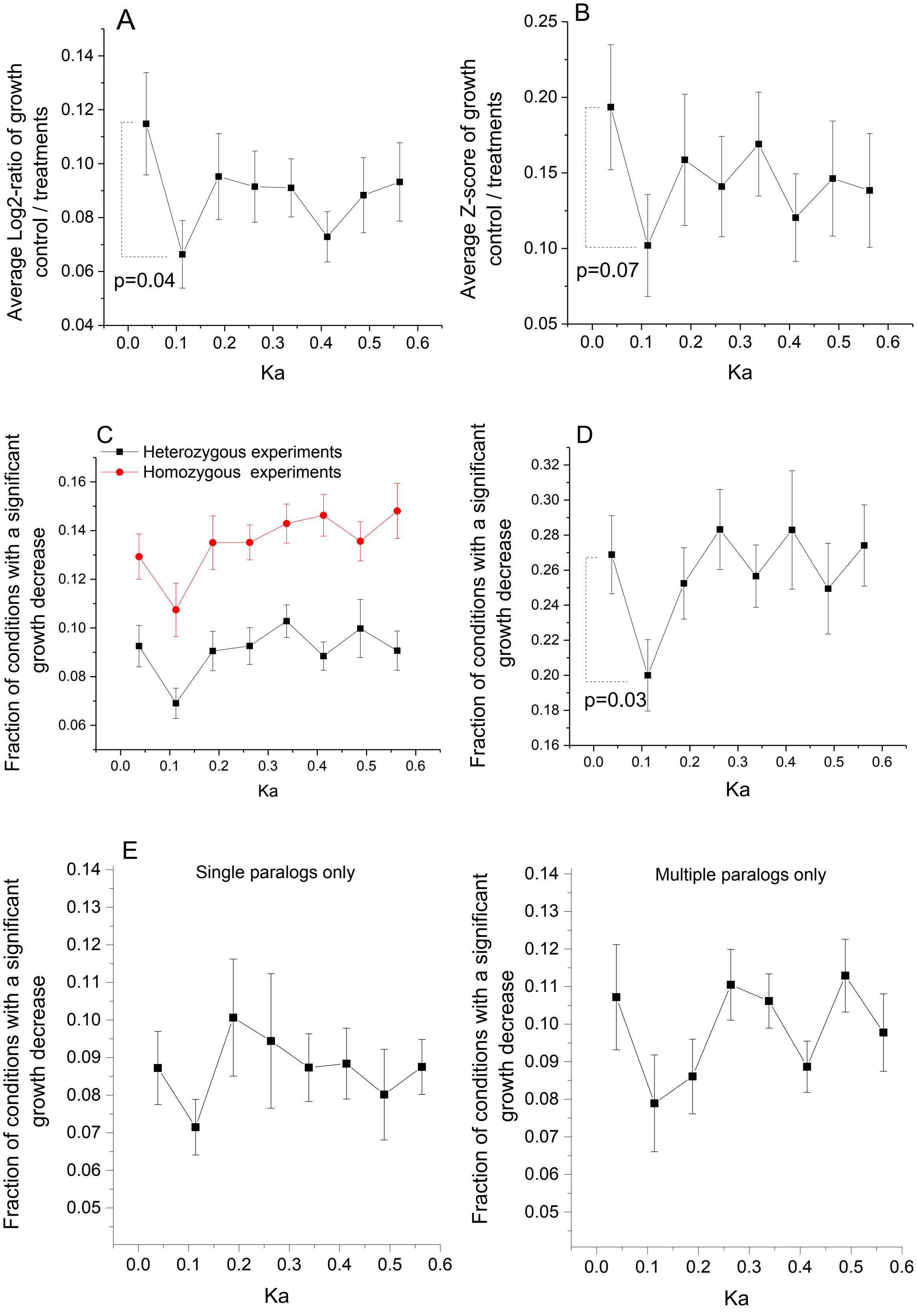
Supplementary figure 7.1. Fraction of conditions with a significant growth decrease for yeast duplicate deletions as a function of their divergence; Ka. **A.** The number of non-synonymous substitutions per site, was calculated using the Maximum Likelihood method implemented in the PAML package [385]. The error bars represent the standard error of the mean (SEM). P-value is for the Mann Whitney U test. **B.** The moving average of the data in Fig. 1A using a window size of Ka=0.05. The red line in panels A-C shows the average fraction of sensitive conditions for a random pair of yeast singletons. **C.** Linear versus quadratic polynomial model comparison for a bin-free fit of the number of sensitive conditions as a function of Ka. The quadratic model is significantly more likely to fit the data, F-test p-value: 0.05.



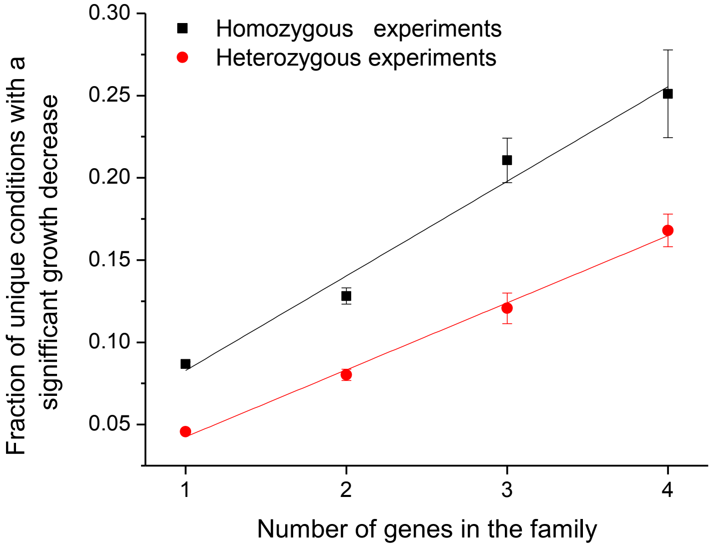
Supplementary figure 7.2. A. Fraction of conditions with a significant growth decrease as a function of Ks, the number of synonymous substitutions per site between duplicates. The error bars represent the SEM. **B.** The sliding window average of the Ks data, using a window size of Ks=0.5. The red line in A and B shows the average fraction of sensitive conditions for a random pair of yeast singletons. **C.** The relative difference in the number of sensitive conditions between duplicates as a function of their initial divergence (Ks). The relative difference was calculated as the absolute difference in the number of sensitive conditions between duplicates normalized to the total number of sensitive conditions for the pair (Spearman's r=-0.71, P-value=0.002; Pearson's r=-0.73, P-value=0.001).



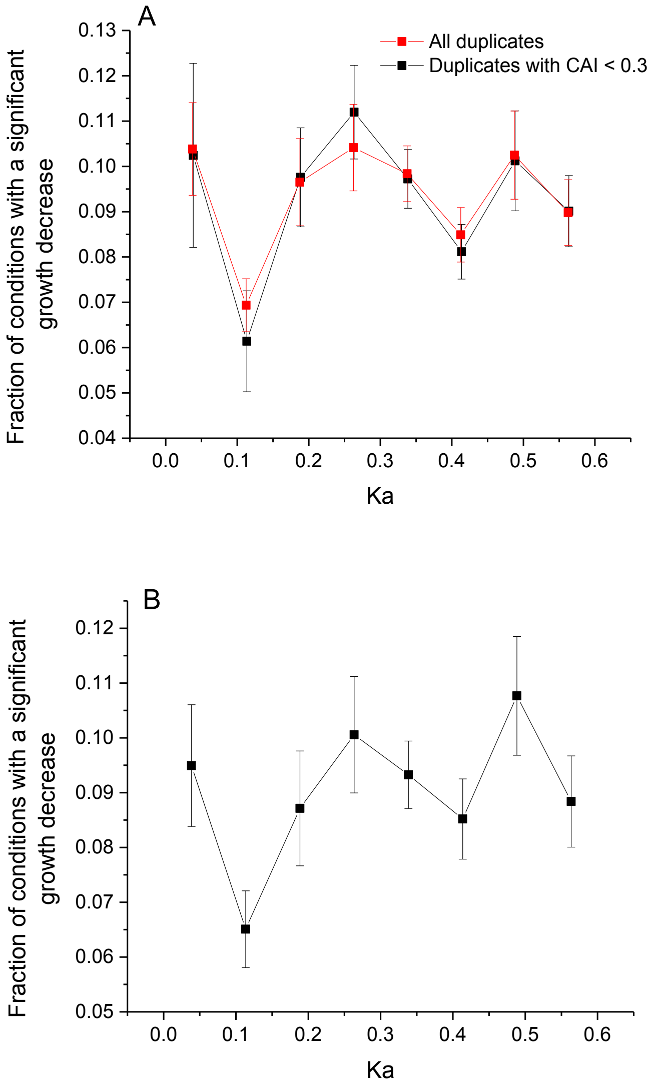
Supplementary figure 7.3. Fraction of conditions with a significant growth decrease for yeast duplicate deletions. In the figures different P-value cutoffs were used to determine whether deletion mutants have significant growth phenotypes: **A.** P-value=0.1, **B.** P-value=0.01, **C.** P-value=10^{-3}, **D.** P-value=10^{-4}, **E.** P-value=$10^{-5}$and **F.** P-value=10^{-6}. The error bars represent the SEM.


A
Average Log2-ratio of growth control / treatments
0.04
0.06
0.08
0.10
0.12
0.14
p=0.04
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Ka
B
Average Z-score of growth control / treatments
0.05
0.10
0.15
0.20
0.25
p=0.07
Ka
C
Heterozygous experiments
Homozygous experiments
Fraction of conditions with a significant growth decrease
0.06
0.08
0.10
0.12
0.14
0.16
Ka
D
Fraction of conditions with a significant growth decrease
0.16
0.18
0.20
0.22
0.24
0.26
0.28
0.30
0.32
p=0.03
Ka
E
Single paralogs only
Fraction of conditions with a significant growth decrease
0.05
0.06
0.07
0.08
0.09
0.10
0.11
0.12
0.13
0.14
Ka
Multiple paralogs only
Fraction of conditions with a significant growth decrease
Ka

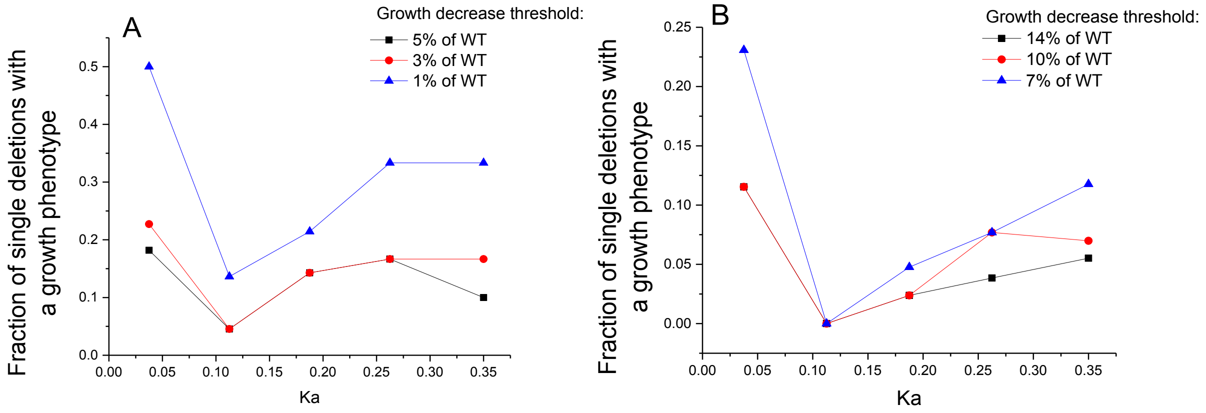
Supplementary figure 7.4. A. The average Log$_2$-ratios across multiple environmental conditions of the wild type growth rates relative to the growth rates of duplicate deletion strains. In the figure larger Log$_2$-ratios correspond to lower growth rates of duplicate deletions compared to the wild type. **B.** The average Z-scores across multiple environmental conditions of the wild type growth rates relative to the growth rates of duplicate deletion strains. **C.** Fraction of conditions with a significant growth decrease for homozygous and heterozygous diploid yeast strains, when multi-drug resistance (MDR) genes, as defined by Hillenmeyer *et al.* [416], are removed from the analysis. **D.** Fraction of conditions with a significant growth decrease for homozygous yeast deletion strains when only 27 environmental perturbations were considered, i.e. perturbations with chemical compounds were not used in the analysis. **E.** Fraction of conditions with a significant growth decrease when only duplicates without additional (i.e. more diverged) paralogs in the yeast genome are considered (Left) or when only duplicates with several paralogs are considered (Right).The error bars represent the SEM.



