Evolutionary systems biology of bacterial metabolic adaptation

Dissertation

To fulfil the requirements for the degree of
“doctor rerum naturalium” (Dr. rer. nat.)

Submitted to the Council of the Faculty
of Biology and Pharmacy
of the Friedrich Schiller University Jena.

By Dipl.-Bioinf. Silvio Waschina
born on 9th December 1986 in Jena, Germany.

Abstract

The metabolic networks of bacteria have been evolving for billions of years. The results of this unceasing evolution are networks, which are incredibly complex, tightly-regulated, strikingly niche-specific, and extremely diverse across species. Albeit it is beyond dispute that the structure and function of metabolic networks have evolved to be well-adapted to the conditions of an organism's lifestyle and natural habitat, surprisingly little is known about the adaptive origin of metabolic traits.

In this thesis, I will address one specific aspect of bacterial metabolic network evolution: the adaptive loss of metabolic capabilities. In this context, a comparative study of the biochemical networks of more than 900 bacterial species revealed that the loss of biosynthetic genes is a much more common trend than previously thought and is not only limited to bacteria living in nutrient-rich and constant environments, such as intracellular symbionts. In fact, more than 75% of analysed bacteria including free-living organisms are lacking the ability to produce one or more metabolites, which are required for cell growth, thus, rendering these organisms auxotrophic.

However, do bacteria lose metabolic reactions due to neutral genome erosion (i.e. drift) or are auxotrophies adaptive and their evolution therefore governed by natural selection? This question was addressed using a synthetic approach, in which biosynthetic genes for amino acid, nucleotide, and vitamin production were deleted from the prototrophic 'wild type' *Escherichia coli* and *Acinetobacter baylyi* strains, thus rendering the engineered mutants auxotrophic. In virtually all cases, the auxotrophs had an increased fitness compared to their respective prototrophic ancestor when the focal metabolite was sufficiently present in the growth medium, even in direct competition.

Analysing the fitness consequences of multiple biosynthetic gene deletions in different biosynthetic pathways per genotype revealed further that both, positive and negative epistatic interactions between loss-of-biosynthetic-function mutations affected

the degree of the fitness advantage of auxotrophs. In line with these results is the observation, that most pairs of auxotrophies in the bacterial metabolic networks analysed occurred more often than expected by chance. These results suggests, that adaptive benefits can explain the loss of metabolic capabilities from bacterial genomes.

Furthermore, measuring the selective benefits of auxotrophies in two different carbon environments showed that the fitness of auxotrophic mutants is strongly environment-dependent. This implies that reductive evolution by the successive loss of biosynthetic genes depends – beyond the presence of the focal metabolites – also on the nature of the resources available in the habitat.

Following the observed strong carbon source-dependent growth- and fitness consequences of metabolic gene loss, the question emerged: What metabolic causes can explain the unexpectedly substantial fitness advantages gained upon biosynthetic gene loss if the focal metabolite is present in the environment? To answer this question, a metabolic network model of *E.coli* was used in combination with f*lux balance analysis* to computationally estimate how much of a provided carbon source is allocated to the production of a given metabolite. The results of this analysis showed indeed that the biosynthetic costs strongly depended on the chemical nature of the carbon source, which was provided for growth. Thus, the architecture of the metabolic network, with which different carbon sources are transformed into metabolites, could explain the observed costs. Using the amino acid metabolism in *E. coli* as a test case, it was further possible to experimentally verify the *in silico* estimations.

In conclusion, resource-efficiency is a major criteria for the metabolic adaptation of bacteria to improve their ability to thrive in their natural environment. The applied systems biological approach using an economic concept of metabolite production costs facilitated new insights into the evolution of bacterial metabolic network structures and metabolic flux coordination. Finally, I argue in this thesis that the principle of metabolic cost minimisation does not only govern the evolution of metabolism, but likely also affects the way bacteria compete and cooperate with other species to optimally utilize limiting resources. The results have further direct medical and biotechnological implications, as they can for example guide the development of new strategies to control the growth of microorganisms.

Zusammenfassung

Bakterien haben sich im Laufe der Evolution an ihre jeweiligen Umweltbedingen angepasst. Viele dieser Anpassungen finden auf der Ebene metabolischer Netzwerke statt. Die metabolischen Netzwerke von Bakterien bestehen aus hunderten bis tausenden biochemischen Reaktionen und sind charakterisiert durch ihre komplexe Struktur sowie durch eine starke Regulierung der Reaktionsflüsse. Verschiedene Arten von Bakterien unterscheiden sich mitunter hochgradig in der Struktur ihrer metabolischen Netzwerke und somit auch in ihren biochemischen Fähigkeiten. Auch wenn oft angenommen wird, dass diese Unterschiede eine Anpassung an die jeweiligen ökologischen Nischen darstellen, ist bisher wenig über die adaptive Evolution von Netzwerkeigenschaften bekannt.

Diese Arbeit beschäftigt sich speziell mit dem adaptiven Verlust von metabolischen Fähigkeiten. Der Vergleich von über 900 metabolischen Netzwerken verschiedener Bakterienarten ergab, dass mehr Arten als ursprünglich angenommen Gene verloren haben, die sonst an der Biosynthese von wachstumsrelevanten Metaboliten beteilig sind. Etwa 75% aller untersuchten Bakterien haben die Fähigkeit verloren, einen oder mehrere chemische Stoffe zu synthetisieren, die für das Wachstum notwendig sind. Dementsprechend sind diese *auxotrophen* Bakterien darauf angewiesen, die Metabolite aus ihrer natürlichen Umgebung aufzunehmen. Dieses Ergebnis trifft nicht nur auf Bakterien zu, die in einer nährstoffreichen Umgebung vorkommen (z. B. intrazelluläre Symbionten) sondern auch auf freilebende Bakterien, die oft in nährstoffärmeren Umgebungen, z. B. aquatischen Lebensräumen, vorkommen.

Bisher ist unklar, ob der häufige Verlust von biosynthetischen Fähigkeiten durch natürliche Selektion oder durch genetischen Drift erklärt werden kann. Diese Frage wurde mit einem synthetisch-biologischen Ansatz untersucht, indem einzelne biosynthetische Gene für die Produktion von Aminosäuren, Nukleotiden oder Vitaminen aus den Genomen von den Bakterien *Escherichia coli* und *Acinetobacter baylyi* entfernt und somit auxotrophe Mutanten erzeugt wurden. In einer Umgebung, in der der entsprechende Metabolit zum Wachstumsmedium hinzugefügt wurde,

hatten die auxotrophen Mutanten in nahezu allen Fällen eine höhere Fitness als die entsprechenden Wildtypstämme, die den Metaboliten weiterhin eigenständig herstellen konnten.

Außerdem wurde die Auswirkung von multiplen Auxotrophien (der Verlust von mehreren biosynthetischen Genen) pro mutierten Genotyp auf die Fitness in gleicher Weise untersucht. Dieses Experiment zeigte, dass die Auswirkung des Verlustes eines biosynthetischen Genes auf die Fitness in über 50% der möglichen Kombinationen davon abhängt, ob und welche weiteren Gene anderer Biosynthesewege der Stamm bereits verloren hat – ein Effekt, der Epistasis genannt wird. Diese Beobachtung stimmt überein mit dem Vergleich von bakteriellen metabolischen Netzwerken, bei denen bestimmte Kombinationen von Auxotrophien häufiger und andere Kombinationen seltener als erwartet vorkommen. Zusammenfassend lassen die Ergebnisse darauf schließen, dass der Verlust von biochemischen Fähigkeiten von Bakterien durch selektive Vorteile von auxotrophen Genotypen erklärt werden kann.

Weiterhin konnte gezeigt werden, dass der Fitnessvorteil von auxotrophen Genotypen stark davon abhängt, welche Kohlenstoffquelle zur Verfügung steht. Ausgehend von dieser Beobachtung und dem bisherigen Wissen, dass verschiedene Kohlenstoffquellen zu verschiedenen Reaktionsflussverteilungen durch das metabolische Netzwerk führen, ergab sich die Frage: Wie können die selektiven Vorteile von auxotrophen Genotypen in Umgebungen, in denen die entsprechenden Metabolite vorkommen, biochemisch erklärt werden? Um diese Frage zu beantworten wurde ein metabolisches Modell von *E. coli* in Kombination mit *Flussbilanzanalyse* (FBA) verwendet, um die biosynthetischen Kosten von Metaboliten theoretisch vorherzusagen. Die biosynthetischen Kosten wurden dabei als Anteil einer gegebenen limitierten Kohlenstoffquelle berechnet, der während des Zellwachstums für die Produktion eines bestimmten Metaboliten benötigt wird. Diese Analyse deutete darauf hin, dass die biosynthetischen Kosten tatsächlich stark von der Art der Kohlenstoffquelle abhängig sind, welche für das Wachstum zur Verfügung steht. Die Unterschiede zwischen Kohlenstoffquellen bezogen auf die Biosynthesekosten konnten dabei durch die Architektur des metabolischen Netzwerks erklärt werden. Die Modellvorhersagen zu den Kostenunterschieden von Metaboliten, abhängig von der Kohlenstoffquelle, konnten weiterhin auch experimentell in *E. coli* für Aminosäuren nachgewiesen werden.

Die Fähigkeit, limitierte Ressourcen möglichst optimal zu nutzen, ist entscheidend für das Wachstum und die Fitness von Bakterien in ihrer natürlichen Umgebung. Durch einen systembiologischen Ansatz konnte gezeigt werden, dass die metabolischen

Kosten, die mit der Produktion von wachstumsrelevanten Metaboliten in Verbindung stehen, die Evolution der metabolischen Netzwerke von Bakterien beeinflussen können. Speziell der adaptive Verlust von metabolischen Fähigkeiten kann vermutlich auch erklären, warum die Mehrheit von bekannten Bakterienarten nicht mit üblichen Techniken im Labor kultivierbar ist und warum viele Bakterien untereinander metabolische Produkte austauschen.

Contents

Abstract **1**

Zusammenfassung **3**

Contents **7**

Glossary of terms **11**

Chapter I – Introduction **15**

1. Evolution of bacterial metabolic networks 17
2. Approaches to reveal and predict bacterial metabolic adaptations 30
3. Mathematical formalisation and analysis of genome-scale metabolic networks 37
4. Objectives of this study 42

Chapter II – Computing autocatalytic sets to unravel inconsistencies in metabolic network reconstructions **45**

1. Abstract 47
2. Introduction 47
3. Material and Methods 50
4. Results and discussion 55
5. Conclusion 64

Chapter III – Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria **67**

1. Abstract 69
2. Introduction 69
3. Material and Methods 71

4. Results 75

5. Discussion 83

6. Conclusions and outlook 87

Chapter IV – Plasticity and epistasis strongly affect bacterial fitness after losing multiple metabolic genes 89

1. Abstract 91

2. Introduction 91

3. Materials and Methods 93

4. Results 98

5. Discussion 105

6. Conclusions and Outlook 108

Chapter V – Metabolic network architecture and carbon source determine metabolite production costs 109

1. Abstract 111

2. Introduction 111

3. Results 113

4. Discussion 118

5. Methods 121

Chapter VI – General discussion 127

1. Reductive evolution of metabolic networks 129

2. The relationship between microbial growth and metabolism 137

3. The evolution of autocatalytic reaction cycles in metabolic networks 140

4. Concluding remarks 141

Acknowledgements 144

References 146

Author contributions to manuscripts 164

Supporting information for chapter II 167

Supporting information for chapter III 173

Supporting information for chapter IV 195

Supporting information for chapter V 211

Curriculum vitae 219

Eigenständigkeitserklärung 223

Glossary of terms

The terms listed here are frequently used in scientific literature, but, depending on the context, sometimes with different definitions. To prevent confusion, this glossary states how the terms are used throughout this thesis.

Biosynthetic cost. The amount of an exogenous resource (e.g. carbon source), which is required to produce a certain metabolite.

Core-metabolic network. The intersection of all reactions of the metabolic networks of a given group of strains (usually from the same species). In other words: the set of reactions, which are part of all metabolic networks of the analysed strains.

Evolvability. The ability of an organism to evolve heritable phenotypic variation [1].

Fitness cost (of a trait). The reduced fitness of a genotype carrying the trait (e.g. antibiotic resistance) compared to genotypes without the trait in environments where the trait is not needed for growth and proliferation (e.g. in the absence of the antibiotic).

Fitness landscape. A map of individual phenotypes (or sometimes genotypes), which are represented as combinations of multiple traits (axes), to the corresponding fitness values (surface) [2].

Growth efficiency. Amount of biomass produced per unit of assimilated carbon source[3].

Growth rate (microbial). Number of progeny cells produced per unit of time.

Growth yield. Same as *growth efficiency.*

Life history trait. A trait which affects the reproduction and/or survival of an organisms and thereby contributes to the organism’s fitness [4].

Metabolic adaptation. A dynamic evolutionary process that increases the fitness of a species in its environment and that relies on beneficial mutations, which affect the performance and regulation of enzymes and transporters.

Metabolic cost. Same as *Biosynthetic cost*.

Metabolic flux. The rate at which an enzyme converts the substrate(s) into the product(s).

Metabolic function. In analogy to a mathematical function, a metabolic function is the transformation of (a) precursor metabolite(s) into (a) chemically different metabolite(s). A metabolic function can be a single reaction or a complete pathway.

Metabolic genotype. The genetic information of an organisms that affects its metabolic capabilities. The metabolic genotype included the structure of the metabolic network (determined by enzyme- and transporter-encoding genes), the regulatory circuits that control *metabolic fluxes* (e.g. transcription factors), and the kinetic properties of enzymes.

Metabolic innovation. A newly acquired metabolic phenotype that provides the corresponding genotype with a qualitative fitness advantage.

Metabolic niche. The combination of an organism's environmental conditions and the organism's lifestyle.

Metabolic phenotype. The distribution of metabolic fluxes throughout the metabolic network and the individual metabolite and enzyme concentrations at a given time point.

Pan-metabolic network. The union of all reactions of the metabolic networks of a given group of strains (usually from the same species).

Plasticity (of metabolic fluxes). The ability of an organism to respond to changing environmental conditions by adjusting metabolic fluxes through the metabolic network.

Synthetic biology. The rational construction of a biological system for a specific purpose; e.g. the commercial production of value-added compounds or to investigate the physiology and behaviour of organisms under defined conditions [5]. The design of the system may include the environmental conditions (e.g. nutrient availability, pH, or temperature), the consortia of involved organisms, and their genotypes.

Chapter I

Introduction

1. Evolution of bacterial metabolic networks

Metabolism is utterly central for every organismal life. It provides building block metabolites and energy for all cellular processes including macromolecule polymerisation, biosynthesis of monomers, transport reactions, cell maintenance, and movement. In order to synthesize all required constituents of a bacterial cell, between 250 and more than 2,000 reactions are required depending on the complexity and variability of the organisms' natural habitats [6,7]. A majority of the reactions are catalysed by enzymes, which are encoded by genes [7]. Hence, the genome contains the information of a complex network consisting of numerous biochemical transformations, which the organism is able to perform. In other words, the metabolic network represents the anabolic (i.e. which metabolites can be produced) and catabolic (which substrates can be degraded and thereby serve as energy source) capabilities of an organism and, at the same time, also the nutritional requirements an organism needs to meet in the environment in order to thrive and proliferate.

All microorganisms allocate resources to various metabolic pathways in response to the nutritional condition of the environment [8]. Furthermore, metabolic processes are also tightly involved in microbial community activities, for example during the assembly of microbial communities [9], the colonisation of new environments [10], and communication between cells of the same- or different species including multicellular host organisms [11].

Due to this intrinsic role of metabolism in determining the evolutionary fate of a species, natural selection should act decisively on the structure and regulation of metabolic networks. Thus, the architecture of microbial metabolic networks and the coordination of its reactions reflect the biotic and abiotic composition of prior selection environments in which an organism evolved over generations [12]. Fundamental questions in the evolution of biochemical networks are: (i) What biophysical and ecological factors govern the evolution of metabolic networks? And, (ii) what evolutionary processes are involved? A better understanding of the metabolic innovations, which led to the adaptation of bacteria to their habitats will facilitate to disclose the factors, which determine how bacteria exploit resources, interact with other organisms, and assemble in microbial communities. Furthermore, it will enable predictions of bacterial evolution in laboratory settings as well as in natural environments. Unravelling the factors that govern the ecology and evolution of bacteria on a metabolic network level can be highly relevant for medical, biotechnological,

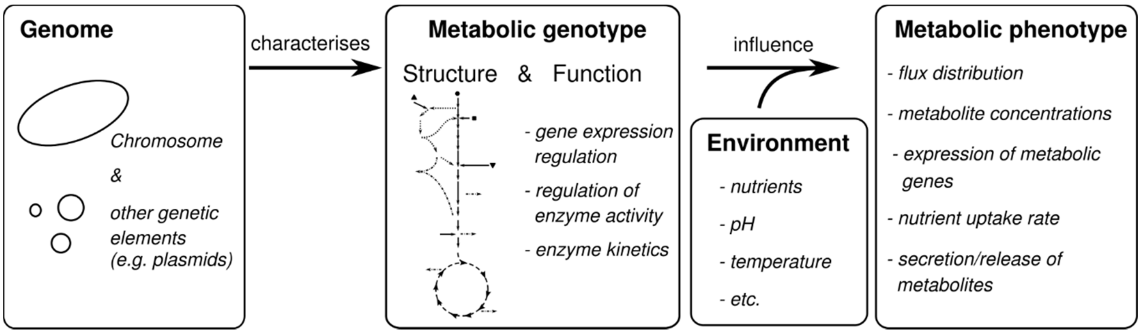


Figure 1. Scheme of the elements of the *metabolic genotype* and *metabolic phenotype*. The genome of an organism determines the *metabolic genotype*, which describes the structure and function of the metabolic network. The function of the metabolic network comprises the regulatory circuits, which control metabolic gene expression and enzyme activities. Moreover, the function includes also the kinetic parameters of enzymes (e.g. reversibility, maximum reaction rates $\boldsymbol{v_{max}}$, and Michaelis-Menten constants $\boldsymbol{K_m}$*). Other factors – besides the environment and the metabolic genotype – that may affect the metabolic phenotypes are biophysical constraints and stochasticity.

bioremediation, and environmental conservation applications involving microorganisms.

1.1. How can metabolic networks change in the course of evolution?

The term *metabolic genotype* refers to the make-up of all genetic traits, which have an impact on the organism's metabolic functions and thereby describes the organism's biochemical capabilities [7]. Mutations, which affect the metabolism of an organism can occur at two different levels of the metabolic genotype: First, a metabolic network can change on a *structural* level by adding or deleting certain reactions. Second, a metabolic network can change on its *functional* level, which includes how individual reactions, transporters, and whole pathways are coordinated in response to environmental conditions and/or the cellular state [8]. The function of a metabolic network further includes the kinetic parameters of enzymes, which determine the dynamics and limits of fluxes through the respective catalysed reaction. In combination with the environmental conditions, the metabolic genotype determines the expressed *metabolic phenotype*, which is characterised for example by the distribution of fluxes, the concentration of metabolites and enzymes, the expression rates of metabolic genes, nutrient uptake- and metabolite secretion/release rates.

---

[*] For Michaelis-Menten enzyme kinetics, K_m is the substrate concentration, at which the reaction rate equals half of the maximal reaction rate v_{max}.

Evolution of the metabolic network structure

In the past decades, new developments in isolation-, screening-, and sequencing methods revealed a tremendous diversity within the prokaryotic kingdom with respect to the species lifestyles, habitats, metabolic phenotypes, or genome sizes [13]. In spite of this rich diversity, the biochemistry for the formation of the cell's constituents is rather conserved across bacteria [14]. The biosynthetic pathways of amino acids, nucleotides, lipids, carbohydrates, and vitamins are frequently recurring across known bacterial species and only very few alternative pathways are known for a given metabolite (see supporting information for chapter III for merged biosynthetic pathways known in bacteria and reference [14]). For instance, there is only a single native pathway found so far for the biosynthesis of tryptophan, which involves five enzymes converting the same substrates into the identical intermediates across all domains of life [15,16]. Despite the limited number of alternative pathways for the production of a given metabolite, combinations of all alternatives for all metabolites required for cell growth create a multi-dimensional space of metabolic reaction sets and thereby a large number of theoretically possible *metabolic genotypes* (Fig. 1.1) [17,18].

Several molecular mechanisms are described, which can result in the acquisition of new reactions or pathways. For example, after the duplication event of an enzyme-encoding gene, one copy can potentially evolve a new catalytic function, which may has been only a weak side activity of the original enzyme, while the other copy retains its original function [19,20]. New reactions can also be acquired by horizontal gene transfer, which is especially prevalent in prokaryotes [21]. In contrast to the evolution of a new enzymatic activity evolving upon gene duplication, a new catabolic activity does not need to evolve *de novo* when an enzymatic reaction is added to a metabolic network via horizontal gene transfer. In the case of *E. coli*, it has been estimated that almost all enzymatic reactions, which were acquired in the past 100 million years were obtained via horizontal gene transfer [22]. In contrast, only a single gene pair (i.e. ornithine carbamoyltransferase 1 and 2) originated due to a duplication event within this time period [22].

Loss of metabolic reactions is the result of loss-of-function mutations that can have diverse molecular causes, e.g. complete or partial gene deletions [7] and pseudogenisation (for example through frame-shift mutations) [23]. The loss of reactions from the metabolic network is especially common in bacteria with symbiotic lifestyles. For example, the networks of the endosymbiotic bacteria *Buchnera aphidicola* and *Blattbacterium cuenoti* comprise each less than 300 reactions [24,25], whereas *E. coli* or *B. subtilis*, which can be found in various and nutritional more unstable environments, are able to catalyse more than 1,400 reactions [26,27]. However, a comprehensive survey of extent of metabolic capability loss in natural bacterial populations is still missing. Yet, the evolutionary reduction of metabolic complexity through the loss of biochemical abilities seems to be a frequent trend for many life forms. For example, even many vertebrate species including anthropoid primates have repeatedly lost the ability to synthesize vitamin C [28].

Although, the sizes of metabolic networks and their biosynthetic capabilities seems to be strongly dependent on the environmental conditions, there are also properties of the global network organisation, which are conserved across all domains of life [29]. One of such structural features are *autocatalytic cycles*. A metabolite is considered *autocatalytic* if it is required for its own biosynthesis [30]. Hence, an initial amount of the autocatalytic metabolite is necessary to enable its production and thereby creates an autocatalytic cycle. Interestingly, the compounds within metabolic networks, which have this self-replicating property, are well-conserved across all domains of life and include typically the main energy currency ATP and reaction cofactors such as NAD^+, coenzyme A, tetrahydrofolate, and quinones [30]. Furthermore, the number of autocatalytic compounds in metabolic networks is smaller than expected in random biochemical reaction networks [31]. It has been proposed, that autocatalytic cycles play a major role in the origin of life [32,33], not only on the level of metabolic networks; e.g. a DNA molecule is required for its own synthesis/replication. Hence, the conservation of autocatalytic cycles embedded in metabolic networks indicate that not only environmental conditions, but also biochemical and biophysical constraints can govern the evolution of network structures.

In summary, the architectures of bacterial metabolic networks are highly diverse with respect to the network size, the ability to utilize different substrates, and biosynthetic capabilities. Furthermore, the network structure is highly dynamic on evolutionary time scales due to various molecular processes.

Evolution of metabolic network function

The function of a metabolic network denotes how the network of biochemical reactions operates in response to environmental and endogenous conditions by adjusting and distributing metabolic fluxes throughout the whole network. Hence, the network function signifies, which metabolic phenotypes an organism can possibly express. In general, cells face the task to simultaneously coordinate the activity of hundreds of different enzymes to provide energy, building block metabolites, and growth factors required for cell maintenance and growth [8]. The regulatory circuits that allow coordinated flux distributions are complex and involve various molecular mechanisms: The flux through an enzymatic reaction can be regulated by adjusting the enzyme's concentration level via transcriptional regulation, e.g. by governing promotor activity; or translational regulation, e.g. by controlling the recruitment of ribosomes to mRNA. Furthermore, the activity of the enzyme itself can be altered by post-translational regulation, namely covalent protein modifications and allosteric regulation by metabolite binding.

The regulatory circuits for the coordination of bacterial metabolism are much more complex than the structure of the metabolic network itself. For instance, the abovementioned perfectly conserved pathway for tryptophan biosynthesis is regulated by very diverse mechanisms across microorganisms [16]. Hence, due to the various levels of metabolic flux coordination there are more targets for natural selection to change functional properties of metabolic networks than structural traits. Moreover, in contrast to the metabolic network structure, which is a qualitative trait (reactions are either present or absent from a given metabolic genotype), the metabolic fluxes that are carried by enzymatic reactions under certain environmental and cellular conditions are quantitative traits and, thus, allow more fine-tuned evolutionary adjustments. This complexity of metabolic regulation provides bacteria with extensive *metabolic plasticity* to respond to different environmental conditions and also with high *evolvability* to metabolically adapt to novel environments [34–36]. Indeed, evolution experiments using bacteria have provided evidence that metabolic regulation is highly flexible [37,38] and new regulatory programs, which steer the activities of a larger number of central metabolic enzymes can evolve within surprisingly few generations [39,40]. For instance, Charusanti *et al.* (2010) deleted the major metabolic gene *pgi,* which encodes a phosphoglucose isomerase, a central enzyme from glycolysis [39]. The gene deletion caused a nearly 5-fold decrease in the growth rate of the mutant compared to the *wild type*. In an evolution experiment over 50 days the mutant regained a 3.6-fold increased growth rate compared to the un-evolved mutant. The authors were able to show that the growth recovery was facilitated by mutations in genes encoding other central enzymes and global metabolic regulators, which re-distributed the metabolic fluxes through the network and thereby allowed a faster conversion of the carbon source glucose into biomass [39].

Several evolution experiments have shown, that adaptation of bacteria is often driven by early mutations in global metabolic regulators [41–44]. Yet, the regulatory consequences of these mutations for the expression of other genes and enzyme activities is challenging to understand and often remain unresolved [45]. A possible way to fill this gap of mechanistic understanding is to ask how metabolic fluxes *should* be distributed according to prior important evolutionary selection criteria [40], while disregarding the molecular mechanisms, which ensure flux coordination. This approach has been successfully applied to predict metabolic phenotypes* (flux distribution) of cells without the need for comprehensive knowledge of the regulatory circuits that underlie metabolic decision making [46]. For example, experimentally evolving *E. coli* for 900 generations in minimal medium with lactate as sole carbon source, which *E. coli* initially utilized with sub-optimal yield (number of cells per unit of utilized lactate), the bacterium increased substantially in growth rate and yield [47]. This derived phenotype was achieved by rewiring the regulatory network in order to redistribute metabolic fluxes towards the predicted optimum [40].

In summary, the regulation of cellular biochemical networks involves multifaceted molecular mechanisms. The regulatory network, which coordinates reaction fluxes adds another dimension to the complexity of metabolic networks. In contrast to the structure of the metabolic network, insufficient information of regulatory circuits exists and thereby hamper a mechanistic understanding of how the whole-cell metabolism is coordinated. Yet, mathematical modelling of how metabolic fluxes through a metabolic network *should be* distributed can provide valuable insights into the regulatory adaptation to different environmental conditions and different selection pressures.

1.2. Metabolic reaction loss through genetic drift

As shown in the previous section, several molecular processes are known, which can genetically cause alterations of the metabolic network architecture and its functions. However, a fundamental question remains: Which evolutionary forces govern these changes and, thus, the evolution of metabolic networks [22]?

In this thesis, the main focus is on the loss of metabolic capabilities through the loss of biosynthetic genes. A comparison of the gene content of 35 groups of closely related bacteria has suggested that the evolutionary rate of gene loss is three times higher than the rate of gene gain [48]. Furthermore, the genome reduction is typically also associated with metabolic network diminution through loss of enzymatic reactions [49,50]. While the selective advantages of reaction additions to the metabolic network can be typically explained by an improved network functionality [22,51,52], the potential benefits of reaction loss are more difficult to understand. So far it remains

---

* The method for the prediction of metabolic flux distribution is explained in section I.3.2.

highly discussed and obscure to what extent the reductive evolution of genomes (e.g. as in the case of endosymbiotic bacteria) and metabolic networks is driven by adaptive or neutral processes.

Metabolic genes loss, and thus network size reduction, from a population can be caused by *genetic drift*. This is because mutations in metabolic genes can accumulate and passed to the next generation when natural selection is not strong enough to remove mutations, i.e. if they do not interfere with the organism's ability to survive and reproduce. The frequencies of such mutations within a population can increase thereby solely due to random sampling of alleles from the population for every new generation [53]. The effect of genetic drift is especially large if the population undergoes repeated bottlenecks of relatively small population sizes. It has been proposed that bacterial endosymbionts of insects experience serious bottlenecks when they are transmitted from one host individual to another and that this may explain the strongly reduced genome- and metabolic network sizes of these bacteria [54–56]. Moreover, experimentally evolving populations of *Salmonella enterica*, which were regularly exposed to severe one-cell bottlenecks also showed extensive genome size reductions within relative short evolutionary time [57].

Taken together, genetic drift can have a strong impact on the evolution of metabolic networks. However, the contribution of genetic drift on the loss of metabolic capabilities in natural populations remains difficult to estimate, because important knowledge on the evolutionary history of the species including potential population size bottlenecks and previous selection pressures is usually lacking.

1.3. Metabolic reaction loss through natural selection

Another evolutionary force that could explain metabolic gene, and thus enzymatic reaction loss is *natural selection*. It has been proposed that natural selection could favour metabolic gene loss by a process termed *genome streamlining* [58]. This hypothesis poses that the loss reduces the metabolic burden of a cell and thereby results in selective advantages of biosynthetic-deficient genotypes over other genotypes, which still have to bear the cost of the focal biosynthetic function [58]. Yet, the contributions of different cell-physiological factors on the metabolic burden of a given biosynthetic function remain difficult to estimate. These factors include (i) the metabolic resource costs that are caused by the flux through the corresponding reaction or pathway to produce one unit of the biosynthetic product, (ii) the costs to produce and maintain the respective enzyme levels, (iii) the costs to maintain the genes within the genome, and

(iv) the resources required for the regulatory circuits that control the expression of the pathway and the activity of its enzymes. In this thesis, I will mainly focus on the first factor: the *metabolite production costs*.

Metabolite production costs

Like all other organisms, bacteria require several nutrients from their environment for cell growth. The major bioelements constituting the organic material of the cell are carbon, hydrogen, oxygen and nitrogen. Another requirement is an energy source, which is utilized to build ATP, the primary energy carrier within cells [14]. Phototrophic bacteria obtain their energy from light to create the high-energy phosphate bonds in adenosine triphosphate (ATP). Chemotrophs enzymatically break down chemical compounds by exergonic reactions, which are coupled to the formation of ATP from ADP and inorganic phosphate P_i. Most of the known bacteria are chemotrophs and a majority of them in turn are chemoorganotrophs, which use organic compounds as energy source [59]. Hence, metabolism fulfils, in most bacteria, the dual function of energy supply (catabolism) and synthesis of monomers (anabolism), which are required to build all cell constituents [3].

Any metabolic function that consumes resources induces an intrinsic burden, a metabolic cost, to the cell because, the used resources are not available anymore for other cellular functions. In general, microbial cells face the problem to allocate resources to several different cellular processes [60]. One particular resource allocation problem is the distribution of fluxes (i.e. the rate at which an enzyme converts a substrate into the product) through the metabolic network to optimally provide building block metabolites (i.e. amino acids, nucleotides, lipids) and growth factors such as vitamins and co-factors for cell growth [46]. The biosynthesis of each metabolite thereby has a metabolic cost depending on the resource requirement of the biosynthetic pathway.

For well-studied microorganisms such as *E. coli*, *B. subtilis*, and *Saccharomyces cerevisiae* most biochemical reactions, which are part of the organism's metabolic network are presumably known [14]. Hence, for those microorganisms it is possible to estimate the costs of certain biosynthetic functions based on the structure of the metabolic network [61–63]. The terms *metabolic costs* and *biosynthetic costs*, are frequently mentioned in the context of the evolution of biochemical networks and denote how much resource an organism needs to invest in the biosynthesis of a specific metabolite. Several different currencies have been used to quantify these costs, which is due to the fact that different types of resources can be considered for cost quantification. For example, the carbon- or nitrogen source can be used to determine the bioelement costs or ATP for the bioenergetic costs. ATP is often considered as the major energy currency of cell. However, quantifying metabolic costs as bioenergetic currency in terms of ATP can be misleading because, (i) the cell itself produces its ATP pool whose production consumes and depends on other resources (e.g. light or organic compounds), and (ii) the transformation of a carbon source via biosynthetic pathways into a given metabolite does not necessarily consume ATP. The biosynthesis of some amino acids for instance even produces ATP from ADP and P_i [63]. Yet, this does not necessarily mean that these biosynthetic functions do not entail bioenergetic costs on the cell, because an organic energy source may also serve as a carbon source, especially since most bacteria derive their energy and carbon from degrading organic compounds. In this case, the chemical transformation of the carbon source into a metabolite diverts also a fraction from the potential energy source and has, thus, also bioenergetic costs. Craig and Weber (1998) combined bioenergetic- and carbon costs by estimating the biosynthetic costs of amino acids as the amount of ATP that is consumed on the biosynthetic pathway plus the amount of ATP that could be produced if the carbon source would be completely oxidised and not transformed into the focal amino acid [62]. An alternative approach is to directly quantify biosynthetic costs as the amount of an organic resource that is needed as 'material'-carbon source to build the focal metabolite plus the resource quantity required as energy source to fuel the reactions of the biosynthetic pathway (chapter V).

Although different approaches exist in the literature to assess biosynthetic costs, three main factors determine the cell's metabolite production expenditures. First, the architecture of the metabolic network designates possible biochemical transformations from an available resource towards the focal metabolite [14]. Second, the structural complexity of the metabolite contributes to the bioelement and energetic requirement for its biosynthesis [64], and third, environmental conditions, especially the nutritional composition of the organism's habitat, affect metabolite production efficiency [65]. In context of the latter factor, chapter V shows how the chemical nature of different carbon sources affect the biosynthetic costs of proteinogenic amino acids.

Fitness consequences of biosynthetic costs

The natural environments of most microbes are limited in bioelement resources [4]. In such habitats, the number of progeny cells that can be produced by a microbial population is constrained by the quantity and quality of available resources [4]. Furthermore, a microorganism is typically also in a constant battle for confined substrates with other species and genotypes [66]. Hence, the growth and survival of an organism is closely connected with the strategic exploitation and allocation of limited resources [67].

Cell growth and reproduction requires the joint operation of various biosynthetic processes. Each process entails metabolic costs (see previous section) depending on the amount of resources needed to produce one unit of the biosynthetic product and the amount of the product required per unit biomass. When resources are limiting, metabolic costs can directly translate into fitness costs, because any mechanism that reduces the resource consumption of a biosynthetic process would increase growth and, hence, also the organism's fitness ([60], chapters III and V). It is therefore not surprising that microorganisms are under strong selective pressure to economize their resource allocation and consumption by metabolic processes [61,68].

One interesting example of how biosynthetic costs can govern genetic changes in the course of evolution of a species is the amino acid composition of highly-expressed proteins [61,62,64]. Usually, only a small fraction of amino acids within a protein has a specific function, which is related to the amino acid's side chain [69]. Hence, non-synonymous mutations in those parts of the protein, which do not necessarily disrupt the function of the protein may only marginally affect fitness [69]. In contrast to this is the observed strongly biased usage of amino acids in highly-expressed proteins in many tested organisms [70]. Richmond (1970) hypothesized, that the amino acid composition of proteins has evolved as a response to natural selection, which was not mainly directed by the protein's function, but by constraints in the synthesis process of the protein itself [68]. Indeed, Akashi and Gojobori (2002) showed that differences in biosynthetic costs between the 20 proteinogenic amino acids constrain their frequencies in the proteome of the bacteria *E. coli* and *B. subtilis* [61]. In other words, less-costly amino acids (e.g. glycine or alanine) are more frequently incorporated into highly-expressed proteins than more costly amino acids (e.g. tryptophan or phenylalanine). Later, similar *cost selection* effects, which direct the amino acid usage in proteomes, were detected in a broader range of bacteria [71] and even in eukaryotes and archaea [65]. These examples illustrate that metabolic costs are endogenous factors contributing to the molecular evolution of genomes through natural selection.

Furthermore, the stringent regulation of biosynthetic pathways implies that the production of metabolites is costly for the organism and affects its fitness. A commonly found regulatory motif is a feedback inhibition loop. Here, the end product metabolite of a pathway controls its own production by inhibiting the flux through the pathway as the concentration of the metabolite increases ([e.g. 16]). Such regulatory circuits ensure that the production rate is tuned to match the rate at which the metabolite is required, thereby avoiding waste of limiting resources [72].

As already mentioned, biosynthetic costs are largely determined by the metabolic network. It is important to note that also the metabolic network itself is shaped by evolutionary processes and so are biosynthetic costs. It is widely acknowledged that biosynthetic cost minimisation is an objective of organisms especially in conditions of nutrient scarcity [73,4]. The cost minimisation for fitness maximisation is even often considered as basic design principle of metabolic networks [12,67]. Such an 'engineering' perspective on the evolution of biochemical networks has been demonstrated to successfully predict evolutionary changes of the network topology [22] and metabolic phenotypes (i.e. metabolic flux distributions; see section 3.2). However, it is essential to recognise the metabolic networks we can find today in nature not as static entities but as the outcome of accumulated metabolic optimisation.

Taken together, biosynthetic costs are a limiting factor of microbial fitness and thus shape the evolution of microorganisms on the level of the metabolite usage, the regulation of biosynthetic pathways, and the structure of the metabolic network.

Metabolic complementarity and the pan-genome

The evolution of the metabolic networks might not only be governed by the abiotic composition of the environment but also by the metabolic capabilities of co-existing organisms [74]. One possible adaptation of a microorganisms to the presence of another microbial species could be the evolution of metabolic strategies to better compete for limiting resources [66,75]. Another possibility is that the metabolic activities of one species benefits another [76], for example the release of metabolic waste products by a community member could serve as a resource for another strain [77,78]. The result of obligatory interactions based on the exchange of metabolites – termed *metabolite cross-feeding* – are interconnected metabolic networks. These networks are characterised by *metabolic complementary*, meaning that some metabolic functions, which are seemingly essential in the environment of the strains are portioned to the different strains residing in the community [76].

In fact, a comparison of 61 *Escherichia coli* and *Shigella* spp. genomes revealed a surprisingly extensive plasticity in gene content: 95% of all predicted gene families were only present in a subset of the 61 genomes (i.e. the *pan-genome;* [79]). Interestingly, the pan-genome contains also metabolic genes, which were deemed essential for the growth of *E. coli* [80–82]. Thus, the strains, which lack these genes must complement the missing metabolic capability either by environmentally available metabolites or by the

Box 1. The Black Queen Hypothesis

The Black Queen Hypothesis (BQH) was formulated by Morris *et al.* (2012) to explain genome reduction and the evolution of dependencies through adaptive gene loss [83]. It uses an analogy to the card game *Hearts*, in which a typical strategy for a player is to avoid to have the queen of spades (black queen) on her/his hands at the end of the game. The BQH posits that, in evolution, certain biological functions are, like the black queen card, costly and therefore provide genotypes with selective advantages, which have lost the genes to perform the function. However, the function also needs to be retained, at least in a subgroup of the community members, because it provides an essential product that is required by all individuals [83]. This requires, that the biological function is also 'leaky', which means that the product of the function is, at least partly, available as public good for neighbouring cells.

provisioning of the focal metabolites by strains, which coexist in the microbial community.

A theory, which could explain the evolution of such metabolic complementary is the *Black Queen Hypothesis* (Box 1; [83]). Briefly, it assumes that metabolic functions are 'leaky', which means that a certain proportion of the product metabolite is released from the producer cell and therefore available as *public good* for other cells, which may lose the genes for the specific metabolic function to reduce their metabolic burden.

Furthermore, the exchange of metabolites upon the loss of metabolic functions may also be the basis for the evolution of cooperative interactions [76,84]. Cooperative metabolic cross-feeding in microbial communities occurs when certain compounds are exchanged between two or more parties in the consortia while each party receives a mutual benefit. Although, it has already been shown that the cooperative exchange of metabolites can be explained by selective advantages of cooperative genotypes over non-cooperators, the metabolic factors that favour the evolution of the division of biosynthetic labour remain so far widely unexplored.

Taken together, microorganisms do not live in pure isolation in nature. In contrast, metabolic interactions between different species, which coexist in the same environment are prevalent in microbial communities [85,86]. Thus, to understand the adaptive evolution of metabolic networks, including the loss of metabolic reactions, it is necessary to consider also the ecological interactions between species.

Epistasis

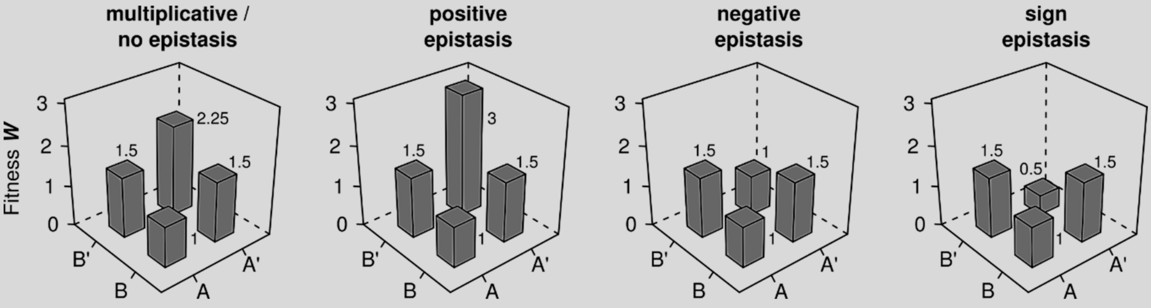
Epistatic interactions between mutations refers to the effect when the phenotypic impact of one mutation depends also on the presence of another mutation [87]. The

Box 2. An epistasis measure for fitness – the multiplicative model

Epistasis describes the effect when the phenotypic outcome of one mutations is altered if combined with another mutation [87]. To quantify epistasis for a quantitative phenotypic trait between two mutations we need to know what phenotypic effect of the combined mutations is expected, if the two mutations are independent of each other. For fitness as phenotypic outcome the most commonly used model to estimate the expected fitness effect of two mutations is the multiplicative model [87,91–93]. This model states that the fitness effect of the combined mutations is given by

$$\boldsymbol{W}_{A'B'} = \boldsymbol{W}_{A'}\boldsymbol{W}_{B'} + \boldsymbol{\varepsilon}$$

Where $\boldsymbol{W}_{\text{A'B'}}$ denotes the observed relative fitness of the double mutant A'B' compared to the un-mutated 'wild type' of fitness $\boldsymbol{W}_{AB} = 1$. $\boldsymbol{W}_{\text{A'}}$ and $\boldsymbol{W}_{\text{B'}}$ are the fitness values of single mutants A' and B', respectively, and $\boldsymbol{\varepsilon}$ is the fitness effect, which is due to epistasis. No epistasis (i.e. $\boldsymbol{\varepsilon} = 0$) ensues if both mutations independently affect the mutant's fitness (see figure below). Positive or negative epistasis occurs if the effect of mutation A' is accelerated (i.e. $\boldsymbol{\varepsilon} > 0$) or diminished (i.e. $\boldsymbol{\varepsilon} < 0$), respectively, in the presence of mutation B'. Sign epistasis denotes the case, in which the sign of the fitness effect of mutation A' is reversed (beneficial → deleterious, or vice versa) if combined with mutation B' [94].



interactions between mutations thereby significantly affect also the shape of the *fitness landscape* (Box 2, [88]). Several studies have already shown, that epistatic interaction between gene deletion mutations, which cause the loss of metabolic functions, are prevalent in bacterial metabolic networks [89,90,88]. Hence, the reductive evolution of microbial genomes, which might be associated with the adaptive reduction of the bacterial network, is likely to be also strongly influenced by epistasis. However, the functional and molecular associations, which cause epistatic effects between metabolic genes often remain unresolved.