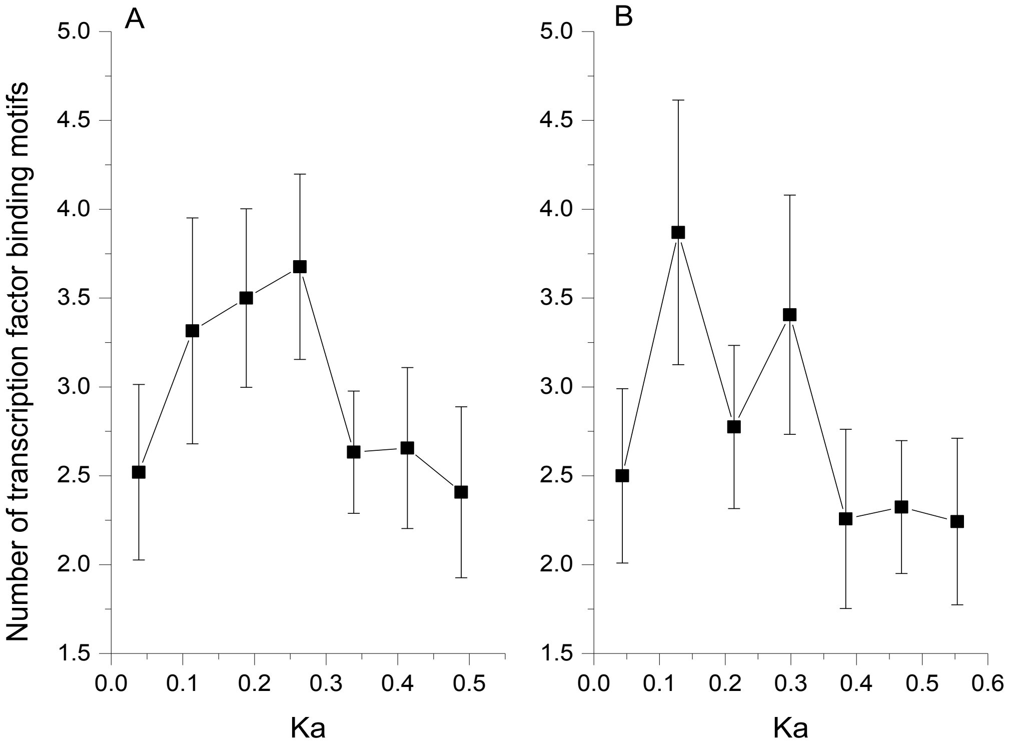
Supplementary figure 7.5. The average fraction of unique conditions with a significant growth decrease, based on Hillenmeyer *et al.* [416], per yeast protein family as a function of the family size, i.e. the number of genes in the family. The error bars represent the SEM, the straight lines represent the linear fits to the data for homozygous (black) and heterozygous (red) deletion experiments.



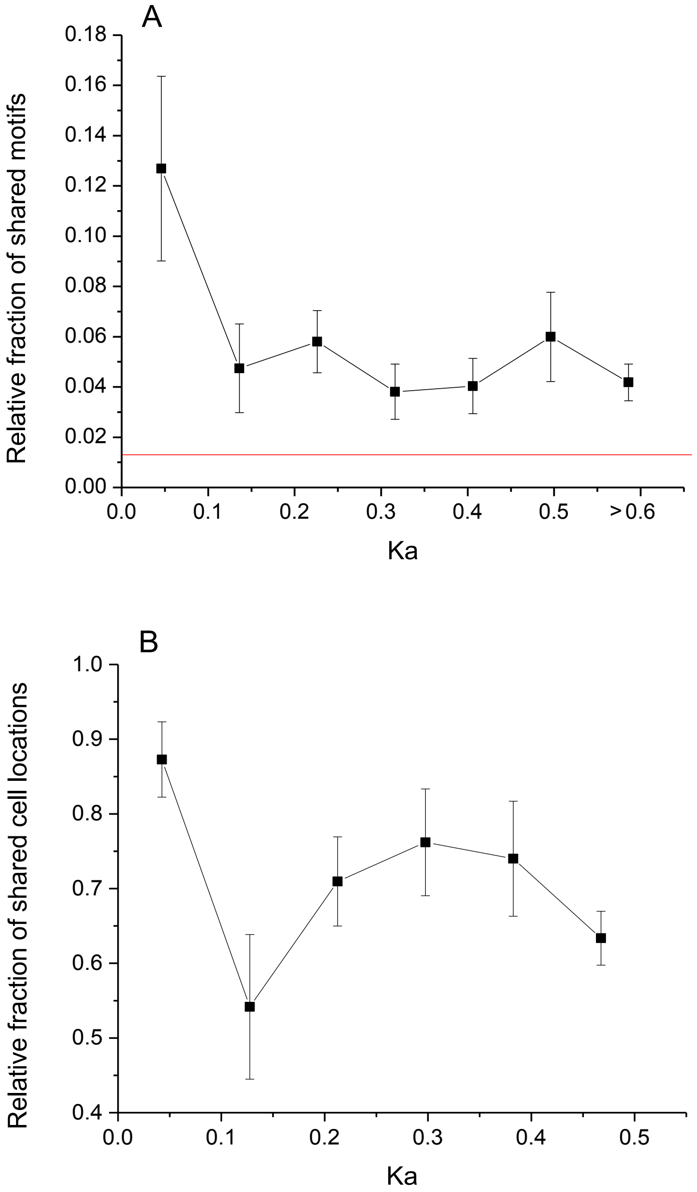
Supplementary figure 7.6. The fraction of conditions with a significant growth phenotype for duplicate deletions. **A.** Excluding pairs with a high Codon Adaptation Index (CAI). High CAI values usually indicate high expression levels of the corresponding genes. In the figure, yeast duplicate pairs for which at least one gene has CAI>0.3 were removed from the analysis. The CAI values for yeast duplicates were obtained from the study by Lu *et al.* [365]. **B.** Excluding gene pairs that participate in known protein complexes. The information about protein complexes was obtained from the Comprehensive Yeast Genome database [420]. The error bars represent the SEM.



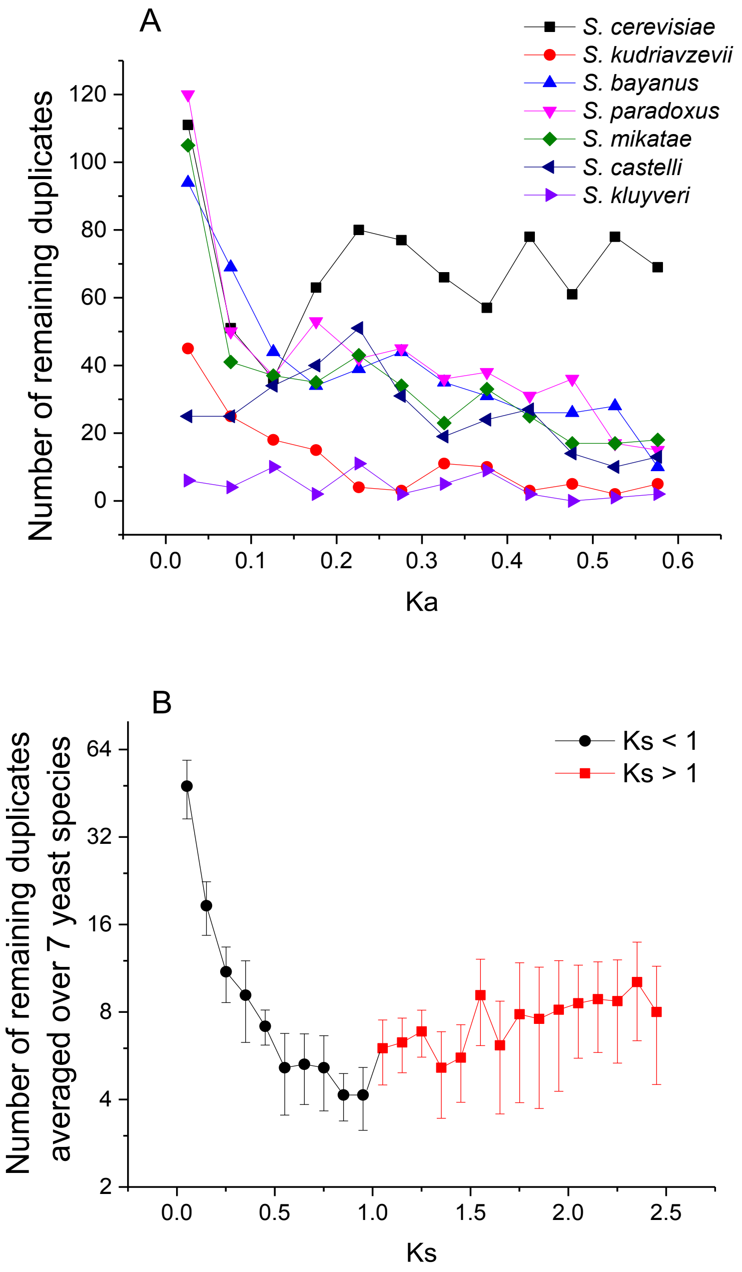
Supplementary figure 7.7. The fraction of single duplicate deletions with a growth decrease below different thresholds. The presented results are based **(A)** on the competition experiments by DeLuna *et al.* [417], and **(B)** on the synthetic genetic arrays by Costanzo *et al.* [418].



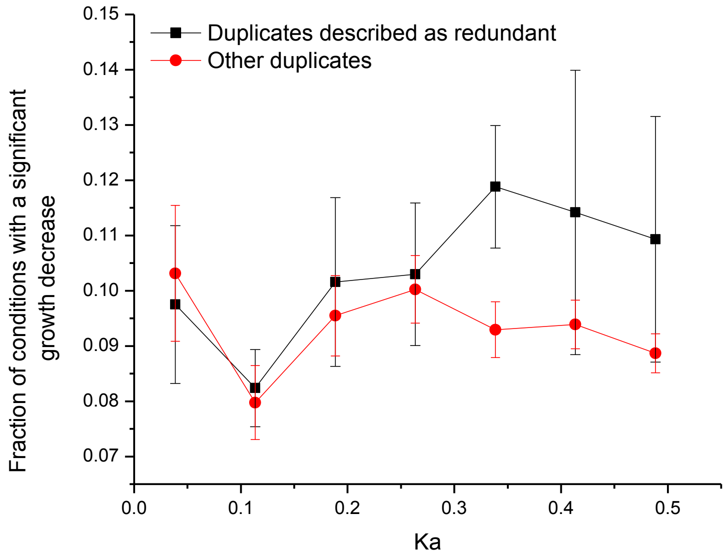
Supplementary figure 7.8. The average number of known and predicted transcription factor binding motifs for yeast duplicates with the largest (**A)** the smallest (**B)** number of sensitive conditions in each pair. The transcription factor binding motifs (150 in total) were obtained from the studies of Kafri *et al.* [421] and Kellis *et al.* [413].



Supplementary figure 7.9. A. The fraction of DNA binding motifs shared between yeast duplicates. The number of shared motifs between each duplicate pair was normalized by the total number of unique motifs for the pair. The DNA binding motifs (150 in total) were obtained from the studies by Kafri *et al.* [421] and Kellis *et al.* [413]. The red line shows the fraction of motifs shared by a random pair of yeast singletons. **B.** The overlap in cellular localization between yeast duplicates. The fraction of shared cellular locations was calculated for each pair as the number of shared locations normalized by the total number of different locations for the pair. The figure is based on the data from Huh *et al.* [422]. The error bars represent the SEM.



Supplementary figure 7.10. **A.** The number of duplicates retained in the genomes of seven yeast species (*S. cerevisiae, S. paradoxus*, *S. bayanus*, *S. castelli*, *S. mikatae*, *S. kudriavzevii* and *S. kluyveri*.) as a function of the duplicate divergence Ka, the number of non-synonymous substitutions per site. **B.** The average number of duplicates retained in the genomes of seven yeast species as a function of the divergence between duplicates (Ks). The number of the remaining duplicates was averaged over the genomes of seven yeast species: *S. cerevisiae, S. paradoxus*, *S. bayanus*, *S. castelli*, *S. mikatae*, *S. kudriavzevii* and *S. kluyveri*. The error bars represent the SEM.



Supplementary figure 7.11. Compensation patterns for yeast duplicates described in research literature as redundant, at least partially, and all other yeast duplicates (red). The set of 112 redundant duplicates was compiled by Kafri *et al.* [432] based on literature analysis. The error bars represent the SEM.

Chapter 8.

CONCLUSIONS

In this dissertation we have introduced comparative systems biology as a means to explore the long term patterns of phenotypic evolution of bacteria. Flux analysis of hundreds of microbial metabolic networks shows that different phenotypes change at different rates but following a well-defined exponential decay-like trend. Our results are in excellent agreement with experimental data and provide a framework to explore several ecological and evolutionary questions. Comparison of species' predicted phenotypic properties can complement gene and genome-based bacterial taxonomy pointing out potential interesting associations between microorganisms. These tools can also be used to measure and understand phenotypic diversity and evolutionary processes of different subsets of bacteria.

Our comparative analysis is good at identifying global trends, but specific phenotypic predictions are limited by the accuracy of the underlying network models. We presented a flexible probabilistic framework (GLOBUS) for metabolic annotations and demonstrated its potential use as an aid in stoichiometric model reconstructions. GLOBUS can integrate multiple sources of evidence including sequence, structure, and context based functional descriptors; moreover, we showed that flux information from a universal mass-balanced metabolic network can provide a global context that significantly improves annotation accuracy. The probabilistic approach allows a direct ranking of annotations based on their quality, and tracing of GLOBUS results to specific sources of evidence.