2. Approaches to reveal and predict bacterial metabolic adaptations

Since the origin of life, evolution has shaped bacterial metabolism towards species-specific and tightly-regulated biochemical networks. The study of the origin and evolution of metabolism is highly restricted in retracing the individual evolutionary changes of metabolic networks mainly due to (i) the absence of natural *metabolic network fossils*[*], (ii) the rather incomplete knowledge on the conditions of ancient environments and their potential metabolic niches, and (iii) the confounding complexity of 'contemporary' networks that hinders to identify the historical sequence in which different metabolic innovations occurred. Especially revealing properties of metabolic networks, which have evolved by means of natural selection has been difficult, because metabolic adaptations rely on beneficial mutations, which are very rare and therefore difficult to detect and retrace within a population [95]. Moreover, the fitness effects of adaptive mutations, which affect the performance of the metabolic network are challenging to quantify – especially in a natural setting.

Papp *et al.* (2009) nicely summarized approaches available to disclose aspects of adaptive metabolic network evolution and classified them into three classes [96]: First, *optimality models* can be applied to predict the optimal design of a given metabolic trait, which subsequently can be compared to the actual conformation of the organism. Second, *inferential analysis* can be used to compare metabolic networks between species with regard to the assumption that different selection pressures caused divergent metabolic network properties. Third, synthetic ecology and experimental evolution techniques allow to discriminate the effects of natural selection and non-adaptive evolutionary processes by experimentally comparing the performance of alternative metabolic strategies. In the following sections, I explain how these three different approaches can been applied to formulate and test hypotheses on the adaptive evolution of bacterial metabolic networks. Moreover, also advantages and limitations of each approach are discussed.

---

[*] Meant is the structure and function of metabolic networks of ancestral nodes in the phylogeny of cellular life.

2.1. Theoretical and computational approaches

Metabolic optimality models

Optimality models predict the state of a given system, which represents the solution that either maximises or minimises an a priori defined objective function. In evolutionary biology, optimality models are used to predict phenotypes, which maximise the organism's fitness [2]. The combinations of several phenotypic traits that contribute to the organism's fitness create a space of individual phenotypes [97] and the mapping of the phenotypic space to the corresponding fitness values is defined as the *fitness-* or *adaptive landscape* [2]. The central assumption behind optimality models is that the evolution of traits is mainly driven by natural selection [98]. The models can be useful to predict the phenotype, which represents a peak in the fitness landscape and to identify how close a given population of organisms is to this fitness optimum [2]. In other words, optimality models can predict how organisms *should* behave to be optimally adapted to a given environment [40].

Flux balance analysis (FBA) is the most frequently used optimality model for the evolution of metabolic networks, which optimises distribution of reaction fluxes through the metabolic network (see section 3.2). Besides metabolic flux states, different optimality models were developed (as reviewed in [96]), which suggest adaptive metabolic network evolution that shaped global network topology [29], functional robustness against mutations [51], pathway regulatory circuits [99,100], and protein assembly strategies for multimere enzymes [101].

One vital limitation of optimality models is the necessity to ideally consider all metabolic trait alternatives that evolution can theoretically attain [96]. For instance, to predict the optimal metabolic flux distribution and assuming steady-state conditions, all theoretically possible metabolic flux states can be mathematically described (see [46,102], and section 3.2) and considered in order to predict the optimal metabolic phenotypes. The situation is more challenging for the topology of biochemical networks: First, albeit ample knowledge is available on a vast spectrum of biochemical reactions across a wide range of species, most likely there are many yet-undiscovered organismal metabolic capabilities. Second, enzymatic reactions or pathways might exist, which were part of ancient metabolic networks, but were lost and replaced by alternative reactions in the course of evolution. Third, to support a hypothesis that states that a given network topology represents the optimal solution it would be necessary to collect all theoretically plausible alternative networks and assess their performance with respect to a defined fitness measure; e.g. mutational robustness or growth rate. Such approaches, which rely on the enumeration of all network alternatives, however, usually fail for genome-scale metabolic networks, because of the combinatorial complexity of theoretically possible biochemical network architectures [7]. In contrast, optimality models can give insight into the adaptive origin of smaller sub-network topologies of metabolism. Hatzimanikatis *et al.* (2005) enumerated all combinations of known biochemical reactions, which, at least theoretically, would enable the production of the aromatic amino acids from the common precursor chorismate [52]. They found around 75,000 possible pathways for phenylalanine; 350,000 for tyrosine, but only 13 for tryptophan. Analysing these pathways for their thermodynamic properties revealed that the biosynthetic pathways that are used in nature are thermodynamically optimal and are, thus, most likely those pathways, which natural selection has favoured [52]. This observation also provides a plausible explanation why all investigated organisms, which are able to produce tryptophan employ the exact same pathway [15].

Another limitation of optimality models is the formulation of the objective function, which is ideally closely related to the species' fitness. However, which metabolic traits contribute to fitness and to what extent is not trivial to judge [103]. As already mentioned above, one of the main traits, which contributes to fitness of bacteria is their ability to grow. Yet, it has been shown that there can be differences in the optimal metabolic strategies depending on whether the selective pressure is favouring optimal *growth rate* or *growth efficiency* [4,104,105]. Hence, to gain meaningful insights into the operation and evolution of metabolic networks using optimality models requires detailed knowledge on the selection pressures that shaped the network.

Taken together, optimality models are powerful tools to predict the adaptive evolution of the metabolic network function and structure. Still, their predictive potential vitally depends on comprehensive consideration of evolutionary plausible metabolic trait alternatives and the selective pressures that governed the evolution of the species' metabolic networks.

Inferential analysis of metabolic network structures

The number of available bacterial whole-genome sequences is constantly increasing. At the time of writing this thesis, the Genomes Online Database (GOLD)* contained more than 42,500 bacterial genomes [59]. Comparative genomics and nucleotide substitution models made it possible to infer signatures of natural selections at the level of

---

* http://gold.jgi-psf.org (retrieved last on July 28th 2015)

nucleotide sequences without the need for fossil DNA, e.g. by using dN/dS ratios[*] in protein-coding genes [106]. Advances in genome-based and (semi-)automatic methods to reconstruct the corresponding biochemical networks further enabled to survey species-to-species variation in bacterial metabolic capabilities [107]. Is it thus possible to also infer adaptive features within metabolic network structures in a phylogenetic context by comparing them between multiple species in a similar fashion as comparing gene sequences?

Differences and similarities between bacterial metabolic networks are a phenotypic manifestation that reflects the past adaptive evolution of bacteria. Forst *et al.* (2006) algebraically compared microbial metabolic networks and showed that the network structures contained similar phylogenetic information as commonly used 16S RNA sequences [108]. However, the difficult task is to disentangle, which features arose due to the adaptation to distinct metabolic niches from those that are simply due to drift.

Pál *et al.* (2005) compared the metabolic capabilities of *E. coli* K12 and *Salmonella typhimurium* to infer, which metabolic reactions were acquired by *E. coli* since the split of the two lineages [22]. The results were combined with flux balance analysis of the metabolic network of *E. coli*, while considering different environmental conditions varying in their carbon source and oxygen availability. In this way, the authors were able to show that the reactions, which were acquired since the split from the *Salmonella* genus provided *E. coli* with the ability to grow (or at least to form biomass) in novel nutritional environments, which suggests that acquisition of biochemical reactions may represent an adaptation of *E. coli* to changing environments [22]. This example shows that a combination of optimisation models and inferential analysis of metabolic networks can provide insights into the adaptive evolution of biochemical networks.

In a similar study, the metabolic networks of the previously mentioned endosymbiont *B. aphidicola* and *Wigglesworthia glossinidia* were compared to the metabolic network of the close free-living relative *E. coli* [109]. The metabolic networks of the endosymbionts are characterised by a severe reduction of network size, which is due to successive losses of biosynthetic genes that are disposable in the intracellular environment. Also in this example, the authors combined the network comparison with flux balance analysis and showed that the different environments and the expendability of certain reactions can account for the observed differences in the metabolic networks between the endosymbionts and *E. coli* [109]. However, whether

---

[*] Rate of non-synonymous nucleotide substitutions (dN) per rate of synonymous substitutions (dS) within a protein-coding gene.

the loss of metabolic reactions is governed by natural selection or drift cannot be inferred using this approach. In fact, is remains obscure to what extent the reductive genome evolution of metabolic networks can be explained by selective advantages, which genotypes gained that lost the biosynthetic functions.

In general, inferential (or comparative) analyses of reconstructed metabolic networks are a rich resource to formulate hypotheses on the evolution of metabolic networks. In combination with other approaches such as optimisation models or *in vivo* experiments, the comparison of metabolic networks between species facilitates a better understanding of the adaptive evolution of microorganisms on a biochemical level. However, it is yet not possible to statistically discriminate the effect of natural selection and drift solely based on network topology comparisons. The reason is the difficulty to construct a probabilistic model for network evolution [96] in a similar way as nucleotide substitution models to identify footprints of natural selection in protein-coding sequences. This is because metabolism fulfils numerous different tasks related to bacterial cell growth and proliferation. Thus, different parts of the metabolic network are likely to be shaped by different selection pressures, which complicated the mathematical formulation of a global objective for network topology evolution. Nonetheless, comparative analyses of metabolic networks can contribute significant insights into the study of metabolic network adaptation, by providing a wide survey of metabolic strategies that exists in nature. For instance, the databases KEGG [110] or MetaCyc [111] collect the knowledge of biochemical reaction with the aim to represent and analyse the global diversity of metabolism and to make metabolic networks comparable between species. An assessment of possible alternatives to perform a certain metabolic function is crucial to decide if a specific metabolic trait has evolved as an adaptation to a microorganism's environment and lifestyle.

2.2. Experimental approaches to unravel metabolic adaptations

Laboratory experiments with microorganisms have the unique advantage to explore the performances of different alternative metabolic strategies *in vivo*. By rationally designing growth environments and even by manipulating genotypic constitution of the focal strains, different scenarios with diverse selection pressures can be systematically tested to specifically address adaptive hypotheses. Moreover, experiments also enable to monitor evolutionary changes on many different levels of organisation: the genome, transcriptome, proteome, fluxome, metabolome, or secretome.

Synthetic biology

In synthetic biology, biological systems are rationally constructed for a specific purpose, for example for the commercial production of value-added compounds by microorganisms [112]. The design of the system may include the environmental conditions (e.g. nutrient availability, pH, or temperature), the consortia of involved organisms, and their genotypes. In evolutionary biology, for instance, such an approach can be used to investigate the fitness consequences of individual mutations under given biotic and abiotic environmental conditions.

For example, Chou *et al.* (2011) studied the effect of four mutations in *Methylobacterium extorquens* EM, which affect the expression of a metabolic pathway that is necessary for the utilisation of methanol as sole carbon source [113]. The authors showed that each individual mutation entailed selective benefits compared to the non-mutated 'wild type' strain but the proportional selective benefits decreased when mutations were introduced in genetic backgrounds, which already carry one of the other beneficial mutations – an effect termed *diminishing returns epistasis* (Box 2; and [113,114]). This example demonstrated the potential of synthetic approaches to study the metabolic adaptation of bacteria and that epistasis also impacts the evolution of metabolic networks [113].

Synthetic ecology is an extension of synthetic biology, where also ecological interactions between different organisms are subject of the rational design. For example, Shou *et al.* (2007) constructed an ecosystem that consisted of two yeast strains, which obligatorily cooperate by the mutual exchange of essential metabolites (i.e. adenine and lysine) [115]. This study has impressively shown that ecological interaction – such as metabolite cross-feeding (see section 1.3) – can be synthetically introduced between different microorganisms by targeted modifications of genomes. Furthermore, such engineered systems can be used to test hypotheses on the origin and stability of cooperative ecological interactions [5,115]. In fact, Pande *et al.* (2014) demonstrated with a conceptually similar approach – an engineered obligate cross-feeding consortia of two *E. coli* strains – that cooperating genotypes gained selective advantages in co-culture compared to the non-cooperating ancestral wild type strain [76]. This example provided evidence that metabolic interdependencies and complementarity might evolve through natural selection and therefore contribute also to the evolution of metabolic networks.

In summary, synthetic biology approaches can be used to assess the fitness consequences of individual mutations, which affect the structure and/or function of metabolic networks. The results gained from synthetic biology model systems can inform the prediction of how bacteria adapt to new environments and to understand how modern bacterial metabolic networks have evolved, by replaying specific evolutionary metabolic transitions. Synthetic biology is still a relatively new approach to test evolutionary hypotheses. However, due to the high flexibility to manipulate several different settings of biological systems, synthetic approaches will facilitate further insights into the origin and evolution of metabolic networks.

Experimental evolution

The strength of experimental evolution lies in the possibility to monitor the genotypic and phenotypic changes over time within a population in a defined experimental setup. Depending on the focal hypothesis, a precisely defined selection pressure can be applied to the evolving population. Most evolution experiments are making use of microorganisms, mainly because of their short generation time. Fast-growing bacteria such as *E. coli* can divide every 20 min under optimal growth conditions [116]. As an illustration, within a 3-year PhD project, one could conduct an evolution experiment with *E. coli* for more than 78,000 generations (assuming no constraints on growth), which corresponds to approximately 1.8 million years of human evolution. Another advantage of microorganisms is the possibility to create a comprehensive library of 'fossil records' of viable cells by regularly freezing population samples. In contrast to inferential analysis of metabolic adaptation (section 2.1), experimental evolution has the strong advantage that is allows one to 'observe' evolutionary processes; including natural selection and genetic drift [41].

Evolution experiments can provide evidence that a specific evolved trait is adaptive if the trait emerged and is maintained in several independent parallel populations and if the trait evolved and is maintained statistically more often in the presence of a specific selection pressure than in the absence of the pressure. However, it needs to be mentioned, that parallel evolution supports an adaptive hypothesis, but it is not a necessary feature of adaptive traits [117].

Furthermore, evolution experiments can give insights into adaptive evolution if the mutation, which caused an evolved trait is known. In this case, the trait can be directly introduced into the genetic background of the ancestral strain and the fitness of the engineered genotype relative to the ancestral genotype can be assessed in competition experiments (a synthetic biology approach). By means of these measures, the dynamics of metabolic adaptation of microorganism has been extensively studied including the evolution towards optimal flux distributions [40,118], adaptive evolution of metabolite cross-feeding [119], the acquisition of metabolic capabilities to utilize novel resources

[117], evolution of resource specialisation [120], sympatric speciation towards distinct metabolic niches [121,122,119], and the adaptive difference between the evolution towards growth rate or growth efficiency [123].

In summary, synthetic biology and experimental evolution approaches are powerful tools to test hypotheses about the evolution of structural and functional traits of metabolic networks. However, it also needs to be noted that experimental studies on metabolic adaptation are typically limited to fast-growing microorganisms, which are culturable in the laboratory. Furthermore, another limitation is that it is difficult to mimic natural environmental conditions in a laboratory setting, e.g. with respect to the stochasticity of the abiotic and biotic composition.

3. Mathematical formalisation and analysis of genome-scale metabolic networks

3.1. Genome-scale metabolic network reconstructions

Systems biology is a pluralistic approach, which brings together different disciplines from biology, chemistry, physics, mathematics, and engineering to test and create new hypothesis on how living systems are organized, function and evolve [124]. A *system* can be defined as a set of components, which interact with each other. Systems biology seeks to identify the components and the types of interactions within a living systems, to formalise the resulting network in mathematical terms, and to draw conclusions from the network structure on the function and dynamics of the system. Considering the cellular metabolism as the biological systems of interest, a central goal of systems biology is to generate models that allow to predict metabolic phenotypes from the genotype of a given organism.

Genome-scale metabolic network reconstructions are large models of cellular metabolism with the aim to integrate all catalytic reactions for which a gene of the corresponding enzyme is present in the genome, all relevant non-enzymatic reactions as well as transport reactions [125]. The main levels of information of such models are: (i) the gene-enzyme relationship states what enzyme is encoded by which gene, or by which set of genes in case of multimer enzymes; (ii) the enzyme-reaction mapping specifies what reaction an enzyme is able to catalyse, and (iii) the cross-linking of reactions through metabolites into a coherent network. In the latter level, each reaction is characterized by its stoichiometry, which is the combination of the compounds involved and their quantities that are consumed or produced. The whole set of reactions, i.e. the metabolic network, can be represented by a so-called stoichiometric matrix $\boldsymbol{S}$, where each column denotes a specific reaction and each row a metabolite. An entry $\boldsymbol{S}_{ij}$ specifies the production (positive value) or consumption (negative value) of the corresponding metabolite i by the respective reactions j . The stoichiometric representation of metabolic networks is a useful concept to computationally analyse the network topology [126]. Moreover, stoichiometric models have been successfully applied to predict metabolic flux distributions [46]. Commonly used computational methods for metabolic flux analysis are elementary flux modes [102,127], flux balance analysis (see section 3.2), and ordinary differential equations [128,129].

The strength of genome-scale metabolic networks lies in their comprehensiveness. For example, one purpose of genome-scale models of metabolism is to simulate microbial growth [46]. Cellular growth involves the production of a large set of different metabolites including amino acids, nucleotides, lipids, vitamins, co-factors, and carbohydrates whose biosynthesis requires numerous of reactions. In combination with constraint-based modelling approaches (see section 3.2), genome-scale metabolic networks have been shown to be a powerful tool to simulate microbial growth and facilitate precise predictions on how growth is affected by genetic perturbations of the metabolic network [130] or different nutritional conditions [131].

Another example for an application of genome-scale models of microbial metabolism is systems metabolic engineering of microorganisms for the production of value-added chemicals [132]. In this context, genome-scale metabolic models are applied to predict genetic modifications, which presumably enhance or enable the economical production of a desired compound. Such genetic modifications include amongst others the addition of biosynthetic genes to introduce new synthetic metabolic pathways [132,133], the overexpression of certain genes to increase the flux through a specific pathway [132,134–136], or the deletion of biosynthetic genes to reroute metabolic fluxes [132,135].

In addition to the simulation of microbial growth and the development of metabolic engineering strategies, genome-scale metabolic network reconstructions have been successfully applied to understand microbial metabolic interactions within ecosystems [10,137], to design cultivation media and processes [138,139], for drug target predictions [140], and to assess the evolutionary plasticity of metabolic networks [108,17]. Taken together, genome-scale metabolic network reconstructions are comprehensive and functional models of an organisms' biochemical capabilities with vast predictive potential to aid genotype-phenotype mapping. While a quantitative understanding of enzyme activities within metabolism remains difficult for genome-scale networks and is still mainly limited to small sub-networks [7], genome-scale metabolic networks are of increasing importance to understand biological processes that involve many or even almost all metabolic pathways of a cell.

3.2. Flux balance analysis

Flux Balance Analysis (FBA) is a computational framework to model metabolism on a genome-scale network level. It predicts the evolutionary optimal distribution of metabolic fluxes by optimising a cellular objective, which is linked to the maximisation of the species' fitness. The most commonly used objective function is the optimisation of biomass formation [103]. Others are the generation of ATP or the minimisation of the sum of fluxes to economise the burden of total enzyme levels needed for metabolism [103]. Independent of the type of the objective function, FBA solves a resource allocation problem on the bases of the structure of the metabolic network by optimally exploiting limited resources and complying with thermodynamic constraints such as the reversibility of reactions [141].

The FBA framework assumes a metabolic *steady-state*, in which each internal metabolite[*] is produced and consumed at the same rate such that reaction fluxes are balanced to keep all metabolite concentrations constant. In a mathematical formulation using the stoichiometric matrix $\boldsymbol{S}$, a distribution of fluxes that complies with the steady state constraint $\boldsymbol{v_{st}}$ is given by

$$\boldsymbol{S}\,\boldsymbol{v_{st}} = \boldsymbol{0}. \tag{1}$$

Besides the stoichiometry and the steady-state assumption, further constraints can be added to the model such as maximal rates for individual reactions and maximal nutrient uptake rates. Linear programming (LP) can be subsequently applied to identify a flux distribution $\boldsymbol{v_{st}}$, which complies with all constraints and reflects the optimal solution for a given objective function. LP is a mathematical optimisation

---

[*] To set boundaries to the model, metabolites are divided into internal and external metabolites. External metabolites are assumed to be buffered, e.g. like nutrients in a chemostat or molecules like water which are presumably by order of magnitudes more abundant than other metabolites. Internal metabolites on the other hand are thought to be limited in their availability and the only source for these metabolites are metabolic reactions forming them.

method, which calculates the best outcome of a given objective function, whose variables contribute each linearly to the functions results and where variables are subject to linear inequality constraints. A flux balance analysis problem is usually represented in the canonical form of LP problems, for example:

$$\text{Maximise} \quad \boldsymbol{c}^T\boldsymbol{v} \quad \text{(Objective function)} \tag{2}$$

$$\text{Subject to} \quad \boldsymbol{Sv} = \boldsymbol{0} \quad \text{(Steady-state constraint)} \tag{3}$$

$$\text{and} \quad \boldsymbol{v}_{min} \leq \boldsymbol{v} \leq \boldsymbol{v}_{max} \quad \text{(Individual flux limits)} \tag{4}$$

Where the vector $\boldsymbol{c}$ specifies how much each reaction linearly contributes to the given objective and $\boldsymbol{v}$ is the vector for the actual flux distribution to be optimised. $\boldsymbol{v}$ states the flow through each reaction of the network. The vectors for the capacity constraints $\boldsymbol{v}_{min}$ and $\boldsymbol{v}_{max}$ denote the lower and upper limits for each reaction. By specifying the constraints for the individual flux limits, the reversibility or irreversibility of reactions can be integrated in the model: If a reaction A → B is irreversible, the direction of the reaction flux can be defined by setting $\boldsymbol{v}_{min} = 0$ and $\boldsymbol{v}_{max} > 0$. In contrast, if the reaction is reversible (A ↔ B) both direction are allowed by setting the lower flux limit to $\boldsymbol{v}_{min} < 0$.

In combination with genome-scale metabolic networks, FBA models have been applied extensively and became a standard tool in systems biology for analysing metabolic networks [141]. The models have made significant contributions in the fields of cellular physiology, synthetic biology, ecology, metabolic network evolution, and network reconstruction [137,142,141].

3.3. Other mathematical techniques to analyse genome-scale metabolic networks

For the project presented here, flux balance analysis to model the allocation of limiting resources in the genome-scale metabolic network of *E. coli* and inferential analysis to systematically compare the metabolic capabilities of a wide range of bacterial species were applied. There are of course more computational techniques to study the structure and function of genome-scale metabolic networks and which have also contributed to a better understanding of network evolution. Two of these methods, i.e. *ordinary differential equations* and *elementary flux modes*, are briefly outlined in the following.

Ordinary differential equations (ODEs) are widely used to model biochemical reaction systems. These models have the unique advantage to be able to predict dynamic changes of reaction rates (fluxes) and metabolite concentrations including oscillations and bi-stability within reaction systems [143]. ODE-models require mechanistic knowledge of the involved enzymes to mathematically formulate the reaction rate kinetics. However, such knowledge is often limited to a small subset of all enzymes encoded in the genome of an organism, which has limited the applicability of ODE-models to small- or medium size metabolic networks [144]. Nevertheless, some approaches now exist, which have combined genome-scale models of metabolism by integrating enzyme kinetics [144,145] and thereby were able to simulate dynamic metabolic responses on a genome-scale metabolic network level.

Elementary flux modes (EFMs) are, as flux balance analysis, a theoretical concept to study reaction networks that assumes steady-state conditions [102]. A *flux mode* $\boldsymbol{v_{st}}$ is defined as non-zero vector of reaction rates $\boldsymbol{v}$, which enures a steady-state of metabolite concentrations (Equation 1). Furthermore, a flux mode is *elementary* if it cannot be represented as a linear combination of other flux modes [102]. The advantage of this definition is that the number of possible elementary flux modes $\boldsymbol{e_i}$ is finite for a given network, while the number of feasible flux modes can be infinite. Every metabolic steady-state (flux mode) $\boldsymbol{v_{st}}$ can thereby be decomposed by the weighted sum of all n EFMs:

$$\boldsymbol{v_{st}} = \sum_{i=1}^{n} \mu_i \boldsymbol{e_i}$$

Where μ_i corresponds to the weight of the EFM $\boldsymbol{e_i}$ in the flux mode $\boldsymbol{v_{st}}$. This representation of all feasible steady-state flux distributions can be the basis for further analysis to characterise cellular metabolism [146]. However, the enumeration of all EFMs in genome-scale metabolic networks is infeasible due to the large number of possible elementary flux modes [147,148]. Nonetheless, there are studies that focused on the enumeration of subsets of EFMs [149,150] and it has been shown that such an approach can, for example, successfully be used to predict reaction knock-out strategies for metabolic engineering [151].

4. Objectives of this study

Our knowledge of the biochemistry of metabolism, which has been constantly increasing over many years and advances in the computational analysis of whole-genome sequences enabled to infer the structure of metabolic networks and the circuits for the regulatory coordination of its reactions based on individual genomes. However, the factors that govern the evolution of metabolic network structures and regulatory circuits remain obscure. The overarching questions of this thesis are: What is the contribution of natural selection on the evolution of metabolic networks? And which properties of bacterial metabolic networks affect the species' fitness and thereby indicate adaptation to specific environmental conditions and/or to biophysical constraints in biochemical networks?

More specifically, I will address the following questions in this thesis: (i) Can the fact that metabolic networks usually involve only relatively few autocatalytic cycles, be used to unravel inconsistencies in genome-scale metabolic network reconstructions in order to improve the accuracy of the *in silico* networks and thereby also the potential of these models to make valuable predictions? (ii) How common is the loss of biosynthetic genes across bacteria? (iii) Can the gene loss and, hence, the loss of metabolic autonomy be explained by selective advantages, which auxotrophic genotypes gain if the focal metabolite can be obtained from the environment? (iv) Do certain combinations of biosynthetic functions tend to be jointly absent from bacterial genomes in nature? (v) Can the co-occurrences of auxotrophies be explained by epistatic interactions between auxotrophy-causing mutations? (vi) How do different carbon sources affect the fitness consequences of biosynthetic gene loss? And (vii) how are the costs, which a cell needs to invest in the performance of a given biosynthetic function, influenced by the metabolic network architecture and the type of carbon source?

To answer these questions, experimental and theoretical approaches were combined. On the theoretical side, the metabolic networks of more than 900 bacterial species were compared for their biosynthetic capabilities and how the differences could be explained by the species' lifestyles. Furthermore flux balance analysis was employed to simulate different resource allocation strategies of *E. coli* under different carbon source environments. Experimentally, different metabolic function-deficient genotypes of the prototrophic *E. coli* and *Acinetobacter baylyi* strains were synthetically generated and tested for their growth kinetics and relative fitness under various carbon environments.

In the end of this thesis, I discuss the main results of the project to elaborate the role of metabolic adaptation in the bacterial biosphere and how the results may explain metabolic complementary, the immense bacterial diversity found in nature, as well as the observation that most known bacteria could not be cultured and characterised under laboratory conditions. Moreover, I highlight practical implications of the results for medical and biotechnological applications.

Chapter II

Computing autocatalytic sets to unravel inconsistencies in metabolic network reconstructions

Published in Bioinformatics, volume 31, issue 3, pages 373-381 in February 2015.

Authors

Ralf Schmidt*, Silvio Waschina*, Daniela Boettger-Schmidt, Christian Kost, and Christoph Kaleta

* These authors contributed equally to this work.

1. Abstract

Motivation: Genome-scale metabolic network reconstructions have been established as a powerful tool for the prediction of cellular phenotypes and metabolic capabilities of organisms. In recent years, the number of network reconstructions has been constantly increasing, mostly because of the availability of novel (semi-)automated procedures, which enabled the reconstruction of metabolic models based on individual genomes and their annotation. The resulting models are widely used in numerous applications. However, the accuracy and predictive power of network reconstructions are commonly limited by inherent inconsistencies and gaps.

Results: Here we present a novel method to validate metabolic network reconstructions based on the concept of autocatalytic sets. Autocatalytic sets correspond to collections of metabolites that, besides enzymes and a growth medium, are required to produce all biomass components in a metabolic model. These autocatalytic sets are well-conserved across all domains of life, and their identification in specific genome-scale reconstructions allows us to draw conclusions about potential inconsistencies in these models. The method is capable of detecting inconsistencies, which are neglected by other gap-finding methods. We tested our method on the Model SEED, which is the largest repository for automatically generated genome-scale network reconstructions. In this way, we were able to identify a significant number of missing pathways in several of these reconstructions. Hence, the method we report represents a powerful tool to identify inconsistencies in large-scale metabolic networks.

Availability and implementation: The method is available as source code on http://users.minet.uni-jena.de/~m3kach/ASBIG/ASBIG.zip.

2. Introduction

In recent years, genome-scale metabolic network reconstructions have become an important tool in systems biology [152]. They have the strong potential to combine distinct experimental data with bibliomic resources to generate a comprehensive knowledge base [142,153]. The resulting network reconstructions have been widely used to simulate metabolic processes or to explore the metabolic capabilities of various species [154,155]. A prominent approach applying network reconstructions is constraint-based modeling, such as flux balance analysis (FBA) [155,156]. The use of metabolic models and associated methods has granted access to diverse scientific subjects, such as analysis of the bacterial metabolism [50], the prediction of growth rates of *Escherichia coli* [46], the comparison of growth rates between wild type and mutant strains of *E. coli* [130] and metabolic engineering [157].

Until recently, the process of network reconstruction was time-consuming and necessarily required laborious curation effort [158]. To cope with the increasing amount of available data, automated methods became necessary to produce high-throughput reconstructions as well. Several approaches are available to address this issue [159–161]. The Model SEED project successfully implemented one of these approaches and contains >190 metabolic reconstructions, which were primarily generated in an automated way [162].

However, the automated process may neglect metabolic capabilities of the focal organism or include wrongly identified functionalities [163]. To minimize this bias, manual refinement and optimization are instrumental in the reconstruction process [152,162], representing the most intricate elements in the workflow. To accelerate model curation, automated and semi-automated methods for the detection and correction of gaps were developed and integrated in the automated reconstruction process by the Model SEED (GapFind and GapFill) [162,164]. The most common approach used to identify gaps in network reconstructions is constraint-based modeling that relies on optimization-based algorithms. The predominant aim is to detect metabolites that cannot be produced at a steady state [164–166].

Here we report ASBIG (Autocatalytic Set-Based Identification of Gaps), a method that detects incomplete parts of network reconstructions based on a novel approach: identifying elements (compounds) of catalytic cycles. ASBIG screens models for essential, self-replicating metabolites, which are pivotal elements of the underlying metabolism. A metabolite is considered to be pivotal if it is required for metabolism to proceed. The compounds are called ‘self-replicating’, or ‘autocatalytic’, as they are usually required for their own biosynthesis [30]. Hence, their production is inaccessible unless an initial amount of the compound is already available. Self-replicating metabolites are energy-currency cofactors, such as ATP and NAD. If these compounds are not present within the cell, even a rich nutritional medium is often insufficient for an organism to produce all of its biomass components. The definition of self-replicating metabolites originates from the work of Eigen and Schuster (1977), in which the concepts of catalytic cycles and autocatalysis were described on a molecular level [167]. The search for autocatalytic compounds is conducted via the method of scope analysis [168].

The set of self-replicators in each network reconstruction is generally small [30]. Furthermore, autocatalytic sets of different organisms very likely contain similar well-conserved compounds. These observations can be used to survey any given metabolic network to detect inherent inconsistencies. If the number of self-replicating compounds is comparably large or unexpected elements and large macromolecules are included in the set, conclusions about possible inconsistencies can be drawn. Hence, gaps in the reconstructed network can lead to the presence of unexpected compounds in the set of self-replicators (owing to the impaired biosynthesis of this compound or its successors).

ASBIG enriches the tool box of gap-finding procedures, as common flux calculating methods are susceptible to overlook gaps associated with autocatalytic cycles. One simple example is a cofactor that is required for the production of a biomass component but itself is not present in the biomass reaction. In such a case, it would be sufficient for constraint-based methods, such as FBA, if reactions that replenish the cofactor exist to produce biomass. A biosynthetic route of this essential cofactor would not be required making the detection of a lack of this route impossible for FBA. ASBIG uses scope analysis to determine such inconsistencies. Although this approach disregards the exact stoichiometric properties of metabolic reactions, our method is not susceptible to problems known to occur when using purely topology-based procedures [169].

To outline the benefit of ASBIG, selected reconstructions from the following organisms were examined: *E. coli*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Bacillus subtilis*, *Chlamydomonas reinhardtii*, and *Zea mays* (Table 1). Using our method, we were able to identify missing pathways in all of these networks except for the *E. coli* reconstruction. Subsequently, the detected inconsistencies were resolved using information from the KEGG database [170] and other resources, like BioCyc, BsubCyc [171] or the Plant metabolic pathway database [172]. Furthermore, all available reconstructions of the Model SEED project [162] have been screened to demonstrate the applicability of the method on a large scale. As one substantial benefit of this screening, common inconsistencies widespread across numerous models were identified. Additionally, ASBIG detected missing reactions that were deleted manually from the model of *E. coli* for validation purposes, demonstrating the reliability and applicability of our method.

Table 1. Closely investigated models.

Organism	Model
E. coli K-12 MG1655	iJO1366 [26]
B. subtilis 168	*i*BSU1103 [27]
A. thaliana	AraGEM [173]
C. reinhardtii	iRC1080 [174]
S. cerevisiae	iMM904 [175]
Z. mays	DTMaize_C4GEM_45632 [176]

In summary, ASBIG is a reliable method for the detection of inconsistencies in any metabolic network reconstruction. It provides an efficient approach to validate metabolic models. By pinpointing discrepancies of network reconstructions, the method supports the improvement of the high number of incomplete models generated by high-throughput methods. Thus, ASBIG can significantly contribute to improve the quality of metabolic network reconstructions.

3. Material and Methods

3.1. Models

ASBIG uses a genome-scale reconstruction of metabolic networks as input. It was applied to investigate the following models: iJO1366 of *E. coli* K-12 MG1655, *i*BSU1103 of *B. subtilis* 168, AraGEM of *A. thaliana*, iRC1080 of *C. reinhardtii*, iMM904 of *S. cerevisiae* and DTMaize_C4GEM_45632 of *Z. mays*. Additionally, >190 models provided by the Model SEED project [162] (as available in October 2012) were screened.

3.2. Identification of autocatalytic metabolic sets

The algorithm was implemented in JAVA. Genome-scale metabolic networks are processed in the standardized Systems Biology Markup Language using the JAVA package jigcell.sbml2 [177].

3.3. Initial set of metabolites (seed set)

A predefined set of metabolites, named ‘initial seed set’, acts as the starting point for ASBIG. It combines the components of a nutritional medium with additional pivotal elements. The set of additional elements is composed of two subsets: metabolites identified as crucial by ASBIG (see section 3.1 and autocatalytic compounds whose biosynthesis depends on their own presence (cf. [30]). These were as follows: Adenosine triphosphate (ATP), nicotinamide adenine dinucleotide (NAD) and coenzyme A (CoA).

Hence, the composition of the initial seed set is assumed to include all essential components necessary for the growth of the organism.

3.4. Scope analysis

ASBIG uses the concept of scopes [168], which can be determined for a given sets of metabolites. The scope is a (large) set of metabolites and corresponds to all compounds that can, in principle, be produced from a (small) predefined set of metabolites. The concept of scope analysis relies on three aspects [168] (i) a reaction is considered to take place if all of its substrates (or products in the case of reversible reactions) have non-zero concentrations, (ii) products of one reaction are immediately considered as potential substrates of another one and (iii) starting with a small set of metabolites, iteratively increasing sets of reactions (and metabolites) are generated by screening the model for further reactions with a non-zero flux. The fundamental principle of ASBIG relies on the calculation of a scope set based on a given seed set of metabolites.

3.5. Biomass reaction

An essential feature among most network reconstructions is the biomass reaction, which is crucial for the ASBIG method. During a run of ASBIG, this reaction is used as a benchmark to evaluate generated metabolite sets. The biomass reaction constitutes a hypothetical reaction within the model that includes all metabolites necessary for the growth of the corresponding organism as substrate (biomass compounds). Usually, it is an essential element of FBA [178]. For ASBIG, it is reasonable to use this reaction as benchmark and compare metabolite sets based on their ability, to provide access to the biomass compounds. A seed set cannot be considered as adequate for the corresponding organisms until all biomass compounds are within the scope of the seed set.

3.6. Expanded seed set

Given the initial seed set, not all biomass compounds are necessarily part of the computed scope. Hence, the initial seed set is insufficient and needs to be expanded to generate a scope comprising all biomass components. Iteratively, metabolites, named

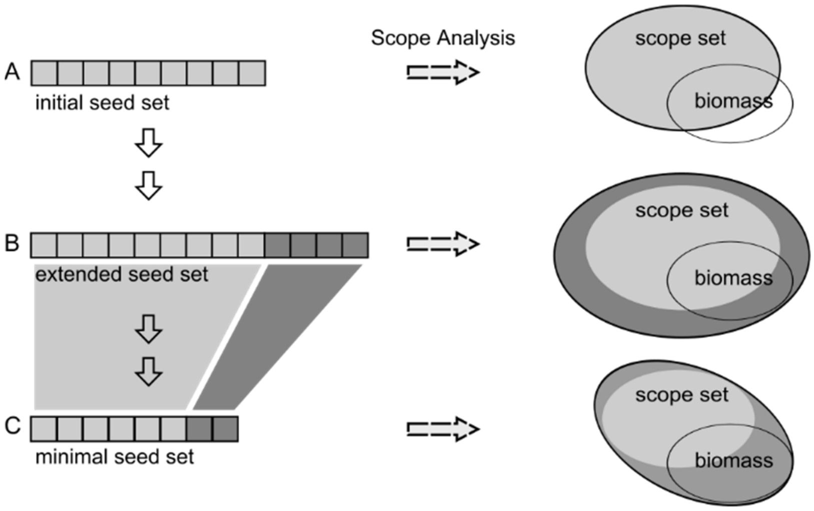


Figure 1. Workflow of ASBIG. (A) Based on an initial seed set of metabolites (light gray squares), a scope set is calculated using scope analysis (this scope set might not include all biomass compounds). (B) The initial seed set is extended with add-on metabolites (dark gray squares) leading to a larger scope set (dark gray circle), until the corresponding scope set contains all biomass compounds. (C) The extended seed set is reduced to the smallest possible seed set that still contains all biomass components (final minimal seed set).

'*add-on metabolites*', are added using a greedy approach. In each iteration, each metabolite that is not within the previously calculated scope is considered and inserted into the extended seed set on a trial basis. Finally, the compound, which results in the biggest increase of scope size represents the add-on metabolite of this iteration and remains in the extended seed set (independent of its impact on the biomass function). To this end, a scope analysis has to be performed for each metabolite, which was not part of the previously computed scope. If two metabolites yield an identical increase in scope size, the smaller metabolite in terms of the number of carbon atoms is chosen. Subsequently, a scope analysis is conducted with the temporary expanded seed set. The expansion process stops when the current expanded seed set results in a scope set that comprises all biomass compounds (Fig. 1 A and B).

3.7. Determining the minimal seed set

Most likely, not every metabolite of the expanded seed set contributes significantly to biomass production. For this reason, we aimed to determine a minimal seed set in terms of size during the second phase of ASBIG (Fig. 1C). To minimize the size of the expanded seed set, each element is removed on trial, and the scope of the remaining seed set is computed. If all biomass compounds are still within the scope, the element remains

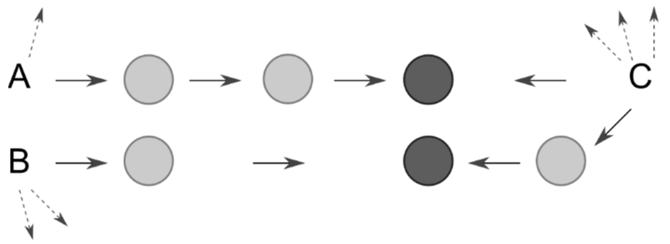


Figure 2. Scenario to illustrate the necessity of a random approach to minimize the extended seed set. A and B represent initial metabolites, whereas C is an add-on compound. Light gray circles symbolize intermediates of the depicted pathways, and the dark gray circles represent biomass compounds. Dashed arrows hint further reaction routes. A and B together lead to the same biomass components as the availability of C. If, for example, A is deleted in the second phase of ASBIG, B alone lacks the potential to maintain the producibility of the two biomass compounds. Consequently, C would emerge in the final minimal seed set.

removed, otherwise it is reinserted. Potential dependencies between compounds of the expanded seed set impede a deterministic strategy to remove elements from the expanded seed set. Alternatively, a strategy accomplishing randomized removal of seed compounds is applied—with respect to one constraint: Elements of the initial seed set (depicted as light gray squares in Fig. 1) were prioritized and were only examined once all other add-on metabolites (dark gray squares in Fig. 1) have been tested (random removal with priorities). Unconstrained randomized removal can be critical, as the following example illustrates (Fig. 2): Three metabolites, A, B, and C, are included in the expanded seed set. A and B are initially given compounds, whereas C was appended as add-on metabolite. C, self-evidently, enlarges the scope size. However, the enlargement may not affect the biomass production (Fig. 2A and B together unlock the same biomass compounds as the presence of C). In other words, C constitutes an add-on metabolite, because of its impact on the scope size, but C is not necessary for biomass production. However, in case of an unconstrained random approach to calculate the minimal seed set, it is two times more likely that C remains in the minimal seed set: if and only if C is deleted first, A and B remain in the minimal seed set. In the two remaining cases of primary deletion of A or B, C will be part of the minimal seed set to ensure the accessibility of the relevant biomass compounds. This would be in contrast to the basic principle of ASBIG to retain the initial metabolites. To minimize this bias, all components of the initial seed set are prioritized.