The significance of building high-quality metabolic reconstructions is demonstrated by our *Plasmodium falciparum* genome-scale metabolic model. Careful review of available literature and reconciliation of model predictions with experimental evidence, allowed us to reproduce many known features of malarial metabolism, including most known gene-knockout phenotypes. This effort led to the identification and further validation of a novel candidate drug target for this important pathogen. We also demonstrated the ability of the network to integrate several sources of high-throughput data –including transcriptomics and metabolomics– which will be important features of metabolic reconstructions as these datasets continue to expand.

Going back to the level of sequence evolution, we experimentally demonstrated that the cost associated with the expression of misfolded proteins in *E. coli* is mainly a consequence of gratuitous protein synthesis and not misfolded protein toxicity. This result calls into question the universality of a recent hypothesis about why highly expressed proteins evolve slowly [331] Complementary computational analyses showed that translational efficiency, or the selection for optimal codons in *E. coli*, is one important determinant of the lower evolutionary rate of abundant proteins.

The above result highlights the importance of even mild fitness differences for sequence evolution in species with large population sizes. On the same vein, we found that the ability of duplicated yeast genes to compensate for the loss of their paralogs is very limited in the context of natural populations. Our results show a tradeoff between the functional load of duplicated gene pairs and their sequence similarity. This way, functional compensation between paralogs was only observed at intermediate divergence distances, with close duplicates being as likely as random singletons to fully buffer for mutations of their paralogs. Our work points out important differences in the patterns of duplicated gene buffering when it is considered from the point of view of genetic interactions, as has typically been the case, and when it is looked at from the point of view of population genetics.

REFERENCES

1. Plata G, Fuhrer T, Hsiao TL, Sauer U, Vitkup D: **Global probabilistic annotation of metabolic networks enables enzyme discovery.** *Nat Chem Biol* 2012, **8:**848-854.

2. Plata G, Hsiao TL, Olszewski KL, Llinas M, Vitkup D: **Reconstruction and flux-balance analysis of the Plasmodium falciparum metabolic network.** *Mol Syst Biol* 2010, **6:**408.

3. Plata G, Gottesman ME, Vitkup D: **The rate of the molecular clock and the cost of gratuitous protein synthesis.** *Genome Biol* 2010, **11:**R98.

4. Ward BB: **How many species of prokaryotes are there?** *Proc Natl Acad Sci USA* 2002, **99:**10234-10236.

5. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome.** *Curr Opin Genet Dev* 2005, **15:**589-594.

6. Lukjancenko O, Wassenaar TM, Ussery DW: **Comparison of 61 sequenced Escherichia coli genomes.** *Microb Ecol* 2010, **60:**708-720.

7. Gripp E, Hlahla D, Didelot X, Kops F, Maurischat S, Tedin K, Alter T, Ellerbroek L, Schreiber K, Schomburg D, et al: **Closely related Campylobacter jejuni strains from different sources reveal a generalist rather than a specialist lifestyle.** *BMC Genomics* 2011, **12:**584.

8. Koonin EV, Wolf YI: **Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world.** *Nucleic Acids Res* 2008, **36:**6688-6719.

9. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC: **The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata.** *Nucleic Acids Res* 2012, **40:**D571-579.

10. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, et al: **Insights into the phylogeny and coding potential of microbial dark matter.** *Nature* 2013.

11. Mason OU, Hazen TC, Borglin S, Chain PS, Dubinsky EA, Fortney JL, Han J, Holman HY, Hultman J, Lamendella R, et al: **Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill.** *ISME J* 2012, **6:**1715-1727.

12. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, et al: **A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea.** *Nature* 2009, **462:**1056-1060.

13. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MC, Parkhill J, Falush D: **Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter.** *Proc Natl Acad Sci USA* 2013.

14. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2013, **41:**D36-42.

15. Siggins A, Gunnigle E, Abram F: **Exploring mixed microbial community functioning: recent advances in metaproteomics.** *FEMS Microbiol Ecol* 2012, **80:**265-280.

16. Madupu R, Richter A, Dodson RJ, Brinkac L, Harkins D, Durkin S, Shrivastava S, Sutton G, Haft D: **CharProtDB: a database of experimentally characterized protein annotations.** *Nucleic Acids Res* 2012, **40:**D237-241.

17. UniProt C: **Reorganizing the protein space at the Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2012, **40:**D71-75.

18. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.

19. Lipman DJ, Pearson WR: **Rapid and sensitive protein similarity searches.** *Science* 1985, **227:**1435-1441.

20. Tian W, Skolnick J: **How well is enzyme function conserved as a function of pairwise sequence identity?** *J Mol Biol* 2003, **333:**863-882.

21. Dini-Andreote F, Andreote FD, Araujo WL, Trevors JT, van Elsas JD: **Bacterial genomes: habitat specificity and uncharted organisms.** *Microb Ecol* 2012, **64:**1-7.

22. Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA: **Modeling the percolation of annotation errors in a database of protein sequences.** *Bioinformatics* 2002, **18:**1641-1649.

23. Karp PD: **What we do not know about sequence analysis and sequence databases.** *Bioinformatics* 1998, **14:**753-754.

24. Schnoes AM, Brown SD, Dodevski I, Babbitt PC: **Annotation error in public databases: misannotation of molecular function in enzyme superfamilies.** *PLoS Comput Biol* 2009, **5:**e1000605.

25. Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I: **New and continuing developments at PROSITE.** *Nucleic Acids Res* 2013, **41:**D344-347.

26. Furnham N, de Beer TA, Thornton JM: **Current challenges in genome annotation through structural biology and bioinformatics.** *Curr Opin Struct Biol* 2012, **22:**594-601.

27. Dale JM, Popescu L, Karp PD: **Machine learning methods for metabolic pathway prediction.** *BMC Bioinformatics* 2010, **11:**15.

28. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L, et al: **Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology.** *Brief Bioinform* 2010, **11:**40-79.

29. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets.** *Nucleic Acids Res* 2012, **40:**D109-114.

30. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, et al: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.** *Nucleic Acids Res* 2005, **33:**5691-5702.

31. Kharchenko P, Chen L, Freund Y, Vitkup D, Church GM: **Identifying metabolic enzymes with multiple types of association evidence.** *BMC Bioinformatics* 2006, **7:**177.

32. Yamada T, Waller AS, Raes J, Zelezniak A, Perchat N, Perret A, Salanoubat M, Patil KR, Weissenbach J, Bork P: **Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours.** *Mol Syst Biol* 2012, **8:**581.

33. Chen L, Vitkup D: **Predicting genes for orphan metabolic activities using phylogenetic profiles.** *Genome Biol* 2006, **7:**R17.

34. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96:**2896-2901.

35. Yanai I, Derti A, DeLisi C: **Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes.** *Proc Natl Acad Sci USA* 2001, **98:**7940-7945.

36. Hsiao TL, Revelles O, Chen L, Sauer U, Vitkup D: **Automatic policing of biochemical annotations using genomic correlations.** *Nat Chem Biol* 2010, **6:**34-40.

37. Tian W, Dong X, Zhou Y, Ren R: **Predicting Gene Function Using Omics Data: From Data Preparation to Data Integration.** In *Protein function prediction for omics era.* 1st edition. Edited by Daisuke K. New York: Springer; 2011: 215-242

38. Kharchenko P, Church GM, Vitkup D: **Expression dynamics of a cellular metabolic network.** *Mol Syst Biol* 2005, **1:**2005 0016.

39. Jiao D, Ye Y, Tang H: **Probabilistic inference of biochemical reactions in microbial communities from metagenomic sequences.** *PLoS Comput Biol* 2013, **9:**e1002981.

40. Peregrin-Alvarez JM, Sanford C, Parkinson J: **The conservation and evolutionary modularity of metabolism.** *Genome Biol* 2009, **10:**R63.

41. Pah AR, Guimera R, Mustoe AM, Amaral LA: **Use of a global metabolic network to curate organismal metabolic networks.** *Sci Rep* 2013, **3:**1695.

42. Galperin MY, Koonin EV: **Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption.** *In Silico Biol* 1998, **1:**55-67.

43. Levy ED, Ouzounis CA, Gilks WR, Audit B: **Probabilistic annotation of protein sequences based on functional classifications.** *BMC Bioinformatics* 2005, **6:**302.

44. Tomar N, De RK: **Comparing methods for metabolic network analysis and an application to metabolic engineering.** *Gene* 2013, **521:**1-14.

45. Orth JD, Thiele I, Palsson BO: **What is flux balance analysis?** *Nat Biotechnol* 2010, **28:**245-248.

46. Joyce AR, Palsson B: **Predicting Gene Essentiality Using Genome-Scale in Silico Models** In *Microbial Gene Essentiality: Protocols and Bioinformatics Volume* 416. Edited by Osterman AL, Gerdes SY. Totowa: Humana Press Inc.; 2008: 433-457: *Methods in Molecular Biology*.

47. Lee DS, Burd H, Liu J, Almaas E, Wiest O, Barabasi AL, Oltvai ZN, Kapatral V: **Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple Staphylococcus aureus genomes identify novel antimicrobial drug targets.** *J Bacteriol* 2009, **191:**4015-4024.

48. Burgard AP, Pharkya P, Maranas CD: **Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.** *Biotechnol Bioeng* 2003, **84:**647-657.

49. Klitgord N, Segre D: **Environments that induce synthetic microbial ecosystems.** *PLoS Comput Biol* 2010, **6:**e1001002.

50. Lewis NE, Nagarajan H, Palsson BO: **Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods.** *Nat Rev Microbiol* 2012, **10:**291-305.

51. Thiele I, Palsson BO: **A protocol for generating a high-quality genome-scale metabolic reconstruction.** *Nat Protoc* 2010, **5:**93-121.

52. Satish Kumar V, Dasika MS, Maranas CD: **Optimization based automated curation of metabolic reconstructions.** *BMC Bioinformatics* 2007, **8:**212.

53. Vitkin E, Shlomi T: **MIRAGE: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks.** *Genome Biol* 2012, **13:**R111.

54. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL: **High-throughput generation, optimization and analysis of genome-scale metabolic models.** *Nat Biotechnol* 2010, **28:**977-982.

55. Latendresse M, Krummenacker M, Trupp M, Karp PD: **Construction and completion of flux balance models from pathway databases.** *Bioinformatics* 2012, **28:**388-396.

56. Oberhardt MA, Puchałka J, Martins dos Santos VAP, Papin JA: **Reconciliation of Genome-Scale Metabolic Reconstructions for Comparative Systems Analysis.** *PLoS Comput Biol* 2011, **7:**e1001116.

57. Alam MT, Medema MH, Takano E, Breitling R: **Comparative genome-scale metabolic modeling of actinomycetes: the topology of essential core metabolism.** *FEBS Lett* 2011, **585:**2389-2394.

58. Barona-Gomez F, Cruz-Morales P, Noda-Garcia L: **What can genome-scale metabolic network reconstructions do for prokaryotic systematics?** *Antonie Van Leeuwenhoek* 2012, **101:**35-43.

59. Mithani A, Hein J, Preston GM: **Comparative analysis of metabolic networks provides insight into the evolution of plant pathogenic and nonpathogenic lifestyles in Pseudomonas.** *Mol Biol Evol* 2011, **28:**483-499.

60. Chavali AK, D'Auria KM, Hewlett EL, Pearson RD, Papin JA: **A metabolic network approach for the identification and prioritization of antimicrobial drug targets.** *Trends Microbiol* 2012, **20:**113-123.

61. Hamilton JJ, Reed JL: **Identification of functional differences in metabolic networks using comparative genomics and constraint-based models.** *PLoS One* 2012, **7:**e34670.

62. Zakrzewski P, Medema MH, Gevorgyan A, Kierzek AM, Breitling R, Takano E: **MultiMetEval: comparative and multi-objective analysis of genome-scale metabolic models.** *PLoS One* 2012, **7:**e51511.

63. Stolyar S, Van Dien S, Hillesland KL, Pinel N, Lie TJ, Leigh JA, Stahl DA: **Metabolic modeling of a mutualistic microbial community.** *Mol Syst Biol* 2007, **3:**92.

64. Zomorrodi AR, Maranas CD: **OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities.** *PLoS Comput Biol* 2012, **8:**e1002363.

65. Zuckerkandl E, Pauling L: **Evolutionary divergence and convergence in proteins.** In *Evolving Genes and Proteins.* Edited by Bryson V, Vogel H. New York: Academic Press; 1965: 97-166

66. Kimura M: **Evolutionary rate at the molecular level.** *Nature* 1968, **217:**624-626.

67. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins.** *EMBO J* 1986, **5:**823-826.

68. Whitman WB, Coleman DC, Wiebe WJ: **Prokaryotes: the unseen majority.** *Proc Natl Acad Sci U S A* 1998, **95:**6578-6583.

69. Philippot L, Andersson SG, Battin TJ, Prosser JI, Schimel JP, Whitman WB, Hallin S: **The ecological coherence of high bacterial taxonomic ranks.** *Nat Rev Microbiol* 2010, **8:**523-529.