This illustrates the problem of dependencies between different compounds of the seed set, which, of course, can occur between any subset of metabolites.

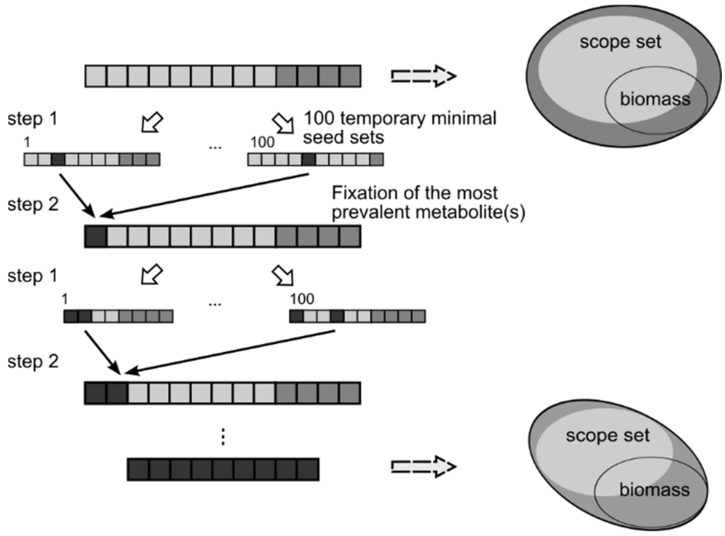

scope set
biomass
step 1
1
...
100
100 temporary minimal seed sets
step 2
Fixation of the most prevalent metabolite(s)
step 1
1
...
100
step 2
scope set
biomass

Figure 3. Visualization of the minimal seed set generation. Beginning with the extended seed set, the method computes 100 (potentially different) minimal seed sets with the constrained random procedure as described. Out of these, the most prevalent compound(s) (depicted in black) are marked as inherent part of the emerging minimal seed set and are not considered in following iterations. Within these, the procedure is repeated until the set of fixed elements is capable of producing a scope that comprises all biomass components (bottom part).

To account for such properties of the expanded seed set, the actual implementation comprises two iteratively repeated steps: During the first step, a distinct number of runs of random removal with priorities is performed to generate a pool of potential minimal seed sets (possibly, all potential minimal seed sets are different). For the analyses presented here, 100 runs were performed in one cycle. In the second step, the temporary minimal seed sets are evaluated and the metabolite(s) with the maximal number of occurrences is (are) marked as definite element(s) of the final minimal seed set (marked in black in step 2 of Fig. 3). Subsequently, fixed metabolites are not removed on trial anymore in the following iterations. Instead, they constitute an invariant part of every temporary minimal seed set once they have been fixed. In the next iteration, only the remaining compounds of the expanded seed set are removed on trial. The procedure is repeated until the set of fixed metabolites is capable of generating a scope that includes all biomass compounds.

To obtain more information about the reasons why a specific metabolite occurs in the seed set, it can be beneficial to repeat the workflow using a slightly different approach (‘*producibility constraint*’). The difference to the default procedure is an additional constraint during seed set expansion restricting the possibility of any considered compound to become part of the expanded seed set. More precisely, each metabolite outside the previously computed scope is considered as putative add-on metabolite only if its corresponding scope contains a reaction that ensures the production of the metabolite.

The described change in the workflow is not part of the default preferences. It is left to the user whether an additional run with the altered approach is required.

3.8. Final minimal seed set

ASBIG computes a final minimal seed set, whose scope set includes all biomass compounds. The minimal seed set consists of two major parts: initially given metabolites and add-on metabolites. The incorporated add-on metabolites can be divided into two major classes: (i) metabolites described in Section 3.1, which are elementary because of the structure of the model and not of principal interest, and (ii) metabolites, which show an autocatalytic behavior and embody a pivotal part of the metabolism. These compounds constitute the crucial result of ASBIG, as their presence in the minimal seed set indicates potential gaps within the metabolic network reconstruction.

4. Results and discussion

We developed a new method to detect inconsistencies in metabolic network reconstructions. The key step in this method is the identification of metabolites, which are necessary to 'unlock' specific metabolic pathways, that is, to make them accessible. Such metabolites usually represent cofactors that are required, besides the substrate of a pathway, to produce the corresponding products. As outlined above, although a wide range of metabolites of a biochemical network can usually be produced from a small set of precursor metabolites, an additional set of cofactors such as ATP and NAD is required to make most of metabolism accessible. The central goal of our approach is to identify these cofactors, which we call autocatalytic compounds, for a specific network and compare them with the usually well-conserved set of autocatalytic compounds of other metabolic models. If uncommon autocatalytic compounds are identified, this can usually be regarded as an indicator of missing metabolic functionality in a network reconstruction.

Figure 1 depicts the methodical procedure. First, the reconstruction of interest is examined via scope analysis with a predefined initial seed set (always containing the same metabolites regardless of the metabolic model investigated). This initial seed set is incrementally expanded by additional compounds until all biomass components are included in the computed metabolic scope (leading to an expanded metabolic seed set, Fig. 1B). Each compound added, named add-on metabolite, unlocks certain pathways in the model and leads to a larger metabolic scope. However, not every add-on metabolite contributes to the biomass production. Hence, in the last step of the ASBIG analysis, the expanded seed set is minimized to the minimal set of compounds required for biomass production (Fig. 1C).

In summary, for every metabolic reconstruction, a minimal set of metabolites is computed. This minimal set allows the production of all biomass components and consists of initially given metabolites and a subset of add-on metabolites (see 2.8). As prerequisite for each model investigation, it is assumed that those parts of the model that are necessary for the biomass production can be unlocked with the initially given metabolites. However, this assumption was not true for most investigated metabolic reconstructions and add-on metabolites had to be included. Each add-on metabolite has the capacity to provide access to priorly blocked parts of metabolism, and thus, uncovers putative gaps within a metabolic network reconstruction.

4.1. Commonly identified inconsistencies in multiple reconstructions

Initial runs of ASBIG indicated the presence of common autocatalytic compounds in all (or at least the majority of) the investigated metabolic models. This observation was a consequence of similar features among different models. Often, protein-derived reaction components, like the cofactor thioredoxin or the cofactor-carrying acyl carrier protein (ACP), were implemented as discrete compounds. However, biosynthesis of ribosomal protein is not included in most metabolic network reconstructions. As a result, these protein-derived compounds were present in the minimal seed set. Occasionally, metabolites that require protein-derived compounds for their biosynthesis appeared in the minimal seed set. Their assignment to the corresponding protein-derived compound involved additional effort, which could be avoided by integrating the protein-derived reaction components in the initial seed set.

Another metabolite identified by ASBIG in several network reconstructions was dihydrolipoamide. In these models, dihydrolipoamide was part of small isolated reaction cycles, excluding its own *de novo* biosynthesis. Hence, dihydrolipoamide and other commonly identified metabolites could be regarded as autocatalytic, although the lack of their biosynthetic pathways represented a common flaw of the corresponding metabolic networks. Consequently, the compounds were added to the initial seed set for subsequent analyses (see supplementary information for a complete list of initial seed set compounds).

4.2. Application to selected reconstructions

Escherichia coli

The metabolism of *E. coli* is well investigated and the associated metabolic reconstruction is one of the best-curated models [26]. Accordingly, no peculiarities or gaps were detected by ASBIG.

Arabidopsis thaliana

ASBIG identified several inconsistencies in the *A. thaliana* model. The presence of cytosolic nicotinamide adenine dinucleotide phosphate (NADP) in the minimal seed set unraveled the lack of a NAD phosphorylation reaction in the cytosol. Similarly, plastidic NADP turned out to be autocatalytic because of its metabolic connection to plastidic hexadecanoic acid (palmitic acid), which was also found in the minimal seed set.

Plant hexadecanoic acid production involves the NADH-dependent reduction of *trans*-2-hexadecenoyl-ACP to hexadecanoyl-ACP. However, the insertion of plastidic NAD, ACP and CoA did not result in the biosynthesis of hexadecanoic acid. Instead, plastidic NADP that is required in its reduced form (NADPH) during the reduction of oxohexadecanoyl-ACP to hydroxypalmitoyl-ACP was present in the minimal seed set. Thus, plastidic CoA, ACP, NADH and NADPH are crucial factors to enable fatty acid biosynthesis in the *A. thaliana* model. However, the requirement of NADPH in addition to NADH was unexpected and had to be resolved through detailed manual network analysis.

The requirement of both plastidic and cytosolic NADPH suggested a gap in the phosphorylation process of NAD. The *A. thaliana* genome contains two genes coding for NAD kinases, which are annotated as NADK1 and NADK2 [179]. The two corresponding enzymes were found to be localized in the cytosol and plastid stoma, respectively [180]. The reactions catalyzed by theses enzymes (EC 2.7.1.23) were absent in the metabolic reconstruction of *A. thaliana*, and NADPH lost its autocatalytic property by adding the reaction EC 2.7.1.23 to the cytosol and to the plastidic stoma compartment. Consequently, plastidic and cytosolic NADPH was not detected in the minimal seed set anymore.

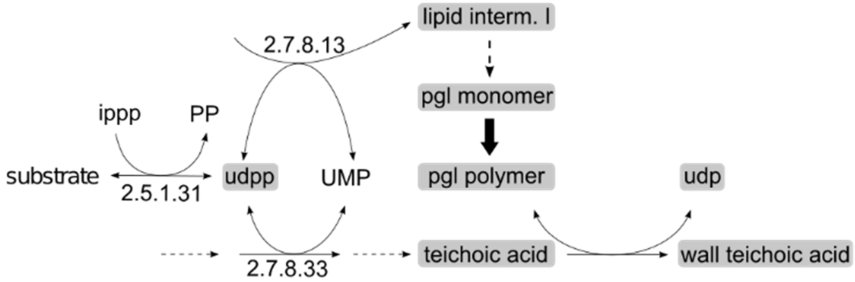


Figure 4. Schematic representation of parts of the PGL and the teichoic acid biosynthetic pathways in the *B.subtilis* model. The conflicting minimal seed substrate (italics) is geranylgeranyl PP in case of the model and, in contrast, *trans,trans*-farnesyl PP according to information from KEGG and BsubCyc. Further, addition of the reaction depicted by the bold arrow would omit the need for PGL as crucial compound. Key intermediates of the pathway are highlighted in gray; the remaining metabolites complete the reactions. Each reaction is labeled with the corresponding EC number (if available). Dashed arrows mark several reaction steps. Abbreviation: IPPP—isopentenyl diphosphate.

Bacillus subtilis

Applying ASBIG to *i*BSU1103, the metabolic network model of *B. subtilis*, revealed two pathways with deficient or no accessibility: cell wall biosynthesis and thiamine biosynthesis.

Although the relevant genes for thiamine biosynthesis in *B. subtilis* are known [181], the corresponding pathway was not present in *i*BSU1103. As a result, thiamine was an inevitable component of the minimal seed set. Adding the thiamine biosynthetic pathway resulted in a gain of functionality for the model and the elimination of thiamine from the minimal seed set.

In addition to thiamine, geranylgeranyl diphosphate (PP) and glycerol teichoic acid were part of the final minimal seed set. In the model, geranylgeranyl PP is the direct precursor of undecaprenyl PP (UDPP), which is an intermediate in the biosynthesis of the cell wall components peptidoglycan (PGL) and teichoic acid (Fig. 4, EC 2.7.8.13 and EC 2.7.8.33) [182]. Hence, geranylgeranyl PP and glycerol teichoic acid were required in the minimal seed set for the cell wall biosynthesis.

The corresponding pathways in the model were examined to resolve the need for the two compounds. The teichoic acid biosynthetic pathway was implemented in *i*BSU1103. However, teichoic acid synthesis remained locked in a scope analysis with the initial seed set owing to the unavailability of UDPP. As mentioned above, geranylgeranyl PP is assigned as the direct precursor of UDPP in the *i*BSU1103 model. In contrast, the database BsubCyc indicates *trans,trans*-farnesyl PP as the precursor of UDPP

(BsubCyc: di-*trans*,poly-*cis*-undecaprenyl phosphate biosynthesis). Furthermore, the gene *uppS* in *B. subtillis* encodes an UDPP synthase [183], which requires *trans,trans*-farnesyl PP to produce UDPP (EC 2.5.1.31, Fig. 4). Apparently, the reaction EC 2.5.1.31 of the model included an incorrect substrate assignment leading to the substrate incorporation into the minimal seed set, a conflicting minimal seed metabolite. The exchange of geranylgeranyl PP with *trans,trans*-farnesyl PP in the reaction EC 2.5.1.31 resolved the need for the conflicting minimal seed metabolite and geranylgeranyl PP was eliminated from the minimal seed set. The identification of conflicting minimal seed metabolites demonstrates the ability of ASBIG to detect wrongly implemented reactants.

Despite the modification of reaction EC 2.5.1.31, teichoic acid remained in the minimal seed set because of its ability to provide the PGL polymer to the metabolic scope (the backward reaction of EC. 3.6.3.40 removes teichoic acid from PGL, Fig. 4). To avoid this property of teichoic acid, a reaction, which transforms multiple PGL monomers to a polymer, was implemented in the model (Fig. 4, light gray dashed arrow). As a result, no polymers had to be provided as self-replicators anymore, and teichoic acid was eliminated from the final minimal seed set. To sum up, the application of ASBIG substantially improved the integrity of the *B. subtilis* model *i*BSU1103 leading to a more comprehensive representation of the *in vivo* metabolism. Furthermore, the investigation of *i*BSU1103 illustrates central characteristics of the method: computed add-on metabolites not necessarily indicate the corresponding pathway or specific location of an inconsistency, and the results generated by ASBIG can suggest the inclusion of further functionality for the model.

Chlamydomonas reinhardtii

The minimal seed set for iRC1080, the metabolic model of *C. reinhardtii*, contained add-on metabolites that indicated the lack of functionalities in some biosynthetic pathways. To access all components of the biomass reaction, thiamine, glutathione, chorismate, magnesium-protoporphyrin IX 13-monomethyl ester and plastidic plastoquinone expanded the initial seed set.

The biosynthesis for thiamine was not implemented in the model, likely because critical parts of the pathway still require experimental confirmation [184]. However, the basic metabolic routes of *de novo* thiamine biosynthesis in *C. reinhardtii* are known, and the corresponding reactions could be added to the model. After the implementation, thiamine lost its autocatalytic property.

The presence of glutathione in the minimal seed set could be linked to the biosynthetic pathway of cysteine. Assimilatory sulfate reduction in *C. reinhardtii* depends on glutathione and leads to the generation of hydrogen sulfide. Subsequently, hydrogen sulfide can be incorporated into *O*-acetyl-l-serine to form cysteine. Hence, it was assumed that hydrogen sulfide, glutathione or cysteine were required to be part of the minimal seed set. Three runs of ASBIG, each with one out of the three components in the initial seed set, confirmed this assumption (and resulted, for hydrogen sulfide and cysteine, in the replacement of glutathione in the minimal seed set).

It was further possible to identify an impaired biosynthesis of chlorophyllide a explaining the presence of protoporphyrin IX 13-monomethyl ester in the minimal seed set. *S*-adenosyl-l-methionine (SAM) was one of the substrates to produce the protoporphyrin ester. Hence, another run of ASBIG was performed with SAM included in the initial seed set. This led to the removal of protoporphyrin IX 13-monomethyl ester, implicating the relevance of SAM.

The essential character for both plastidic SAM and cytosolic chorismate could be explained with a missing transport reaction between plastid and cytosol or missing biosynthetic pathways in one of the compartments. A putative transport enzyme could overcome this lack in both cases, even though no such transporters have been identified in *C. reinhardtii* yet. However, an enzyme to import SAM into the chloroplast was reported for *A. thaliana* [185]. Therefore, the findings of ASBIG highlight the need for further investigations of the metabolic capabilities of *C. reinhardtii*.

The putative plastoquinone biosynthetic pathway for *C. reinhardtii* is available in the KEGG database (KEGG: ubiquinone and other terpenoid-quinone biosynthesis). As it was not integrated in the investigated *C. reinhardtii* metabolic model, plastoquinone possessed autocatalytic property. Despite the lack of experimental evidence, the pathway was added to the model. First, the network reconstruction had to be expanded by two metabolites previously not implemented: nonaprenyl PP and 2-methyl-6-solanyl-1,4-benzoquinol. Nonaprenyl PP is the product of the reaction of geranylgeranyl PP with five molecules of isopentenyl PP (catalyzed by EC 2.5.1.85). Subsequently, nonaprenyl PP reacts with homogentisate to form 2-methyl-6-solanyl-1,4-benzoquinol [186], which is further converted to plastoquinone in a SAM-dependent reaction. The enzyme involved in the plastoquinone formation is not yet characterized, however, it is assumed that *VTE3* codes for the corresponding enzyme [186]. By adding the aforementioned metabolic reactions to the *C. reinhardtii* model, plastoquinone lost its autocatalytic property and was eliminated from the minimal seed set.

The results for the examined network reconstruction of *C. reinhardtii*, iRC1080, show the versatile capabilities of ASBIG, for example, to identify inconsistencies that require additional experimental investigation as shown for the case of SAM or chorismate.

Saccharomyces cerevisiae

In the *S. cerevisiae* model iMM904, dolichol was identified as a crucial add-on metabolite of the minimal seed set, implying an impaired biosynthesis of the compound. A survey of the metabolic model confirmed the absence of the corresponding pathway, although the biosynthesis of dolichol in *S. cerevisiae* has been previously described [187]. To implement the metabolic route in the network reconstruction, several additions were required: (i) the condensation of farnesyl PP with 13 units of isopentenyl pyrophosphate to form polyprenyl PP (EC 2.5.1.87), (ii) the formation of polyprenol from polyprenyl PP and water (both polyprenyl PP and polyprenol had to be implemented as novel compounds in the model) and (iii) the conversion of polyprenol to dolichol. Manual expansion of the network reconstruction of *S. cerevisiae* provided access to the novel biosynthetic pathway of dolichol, thus eliminating the compound from the minimal seed set.

Another unexpected metabolite in the minimal seed set was NADP located in the endoplasmatic reticulum (ER). This compound remained in the minimal seed set because of the lack of information on a potential pathway providing NADP in the ER.

Even though dolichol is the only gap-indicating add-on metabolite, the identification of NADP as an autocatalytic compound suggests general inconsistencies of the model.

Zea mays

Application of ASBIG to the *Z. mays* model DTMaize_C4GEM_45632 revealed four add-on metabolites in the final minimal seed set that required further investigation: thiamine, NADP, undecaprenyl phosphate (UDP), and 4-methylthio-2-oxobutanoate (all localized in the cytosol).

Although maize is capable of thiamine biosynthesis, the pathway has not yet been fully characterized [188], and thus, is not entirely represented within the metabolic network model. As a result, the biosynthetic route of thiamine was blocked and thiamine had to be incorporated in the minimal seed set. With the future availability of novel information, the model can be improved by additionally including this specific biosynthetic pathway.

The absence of a cytosolic NAD kinase necessitated NADP to be an element of the minimal seed set. However, transcriptome analysis and automated annotation suggested the existence of a functional NAD kinase in *Z. mays* (UniProt ID: B6TAB2_MAIIZE) [189]. Therefore, the reaction (EC 2.7.1.23) was added leading to the elimination of NADP from the minimal seed set.

Subsequently, the crucial character of UDP was examined and a missing reaction to phosphorylate uridine 5′-phosphate (UMP) to UDP in UTP biosynthetic pathway was identified. As for the NAD kinase, transcriptome data and electronic annotation (UniProt ID: B6T904_MAIZE) [189] suggested the presence of a functional UMP kinase. Hence, the detected gap could be resolved by implementing the additional reaction, thus eliminating UDP from the minimal seed set.

The application of the more restrictive 'producibility constraint' during the seed set expansion (see section 2) allowed to link the autocatalytic metabolite 4-methylthio-2-oxobutanoate with the biosynthetic pathway of methionine. 4-Methylthio-2-oxobutanoate was part of the minimal seed set owing to its ability to bypass a blocked canonical methionine biosynthetic pathway (KEGG: map00270). It was assumed that a limited availability of cysteine, an essential intermediate in the methionine biosynthesis, caused the blocked methionine production (see also cysteine biosynthesis in *C. reinhardtii*). By adding cysteine to the initial seed set, folate instead of 4-methylthio-2-oxobutanoate was represented as an essential add-on metabolite of the minimal seed set. Evidently, the canonical methionine biosynthetic pathway was inaccessible unless cysteine and folate (or precursors/intermediates of the corresponding pathways) were part of the minimal seed set. A derivative of folate, the methyl group donor methyltetrahydropteroyltri-l-glutamate, is vital for the last step of the methionine biosynthesis [190]. However, the biosynthesis of folate and its derivatives was impaired in the model, likely because the *in vivo* pathway is not yet fully understood [191], leading to the inevitable role of folate. Comparable with thiamine, the implementation of the biosynthetic pathway in the model depends on additional experiments to elucidate folate biosynthesis in *Z. mays*.

Another inconsistency in the last step of the methionine biosynthesis was identified: According to literature, the triglutamate of methyltetrahydrofolate (MeTHF), methyltetrahydropteroyltri-l-glutamate, acts as the substrate for the methionine synthase (EC 2.1.1.14) [192,193]. In the *Z. mays* model, this enzyme used 'pure' MeTHF as methyl donor (Fig. 4), indicating a conflicting minimal seed metabolite. Additionally, another methionine-producing reaction (EC 2.4.2.7), which involved methyltetrahydropteroyltri-l-glutamate as substrate, was available. However, the latter reaction was misassigned, as the denoted EC number characterizes an adenine phosphoribosyltransferase (EC 2.4.2.7). Thus, methyltetrahydropteroyltri-l-glutamate was implemented as substrate for the methionine synthase (EC 2.1.1.14), and the redundant misassigned reaction (EC 2.4.2.7) was deleted. As the *Z. mays* model lacked the reactions necessary for the transfer of the single carbon to tetrahydropteroyltri-l-glutamate, the triglutamate of MeTHF replaced folate (precursor of MeTHF) in the final minimal seed set. Although the exact mechanism of the methyl group transfer is not completely characterized for plants, the basic pathway composed of two reactions is known [194]. Hence, the pathway was implemented in the model to connect methyltetrahydropteroyltri-l-glutamate with the folate metabolism. Subsequently, folate and cysteine were part of the minimal seed set again and enabled methionine biosynthesis.

Each of the four identified add-on metabolites indicates substantial restrictions in the *Z. mays* model DTMaize_C4GEM_45632. Even though not each issue could be resolved, the results of ASBIG provide valuable indications to the *Z. mays* model for improvements.

4.3. Application to automatically reconstructed networks

To further scrutinize the ability of ASBIG to improve the consistency of metabolic network reconstructions, we analyzed 193 genome-scale metabolic networks, which have been reconstructed automatically based on the organisms' genome sequences and genomic annotations [162]. Among all models, 234 different metabolites were identified as crucial elements in the final minimal seed set after adding them during scope analysis. This list of compounds suggested that reactions were possibly missing in the biosynthetic pathways linked to these metabolites (Supplementary Table S1). A median of 6 additional compounds per model, necessary to facilitate biomass production, was identified (Supplementary Fig. S2). In 69% of all models, PGL polymer (n–1 subunits) was identified as an inevitable compound. Peptidoglycan polymers are found in the membranes of bacteria and occur with a varying number of subunits [195]. The elongation and shortening of this polymer are included in the models, but not the *de novo* biosynthesis. Furthermore, in multiple models, which were analyzed by ASBIG, external spermidine (47%), the glutamate-accepting tRNA (26%), thiamine (24%), external alanylhistidine (22%) and glycyl-l-asparagine (21%) were frequently identified as compounds that could not be synthesized. The recurrence of these compounds might be owing to missing knowledge of possible biosynthetic routes or owing to properties of the reconstruction process (all 193 automated reconstructions analyzed here were generated using the same procedure) [162]. On the other hand, 194 of the 243 compounds were identified in <5% of all automated reconstructions analyzed, suggesting model-specific gaps or errors in the reconstruction or actual auxotrophies of the organisms.

Taken together, our analysis of a wide range of different automatically reconstructed metabolic networks reveals that several metabolites have an autocatalytic property or are inevitable for biomass production. The identification of these autocatalytic metabolites using ASBIG can improve the quality of the reconstruction by suggesting that the *de novo* biosynthetic pathways of the identified metabolites are either incomplete or missing.

4.4. Validation of ASBIG on auxotrophic mutants of E. coli

As mentioned above, no gaps were identified in the *E. coli* model iJO1366. Using this network, we knocked out reactions, recalculated the minimal seed set and compared the changes with reported experimental phenotypes of single-gene deletion mutants of *E. coli* K12 derivatives [196]. Using this approach, we were able to validate ASBIG in terms of the method's potential to predict gene essentiality as well as altered nutritional requirements of mutants compared with the wild type strain. We deleted 10 reactions, one at the time, each one in the biosynthetic pathways of a specific amino acid. The deleted reactions correspond to gene deletions, which are known to cause a specific auxotrophy for the focal amino acid of the mutant strain [196]. For each of the mutated networks, ASBIG identified an add-on metabolite, which was necessary to ensure biomass production. The add-on metabolite was the final product (the amino acid) of the affected pathway, an intermediate within the pathway or a derivative of the focal amino acid (for an overview see Supplementary Table S2). Hence, the output of ASBIG reflects the auxotrophy of the mutant strain impressively demonstrating the capability of ASBIG to detect gaps within primary metabolic pathways.

5. Conclusion

In conclusion, ASBIG is an efficient method for detecting inconsistencies in existing genome-scale metabolic network reconstructions. The method facilitates network validation and automated gap detection in primary metabolism contributing considerably to the quality improvement of metabolic models. ASBIG combines the conceptual approaches of autocatalytic metabolites and scope analysis, thereby allowing to test network reconstructions for their integrity. For each investigated metabolic model, a minimal seed set of metabolites, which provides access to all fundamental metabolic pathways, is computed. Putative flaws of the model can be deduced directly from identified minimal seed set.

The benefit of ASBIG was demonstrated by investigating different metabolic models. In the closely examined models, numerous kinds of errors could be identified including i) missing reactions, ii) missing pathways or iii) gaps caused by insufficient knowledge of metabolic processes. Furthermore, not only single gaps in individual models were identified, but also common flaws simultaneously present in several network reconstructions, could be detected in a large screen of 190 network reconstructions.

In contrast to the commonly applied gap-finding approach (i.e. deriving putative gaps from the calculated flux distributions), ASBIG is a simple method and detects other types of inconsistencies. Hence, the application of ASBIG contributes significantly to model refinement and validation representing a complementary approach to existing gap-finding methods.

Funding: This work was supported by the Jena School for Microbial Communication (JSMC) to S.W. and the Volkswagen Foundation to C.Ko.
Conflict of interest: none declared.

Chapter III

Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria

Published in Evolution, volume 68, issue 9, pages 2559-2570 in September 2014.

Authors

Glen D'Souza, Silvio Waschina, Samay Pande, Katrin Bohl, Christoph Kaleta, and Christian Kost

1. Abstract

Bacteria that have adapted to nutrient-rich, stable environments are typically characterized by reduced genomes. The loss of biosynthetic genes frequently renders these lineages auxotroph, hinging their survival on an environmental uptake of certain metabolites. The evolutionary forces that drive this genome degradation, however, remain elusive. Our analysis of 949 metabolic networks revealed auxotrophies are likely highly prevalent in both symbiotic and free-living bacteria. To unravel whether selective advantages can account for the rampant loss of anabolic genes, we systematically determined the fitness consequences that result from deleting conditionally essential biosynthetic genes from the genomes of *Escherichia coli* and *Acinetobacter baylyi* in the presence of the focal nutrient. Pairwise competition experiments with each of 20 mutants auxotrophic for different amino acids, vitamins, and nucleobases against the prototrophic wild type unveiled a pronounced, concentration-dependent growth advantage of around 13% for virtually all mutants tested. Individually deleting different genes from the same biosynthesis pathway entailed gene-specific fitness consequences and loss of the same biosynthetic genes from the genomes of *E. coli* and *A. baylyi* differentially affected the fitness of the resulting auxotrophic mutants. Taken together, our findings suggest adaptive benefits could drive the loss of conditionally essential biosynthetic genes.

2. Introduction

Although it has been known for a long time that factors such as deletions, duplications, and horizontal gene transfer can drastically shape the size and information content of bacterial genomes, one of the most surprising insights that resulted from sequencing multiple isolates of the same, seemingly identical species was the enormous plasticity that characterized all genomes analyzed so far [197–199]. For example, a comparison of 61 publicly available *Escherichia coli* and *Shigella* spp. genome sequences revealed that only 6% of the predicted gene families were represented in every genome (i.e., the "*core genome*"), whereas all others were present only in a subset of strains (i.e., the "*accessory*" or "*pan-genome*"; [79]). Interestingly, even the gene repertoire that constituted the core genome lacked genes that were otherwise deemed essential for the growth of *E. coli* [80–82]. These observations raise the question what major forces drive the loss of genes that essentially contribute to cellular fitness.

Genome reduction is a typifying feature of bacteria that occur in nutrient-rich or constant environments such as lactic acid bacteria [200], endosymbionts [56], or pathogens [54], respectively. Under these conditions, coding regions that provide little or no adaptive value in a given environment may be lost [54,201]. This so-called "*genome streamlining*" is thought to reduce the metabolic burden for basic cellular processes and could thus provide the resulting genotype with selective advantages over other genotypes that still bear these costs [58]. Adaptive benefits as a consequence of losing essential biosynthetic functions may arise when the corresponding metabolite is sufficiently present in the bacterial growth environment or provided by a co-occurring organism [83]. The latter scenario likely explains why amino acid biosynthesis pathways are sometimes partitioned between a eukaryotic host and its prokaryotic endosymbiont [202] or between multiple co-symbionts [56].

A second factor that could explain the loss of genetic information from bacterial genomes is genetic drift [49,203,204]. When bacteria transition from a free-living to a symbiotic lifestyle such as the bacterial endosymbionts of insects [54–56], repeated bottlenecks of relatively small populations may result in a weakened selection even for required genes, thus resulting in an elimination of dispensable genes [205]. Indeed, experimentally evolving *Salmonella enterica* by subjecting it to regular population bottlenecks resulted in a reduction of genome size and a concomitant loss of essential genes [57]. Similar mechanisms might act on obligate bacterial endosymbionts, thus explaining their typically extremely reduced genomes that retain few essential biosynthetic genes. However, it is generally difficult to infer from genomic analyses whether drift or selection was the main force to explain genome degradation. Hence, alternative approaches are necessary to determine the drivers of bacterial gene loss.

In vitro approaches are ideal for this purpose, because experiments can be purposefully designed and environmental conditions rigorously controlled. Long-term evolution experiments, in which different bacterial strains were serially propagated and thus allowed to adapt to the respective environments have shown that large genomic deletions are indeed prevalent under these conditions [206–208]. Moreover, fitness advantages accompanied some of these deletions, suggesting selection rather than drift drove the loss of these genes [114,207,208]. Interestingly, in a study with the bacterium *Methylobacterium extorquens* AM1 [208], the observed fitness advantage did not seem to result from a general shortening of the genome, but was rather due to the loss of specific genes. However, determining whether the deletion of a metabolic gene has a negative, neutral, or beneficial effect on the fitness of the resulting mutant is a nontrivial task. Problems that arise when naturally evolved mutants are being analyzed are first that very often multiple genes are lost simultaneously [207], thus making it difficult to link an observed change in fitness to the loss of a particular gene. Second, multiple auxotrophies often hamper the culturability and hence experimental amenability of a given strain (e.g., bacterial endosymbionts). Third, genetic interactions among different mutations that arose independently impede the determination of fitness consequences of a single mutation. These problems can be circumvented by analyzing genetically well-characterized natural or engineered mutants. Indeed, *Bacillus subtilis* [209] and *E. coli* [210] mutants impaired in tryptophan biosynthesis revealed significant fitness advantages in the presence of the amino acid relative to prototrophic cells. However, whether the loss of essential metabolic genes always results in selective advantages when the required metabolite is present in the environment as well as which causal mechanisms explain this observation remain obscure.

Here, we combine in silico analyses with systematic laboratory experiments of genetically engineered mutants to address the following questions: (1) How widespread is the loss of conditionally essential metabolic genes among bacteria in nature? (2) What fitness consequences result from the loss of a metabolic gene? (3) Does the fitness effect depend on (a) the gene analyzed, (b) the concentration of the corresponding metabolite in the growth environment, or (c) the position of the catalytic enzyme within a metabolic pathway?, and (4) Do different species differ in their fitness consequences upon gene loss?

3. Material and Methods

3.1. Prediction of auxotrophies and bioinformatics analysis

To predict putative auxotrophies in different bacterial species, the metabolic networks of 949 bacteria were examined for the presence of metabolic routes leading to the formation of amino acids, vitamins, or nucleobases. As a first step, all biosynthetic pathways known in bacteria to be involved in the formation of each of 20 amino acids, three vitamins, and two nucleobases (Table S1) were collected from the manually curated metabolic pathway database MetaCyc [211]. The pathways were consolidated (Fig. S1) to identify alternative biosynthetic routes and pathway dependencies (e.g., a pathway that provides the precursor metabolite). In a second step, the existence of the individual pathway reactions in 949 bacterial species was inferred using the

MicroScope genome annotation and analysis platform [212]. Briefly, the MicroScope platform is a collection of microbial metabolic networks, which consist of a subset of those reactions from the MetaCyc database, for which a genome segment (including plasmids) was identified or predicted, that is part of a gene for an enzyme that can catalyze the corresponding reaction. In a third step, an organism was predicted to be auxotrophic for a given metabolite if all possible metabolite-forming routes (Fig. S1) lacked more than 50% of the pathway's reactions as indicated by the absence of the corresponding annotated genes from the organism's genome sequence. This 50% cut-off was chosen, to increase robustness of the predictions against sequencing errors (i.e., missing annotations) and errors during the process of the metabolic network reconstruction. Genes annotated as pseudogenes were excluded from the analysis because pseudogenes are often a transitional stage of the gene from a functional gene toward complete gene loss [213]. Therefore, all reactions that depended on pseudogenes were classified as “not present.” The observed results thus represent a conservative estimate of the frequency of auxotrophies in bacteria with a sequenced genome.

All bacterial strains were categorized as “free-living,” “gut-inhabiting,” or “endosymbiotic” based on the genome meta-information stored in the Genomes OnLine Database [214].

The total mass of each protein in Mega Dalton (i.e., mass of the individual protein multiplied with the abundance of protein copies per cell) that was involved in the biosynthesis of Arg, His, and Trp was obtained from Wessely et al. (2011) [153].

3.2. Culture conditions

All cultures were incubated at 30°C under shaking conditions and experiments were performed in minimal medium for Azospirillium brasillense (MMAB) [215] without biotin and using fructose (5 g L^{-1}) instead of malate as carbon source. Growth kinetic studies were performed in 96-microwell plates (Nunc, Denmark) with a culture volume of 0.2 mL. Competitive fitness experiments were performed in 96-deepwell plates (Eppendorf, Germany) with a culture volume of 1 mL.

3.3. Construction of strains

Single-gene deletions in *E. coli* that would lead to auxotrophy for a single amino acid, nucleobase, or vitamin were identified using the KEGG pathway [216] and the Ecocyc database [217]. All deletions were transferred from existing strains [80] using P1 phage-mediated transduction [218] into *E. coli* BW 25113 [80]. To distinguish different strains in competition experiments, the arabinose utilization locus (Ara$^+$) of *E. coli* strain REL 607 [219] was introduced into all auxotrophs by P1 transduction. Potential genetic targets to construct auxotrophs for Arg, His, Leu, and Trp in *A. baylyi* were identified using the KEGG pathway database [216] and deletion mutants were constructed as described ([220]; see Supporting Methods for details). Conditional lethality of these mutations in MMAB medium was verified in previous studies [80,81,221] as well as by inoculating 10^5 colony-forming units (CFUs) mL^{-1} of these strains into 1 mL MMAB medium. After 24 h, their optical density (OD) was determined spectrophotometrically at 600 nm using a Tecan Infinite F200 Pro platereader (Tecan Austria GmBh, Austria) and the mutation was deemed conditionally essential when the auxotroph's growth did not exceed the OD_{600nm} of uninoculated minimal medium. In contrast, when the mutant was able to grow (i.e., exceed the OD_{600nm} of uninoculated minimal medium), the strain was excluded from further analysis and the next gene upstream the biosynthetic pathway was deleted until a mutant was found that satisfied the criterion of conditional essentiality. Gene deletions were in all cases confirmed by sequencing the corresponding genomic regions.

3.4. Growth kinetic and fitness assays

For all experiments, auxotrophs were precultured at 30°C in MMAB medium supplemented with 200 µM of the required nutrient. Growth kinetics of auxotrophic strains and a matching prototrophic wild type (WT) were recorded in MMAB medium supplemented with the focal nutrient at the respective concentration. The pH of the medium did not change significantly over the course of the experiments. The medium was inoculated with ~10^5 CFUs mL^{-1} of an overnight culture (i.e., 16 h). Growth kinetic experiments were performed in a Tecan Infinite Pro 200 plate reader (Tecan Austria GmBh). Growth was measured as absorbance at 600 nm (i.e., OD) every 8 min for 24 h with 3 min of shaking between measurements. The maximum population density (i.e., OD) reached was calculated using the Magellan 7.1 software (Tecan Austria GmBh). The relative maximum OD was calculated by dividing the OD of the auxotroph by the OD of the WT grown at the same metabolite concentration. Monoculture experiments of every auxotroph and its cognate WT control were replicated four times for each metabolite concentration tested.

For competitive fitness assays, ~10^5 CFUs mL^{-1} of either WT or auxotrophs were inoculated into 1 mL MMAB medium with the requisite nutrient concentration and cell numbers were determined at 0 and 24 h by plating. *Escherichia coli* auxotrophs were differentiated from WT using the arabinose utilization marker (Ara^+/Ara^-) as described [219] and *A. baylyi* strains were differentiated using an antibiotic marker (kanamycin). The ara marker was swapped between competitors. None of the two markers used incurred detectable fitness costs (paired-samples *t*-test: $P > 0.05$, $n = 8$). Competitive fitness of auxotrophs versus WT was determined by calculating the Malthusian parameter (M) of both genotypes: $M = (\ln (N_f/N_i)/24)$, where N_i is initial number of CFUs at 0 h and N_f is the final CFU count after 24 h [219]. Relative fitness was calculated as the ratio of Malthusian parameters. Coculture experiments were replicated eight times (i.e., comparison of [1] WT and deletion mutants within the same biosynthetic pathway, and [2] WT and auxotrophic *A. baylyi* mutants) or four times for each metabolite concentration tested (all others).

The two methods used to quantify bacterial productivity were quantitatively comparable, as indicated by a significant correlation between CFU plate counts and OD readings (Spearman rank correlation: $\rho = 0.76$, $P = 4.4 \times 10^{-26}$, $n = 128$).

3.5. Statistical analysis

Frequency distributions of auxotrophic bacteria with different lifestyles were compared with a Pearson's Chi-squared test with Yates' continuity correction and the distributions of the number of auxotrophies per organism with the Wilcoxon rank sum test with continuity correction. The Levene's test was used to assess homogeneity of variances and variances were assumed to be inhomogeneous when $P > 0.05$.

Statistical differences in the growth parameters (i.e., OD, relative fitness) of WT and auxotrophs were determined by independent sample *t*-tests (monoculture growth experiments) or paired-sample *t*-tests (coculture competition experiments). Brown–Forsythe tests followed by either Tamhane's T2 (nonhomogenous variances) or LSD (homogenous variances) post-hoc tests were used to infer statistical differences in the relative fitness of mutants lacking different genes of the same biosynthetic pathway. The false discovery rate (FDR) procedure of Benjamini et al. (2006) was applied to correct P values after multiple testing [222]. The concentrations required by auxotrophs to exceed WT fitness were compared with Mann–Whitney *U*-tests. Two-sample *t*-tests were used to detect fitness costs of genetic markers. The relationship between monoculture OD and plate counts as well as between protein mass invested and the relative position within the three biosynthetic pathways was investigated by applying Spearman's rank correlations. All statistical analyses were performed using the R software ([223], version 2.15.3) and the SPSS package (version 17.0, IBM, Rochester, MN).

4. Results

4.1. Loss of conditionally essential biosynthetic functions is common in bacteria

To determine how common the loss of conditionally essential biosynthetic functions is among natural bacterial isolates, we investigated the frequency with which auxotrophies occurred in each of 949 sequenced eubacterial genomes. The set of genomes analyzed covered a phylogenetically diverse spectrum of bacterial phyla (Fig. S2), yet was biased in its composition toward bacteria of biotechnological or medical relevance. Taking advantage of genome sequences, pathway information, and genome annotation, we focused our analysis on all 20 proteinogenic amino acids, two nucleosides, as well as three vitamins (Table S1). A majority of Eubacteria (i.e., 76%) were predicted to be auxotrophic for between one and 25 different metabolites that are needed for growth and metabolism (Fig. 1 A). The most commonly predicted compounds that could not be synthesized by the organisms analyzed were biotin (36%), phenylalanine (36%), and asparagine (37%; Fig. 1 C). In contrast, very few bacteria (i.e., 7%) were auxotrophic for proline and isoleucine. Notably, three-fourths of all strains predicted to be auxotrophic had lost more than 85% of the genes involved in the biosynthesis of tryptophan, histidine, leucine, pyrimidines, and purines (Fig. S3), which are the longest linear pathways analyzed (Fig. S1). This finding suggests that auxotrophic strains tend to lose entire pathways once a biosynthetic function has been lost.