70. Koonin EV: **Systemic determinants of gene evolution and function.** *Mol Syst Biol* 2005, **1:**2005 0021.

71. Pal C, Papp B, Lercher MJ: **An integrated view of protein evolution.** *Nat Rev Genet* 2006, **7:**337-348.

72. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW: **Evolutionary rate in the protein interaction network.** *Science* 2002, **296:**750-752.

73. Pal C, Papp B, Hurst LD: **Highly expressed genes in yeast evolve slowly.** *Genetics* 2001, **158:**927-931.

74. Tamames J: **Evolution of gene order conservation in prokaryotes.** *Genome Biol* 2001, **2:**RESEARCH0020.

75. Snir S, Wolf YI, Koonin EV: **Universal pacemaker of genome evolution.** *PLoS Comput Biol* 2012, **8:**e1002785.

76. Oberhardt MA, Palsson BO, Papin JA: **Applications of genome-scale metabolic reconstructions.** *Mol Syst Biol* 2009, **5:**320.

77. Feist AM, Palsson BO: **The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli.** *Nat Biotechnol* 2008, **26:**659-667.

78. Varma A, Palsson BO: **Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110.** *Appl Environ Microbiol* 1994, **60:**3724-3731.

79. Shlomi T, Berkman O, Ruppin E: **Regulatory on/off minimization of metabolic flux changes after genetic perturbations.** *Proc Natl Acad Sci USA* 2005, **102:**7695-7700.

80. Segre D, Vitkup D, Church GM: **Analysis of optimality in natural and perturbed metabolic networks.** *Proc Natl Acad Sci USA* 2002, **99:**15112-15117.

81. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D, et al: **Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models.** *Mol Syst Biol* 2010, **6:**390.

82. Ibarra RU, Edwards JS, Palsson BO: **Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth.** *Nature* 2002, **420:**186-189.

83. Fong SS, Palsson BO: **Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes.** *Nat Genet* 2004, **36:**1056-1058.

84. Kim TY, Sohn SB, Kim YB, Kim WJ, Lee SY: **Recent advances in reconstruction and applications of genome-scale metabolic models.** *Curr Opin Biotechnol* 2011.

85. Notebaart RA, van Enckevort FH, Francke C, Siezen RJ, Teusink B: **Accelerating the reconstruction of genome-scale metabolic networks.** *BMC Bioinformatics* 2006, **7:**296.

86. DeJongh M, Formsma K, Boillot P, Gould J, Rycenga M, Best A: **Toward the automated generation of genome-scale metabolic networks in the SEED.** *BMC Bioinformatics* 2007, **8:**139.

87. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4:**406-425.

88. Kuo CH, Ochman H: **Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria.** *Biol Direct* 2009, **4:**35.

89. Moran NA, Munson MA, Baumann P, Ishikawa H: **A Molecular Clock in Endosymbiotic Bacteria Is Calibrated Using the Insect Hosts.** *Proceedings of the Royal Society of London Series B-Biological Sciences* 1993, **253:**167-171.

90. Bochner BR: **Global phenotypic characterization of bacteria.** *FEMS Microbiol Rev* 2009, **33:**191-205.

91. Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U: **Multidimensional optimality of microbial metabolism.** *Science* 2012, **336:**601-604.

92. Koeppel AF, Wertheim JO, Barone L, Gentile N, Krizanc D, Cohan FM: **Speedy speciation in a bacterial microcosm: new species can arise as frequently as adaptations within a species.** *ISME J* 2013, **7:**1080-1091.

93. Doolittle WF, Zhaxybayeva O: **On the origin of prokaryotic species.** *Genome Res* 2009, **19:**744-756.

94. Koonin EV: **The Biological Big Bang model for the major transitions in evolution.** *Biol Direct* 2007, **2:**21.

95. Lawrence JG, Retchless AC: **The interplay of homologous recombination and horizontal gene transfer in bacterial speciation.** *Methods Mol Biol* 2009, **532:**29-53.

96. Riley MA, Lizotte-Waniewski M: **Population genomics and the bacterial species concept.** *Methods Mol Biol* 2009, **532:**367-377.

97. Bochner BR, Gadzinski P, Panomitros E: **Phenotype microarrays for high-throughput phenotypic testing and assay of gene function.** *Genome Res* 2001, **11:**1246-1255.

98. Shea A, Wolcott M, Daefler S, Rozak DA: **Biolog phenotype microarrays.** *Methods Mol Biol* 2012, **881:**331-373.

99. Borglin S, Joyner D, DeAngelis KM, Khudyakov J, D'Haeseleer P, Joachimiak MP, Hazen T: **Application of phenotypic microarrays to environmental microbiology.** *Curr Opin Biotechnol* 2012, **23:**41-48.

100. Dufresne A, Garczarek L, Partensky F: **Accelerated evolution associated with genome reduction in a free-living prokaryote.** *Genome Biol* 2005, **6:**R14.

101. Koskiniemi S, Sun S, Berg OG, Andersson DI: **Selection-driven gene loss in bacteria.** *PLoS Genet* 2012, **8:**e1002787.

102. van Ham RC, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernandez JM, Jimenez L, Postigo M, Silva FJ, et al: **Reductive genome evolution in Buchnera aphidicola.** *Proc Natl Acad Sci USA* 2003, **100:**581-586.

103. Mendonca AG, Alves RJ, Pereira-Leal JB: **Loss of genetic redundancy in reductive genome evolution.** *PLoS Comput Biol* 2011, **7:**e1001082.

104. Rosinski-Chupin I, Sauvage E, Mairey B, Mangenot S, Ma L, Da Cunha V, Rusniok C, Bouchier C, Barbe V, Glaser P: **Reductive evolution in Streptococcus agalactiae and the emergence of a host adapted lineage.** *BMC Genomics* 2013, **14:**252.

105. Pal C, Papp B, Lercher MJ: **Adaptive evolution of bacterial metabolic networks by horizontal gene transfer.** *Nat Genet* 2005, **37:**1372-1375.

106. Stackebrandt E, Frederiksen W, Garrity GM, Grimont PA, Kampfer P, Maiden MC, Nesme X, Rossello-Mora R, Swings J, Truper HG, et al: **Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology.** *Int J Syst Evol Microbiol* 2002, **52:**1043-1047.

107. Konstantinidis KT, Ramette A, Tiedje JM: **The bacterial species definition in the genomic era.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361:**1929-1940.

108. Snel B, Huynen MA, Dutilh BE: **Genome trees and the nature of genome evolution.** *Annu Rev Microbiol* 2005, **59:**191-209.

109. Yabuuchi E, Kosako Y, Oyaizu H, Yano I, Hotta H, Hashimoto Y, Ezaki T, Arakawa M: **Proposal of Burkholderia gen. nov. and transfer of seven species of the genus Pseudomonas homology group II to the new genus, with the type species Burkholderia cepacia (Palleroni and Holmes 1981) comb. nov.** *Microbiol Immunol* 1992, **36:**1251-1275.

110. Coenye T, Vandamme P, Govan JR, LiPuma JJ: **Taxonomy and identification of the Burkholderia cepacia complex.** *J Clin Microbiol* 2001, **39:**3427-3436.

111. Bordiec S, Paquis S, Lacroix H, Dhondt S, Ait Barka E, Kauffmann S, Jeandet P, Mazeyrat-Gourbeyre F, Clement C, Baillieul F, Dorey S: **Comparative analysis of defence responses induced by the endophytic plant growth-promoting rhizobacterium Burkholderia phytofirmans strain PsJN and the non-host bacterium Pseudomonas syringae pv. pisi in grapevine cell suspensions.** *J Exp Bot* 2011, **62:**595-603.

112. Deng L, Ren Y, Wei C: **Pyrene degradation by Pseudomonas sp. and Burkholderia sp. enriched from coking wastewater sludge.** *J Environ Sci Health A Tox Hazard Subst Environ Eng* 2012, **47:**1984-1991.

113. Dixon SJ, Fedyshyn Y, Koh JL, Prasad TS, Chahwan C, Chua G, Toufighi K, Baryshnikova A, Hayles J, Hoe KL, et al: **Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes.** *Proc Natl Acad Sci USA* 2008, **105:**16653-16658.

114. Tischler J, Lehner B, Fraser AG: **Evolutionary plasticity of genetic interaction networks.** *Nat Genet* 2008, **40:**390-391.

115. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H: **Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection.** *Mol Syst Biol* 2006, **2:**2006 0008.

116. Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, et al: **Essential Bacillus subtilis genes.** *Proc Natl Acad Sci U S A* 2003, **100:**4678-4683.

117. Xu P, Ge X, Chen L, Wang X, Dou Y, Xu JZ, Patel JR, Stone V, Trinh M, Evans K, et al: **Genome-wide essential gene identification in Streptococcus sanguinis.** *Sci Rep* 2011, **1:**125.

118. Langridge GC, Phan MD, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G, et al: **Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants.** *Genome Res* 2009, **19:**2308-2316.

119. Christen B, Abeliuk E, Collier JM, Kalogeraki VS, Passarelli B, Coller JA, Fero MJ, McAdams HH, Shapiro L: **The essential genome of a bacterium.** *Mol Syst Biol* 2011, **7:**528.

120. Darwin C: *The origin of species.* New York, NY: Barnes & Noble Classics; 2008.

121. Wagner A: **The molecular origins of evolutionary innovations.** *Trends Genet* 2011, **27:**397-410.

122. Hindre T, Knibbe C, Beslon G, Schneider D: **New insights into bacterial adaptation through in vivo and in silico experimental evolution.** *Nat Rev Microbiol* 2012, **10:**352-365.

123. Pommerenke C, Musken M, Becker T, Dotsch A, Klawonn F, Haussler S: **Global genotype-phenotype correlations in Pseudomonas aeruginosa.** *PLoS Pathog* 2010, **6:**e1001074.

124. Lobkovsky AE, Wolf YI, Koonin EV: **Gene frequency distributions reject a neutral model of genome evolution.** *Genome Biol Evol* 2013, **5:**233-242.

125. Haegeman B, Weitz JS: **A neutral theory of genome evolution and the frequency distribution of genes.** *BMC Genomics* 2012, **13:**196.

126. Barve A, Wagner A: **A latent capacity for evolutionary innovation through exaptation in metabolic systems.** *Nature* 2013, **500:**203-206.

127. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox.** *Nat Protoc* 2007, **2:**727-738.

128. Real R, Vargas JM: **The probabilistic basis of Jaccard's index of similarity.** *Systematic Biology* 1996, **45:**380-385.

129. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL: **Global organization of metabolic fluxes in the bacterium Escherichia coli.** *Nature* 2004, **427:**839-843.

130. Almaas E, Oltvai ZN, Barabasi AL: **The activity reaction core and plasticity of metabolic networks.** *PLoS Comput Biol* 2005, **1:**e68.

131. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Res* 2008, **36:**D480-484.

132. Karp PD, Paley S, Romero P: **The Pathway Tools software.** *Bioinformatics* 2002, **18 Suppl 1:**S225-232.

133. Lee D, Redfern O, Orengo C: **Predicting protein function from sequence and structure.** *Nat Rev Mol Cell Biol* 2007, **8:**995-1005.

134. Iliopoulos I, Tsoka S, Andrade MA, Janssen P, Audit B, Tramontano A, Valencia A, Leroy C, Sander C, Ouzounis CA: **Genome sequences and great expectations.** *Genome Biol* 2001, **2:**interactions0001-interactions0001.0003.

135. Devos D, Valencia A: **Intrinsic errors in genome annotation.** *Trends Genet* 2001, **17:**429-431.

136. Chang A, Scheer M, Grote A, Schomburg I, Schomburg D: **BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009.** *Nucleic Acids Res* 2009, **37:**D588-592.

137. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D: **GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support.** *Bioinformatics* 1998, **14:**656-664.

138. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, et al: **The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.** *Nucleic Acids Res* 2008, **36:**D623-631.

139. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its new supplement TREMBL.** *Nucleic Acids Res* 1996, **24:**21-25.

140. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96:**4285-4288.

141. Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ: **Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters.** *Nat Genet* 2002, **31:**255-265.

142. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23:**324-328.

143. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: **Protein function in the post-genomic era.** *Nature* 2000, **405:**823-826.

144. Korbel JO, Jensen LJ, von Mering C, Bork P: **Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs.** *Nat Biotechnol* 2004, **22:**911-917.

145. von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, Ouzounis CA, Bork P: **Genome evolution reveals biochemical networks and functional modules.** *Proc Natl Acad Sci USA* 2003, **100:**15428-15433.

146. Green ML, Karp PD: **A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases.** *BMC Bioinformatics* 2004, **5:**76.

147. Koller D, Friedman N: *Probabilistic graphical models : principles and techniques.* Cambridge, MA: MIT Press; 2009.

148. Price ND, Shmulevich I: **Biochemical and statistical network models for systems biology.** *Curr Opin Biotechnol* 2007, **18:**365-370.

149. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31:**365-370.

150. Li SZ: *Markov random field modeling in image analysis.* Tokyo: Springer; 2001.

151. Casella G, George EI: **Explaining the Gibbs sampler.** *Am Stat* 1992, **46:**167-174.

152. Hastings WK: **Monte Carlo sampling methods using Markov chains and their applications.** *Biometrika* 1970, **57:**97-109.

153. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262:**208-214.

154. Kuepfer L, Sauer U, Blank LM: **Metabolic functions of duplicate genes in Saccharomyces cerevisiae.** *Genome Res* 2005, **15:**1421-1430.

155. Kirkpatrick S, Gelatt CD, Jr., Vecchi MP: **Optimization by Simulated Annealing.** *Science* 1983, **220:**671-680.

156. Mo ML, Palsson BO, Herrgard MJ: **Connecting extracellular metabolomic measurements to intracellular flux states in yeast.** *BMC Syst Biol* 2009, **3:**37.

157. Henry CS, Zinner JF, Cohoon MP, Stevens RL: **iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations.** *Genome Biol* 2009, **10:**R69.

158. Oh YK, Palsson BO, Park SM, Schilling CH, Mahadevan R: **Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data.** *J Biol Chem* 2007, **282:**28791-28799.

159. Becker SA, Palsson BO: **Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation.** *BMC Microbiol* 2005, **5:**8.

160. Stamford NP, Capretta A, Battersby AR: **Expression, purification and characterisation of the product from the Bacillus subtilis hemD gene, uroporphyrinogen III synthase.** *Eur J Biochem* 1995, **231:**236-241.

161. Bower S, Perkins JB, Yocum RR, Howitt CL, Rahaim P, Pero J: **Cloning, sequencing, and characterization of the Bacillus subtilis biotin biosynthetic operon.** *J Bacteriol* 1996, **178:**4122-4130.

162. Faille C, Lequette Y, Ronse A, Slomianny C, Garenaux E, Guerardel Y: **Morphology and physico-chemical properties of Bacillus spores surrounded or not with an exosporium: consequences on their ability to adhere to stainless steel.** *Int J Food Microbiol* 2010, **143:**125-135.

163. Eichenberger P, Fujita M, Jensen ST, Conlon EM, Rudner DZ, Wang ST, Ferguson C, Haga K, Sato T, Liu JS, Losick R: **The program of gene transcription for a single differentiating cell type during sporulation in Bacillus subtilis.** *PLoS Biol* 2004, **2:**e328.

164. Timmons SC, Mosher RH, Knowles SA, Jakeman DL: **Exploiting nucleotidylyltransferases to prepare sugar nucleotides.** *Org Lett* 2007, **9:**857-860.

165. Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, Kerner MJ, Frishman D: **Protein abundance profiling of the Escherichia coli cytosol.** *BMC Genomics* 2008, **9:**102.

166. Hecker M, Reder A, Fuchs S, Pagels M, Engelmann S: **Physiological proteomics and stress/starvation responses in Bacillus subtilis and Staphylococcus aureus.** *Res Microbiol* 2009, **160:**245-258.

167. Tam le T, Antelmann H, Eymann C, Albrecht D, Bernhardt J, Hecker M: **Proteome signatures for stress and starvation in Bacillus subtilis as revealed by a 2-D gel image color coding approach.** *Proteomics* 2006, **6:**4565-4585.

168. Galperin MY, Moroz OV, Wilson KS, Murzin AG: **House cleaning, a part of good housekeeping.** *Mol Microbiol* 2006, **59:**5-19.

169. Bairoch A: **The ENZYME database in 2000.** *Nucleic Acids Res* 2000, **28:**304-305.

170. Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P: **iPath2.0: interactive pathway explorer.** *Nucleic Acids Res* 2011, **39:**W412-415.

171. Breitling R, Vitkup D, Barrett MP: **New surveyor tools for charting microbial metabolic maps.** *Nat Rev Microbiol* 2008, **6:**156-161.

172. Zamboni N, Sauer U: **Novel biological insights through metabolomics and 13C-flux analysis.** *Curr Opin Microbiol* 2009, **12:**553-558.

173. Rost B: **Enzyme function less conserved than anticipated.** *J Mol Biol* 2002, **318:**595-608.

174. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci USA* 1988, **85:**2444-2448.

175. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, Marcotte EM: **Protein interaction networks from yeast to human.** *Curr Opin Struct Biol* 2004, **14:**292-299.

176. Spirin V, Gelfand MS, Mironov AA, Mirny LA: **A metabolic network in the evolutionary context: multiscale structure and modularity.** *Proc Natl Acad Sci USA* 2006, **103:**8774-8779.

177. Kharchenko P, Vitkup D, Church GM: **Filling gaps in a metabolic network using expression information.** *Bioinformatics* 2004, **20 Suppl 1:**i178-185.

178. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS: **Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.** *PLoS Biol* 2007, **5:**e8.

179. Huynen MA, Bork P: **Measuring genome evolution.** *Proc Natl Acad Sci USA* 1998, **95:**5849-5856.

180. Lee JM, Sonnhammer EL: **Genomic gene clustering analysis of pathways in eukaryotes.** *Genome Res* 2003, **13:**875-882.

181. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278:**680-686.

182. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, et al: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102:**109-126.

183. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R: **NCBI GEO: mining millions of expression profiles--database and tools.** *Nucleic Acids Res* 2005, **33:**D562-566.

184. Geman S, Geman D: **Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images.** *IEEE T Pattern Anal* 1984, **6:**721-741.

185. Gelfand AE, Smith AFM: **Sampling-based approches to calculating marginal densities.** *J Am Stat Assoc* 1990, **85:**398-409.

186. Metropolis N, Rosenbluth AW, Rosenbluth MN, H. TA, E. T: **Equations of state calculations by fast computing machines.** *J Chem Phys* 1953, **21:**1087-1091.

187. Fuhrer T, Heer D, Begemann B, Zamboni N: **High-throughput, accurate mass metabolome profiling of cellular extracts by flow injection-time-of-flight mass spectrometry.** *Anal Chem* 2011, **83:**7074-7080.

188. Gudmundsson S, Thiele I: **Computationally efficient flux variability analysis.** *BMC Bioinformatics* 2010, **11:**489.

189. Nerlich AG, Schraut B, Dittrich S, Jelinek T, Zink AR: **Plasmodium falciparum in ancient Egypt.** *Emerg Infect Dis* 2008, **14:**1317-1319.

190. W.H.O.: *World malaria report 2008.* Geneva: World Health Organization; 2008.

191. Baird JK: **Effectiveness of antimalarial drugs.** *N Engl J Med* 2005, **352:**1565-1577.

192. Aly AS, Vaughan AM, Kappe SH: **Malaria parasite development in the mosquito and infection of the mammalian host.** *Annu Rev Microbiol* 2009, **63:**195-221.

193. Haldar K, Mohandas N: **Malaria, erythrocytic infection, and anemia.** *Hematology Am Soc Hematol Educ Program* 2009**:**87-93.

194. Wongsrichanalai C, Pickard AL, Wernsdorfer WH, Meshnick SR: **Epidemiology of drug-resistant malaria.** *Lancet Infect Dis* 2002, **2:**209-218.

195. Mackinnon MJ, Marsh K: **The selection landscape of malaria parasites.** *Science* 2010, **328:**866-871.

196. Hyde JE: **Drug-resistant malaria - an insight.** *FEBS J* 2007, **274:**4688-4698.

197. Liu S, Mu J, Jiang H, Su XZ: **Effects of Plasmodium falciparum mixed infections on in vitro antimalarial drug tests and genotyping.** *Am J Trop Med Hyg* 2008, **79:**178-184.

198. Dharia NV, Chatterjee A, Winzeler EA: **Genomics and systems biology in malaria drug discovery.** *Curr Opin Investig Drugs* 2010, **11:**131-138.

199. Carvalho TG, Menard R: **Manipulating the Plasmodium genome.** *Curr Issues Mol Biol* 2005, **7:**39-55.

200. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al: **Genome sequence of the human malaria parasite Plasmodium falciparum.** *Nature* 2002, **419:**498-511.

201. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL: **The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum.** *PLoS Biol* 2003, **1:**E5.

202. Hu G, Cabrera A, Kono M, Mok S, Chaal BK, Haase S, Engelberg K, Cheemadan S, Spielmann T, Preiser PR, et al: **Transcriptional profiling of growth perturbations of the human malaria parasite Plasmodium falciparum.** *Nat Biotechnol* 2010, **28:**91-98.

203. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder AA, Batalov S, Carucci DJ, Winzeler EA: **Discovery of gene function by expression profiling of the malaria parasite life cycle.** *Science* 2003, **301:**1503-1508.

204. Llinas M, Bozdech Z, Wong ED, Adai AT, DeRisi JL: **Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains.** *Nucleic Acids Res* 2006, **34:**1166-1173.

205. Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, Pain A, Sauerwein RW, Eling WM, Hall N, Waters AP, Stunnenberg HG, Mann M: **Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry.** *Nature* 2002, **419:**537-542.

206. Lasonder E, Janse CJ, van Gemert GJ, Mair GR, Vermunt AM, Douradinha BG, van Noort V, Huynen MA, Luty AJ, Kroeze H, et al: **Proteomic profiling of Plasmodium sporozoite maturation identifies new proteins essential for parasite development and infectivity.** *PLoS Pathog* 2008, **4:**e1000195.

207. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, et al: **A proteomic view of the Plasmodium falciparum life cycle.** *Nature* 2002, **419:**520-526.

208. Olszewski KL, Morrisey JM, Wilinski D, Burns JM, Vaidya AB, Rabinowitz JD, Llinas M: **Host-parasite interactions revealed by Plasmodium falciparum metabolomics.** *Cell Host Microbe* 2009, **5:**191-199.

209. Yeh I, Hanekamp T, Tsoka S, Karp PD, Altman RB: **Computational analysis of Plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery.** *Genome Res* 2004, **14:**917-924.

210. **Malaria Parasite Metabolic Pathways** [http://sites.huji.ac.il/malaria/]

211. Price ND, Reed JL, Palsson BO: **Genome-scale models of microbial cells: evaluating the consequences of constraints.** *Nat Rev Microbiol* 2004, **2:**886-897.

212. Chavali AK, Whittemore JD, Eddy JA, Williams KT, Papin JA: **Systems analysis of metabolism in the pathogenic trypanosomatid Leishmania major.** *Mol Syst Biol* 2008, **4:**177.

213. Navid A, Almaas E: **Genome-scale reconstruction of the metabolic network in Yersinia pestis, strain 91001.** *Mol Biosyst* 2009, **5:**368-375.

214. Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, Farhat MR, Cheng TY, Moody DB, Murray M, Galagan JE: **Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production.** *PLoS Comput Biol* 2009, **5:**e1000489.

215. Schuetz R, Kuepfer L, Sauer U: **Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli.** *Mol Syst Biol* 2007, **3:**119.

216. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO: **Reconstruction of biochemical networks in microorganisms.** *Nat Rev Microbiol* 2009, **7:**129-143.

217. Ralph SA, van Dooren GG, Waller RF, Crawford MJ, Fraunholz MJ, Foth BJ, Tonkin CJ, Roos DS, McFadden GI: **Tropical infectious diseases: metabolic maps and functions of the Plasmodium falciparum apicoplast.** *Nat Rev Microbiol* 2004, **2:**203-216.

218. Duarte NC, Herrgard MJ, Palsson BO: **Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model.** *Genome Res* 2004, **14:**1298-1309.

219. Kirk K, Staines HM, Martin RE, Saliba KJ: **Transport properties of the host cell membrane.** *Novartis Found Symp* 1999, **226:**55-66; discussion 66-73.

220. Martin RE, Henry RI, Abbey JL, Clements JD, Kirk K: **The 'permeome' of the malaria parasite: an overview of the membrane transport proteins of Plasmodium falciparum.** *Genome Biol* 2005, **6:**R26.

221. Koncarevic S, Rohrbach P, Deponte M, Krohne G, Prieto JH, Yates J, 3rd, Rahlfs S, Becker K: **The malarial parasite Plasmodium falciparum imports the human protein peroxiredoxin 2 for peroxide detoxification.** *Proc Natl Acad Sci USA* 2009, **106:**13323-13328.

222. Bonday ZQ, Taketani S, Gupta PD, Padmanaban G: **Heme biosynthesis by the malarial parasite. Import of delta-aminolevulinate dehydrase from the host red cell.** *J Biol Chem* 1997, **272:**21839-21846.

223. Francis SE, Sullivan DJ, Jr., Goldberg DE: **Hemoglobin metabolism in the malaria parasite Plasmodium falciparum.** *Annu Rev Microbiol* 1997, **51:**97-123.

224. Saito T, Maeda T, Nakazawa M, Takeuchi T, Nozaki T, Asai T: **Characterisation of hexokinase in Toxoplasma gondii tachyzoites.** *Int J Parasitol* 2002, **32:**961-967.

225. LeRoux M, Lakshmanan V, Daily JP: **Plasmodium falciparum biology: analysis of in vitro versus in vivo growth conditions.** *Trends Parasitol* 2009, **25:**474-481.

226. Janse CJ, Ramesar J, Waters AP: **High-efficiency transfection and drug selection of genetically transformed blood stages of the rodent malaria parasite Plasmodium berghei.** *Nat Protoc* 2006, **1:**346-356.

227. Janse CJ, Waters AP: **Plasmodium berghei: the application of cultivation and purification techniques to molecular studies of malaria parasites.** *Parasitol Today* 1995, **11:**138-143.