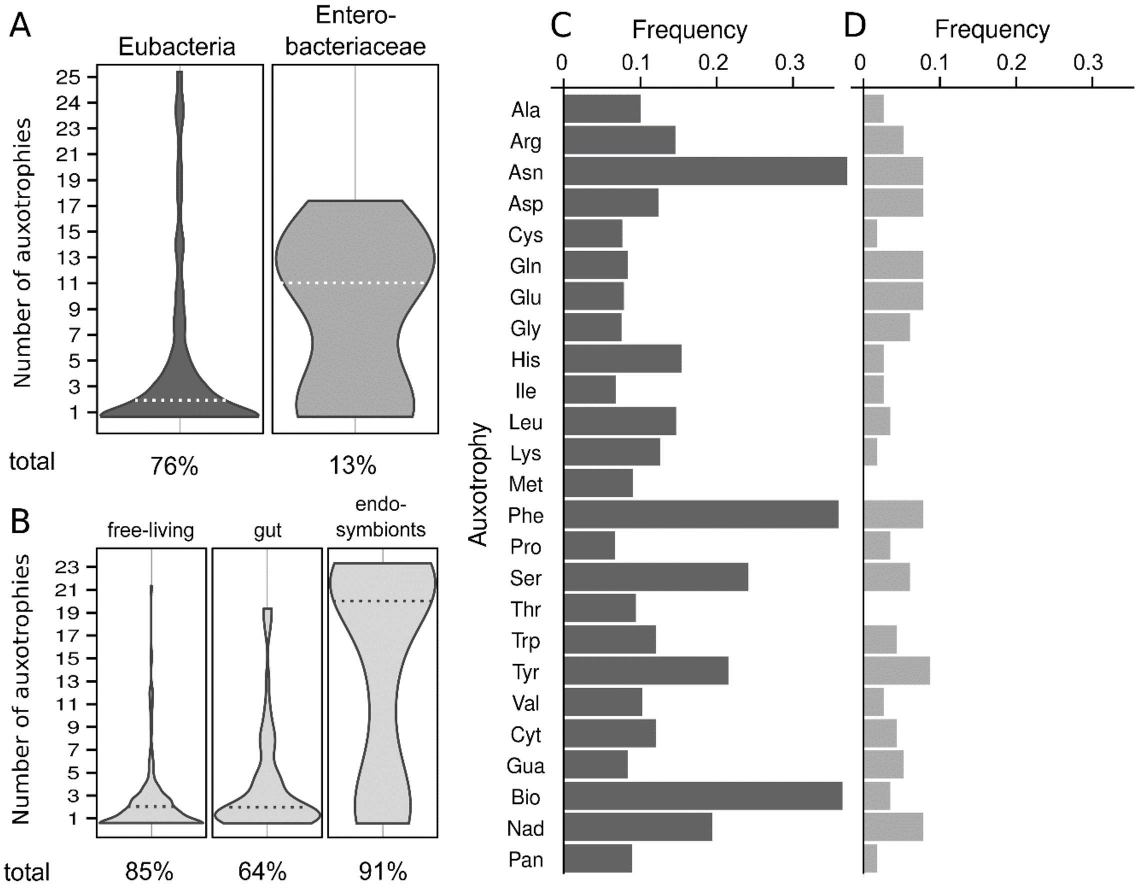


Figure 1. Distribution of metabolic auxotrophies in bacteria. Loss of a given biosynthetic function was predicted in silico using 949 eubacterial, genome-annotated taxa [211,212]. (A) Distribution and median (dashed line) of the number of predicted auxotrophies per auxotrophic organism for Eubacteria (dark gray, *n* = 949) and Enterobacteriaceae (light gray, *n* = 116) for all 25 metabolites analyzed. Percentages indicate the fractions of predicted auxotrophic organisms. Violin plots are scaled to the same maximum width. (B) Distribution and median (dashed line) of the number of auxotrophies for all auxotrophic Eubacteria depending on their lifestyle. Percentages indicate the fractions of predicted auxotrophic organisms within each lifestyle group. The lifestyle group sizes are as follows: free-living (*n* = 246), intestinal microflora-associated (*n* = 111), and endosymbiotic organisms (*n* = 57). Violin plots are scaled to the same maximum width. Frequencies of auxotrophies within (C) Eubacteria (*n* = 949), and (D) *Enterbacteriaceae* (*n* = 116). See Table S1 for abbreviations of metabolite names.

When putative auxotrophy frequencies were determined for the phylum Enterobacteriaceae (i.e., 116 organisms), 13% of all strains in this subset were predicted to be auxotrophic (Fig. 1 A). Here, the most commonly found auxotrophy was tyrosine (Fig. 1 D), which could not be synthesized by 9% of the tested Enterobacteriaceae. None of the enterobacterial genomes analyzed had lost the ability to produce methionine or threonine.

Mapping all detected auxotrophies onto the lifestyles of the 949 eubacterial species analyzed [214] indicated that 85% free-living, 64% gut-inhabiting, and 91% endosymbiotic bacteria were predicted to be auxotrophic for at least one metabolite (Fig. 1 B). Bacteria of the intestinal microflora were less frequently auxotrophic than free-living bacteria and endosymbionts (Chi-squared test with Yates correction: $\chi^2 = 19$, $P = 1.5 \times 10^{-5}$, $n = 111$ and 246 and $\chi^2 = 13$, $P = 3.3 \times 10^{-4}$, $n = 111$ and 57, respectively; Fig. 1 B). Furthermore, auxotrophic endosymbionts were predicted to be auxotrophic for 20 metabolites per organism (median), which is significantly more than was predicted for auxotrophic free-living and gut-inhabiting bacteria (both groups: median of two auxotrophies per organism; Mann–Whitney U-test with continuity correction: $W = 9541$, $P < 2.2 \times 10^{-16}$, $n = 52$ and 209 and $W = 3033.5$, $P = 8.8 \times 10^{-10}$, $n = 52$ and 71, respectively; Fig. 1 B). The phylogenetic distribution of lifestyles among the 949 analyzed organisms strikingly matched the phylogenetic distribution of all known bacteria with a completely sequenced genome (Fig. S4).

Taken together, our in silico analysis of eubacterial genomes predicted a surprisingly pervasive loss of multiple conditionally essential metabolic functions including the biosynthesis of amino acids, nucleosides, and vitamins. Furthermore, the distribution and frequency of auxotrophies was strongly dependent on the lifestyle of the bacterial species analyzed.

4.2. Auxotrophy-causing mutations are beneficial when the focal metabolite is present in the environment

The growth of *E. coli* WT in monocultures was significantly enhanced by the supplementation of five compounds (i.e., His, Met, Phe, Trp, and Nad; FDR-corrected independent sample t-tests: $P \leq 0.05$, $n = 4$; Fig. S5A, C) although different concentrations of each metabolite were required to achieve this effect. In contrast, growth was unaffected by the addition of eight metabolites (i.e., Ile, Leu, Lys, Pro, Tyr Cyt, Bio, and Pan; FDR-corrected independent sample t-tests: $P > 0.05$, $n = 4$; Fig. S5) and even inhibited by three of the 16 metabolites tested (i.e., Arg, Thr, and Gua; FDR-corrected independent sample t-tests: $P \leq 0.05$, $n = 4$; Fig. S5A, B).

When each of these metabolites was supplied in increasing concentrations to the corresponding auxotrophs, growth was strongly dependent on the concentration of the respective nutrient (Fig. S6). Half of all gene deletions tested resulted in a maximum population density (i.e., OD) that was significantly increased over WT levels (FDR-corrected independent sample t-tests: $P \leq 0.05$, $n = 4$; Figs. 2 A, S6) at some concentration of the focal metabolite. Exceptions were the auxotrophs for His, Lys, Phe, Tyr, Cyt, and Nad, whose maximum population density did not exceed WT levels (FDR-corrected independent sample *t*-tests: $P > 0.05$, $n = 4$, Figs. 2 A, S6) as well as the Pro and Thr auxotrophs that did not even reach WT levels under the range of concentrations tested (FDR-corrected independent sample *t*-tests: $P < 0.05$, $n = 4$; Figs. 2 A, S6A). The growth advantage over WT of the Arg and Gua auxotrophs was probably attributable to a significant inhibitory effect of the metabolites added on the growth of the WT (FDR-corrected independent sample *t*-tests: $P \leq 0.05$, $n = 4$; Figs. S5A, B and S6A, B), rather than an enhanced growth of the auxotrophic strains. Notably, growth of vitamin auxotrophs exceeded WT levels at much lower concentrations (0.2–0.5 µM) than was the case for nucleobase and amino acid auxotrophs (25–200 µM; Mann–Whitney *U*-test: $P = 1.04 \times 10^{-9}$, $n = 12$ and 28; Figs. 2 A, S6).

To verify whether the observed fitness advantages also manifest when an auxotrophic mutant directly competes against its prototrophic ancestor, pairwise competition experiments were performed, in which each of 16 auxotrophs were directly competed against the prototrophic WT in environments that contained different concentrations of the focal metabolites. Under these conditions, all auxotrophs except the Cyt and Nad auxotrophs reached fitness values that significantly exceeded WT levels (FDR-corrected paired-sample *t*-tests: $P \leq 0.05$, $n = 4$; Figs. 2 B, S6). This included also auxotrophs, whose fitness did not increase over WT levels in monocultures (i.e., auxotrophs for His, Lys, Phe, Pro, Thr, Tyr). Similar to monocultures, vitamin auxotrophs achieved their maximum relative fitness at much lower concentrations (0.05–0.5 µM) than amino acid- and nucleobase-deficient strains (25–200 µM) required to exceed WT growth (Mann–Whitney *U*-test: $P = 3.58 \times 10^{-10}$, $n = 16$ and 52; Figs. 2 B, S6). Biotin, one of the compounds for which biosynthesis genes were most frequently lost in natural bacterial isolates (Fig. 1) was required in the lowest concentrations of all metabolites analyzed in both mono- and coculture experiments (Figs. 2 A, B, and S6).

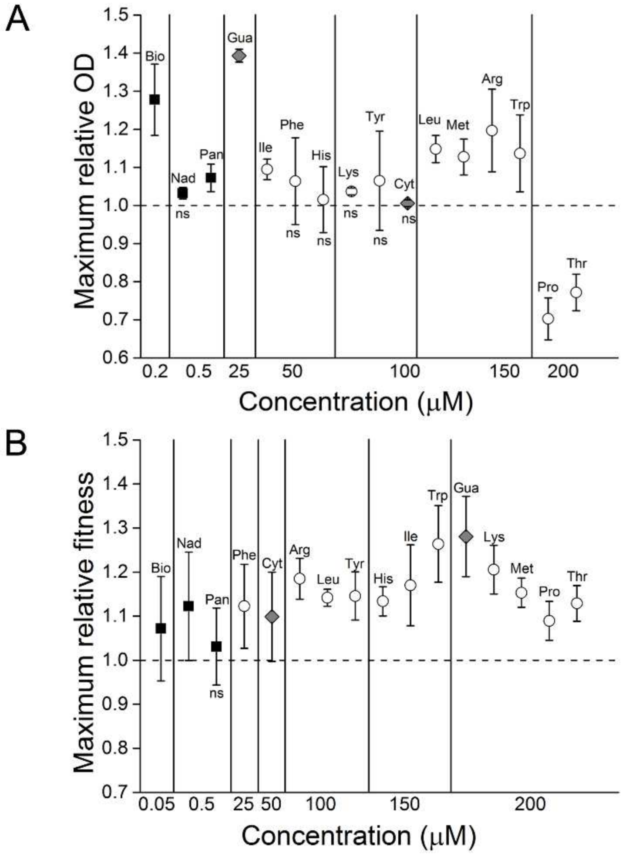


Figure 2. Maximum productivity and competitive fitness of *Escherichia coli* auxotrophs relative to WT. (A) Maximum OD in monoculture and (B) maximum fitness in coculture of the amino acid (circles), vitamin (squares), and nucleobase auxotrophs (diamonds) relative to WT. All values are medians (±95% CI) of four replicates and are significantly different from WT levels (i.e., dashed line; FDR-corrected independent sample *t*-tests (monoculture) and paired-sample *t*-tests (coculture): $P \leq 0.05$, $n = 4$), except those marked by "ns." See Table S1 for abbreviations of metabolite names.

Taken together, these results indicate that the loss of essential biosynthetic genes from the genome of *E. coli* generally resulted in strong and significant fitness advantages over the prototrophic WT when the required compounds were sufficiently present in the environment. The extent of fitness advantage, however, was context-dependent and strongly affected by (1) the concentration of the metabolite in the environment, (2) the identity of the metabolite, and (3) the absence/presence of a competitor.

4.3. Fitness benefits depend on which gene of a biosynthetic pathway is lost

Amino acid biosynthesis involves the action of multiple enzymes that are encoded by different genes. Thus, the fitness benefit a strain gains by not having to carry out a certain biosynthetic step may differ depending on which gene has been lost. Observing different fitness benefits when different genes of the same pathway are lost may reflect differences in the biosynthetic costs incurred at each step or regulatory interactions among genes.

To verify this possibility, several genes were individually deleted from the biosynthetic pathways for Arg (six genes), His (four genes), and Trp (four genes; Fig. 3 and Table S2) whose deletion renders the resulting mutant auxotroph for the corresponding amino acid. All generated mutants were individually competed against WT in three environments, which differed in the concentration of the focal amino acid. The range of these concentrations covered a span (±50 µM) around the concentration at which the terminal deletion mutant of each pathway had reached maximum fitness relative to cocultured WT (Figs. 2 B, S6).

Fitness consequences resulting from the loss of a conditionally essential gene from one of the three multistep pathways analyzed strongly depended on both the identity of the lost gene as well as the concentration of amino acids available (Fig. 3). A pattern that seemed to emerge was that as the amino acid concentration in the environment increased, deletion of terminal genes tended to be more advantageous than the loss of more anterior genes (FDR-corrected paired-sample *t*-tests and Brown–Forsythe tests followed by an LSD or Tamhane's T2 post-hoc test: $P < 0.05$, $n = 8$; Fig. 3). This trend was evident in two out of the three amino acid concentrations assayed for each of the three pathways analyzed (Fig. 3). Furthermore, one of the three amino acid concentrations tested for each biosynthetic pathway caused a significant positive correlation between the mutants' relative fitness and the position of the deleted gene within the pathway (Pearson product-moment correlation: Arg pathway, 100 µM: $r = 0.33$, $P = 0.012$, $n = 48$; His pathway, 200 µM: $r = 0.5$, $P = 0.05$, $n = 28$; Trp pathway, 200 µM: $r = 0.55$, $P < 0.001$, $n = 32$). Interestingly, when the amount of protein invested by *E. coli* to catalyze different steps of these biosynthetic pathways was taken into account, the protein investment also increased toward the end of these pathways (Spearman's rank correlation: $\rho = 0.55$, $P = 0.02$, $n = 17$; Fig. S7). Calculating the energetic cost for the individual coding sequences of these three pathways as well as the corresponding protein machinery in *E. coli* (Supporting Information Methods) revealed a significant greater protein cost (Wilcoxon signed rank test: $P = 0.002$, $n = 10$) that exceeded DNA biosynthesis costs by factor 34 (Table S3). Hence, these results suggest that a saving of protein costs may be involved in explaining the observed gain in fitness.

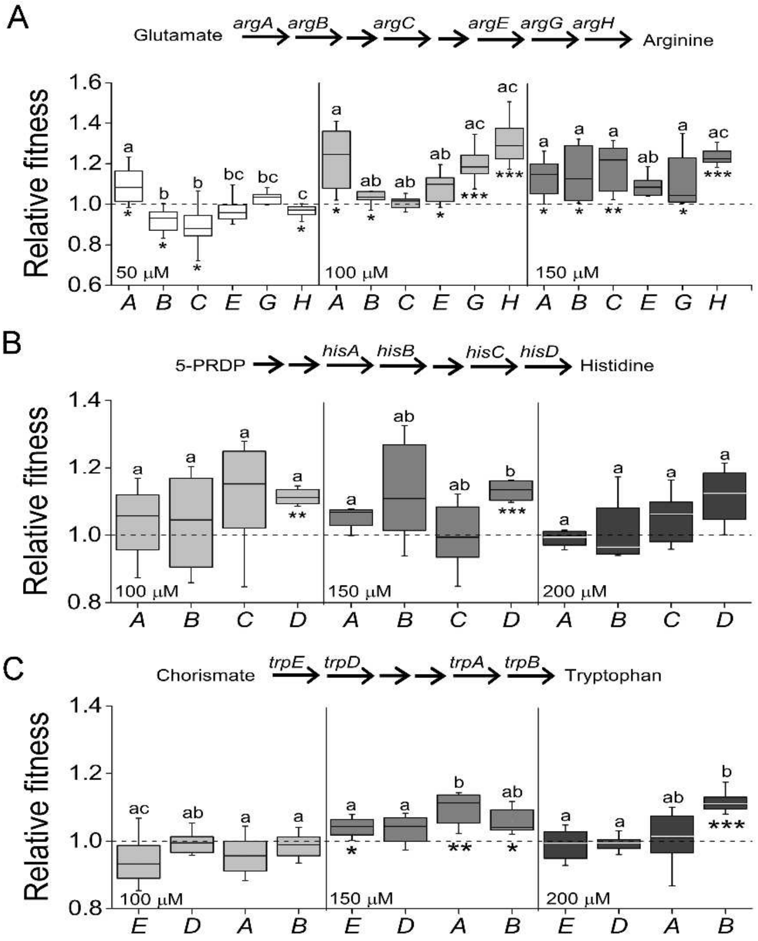


Figure 3. Competitive fitness of auxotrophic *Escherichia coli* mutants that lack different genes of the same biosynthetic pathway. Fitness of different deletion mutants that are auxotrophic for (A) Arg, (B) His, and (C) Trp was determined relative to WT. Experiments were conducted in minimal medium to which 50, 100, and 150 µM of Arg (A) or 100, 150, and 200 µM of either His (B) and Trp (C) has been supplemented. *X*-axes are labeled with the last letter of the focal gene's name (e.g., A for *argA*, *hisA*, *or trpA*). Asterisks denote significant differences from WT levels (i.e., dashed line; FDR-corrected paired-sample *t*-tests: $^{*}P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$). Different letters indicate significant differences among deletion mutants (univariate ANOVA followed by an LSD or Tamhane's T2 post-hoc test: $P < 0.05$; $n = 8$). Boxplot: median (horizontal lines in boxes), interquartile range (boxes, 1.5×-interquartile range (whiskers). Pathway insert: The flow of biosynthetic steps in each pathway. Unlabeled arrows represent nonessential genes. 5-PRDP: 5-phospho-α-d-ribose 1-diphosphate.

In case of the Arg biosynthetic pathway, the *argA* deletion mutant displayed a particularly strong fitness increase over WT in two of the three Arg concentrations tested (FDR-corrected paired-sample *t*-test: $P < 0.05$, $n = 8$; Fig. 3 A). Interestingly, the gene product of *argA* (i.e., *N*-acetylglutamate synthase) catalyzes the first step in the Arg biosynthesis pathway and is the target enzyme for feedback inhibition by arginine

[224]. However, deletion of *trpE* and *trpD* (i.e., anthranilate synthase), which fulfill the same function in the Trp biosynthesis pathway [225], did not result in a similar effect (Fig. 3 C). Thus, the particularly strong fitness advantage gained by *argA* deletion mutants in the presence of sufficient amounts of Arg points to a special regulatory role this gene plays within the Arg biosynthesis pathway.

Together, these results demonstrate significant gene-specific fitness effects that arise upon deletion of different genes of the same metabolic pathway and suggest the saving of protein costs may be involved in explaining these differences.

4.4. Also Acinetobacter baylyi *auxotrophs gain a fitness advantage upon gene loss*

All *A. baylyi* auxotrophs (i.e., *ΔhisD*, *ΔleuB*, and *ΔtrpB*) except the *ΔargH* mutant gained a significant fitness advantage upon gene loss when the corresponding amino acid was present (FDR-corrected paired-sample *t*-tests: $P \leq 0.05$, $n = 8$; Fig. 4). As previously observed in *E. coli*, the fitness advantage gained by *A. baylyi* auxotrophs was strongly dependent on the concentration of the focal metabolite, yet followed a completely different, downright opposite pattern (Fig. 4). Interestingly, only one of the four *A. baylyi* auxotrophs tested (i.e., *ΔhisD*) gained an advantage in relative fitness that was significantly increased over the fitness levels that the corresponding *E. coli* auxotrophs achieved under the same conditions (FDR-corrected independent sample *t*-tests: $P \leq 0.05$, $n \geq 4$). In sum, these results corroborate that the loss of essential biosynthetic genes can be selected for when the required metabolite is present in the environment, yet point to significant, species-specific differences.

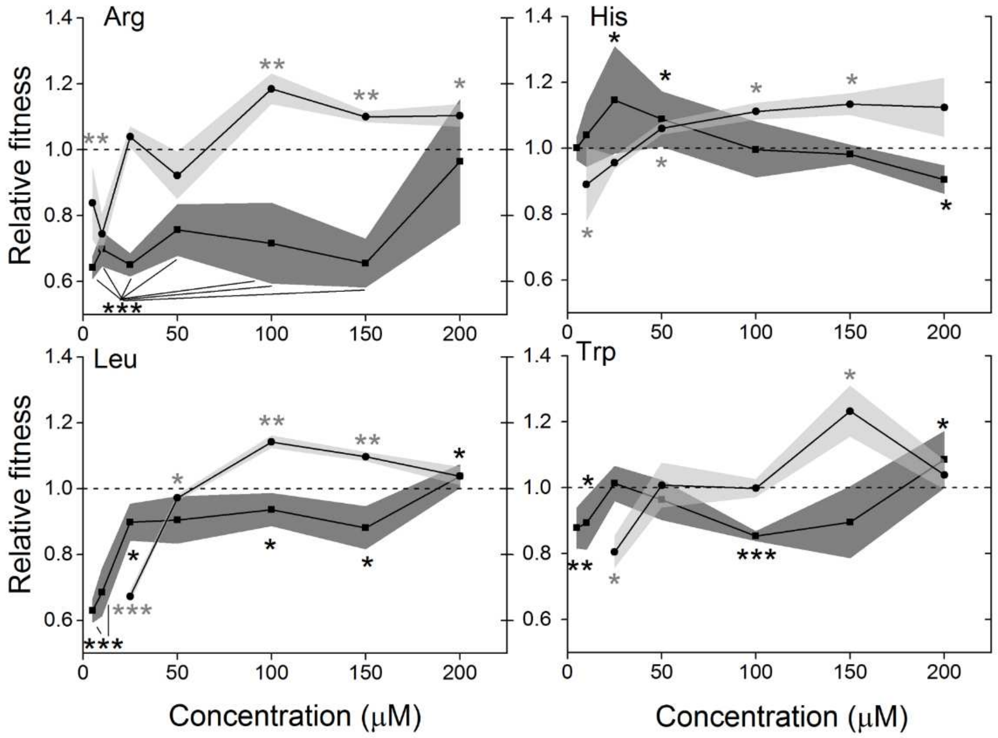


Figure 4. Competitive fitness of *Acinetobacter baylyi* and *Escherichia coli* auxotrophs relative to WT in increasing concentrations of the focal amino acids. Fitness of *E. coli* (circles) and *A. baylyi* (squares) mutants auxotrophic for Arg, His, Leu, and Trp relative to the corresponding WT. All values are medians of four replicates for *E. coli* and eight replicates for *A. baylyi*. The gray and dark gray regions mark the 95% confidence intervals for *E. coli* and *A. baylyi*, respectively, and the gray and dark gray asterisks mark significant differences of the *E. coli* and *A. baylyi* auxotrophs to WT levels (i.e., dashed line; FDR-corrected paired-sample *t*-tests: *$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$, $n \geq 4$).