228. Hanada K, Palacpac NM, Magistrado PA, Kurokawa K, Rai G, Sakata D, Hara T, Horii T, Nishijima M, Mitamura T: **Plasmodium falciparum phospholipase C hydrolyzing**

sphingomyelin and lysocholinephospholipids is a possible target for malaria chemotherapy. *J Exp Med* 2002, **195:**23-34.

229. Yu M, Kumar TR, Nkrumah LJ, Coppi A, Retzlaff S, Li CD, Kelly BJ, Moura PA, Lakshmanan V, Freundlich JS, et al: **The fatty acid biosynthesis enzyme FabI plays a key role in the development of liver-stage malarial parasites.** *Cell Host Microbe* 2008, **4:**567-578.

230. Surolia N, Surolia A: **Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of Plasmodium falciparum.** *Nat Med* 2001, **7:**167-173.

231. Vaughan AM, O'Neill MT, Tarun AS, Camargo N, Phuong TM, Aly AS, Cowman AF, Kappe SH: **Type II fatty acid synthesis is essential only for malaria parasite late liver stage development.** *Cell Microbiol* 2009, **11:**506-520.

232. Dessens JT, Mendoza J, Claudianos C, Vinetz JM, Khater E, Hassard S, Ranawaka GR, Sinden RE: **Knockout of the rodent malaria parasite chitinase pbCHT1 reduces infectivity to mosquitoes.** *Infect Immun* 2001, **69:**4041-4047.

233. Spry C, Chai CL, Kirk K, Saliba KJ: **A class of pantothenic acid analogs inhibits Plasmodium falciparum pantothenate kinase and represses the proliferation of malaria parasites.** *Antimicrob Agents Chemother* 2005, **49:**4649-4657.

234. Zhang Y, Meshnick SR: **Inhibition of Plasmodium falciparum dihydropteroate synthetase and growth in vitro by sulfa drugs.** *Antimicrob Agents Chemother* 1991, **35:**267-271.

235. Jiang L, Lee PC, White J, Rathod PK: **Potent and selective activity of a combination of thymidine and 1843U89, a folate-based thymidylate synthase inhibitor, against Plasmodium falciparum.** *Antimicrob Agents Chemother* 2000, **44:**1047-1050.

236. Schnick C, Polley SD, Fivelman QL, Ranford-Cartwright LC, Wilkinson SR, Brannigan JA, Wilkinson AJ, Baker DA: **Structure and non-essential function of glycerol kinase in Plasmodium falciparum blood stages.** *Mol Microbiol* 2009, **71:**533-545.

237. Wanidworanun C, Nagel RL, Shear HL: **Antisense oligonucleotides targeting malarial aldolase inhibit the asexual erythrocytic stages of Plasmodium falciparum.** *Mol Biochem Parasitol* 1999, **102:**91-101.

238. Razakantoanina V, Nguyen Kim PP, Jaureguiberry G: **Antimalarial activity of new gossypol derivatives.** *Parasitol Res* 2000, **86:**665-668.

239. Cassera MB, Merino EF, Peres VJ, Kimura EA, Wunderlich G, Katzin AM: **Effect of fosmidomycin on metabolic and transcript profiles of the methylerythritol phosphate pathway in Plasmodium falciparum.** *Mem Inst Oswaldo Cruz* 2007, **102:**377-383.

240. Ono T, Cabrita-Santos L, Leitao R, Bettiol E, Purcell LA, Diaz-Pulido O, Andrews LB, Tadakuma T, Bhanot P, Mota MM, Rodriguez A: **Adenylyl cyclase alpha and cAMP**

signaling mediate Plasmodium sporozoite apical regulated exocytosis and hepatocyte infection. *PLoS Pathog* 2008, **4:**e1000008.

241. Das Gupta R, Krause-Ihle T, Bergmann B, Muller IB, Khomutov AR, Muller S, Walter RD, Luersen K: **3-Aminooxy-1-aminopropane and derivatives have an antiproliferative effect on cultured Plasmodium falciparum by decreasing intracellular polyamine concentrations.** *Antimicrob Agents Chemother* 2005, **49:**2857-2864.

242. Muller IB, Das Gupta R, Luersen K, Wrenger C, Walter RD: **Assessing the polyamine metabolism of Plasmodium falciparum as chemotherapeutic target.** *Mol Biochem Parasitol* 2008, **160:**1-7.

243. Ramya TN, Surolia N, Surolia A: **Polyamine synthesis and salvage pathways in the malaria parasite Plasmodium falciparum.** *Biochem Biophys Res Commun* 2006, **348:**579-584.

244. Haider N, Eschbach ML, Dias Sde S, Gilberger TW, Walter RD, Luersen K: **The spermidine synthase of the malaria parasite Plasmodium falciparum: molecular and biochemical characterisation of the polyamine synthesis enzyme.** *Mol Biochem Parasitol* 2005, **142:**224-236.

245. Ramya TN, Mishra S, Karmodiya K, Surolia N, Surolia A: **Inhibitors of nonhousekeeping functions of the apicoplast defy delayed death in Plasmodium falciparum.** *Antimicrob Agents Chemother* 2007, **51:**307-316.

246. Dawson PA, Cochran DA, Emmerson BT, Gordon RB: **Inhibition of Plasmodium falciparum hypoxanthine-guanine phosphoribosyltransferase mRNA by antisense oligodeoxynucleotide sequence.** *Mol Biochem Parasitol* 1993, **60:**153-156.

247. Webster HK, Whaun JM, Walker MD, Bean TL: **Synthesis of adenosine nucleotides from hypoxanthine by human malaria parasites (Plasmodium falciparum) in continuous erythrocyte culture: inhibition by hadacidin but not alanosine.** *Biochem Pharmacol* 1984, **33:**1555-1557.

248. Bulusu V, Srinivasan B, Bopanna MP, Balaram H: **Elucidation of the substrate specificity, kinetic and catalytic mechanism of adenylosuccinate lyase from Plasmodium falciparum.** *Biochim Biophys Acta* 2009, **1794:**642-654.

249. Hirai M, Arai M, Kawai S, Matsuoka H: **PbGCbeta is essential for Plasmodium ookinete motility to invade midgut cell and for successful completion of parasite life cycle in mosquitoes.** *J Biochem* 2006, **140:**747-757.

250. Deng X, Gujjar R, El Mazouni F, Kaminsky W, Malmquist NA, Goldsmith EJ, Rathod PK, Phillips MA: **Structural plasticity of malaria dihydroorotate dehydrogenase allows selective binding of diverse chemical scaffolds.** *J Biol Chem* 2009, **284:**26999-27009.

251. Ho MC, Cassera MB, Madrid DC, Ting LM, Tyler PC, Kim K, Almo SC, Schramm VL: **Structural and metabolic specificity of methylthiocoformycin for malarial adenosine deaminases.** *Biochemistry* 2009, **48:**9618-9626.

252. Kicska GA, Tyler PC, Evans GB, Furneaux RH, Schramm VL, Kim K: **Purine-less death in Plasmodium falciparum induced by immucillin-H, a transition state analogue of purine nucleoside phosphorylase.** *J Biol Chem* 2002, **277:**3226-3231.

253. Chakrabarti D, Schuster SM, Chakrabarti R: **Cloning and characterization of subunit genes of ribonucleotide reductase, a cell-cycle-regulated enzyme, from Plasmodium falciparum.** *Proc Natl Acad Sci U S A* 1993, **90:**12020-12024.

254. Krungkrai J, Krungkrai SR, Supuran CT: **Carbonic anhydrase inhibitors: inhibition of Plasmodium falciparum carbonic anhydrase with aromatic/heterocyclic sulfonamides-in vitro and in vivo studies.** *Bioorg Med Chem Lett* 2008, **18:**5466-5471.

255. Flores MV, Atkins D, Wade D, O'Sullivan WJ, Stewart TS: **Inhibition of Plasmodium falciparum proliferation in vitro by ribozymes.** *J Biol Chem* 1997, **272:**16940-16945.

256. Nguyen C, Kasinathan G, Leal-Cortijo I, Musso-Buendia A, Kaiser M, Brun R, Ruiz-Perez LM, Johansson NG, Gonzalez-Pacanowska D, Gilbert IH: **Deoxyuridine triphosphate nucleotidohydrolase as a potential antiparasitic drug target.** *J Med Chem* 2005, **48:**5942-5954.

257. Thornalley PJ, Strath M, Wilson RJ: **Antimalarial activity in vitro of the glyoxalase I inhibitor diester, S-p-bromobenzylglutathione diethyl ester.** *Biochem Pharmacol* 1994, **47:**418-420.

258. Yano K, Komaki-Yasuda K, Tsuboi T, Torii M, Kano S, Kawazu S: **2-Cys Peroxiredoxin TPx-1 is involved in gametocyte development in Plasmodium berghei.** *Mol Biochem Parasitol* 2006, **148:**44-51.

259. Yano K, Otsuki H, Arai M, Komaki-Yasuda K, Tsuboi T, Torii M, Kano S, Kawazu S: **Disruption of the Plasmodium berghei 2-Cys peroxiredoxin TPx-1 gene hinders the sporozoite development in the vector mosquito.** *Mol Biochem Parasitol* 2008, **159:**142-145.

260. Vega-Rodriguez J, Franke-Fayard B, Dinglasan RR, Janse CJ, Pastrana-Mena R, Waters AP, Coppens I, Rodriguez-Orengo JF, Srinivasan P, Jacobs-Lorena M, Serrano AE: **The glutathione biosynthetic pathway of Plasmodium is essential for mosquito transmission.** *PLoS Pathog* 2009, **5:**e1000302.

261. Krnajski Z, Gilberger TW, Walter RD, Cowman AF, Muller S: **Thioredoxin reductase is essential for the survival of Plasmodium falciparum erythrocytic stages.** *J Biol Chem* 2002, **277:**25970-25975.

262. Roberts F, Roberts CW, Johnson JJ, Kyle DE, Krell T, Coggins JR, Coombs GH, Milhous WK, Tzipori S, Ferguson DJ, et al: **Evidence for the shikimate pathway in apicomplexan parasites.** *Nature* 1998, **393:**801-805.

263. McRobert L, McConkey GA: **RNA interference (RNAi) inhibits growth of Plasmodium falciparum.** *Mol Biochem Parasitol* 2002, **119:**273-278.

264. Mukkamala D, No JH, Cass LM, Chang TK, Oldfield E: **Bisphosphonate inhibition of a Plasmodium farnesyl diphosphate synthase and a general method for predicting cell-based activity from enzyme data.** *J Med Chem* 2008, **51:**7827-7833.

265. Promeneur D, Liu Y, Maciel J, Agre P, King LS, Kumar N: **Aquaglyceroporin PbAQP during intraerythrocytic development of the malaria parasite Plasmodium berghei.** *Proc Natl Acad Sci USA* 2007, **104:**2211-2216.

266. Bhanot P, Schauer K, Coppens I, Nussenzweig V: **A surface phospholipase is involved in the migration of plasmodium sporozoites through cells.** *J Biol Chem* 2005, **280:**6752-6760.

267. Lim L, McFadden GI: **The evolution, metabolism and functions of the apicoplast.** *Philos Trans R Soc Lond B Biol Sci* 2010, **365:**749-763.

268. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD: **Computational prediction of human metabolic pathways from the complete human genome.** *Genome Biol* 2005, **6:**R2.

269. van Dooren GG, Stimmler LM, McFadden GI: **Metabolic maps and functions of the Plasmodium mitochondrion.** *FEMS Microbiol Rev* 2006, **30:**596-630.

270. Foth BJ, Stimmler LM, Handman E, Crabb BS, Hodder AN, McFadden GI: **The malaria parasite Plasmodium falciparum has only one pyruvate dehydrogenase complex, which is located in the apicoplast.** *Mol Microbiol* 2005, **55:**39-53.

271. Dixon SJ, Costanzo M, Baryshnikova A, Andrews B, Boone C: **Systematic mapping of genetic interaction networks.** *Annu Rev Genet* 2009, **43:**601-625.

272. Wagner A: **Robustness, evolvability, and neutrality.** *FEBS Lett* 2005, **579:**1772-1778.

273. Vaidya AB, Mather MW: **Mitochondrial evolution and functions in malaria parasites.** *Annu Rev Microbiol* 2009, **63:**249-267.

274. Berwal R, Gopalan N, Chandel K, Prasad GB, Prakash S: **Plasmodium falciparum: enhanced soluble expression, purification and biochemical characterization of lactate dehydrogenase.** *Exp Parasitol* 2008, **120:**135-141.

275. Reed JL, Vo TD, Schilling CH, Palsson BO: **An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR).** *Genome Biol* 2003, **4:**R54.

276. Ting LM, Shi W, Lewandowicz A, Singh V, Mwakingwe A, Birck MR, Ringia EA, Bench G, Madrid DC, Tyler PC, et al: **Targeting a novel Plasmodium falciparum purine recycling pathway with specific immucillins.** *J Biol Chem* 2005, **280:**9547-9554.

277. Hyde JE, Dittrich S, Wang P, Sims PF, de Crecy-Lagard V, Hanson AD: **Plasmodium falciparum: a paradigm for alternative folate biosynthesis in diverse microorganisms?** *Trends Parasitol* 2008, **24:**502-508.

278. Mi-Ichi F, Kita K, Mitamura T: **Intraerythrocytic Plasmodium falciparum utilize a broad range of serum-derived fatty acids with limited modification for their growth.** *Parasitology* 2006, **133:**399-410.

279. Hsiao LL, Howard RJ, Aikawa M, Taraschi TF: **Modification of host cell membrane lipid composition by the intra-erythrocytic human malaria parasite Plasmodium falciparum.** *Biochem J* 1991, **274 (Pt 1):**121-132.

280. Magni G, Di Stefano M, Orsomando G, Raffaelli N, Ruggieri S: **NAD(P) biosynthesis enzymes as potential targets for selective drug design.** *Curr Med Chem* 2009, **16:**1372-1390.

281. Merrick CJ, Duraisingh MT: **Plasmodium falciparum Sir2: an unusual sirtuin with dual histone deacetylase and ADP-ribosyltransferase activity.** *Eukaryot Cell* 2007, **6:**2081-2091.

282. Sorci L, Pan Y, Eyobo Y, Rodionova I, Huang N, Kurnasov O, Zhong S, MacKerell AD, Jr., Zhang H, Osterman AL: **Targeting NAD biosynthesis in bacterial pathogens: Structure-based development of inhibitors of nicotinate mononucleotide adenylyltransferase NadD.** *Chem Biol* 2009, **16:**849-861.

283. Smilkstein M, Sriwilaijaroen N, Kelly JX, Wilairat P, Riscoe M: **Simple and inexpensive fluorescence-based technique for high-throughput antimalarial drug screening.** *Antimicrob Agents Chemother* 2004, **48:**1803-1806.

284. Divo AA, Geary TG, Davis NL, Jensen JB: **Nutritional requirements of Plasmodium falciparum in culture. I. Exogenously supplied dialyzable components necessary for continuous growth.** *J Protozool* 1985, **32:**59-64.

285. Daran-Lapujade P, Jansen ML, Daran JM, van Gulik W, de Winde JH, Pronk JT: **Role of transcriptional regulation in controlling fluxes in central carbon metabolism of Saccharomyces cerevisiae. A chemostat culture study.** *J Biol Chem* 2004, **279:**9125-9138.

286. Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E: **Network-based prediction of human tissue-specific metabolism.** *Nat Biotechnol* 2008, **26:**1003-1010.

287. Ginsburg H: **Progress in in silico functional genomics: the malaria Metabolic Pathways database.** *Trends Parasitol* 2006, **22:**238-240.

288. Ginsburg H: **Caveat emptor: limitations of the automated reconstruction of metabolic pathways in Plasmodium.** *Trends Parasitol* 2009, **25:**37-43.

289. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, et al: **Functional profiling of the Saccharomyces cerevisiae genome.** *Nature* 2002, **418:**387-391.

290. Covert MW, Palsson BO: **Transcriptional regulation in constraints-based metabolic models of Escherichia coli.** *J Biol Chem* 2002, **277:**28058-28064.

291. De Silva EK, Gehrke AR, Olszewski K, Leon I, Chahal JS, Bulyk ML, Llinas M: **Specific DNA-binding by apicomplexan AP2 transcription factors.** *Proc Natl Acad Sci U S A* 2008, **105:**8393-8398.

292. Chung DW, Ponts N, Cervantes S, Le Roch KG: **Post-translational modifications in Plasmodium: more than you think!** *Mol Biochem Parasitol* 2009, **168:**123-134.

293. Dhanasekaran S, Chandra NR, Chandrasekhar Sagar BK, Rangarajan PN, Padmanaban G: **Delta-aminolevulinic acid dehydratase from Plasmodium falciparum: indigenous versus imported.** *J Biol Chem* 2004, **279:**6934-6942.

294. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO: **Global reconstruction of the human metabolic network based on genomic and bibliomic data.** *Proc Natl Acad Sci USA* 2007, **104:**1777-1782.

295. Zhao J, Geng C, Tao L, Zhang D, Jiang Y, Tang K, Zhu R, Yu H, Zhang WD, He F, et al: **Reconstruction and analysis of human liver-specific metabolic network based on CNHLPP data.** *J Proteome Res* 2010.

296. Jamshidi N, Edwards JS, Fahland T, Church GM, Palsson BO: **Dynamic simulation of the human red blood cell metabolic network.** *Bioinformatics* 2001, **17:**286-287.

297. Joshi A, Palsson BO: **Metabolic dynamics in the human red cell. Part III--Metabolic reaction rates.** *J Theor Biol* 1990, **142:**41-68.

298. Joshi A, Palsson BO: **Metabolic dynamics in the human red cell. Part IV--Data prediction and some model computations.** *J Theor Biol* 1990, **142:**69-85.

299. Joshi A, Palsson BO: **Metabolic dynamics in the human red cell. Part I--A comprehensive kinetic model.** *J Theor Biol* 1989, **141:**515-528.

300. Joshi A, Palsson BO: **Metabolic dynamics in the human red cell. Part II--Interactions with the environment.** *J Theor Biol* 1989, **141:**529-545.

301. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P: **A side effect resource to capture phenotypic effects of drugs.** *Mol Syst Biol* 2010, **6:**343.

302. Yao L, Rzhetsky A: **Quantitative systems-level determinants of human genes targeted by successful drugs.** *Genome Res* 2008, **18:**206-213.

303. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, et al: **PlasmoDB: a functional genomic database for malaria parasites.** *Nucleic Acids Res* 2009, **37:**D539-543.

304. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO: **A genome-scale metabolic reconstruction for**

Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 2007, **3:**121.

305. Chan M, Tan DS, Wong SH, Sim TS: **A relevant in vitro eukaryotic live-cell system for the evaluation of plasmodial protein localization.** *Biochimie* 2006, **88:**1367-1375.

306. Waller RF, Keeling PJ, Donald RG, Striepen B, Handman E, Lang-Unnasch N, Cowman AF, Besra GS, Roos DS, McFadden GI: **Nuclear-encoded proteins target to the plastid in Toxoplasma gondii and Plasmodium falciparum.** *Proc Natl Acad Sci USA* 1998, **95:**12352-12357.

307. Edwards JS, Ramakrishna R, Schilling CH, B.O. P: **Metabolic flux balance analysis.** In *Metabolic engineering.* Edited by Lee SYPET. New York, NY: Marcel Dekker Inc.; 1999: pp. 13-57

308. Mullin KA, Lim L, Ralph SA, Spurck TP, Handman E, McFadden GI: **Membrane transporters in the relict plastid of malaria parasites.** *Proc Natl Acad Sci USA* 2006, **103:**9572-9577.

309. Quashie NB, Dorin-Semblat D, Bray PG, Biagini GA, Doerig C, Ranford-Cartwright LC, De Koning HP: **A comprehensive model of purine uptake by the malaria parasite Plasmodium falciparum: identification of four purine transport activities in intraerythrocytic parasites.** *Biochem J* 2008, **411:**287-295.

310. Wright J, Wagner A: **The Systems Biology Research Tool: evolvable open-source software.** *BMC Syst Biol* 2008, **2:**55.

311. Le Novere N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, et al: **BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems.** *Nucleic Acids Res* 2006, **34:**D689-691.

312. Trager W, Jensen JB: **Human malaria parasites in continuous culture. 1976.** *J Parasitol* 2005, **91:**484-486.

313. Lambros C, Vanderberg JP: **Synchronization of Plasmodium falciparum erythrocytic stages in culture.** *J Parasitol* 1979, **65:**418-420.

314. Kaufman DE, Smith RL: **Direction choice for accelerated convergence in hit-and-run sampling.** *Operations Research* 1998, **46:**84-95.

315. Margoliash E: **Primary structure and evolution of cytochrome C.** *Proc Natl Acad Sci USA* 1963, **50:**672-679.

316. Benton MJ, Donoghue PC: **Paleontological evidence to date the tree of life.** *Mol Biol Evol* 2007, **24:**26-53.

317. Gillespie JH: *The Causes of Molecular Evolution.* New York: Oxford University Press; 1991.

318. Ayala FJ: **Molecular clock mirages.** *Bioessays* 1999, **21:**71-75.

319. Zuckerkandl E: **Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins.** *J Mol Evol* 1976, **7:**167-183.

320. Wilson AC, Carlson SS, White TJ: **Biochemical evolution.** *Annu Rev Biochem* 1977, **46:**573-639.

321. Wang Z, Zhang JZ: **Why Is the correlation between gene importance and gene evolutionary rate so weak?** *PLoS Genet* 2009, **5:**e1000329.

322. Vitkup D, Kharchenko P, Wagner A: **Influence of metabolic network structure and function on enzyme evolution.** *Genome Biol* 2006, **7:**R39.

323. Jordan IK, Rogozin IB, Wolf YI, Koonin EV: **Essential genes are more evolutionarily conserved than are nonessential genes in bacteria.** *Genome Res* 2002, **12:**962-968.

324. Hirsh AE, Fraser HB: **Protein dispensability and rate of evolution.** *Nature* 2001, **411:**1046-1049.

325. Jordan IK, Wolf YI, Koonin EV: **No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly.** *BMC Evol Biol* 2003, **3:**5.

326. Rocha EP, Danchin A: **An analysis of determinants of amino acids substitution rates in bacterial proteins.** *Mol Biol Evol* 2004, **21:**108-116.

327. Xia Y, Franzosa EA, Gerstein MB: **Integrated assessment of genomic correlates of protein evolutionary rate.** *PLoS Comput Biol* 2009, **5:**e1000413.

328. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH: **Why highly expressed proteins evolve slowly.** *Proc Natl Acad Sci USA* 2005, **102:**14338-14343.

329. Marais G, Duret L: **Synonymous codon usage, accuracy of translation, and gene length in Caenorhabditis elegans.** *J Mol Evol* 2001, **52:**275-280.

330. Rocha EP: **The quest for the universals of protein evolution.** *Trends Genet* 2006, **22:**412-416.

331. Drummond DA, Wilke CO: **Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution.** *Cell* 2008, **134:**341-352.

332. Drummond DA, Wilke CO: **The evolutionary consequences of erroneous protein synthesis.** *Nat Rev Genet* 2009, **10:**715-724.

333. Kramer EB, Farabaugh PJ: **The frequency of translational misreading errors in E. coli is largely determined by tRNA competition.** *RNA* 2007, **13:**87-96.

334. Stansfield I, Jones KM, Herbert P, Lewendon A, Shaw WV, Tuite MF: **Missense translation errors in Saccharomyces cerevisiae.** *J Mol Biol* 1998, **282:**13-24.

335. Müller-Hill B: *The lac Operon: A Short History of a Genetic Paradigm.* New York: Walter de Gruyter; 1996.

336. Dong HJ, Nilsson L, Kurland CG: **Gratuitous overexpression of genes in Escherichia coli leads to growth inhibition and ribosome destruction.** *J Bacteriol* 1997, **179:**2096-2096.

337. Dekel E, Alon U: **Optimality and evolutionary tuning of the expression level of a protein.** *Nature* 2005, **436:**588-592.

338. Pakula AA, Sauer RT: **Genetic analysis of protein stability and function.** *Annu Rev Genet* 1989, **23:**289-310.

339. Matthews BW: **Structural and genetic analysis of the folding and function of T4 lysozyme.** *FASEB J* 1996, **10:**35-41.

340. Vlahovicek K, Pintar A, Parthasarathi L, Carugo O, Pongor S: **CX, DPX and PRIDE: WWW servers for the analysis and comparison of protein 3D structures.** *Nucleic Acids Res* 2005, **33:**W252-W254.

341. Capriotti E, Fariselli P, Casadio R: **I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure.** *Nucleic Acids Res* 2005, **33:**W306-W310.

342. Vind J, Sorensen MA, Rasmussen MD, Pedersen S: **Synthesis of proteins in Escherichia coli is limited by the concentration of free ribosomes. Expression from reporter genes does not always reflect functional mRNA levels.** *J Mol Biol* 1993, **231:**678-688.

343. Sambrook J, Russell DW: *Molecular Cloning: A laboratory manual.* CSHL Press; 2001.

344. Lesley SA, Graziano J, Cho CY, Knuth MW, Klock HE: **Gene expression response to misfolded protein as a screen for soluble recombinant protein.** *Protein Eng* 2002, **15:**153-160.

345. Parsell DA, Sauer RT: **Induction of a heat shock-like response by unfolded protein in Escherichia Coli: Dependence on protein level not protein degradation.** *Genes Dev* 1989, **3:**1226-1232.

346. Wang IN, Deaton J, Young R: **Sizing the holin lesion with an endolysin-beta-galactosidase fusion.** *J Bacteriol* 2003, **185:**779-787.

347. Andersson SG, Kurland CG: **Codon preferences in free-living microorganisms.** *Microbiol Rev* 1990, **54:**198-210.

348. Bulmer M: **The selection-mutation-drift theory of synonymous codon usage.** *Genetics* 1991, **129:**897-907.

349. Akashi H: **Gene expression and molecular evolution.** *Curr Opin Genet Dev* 2001, **11:**660-666.

350. Kudla G, Murray AW, Tollervey D, Plotkin JB: **Coding-sequence determinants of gene expression in Escherichia coli.** *Science* 2009, **324:**255-258.

351. Niwa T, Ying BW, Saito K, Jin W, Takada S, Ueda T, Taguchi H: **Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins.** *Proc Natl Acad Sci USA* 2009, **106:**4201-4206.

352. Koonin EV, Wolf YI: **Constraints and plasticity in genome and molecular-phenome evolution.** *Nat Rev Genet* 2010, **11:**487-498.

353. Dobson CM: **Protein misfolding, evolution and disease.** *Trends Biochem Sci* 1999, **24:**329-332.

354. Tartaglia GG, Pechmann S, Dobson CM, Vendruscolo M: **A relationship between mRNA expression levels and protein solubility in E. coli.** *J Mol Biol* 2009, **388:**381-389.

355. de Groot NS, Ventura S: **Protein aggregation profile of the bacterial cytosol.** *PLoS One* 2010, **5:**e9383.

356. Zhang J, Maslov S, Shakhnovich EI: **Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size.** *Mol Syst Biol* 2008, **4:**210.

357. Strub C, Alies C, Lougarre A, Ladurantie C, Czaplicki J, Fournier D: **Mutation of exposed hydrophobic amino acids to arginine to increase protein stability.** *BMC Biochem* 2004, **5:**9.

358. Baldwin RL: **Energetics of protein folding.** *J Mol Biol* 2007, **371:**283-301.

359. Dill KA: **Dominant forces in protein folding.** *Biochemistry* 1990, **29:**7133-7155.

360. Honig B, Yang AS: **Free energy balance in protein folding.** *Adv Protein Chem* 1995, **46:**27-58.

361. Thompson MJ, Eisenberg D: **Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability.** *J Mol Biol* 1999, **290:**595-604.

362. Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A: **ProTherm, version 4.0: thermodynamic database for proteins and mutants.** *Nucleic Acids Res* 2004, **32:**D120-121.

363. Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, Finkelstein AV: **Contact order revisited: influence of protein size on the folding rate.** *Protein Sci* 2003, **12:**2057-2062.

364. Plaxco KW, Simons KT, Baker D: **Contact order, transition state placement and the refolding rates of single domain proteins.** *J Mol Biol* 1998, **277:**985-994.

365. Lu P, Vogel C, Wang R, Yao X, Marcotte EM: **Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation.** *Nat Biotechnol* 2007, **25:**117-124.

366. Thatcher JW, Shaw JM, Dickinson WJ: **Marginal fitness contributions of nonessential genes in yeast.** *Proc Natl Acad Sci USA* 1998, **95:**253-257.

367. Lindner AB, Madden R, Demarez A, Stewart EJ, Taddei F: **Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation.** *Proc Natl Acad Sci USA* 2008, **105:**3076-3081.

368. Stoebel DM, Dean AM, Dykhuizen DE: **The cost of expression of Escherichia coli lac operon proteins is in the process, not in the products.** *Genetics* 2008, **178:**1653-1660.

369. Tuller T, Waldman YY, Kupiec M, Ruppin E: **Translation efficiency is determined by both codon bias and folding energy.** *Proc Natl Acad Sci USA* 2010, **107:**3645-3650.

370. Akashi H: **Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy.** *Genetics* 1994, **136:**927-935.

371. Huang Y, Koonin EV, Lipman DJ, Przytycka TM: **Selection for minimization of translational frameshifting errors as a factor in the evolution of codon usage.** *Nucleic Acids Res* 2009, **37:**6799-6810.

372. Stoletzki N, Eyre-Walker A: **Synonymous codon usage in Escherichia coli: selection for translational accuracy.** *Mol Biol Evol* 2007, **24:**374-381.

373. Liao BY, Scott NM, Zhang J: **Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins.** *Mol Biol Evol* 2006, **23:**2072-2080.

374. Wolf MY, Wolf YI, Koonin EV: **Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution.** *Biol Direct* 2008, **3:**40.

375. Bucciantini M, Giannoni E, Chiti F, Baroni F, Formigli L, Zurdo J, Taddei N, Ramponi G, Dobson CM, Stefani M: **Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases.** *Nature* 2002, **416:**507-511.

376. Gidalevitz T, Ben-Zvi A, Ho KH, Brignull HR, Morimoto RI: **Progressive disruption of cellular protein folding in models of polyglutamine diseases.** *Science* 2006, **311:**1471-1474.

377. Munch C, Bertolotti A: **Exposure of Hydrophobic Surfaces Initiates Aggregation of Diverse ALS-Causing Superoxide Dismutase-1 Mutants.** *J Mol Biol* 2010.

378. Link CD, Fonte V, Hiester B, Yerg J, Ferguson J, Csontos S, Silverman MA, Stein GH: **Conversion of green fluorescent protein into a toxic, aggregation-prone protein by C-terminal addition of a short peptide.** *Journal of Biological Chemistry* 2006, **281:**1808-1816.

379. Abramoff MD, Magelhaes PJ, Ram SJ: **Image Processing with ImageJ.** *Biophotonics International* 2004, **11:**36-42.

380. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-242.

381. Momen-Roknabadi A, Sadeghi M, Pezeshk H, Marashi SA: **Impact of residue accessible surface area on the prediction of protein secondary structures.** *BMC Bioinformatics* 2008, **9:**357.

382. Rost B, Sander C: **Conservation and prediction of solvent accessibility in protein families.** *Proteins* 1994, **20:**216-226.

383. Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers* 1983, **22:**2577-2637.

384. Bloom JD, Drummond DA, Arnold FH, Wilke CO: **Structural determinants of the rate of protein evolution in yeast.** *Mol Biol Evol* 2006, **23:**1751-1761.

385. Yang ZH: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13:**555-556.

386. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS: **Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata.** *Nucleic Acids Res* 2008, **36:**D866-870.

387. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH: **Role of duplicate genes in genetic robustness against null mutations.** *Nature* 2003, **421:**63-66.

388. Stelling J, Sauer U, Szallasi Z, Doyle FJ, 3rd, Doyle J: **Robustness of cellular functions.** *Cell* 2004, **118:**675-685.

389. Wagner A: *Robustness and Evolvability in Living Systems.* Princeton: Princeton Univ. Press; 2005.

390. Wagner A: **Robustness against mutations in genetic networks of yeast.** *Nat Genet* 2000, **24:**355-361.

391. Papp B, Pal C, Hurst LD: **Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast.** *Nature* 2004, **429:**661-664.

392. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290:**1151-1155.

393. Wapinski I, Pfeffer A, Friedman N, Regev A: **Natural history and evolutionary principles of gene duplication in fungi.** *Nature* 2007, **449:**54-61.

394. Kellis M, Birren BW, Lander ES: **Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae.** *Nature* 2004, **428:**617-624.

395. Hsiao TL, Vitkup D: **Role of duplicate genes in robustness against deleterious human mutations.** *PLoS Genet* 2008, **4:**e1000014.

396. Ohno S: *Evolution by gene duplication.* Berlin, New York: Springer-Verlag; 1970.

397. Conant GC, Wolfe KH: **Turning a hobby into a job: how duplicated genes find new functions.** *Nat Rev Genet* 2008, **9:**938-950.

398. Innan H, Kondrashov F: **The evolution of gene duplications: classifying and distinguishing between models.** *Nat Rev Genet* 2010, **11:**97-108.

399. Nadeau JH, Sankoff D: **Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution.** *Genetics* 1997, **147:**1259-1266.

400. Sidow A: **Genome duplications in the evolution of early vertebrates.** *Curr Opin Genet Dev* 1996, **6:**715-722.

401. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations.** *Genetics* 1999, **151:**1531-1545.

402. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization.** *Genetics* 2000, **154:**459-473.

403. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV: **Selection in the evolution of gene duplications.** *Genome Biol* 2002, **3:**RESEARCH0008.

404. Bergthorsson U, Andersson DI, Roth JR: **Ohno's dilemma: evolution of new genes under continuous selection.** *Proc Natl Acad Sci U S A* 2007, **104:**17004-17009.

405. VanderSluis B, Bellay J, Musso G, Costanzo M, Papp B, Vizeacoumar FJ, Baryshnikova A, Andrews B, Boone C, Myers CL: **Genetic interactions reveal the evolutionary trajectories of duplicate genes.** *Mol Syst Biol* 2010, **6:**429.

406. Ihmels J, Collins SR, Schuldiner M, Krogan NJ, Weissman JS: **Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss.** *Mol Syst Biol* 2007, **3:**86.

407. Li J, Yuan Z, Zhang Z: **The cellular robustness by genetic redundancy in budding yeast.** *PLoS Genet* 2010, **6:**e1001187.

408. Guan Y, Dunham MJ, Troyanskaya OG: **Functional analysis of gene duplications in Saccharomyces cerevisiae.** *Genetics* 2007, **175:**933-943.

409. Conant GC, Wagner A: **Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans.** *Proc Biol Sci* 2004, **271:**89-96.

410. Gillespie JH: *Population Genetics, A Concise Guide.* Baltimore: Johns Hopkins Univ. Press; 1998.

411. Hartl D, Clark A: *Principles of Population Genetics.* 3 edn. Sunderland: Sinauer Associates; 1997.

412. Lynch M: **Streamlining and simplification of microbial genome architecture.** *Annu Rev Microbiol* 2006, **60:**327-349.

413. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423:**241-254.

414. Gu Z, Cavalcanti A, Chen FC, Bouman P, Li WH: **Extent of gene duplication in the genomes of Drosophila, nematode, and yeast.** *Mol Biol Evol* 2002, **19:**256-262.

415. Yang Z, Nielsen R: **Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models.** *Mol Biol Evol* 2000, **17:**32-43.

416. Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St Onge RP, Tyers M, Koller D, et al: **The chemical genomic portrait of yeast: uncovering a phenotype for all genes.** *Science* 2008, **320:**362-365.

417. DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colon-Gonzalez M, Chao S, Kishony R: **Exposing the fitness contribution of duplicated genes.** *Nat Genet* 2008, **40:**676-681.

418. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, et al: **The genetic landscape of a cell.** *Science* 2010, **327:**425-431.

419. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.

420. Guldener U, Munsterkotter M, Kastenmuller G, Strack N, van Helden J, Lemer C, Richelles J, Wodak SJ, Garcia-Martinez J, Perez-Ortin JE, et al: **CYGD: the Comprehensive Yeast Genome Database.** *Nucleic Acids Res* 2005, **33:**D364-368.

421. Kafri R, Bar-Even A, Pilpel Y: **Transcription control reprogramming in genetic backup circuits.** *Nat Genet* 2005, **37:**295-299.

422. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425:**686-691.

423. Gimeno CJ, Fink GR: **The logic of cell division in the life cycle of yeast.** *Science* 1992, **257:**626.

424. Conant GC, Wolfe KH: **Functional partitioning of yeast co-expression networks after genome duplication.** *PLoS Biol* 2006, **4:**e109.

425. Wagner A: **Asymmetric functional divergence of duplicate genes in yeast.** *Mol Biol Evol* 2002, **19:**1760-1768.

426. He X, Zhang J: **Gene complexity and gene duplicability.** *Curr Biol* 2005, **15:**1016-1021.

427. Hu Z, Killion PJ, Iyer VR: **Genetic reconstruction of a functional transcriptional regulatory network.** *Nat Genet* 2007, **39:**683-687.

428. Papp B, Pal C, Hurst LD: **Evolution of cis-regulatory elements in duplicated genes of yeast.** *Trends Genet* 2003, **19:**417-422.

429. Kafri R, Levy M, Pilpel Y: **The regulatory utilization of genetic redundancy through responsive backup circuits.** *Proc Natl Acad Sci USA* 2006, **103:**11653-11658.

430. DeLuna A, Springer M, Kirschner MW, Kishony R: **Need-based up-regulation of protein levels in response to deletion of their duplicate genes.** *PLoS Biol* 2010, **8:**e1000347.

431. Springer M, Weissman JS, Kirschner MW: **A general lack of compensation for gene dosage in yeast.** *Mol Syst Biol* 2010, **6:**368.

432. Kafri R, Dahan O, Levy J, Pilpel Y: **Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy.** *Proc Natl Acad Sci USA* 2008, **105:**1243-1248.

433. Boucherie H, Bataille N, Fitch IT, Perrot M, Tuite MF: **Differential synthesis of glyceraldehyde-3-phosphate dehydrogenase polypeptides in stressed yeast cells.** *FEMS Microbiol Lett* 1995, **125:**127-133.

434. Seufert W, Jentsch S: **Ubiquitin-conjugating enzymes UBC4 and UBC5 mediate selective degradation of short-lived and abnormal proteins.** *EMBO J* 1990, **9:**543-550.

435. Zou C, Lehti-Shiu MD, Thomashow M, Shiu SH: **Evolution of stress-regulated gene expression in duplicate genes of Arabidopsis thaliana.** *PLoS Genet* 2009, **5:**e1000581.

436. Chung WY, Albert R, Albert I, Nekrutenko A, Makova KD: **Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network.** *BMC Bioinformatics* 2006, **7:**46.