5. Discussion

Our analysis revealed that the loss of conditionally essential genes, which likely results in metabolic auxotrophies, is not limited to bacterial endosymbionts, but equally prevalent among free-living bacteria. However, why do microorganisms loose genes at the expense of their metabolic autonomy? For endosymbiotic bacteria, this question is commonly answered by pointing to their small population sizes and a lack of genetic recombination. These factors should result in a relaxed selection even for essential genes and—combined with a strong effect of genetic drift—could explain the rapid erosion of symbiont genomes [49,54,55,202–204,226]. However, free-living and gut-dwelling bacteria drastically differ from bacteria with an intracellular lifestyle in terms of their population biology as well as the selective environment they experience. Also, the high degree of metabolic complementarity and mutual interdependency that has been frequently observed among co-occurring endosymbionts [56,205,227] is likely favored and maintained by natural selection.

To experimentally determine the potential role of selection in favoring mutants that lack essential genes, different biosynthetic genes were individually deleted from the genomes of two bacterial species and the resulting auxotrophic mutants systematically analyzed. This analysis revealed that (1) the loss of essential biosynthetic genes was generally beneficial when the required metabolite was sufficiently present in the cells' growth environment, (2) the metabolite concentration an auxotroph required to attain WT growth levels differed significantly depending on the metabolite as well as the species analyzed, (3) the loss of different genes from the same metabolic pathway resulted in differential fitness consequences for the corresponding mutants, and (4) auxotrophs of two species that lacked the same biosynthetic gene responded very differently when exposed to the same concentrations of the required amino acid.

5.1. What causes the unexpectedly strong fitness advantage?

A key finding of this study is that the loss of different biosynthetic genes gave rise to different fitness benefits when the focal metabolite was sufficiently present in the mutants' growth environment. This was not only true for genes of different metabolic pathways, but also when genes of the same biosynthetic pathway were considered. A number of relevant insights emerge from this analysis. First, it made a significant difference whether a mutant's phenotype was indirectly compared to WT (monoculture) or directly competed against WT (coculture). Here, both (1) the minimally required metabolite concentration as well as (2) the maximally achieved advantage over WT differed between the two perspectives. These findings cannot be exclusively explained by the costs auxotrophs save for the production of the focal metabolite relative to WT. Instead, other factors such as the cells' requirement for a given metabolite and/or the auxotrophs' transport efficiency with which they can take up different metabolites may have caused this pattern. Second, the finding that the deletion of different genes from the same biosynthetic pathway engendered different fitness consequences for the resulting auxotrophic mutants, suggests the unexpectedly strong fitness advantage of auxotrophs is at least partially caused by effects emanating from the loss of individual genes rather than a systemic response. The seeming increase of the fitness advantage auxotrophic mutants gained when terminal genes of a given pathway were deleted together with the concurrent enlarged investment of protein mass toward the end of these pathways implies the saving of protein costs may contribute to the observed gain in fitness. This interpretation is in line with empirical evidence, which suggests protein costs can significantly limit bacterial growth [228–230] or cause redistributions of metabolic fluxes to less expensive pathways [231]. However, Dykhuizen (1978), who addressed this question previously in *E. coli* did not find evidence for a cost-saving of Trp auxotrophs relative to prototrophic revertants [210]. Another possibility is a metabolic or regulatory rewiring that renders auxotrophs more efficient in coping with amino acid-deficient conditions [232]. This could be achieved by an enhanced uptake of amino acids or a reallocation of the cell-internal protein pool. Future work should scrutinize these hypotheses.

5.2. Distribution of metabolic auxotrophies in nature

Our in silico analysis provides a first systematic assessment of the prevalence of putative metabolic auxotrophies among eubacteria. Even our conservative estimation indicated that the vast majority of genomes analyzed lacked conditionally essential biosynthetic genes. A recent study corroborates these findings: reconstructing metabolic models of 55 sequenced *E. coli* and *Shigella* strains revealed multiple auxotrophies for vitamins and amino acids in 12 of these strains [233]. Taken together, these analyzes suggest metabolic auxotrophies may be more widespread than previously thought.

However, can the distribution pattern of auxotrophies predicted for Enterobacteriaceae (Fig. 1 D) be explained with the different fitness advantages observed in this study (Figs. 2, S6)? A series of statistical tests in which different kinetic parameters determined in this study (i.e., maximum OD/relative fitness reached after 24 h, slope of metabolite dependency curve [i.e., metabolite concentrations vs. OD/relative fitness after 24 h]) was correlated to the predicted enterobacterial auxotrophy frequencies did not detect significant relationships between these parameters (Spearman rank correlation: $P > 0.05$ in all cases). This result is likely caused by fact that the frequency with which certain biosynthetic genes are lost is due to the availability of the corresponding metabolites in the strains' natural environments and not the potentially gained fitness advantage. Moreover, because strains are likely auxotrophic for more than one metabolite (Fig. 1 A), epistatic interactions among these mutations may affect the fitness consequences of individual mutations.

5.3. Adaptive gene loss and the formation of interorganismal networks

Our results imply that whenever local metabolite concentrations exceed certain threshold levels, strong selection pressures build up that favor the loss of the corresponding biosynthetic functions in bacteria. Thus, our analysis provides a plausible adaptive explanation for the widespread loss of conditionally essential biosynthetic genes (Fig. 1). Accordingly, amino acid concentrations in certain bacteria-inhabited environments such as soil or insect guts generally exceeded the levels required for auxotrophic mutants to outcompete prototrophic cells by orders of magnitude (Fig. S8). In contrast, freshwater lakes exhibited only meager amounts of free amino acids (i.e., 2.6–4124 nM; [234]), which may explain why prototrophic *E. coli* strains seem to dominate in these environments [235]. Finally, metabolic auxotrophies have been observed to readily emerge in laboratory evolution experiments with, for example, *Pseudomonas aeruginosa* [236], mutator strains of *E. coli* [237] that adapted to the mouse gut or *Legionella pneumophila* parasites adapting to mouse macrophages [238].

As these examples illustrate, metabolites can either originate from the growth environment or be produced by another organism [83,239]. The "compensated gene loss" resulting from the latter [239] can account for the rapid reduction of genome size of both parasitic [240] and mutualistic bacterial symbionts and is likely also driving the formation of tightly integrated metabolic networks of co-occurring bacterial endosymbionts [56,241]. Our finding that this phenomenon is not restricted to organisms that interact over long periods of time, but also occurs among seemingly independent and free-living bacteria (Fig. 1) implies a pervasive role of adaptive gene loss for driving the evolution within microbial communities. An unavoidable leakage of vital metabolites during bacterial growth and subsistence [83] combined with the enormous and prevalent fitness advantages gained upon gene loss as observed in this and other studies [76,209,210], should result in the formation of intricately connected, intercellular networks. By mutually exchanging metabolites as "public goods," while at the same time specializing in the production of a reduced subset of metabolites, both the individual genotype and the whole bacterial community might benefit [76,242]. In particular, the difference in the concentration-dependent fitness advantage observed between two bacterial species (Fig. 4) may facilitate interspecific cross-feeding interactions. The general difficulty to isolate bacterial species from the wild [243,244] may be a reflection of this pattern.

6. Conclusions and outlook

Our study provides strong empirical support for the hypothesis that adaptive fitness advantages can account for the frequently observed loss of biosynthetic functions in bacteria. Our findings have a number of significant ramifications that should be investigated in future studies. First, the molecular causes underlying the unexpectedly strong fitness advantage upon gene loss should be identified. Second, as evidenced in our study, the natural bacterial isolates analyzed were rarely auxotrophic for just one metabolite, but commonly lacked multiple biosynthetic capabilities simultaneously (Fig. 1 A). Hence, future studies should address the question whether fitness effects combine additively when multiple auxotrophies are combined in one genetic background, or whether epistatic interactions limit an even further increase of the auxotrophs' fitness. Third, the "black queen hypothesis" [83] predicts for bacterial strains, which loose costly metabolites by leakage and that coevolve within a microbial community, to continuously loose biosynthetic genes until an equilibrium is reached, at which the benefit of gene loss is outweighed by its costs. Our study provides a first estimate of these benefits, thus allowing to further explore how they affect the race for biosynthetic disarmament within microbial communities.

Acknowledgments

We thank U. Gerischer and I. Steiner for sharing the *A. baylyi* ADP1 wild type strain as well as D. Refardt for providing us with strain *E. coli* REL 607. Constructive criticism by E. Kothe and the EEE group as well as support by W. Boland is gratefully acknowledged. This work was supported by the Jena School for Microbial Communication (JSMC), the Volkswagen Foundation, and the Max Planck Society.

Data archiving

The doi for our data is 10.5061/dryad.b7sp7.

Chapter IV

Plasticity and epistasis strongly affect bacterial fitness after losing multiple metabolic genes

Published in Evolution, volume 69, issue 5, pages 1244-1254 in May 2015.

Authors
Glen D'Souza*, Silvio Waschina*, Christoph Kaleta, and Christian Kost

* These authors contributed equally to this work.

1. Abstract

Many bacterial lineages lack seemingly essential metabolic genes. Previous work suggested selective benefits could drive the loss of biosynthetic functions from bacterial genomes when the corresponding metabolites are sufficiently available in the environment. However, the factors that govern this "*genome streamlining*" remain poorly understood. Here we determine the effect of plasticity and epistasis on the fitness of *Escherichia coli* genotypes from whose genome biosynthetic genes for one, two, or three different amino acids have been deleted. Competitive fitness experiments between auxotrophic mutants and prototrophic wild-type cells in one of two carbon environments revealed that plasticity and epistasis strongly affected the mutants' fitness individually and interactively. Positive and negative epistatic interactions were prevalent, yet on average cancelled each other out. Moreover, epistasis correlated negatively with the expected effects of combined auxotrophy-causing mutations, thus producing a pattern of diminishing returns. Moreover, computationally analyzing 1,432 eubacterial metabolic networks revealed that most pairs of auxotrophies co-occurred significantly more often than expected by chance, suggesting epistatic interactions and/or environmental factors favored these combinations. Our results demonstrate that both the genetic background and environmental conditions determine the adaptive value of a loss-of-biochemical-function mutation and that fitness gains decelerate, as more biochemical functions are lost.

2. Introduction

Bacterial genomes are not static entities, but are highly dynamic on evolutionary time scales in terms of both size and composition [48]. Variation in the size of prokaryotic genomes can be caused by the duplication of existing genes, the acquisition of new genetic information from the environment (i.e. *horizontal gene transfer*), or, alternatively, by gene loss. Reductive genome evolution is a feature that characterizes many bacterial taxa, and comparative genomics indicates that gene loss appears to be more important for shaping prokaryotic genomes than gene gain by horizontal gene transfer [48]. In many cases, one or more essential biosynthetic genes are lost, thus rendering the resulting auxotrophic bacteria dependent on an environmental uptake of the required metabolites [233,245,246]. Surprisingly, the loss of essential biosynthetic functions is not limited to endosymbiotic bacteria or intracellular parasites where essential nutrients can potentially be obtained from the host, but also prevails in free-living taxa such as saprophytes [247,248] or marine bacteria [249,250].

Two main scenarios can account for the frequently observed loss of conditionally essential biosynthetic genes from prokaryotic genomes: First, genetic drift may weaken selection even for essential genes and could thus explain the fixation of maladaptive mutations. This effect is likely strongest in small bacterial populations [57,203,204] such as endosymbiotic bacteria that repeatedly undergo severe population bottlenecks during host-to-host transmission [205]. Second, the loss of biosynthesis genes may be selectively favored when the required metabolite is either sufficiently present in the growth environment or provided by co-occurring organisms [83,250]. Under these conditions, mutations that deactivate the biosynthetic machinery for a certain metabolite may result in the saving of production costs or could induce regulatory changes to economize the cell's resources, for example by rerouting metabolic fluxes, which allow the bacterial cell to better cope with starvation for the required metabolite.

Several studies using different bacterial species support the hypothesis that adaptive benefits may drive the loss of essential biosynthetic functions. In these cases, pairwise competition experiments between prototrophic bacterial cells and mutants lacking the ability to biosynthesize a certain metabolite pointed to a significant fitness advantage auxotrophs gain over prototrophic genotypes when the required metabolite is sufficiently present in the cells' growth environment [209,210,245,245]. Even though these studies suggest that metabolic loss-of-function mutants can be selectively favored, very little is known on how metabolic auxotrophies evolve.

Given that theoretical evidence predicts multiple auxotrophy-causing mutations are frequently co-occurring in the same genetic background [245], the extent to which these mutations interact with each other (i.e., *epistasis*) remains poorly understood. In other words, do the previously observed positive fitness effects combine additively as more loss-of-function mutations accumulate in the same genome, or do epistatic effects constrain the fitness achievable by a multiply auxotrophic genotype? Moreover, natural habitats of bacteria are usually quite complex and may not only contain several primary metabolites (e.g. amino acids or vitamins), but also differ in the available carbon source. Because fluxes through metabolic networks change depending on the carbon source used [88,251], fitness and ultimately also epistatic interactions among mutations are expected to depend on the carbon source metabolized (i.e. *plasticity*).

An increasing number of studies suggest both epistasis [114,252,253] and plasticity [93,254] can significantly influence the trajectories of beneficial mutations accessible to evolving bacteria. The general pattern that seems to emerge from these analyses is that negative epistasis is prevalent and often results in diminishing returns as more beneficial mutations accrue in a single genetic background [93,114,253].

Understanding the constraints that determine evolutionary routes leading to multiply auxotrophic bacteria requires insight into how plasticity and epistasis influence the fitness consequences upon loss of metabolic genes. However, examining these effects in natural isolates is hampered by difficulties of cultivating auxotrophic genotypes under laboratory conditions or to manipulate the genome of nonmodel organisms. Thus, deleting a defined number of genes from the genome of a well-characterized model organism and evaluating the fitness consequences under carefully controlled growth conditions provides a tractable approach to quantify how environmental and genetic effects determine the fitness of multiply auxotrophic genotypes. Here we use a combination of computational and experimental approaches to address the following questions: (1) Do certain combinations of biosynthetic genes show an increased propensity to be jointly lost from bacterial genomes in nature? (2) How do fitness effects combine as multiple auxotrophy-causing mutations accumulate in the same genome? (3) Does the available carbon source affect fitness consequences of auxotrophy-causing mutations? (4) Do epistasis and plasticity interactively influence the effects of auxotrophy-causing mutations?

3. Materials and Methods

3.1. Co-occurrence prediction of multiple auxotrophies

A previously published dataset of amino acid auxotrophies that were predicted for different bacterial species [245] was updated to include the most recently available sequenced genomes and the resulting 1,432 eubacterial metabolic networks were subjected to further examination. To test, whether pairs of auxotrophies were statistically over- or underrepresented, the presence of reactions required for amino acid biosynthesis [245] was randomized, while controlling for the number of deletions (i.e. absence of a particular reaction) per species and the number of species possessing a particular reaction. A total of 8,000 samples were randomly drawn from the [species × reaction existence] space using the Rasch Sampler [255] and auxotrophy frequencies were recalculated from these random samples. Then, the frequencies of auxotrophy pairs predicted in the original dataset to co-occur were compared to the expected distribution of double-auxotrophies inferred from the randomized dataset. This approach allowed correcting the frequencies of predicted double-auxotrophies by the expected co-occurrence pattern of auxotrophies, which are simply due to chance (e.g. genetic drift) or the structure of the metabolic network (i.e. shared reactions in the biosynthetic pathways of two or more amino acids).

To test whether the observed co-occurrence pattern was reflecting the distribution of amino acids in natural environments, predicted auxotrophy frequencies were correlated with a published dataset of 69 different aquatic, terrestrial, and host-associated environments [256]. For this, the medians of the pairwise products of relative amino acid abundances were correlated with the pairwise co-occurrence of predicted amino acid auxotrophies. In this analysis, the amino acids Glu/ Gln and Asp/ Asn were not considered, because the dataset used did not allow distinguishing these pairs of amino acids.

3.2. Bacterial strains and their construction

Eleven different single gene deletions that each would render *Escherichia coli* auxotrophic for a single amino acid were identified and constructed as described ([196,245]; Table S1). All deletion alleles were transferred from existing strains [80] into *E. coli* BW 25113 [80] using P1 phage-mediated transduction [218] and recombinants were selected for their ability to grow on kanamycin-containing LB plates (50 µg ml^{-1}). In addition, 50 of the 55 possible combinations of double deletion mutants and 16 of 165 possible triple deletion mutants were successfully generated (Table S1). For this, single gene deletion mutants were first cured of the kanamycin resistance by excising the kanamycin cassette from the mutant's genomes using the pCP20 plasmid that harbors the FLP recombinase [257]. Subsequently, the second deletion allele was transferred into the resulting strains and successful recombinants were again selected for their resistance to kanamycin. A subset of double deletion mutants was cured of the kanamycin resistance cassette using the above-mentioned approach to yield triple deletion mutants. All generated genotypes used for subsequent experiments thus contained one copy of the kanamycin cassette in their genome, although at different chromosomal locations.

To examine the possibility that unintended, secondary mutations have been co-transduced, a control experiment was performed where the same deletion allele was repeatedly reintroduced into the same recipient genotype via P1 phage transduction. For this, three different deletion alleles were randomly selected (i.e. *ΔasnB*, *Δmdh*, and *ΔargH*). After transduction of *E. coli* BW 25113, the kanamycin cassette was cured and the same phage lysates were used to re-infect the recipient now carrying a deletion allele. This procedure was repeated three times to mimic the number of transduction steps required to construct triple auxotrophic mutants. Fitness of the genotypes resulting from each mutational step was determined relative to the ancestral WT as described next. In none of the three cases was the fitness of the resulting mutants significantly affected by the number of transduction rounds (one-way ANOVA: $P > 0.05$, $n = 8$ for each genotype). Thus, the phage transduction procedure used is very unlikely to have produced unintended, secondary mutations.

Conditional lethality of multiple auxotrophies was verified by inoculating 10^5 colony-forming units (CFUs) of these genotypes into 1 mL minimal medium for *Azospirillium brasilense* (MMAB; [215]) without biotin and using fructose (5 g l^{-1}) as a carbon source. The optical density (OD) the corresponding mutant strain achieved during 24 h of growth was determined spectrophotometrically at 600 nm using a Tecan Infinite F200 Pro platereader (Tecan Group Ltd., Switzerland). The mutation was deemed conditionally essential when the auxotroph's growth did not exceed the OD_{600nm} of uninoculated minimal medium. This was the case for all double- and triple-gene deletion mutants generated. Gene deletions were confirmed by sequencing the corresponding genomic regions. To phenotypically distinguish genotypes in fitness experiments, the arabinose utilization locus (Ara^+) from strain REL 607 [219] was introduced into BW 25113 using P1 phage-mediated transduction [218].

3.3. Culture conditions

Cultures were incubated at 30°C under shaking conditions and experiments were performed in MMAB minimal medium [215] without biotin and using either fructose (5 g l^{-1}) or disodium succinate (8.86 g/l) instead of malate as a carbon source. The concentration of fructose and succinate was chosen such that, at least theoretically, the same amount of biomass could be produced under both carbon sources (see Supporting Information Methods for details). In addition, both media were supplemented with a mixture of all 11 amino acids, each at a concentration of 100 µM.

3.4. Competitive fitness assays

Competitive fitness experiments were performed in 96-deepwell plates (Eppendorf, Germany) with a culture volume of 1 ml. Auxotrophs were pre-cultured at 30°C in MMAB medium supplemented with amino acids and the corresponding carbon source.

For competitive fitness assays, ~10^5 cfus ml^{-1} of WT and a focal auxotrophic mutant were co-inoculated into 1 ml MMAB medium (ratio: 1:1) supplemented with amino acids and the respective carbon source (i.e., fructose or succinate) and cell numbers were determined at 0 h and 24 h by dilution plating. *Escherichia coli* auxotrophs (Ara^-) were differentiated from WT (Ara^+) using the arabinose utilization marker as described [219]. Competitive fitness of auxotrophs versus WT was determined by calculating the Malthusian parameter (M) of both genotypes: $M = (\ln (N_f/N_i)/ 24)$, where N_i is initial number of CFUs at 0 h and N_f is the final CFU count after 24 h [219]. Relative fitness was calculated as the ratio of Malthusian parameters. Each competition assay was replicated eight times. Competition experiments between WT that did or did not contain the Ara marker provided no evidence for a fitness cost of the marker in either environment (i.e., fructose and succinate; independent sample *t*-test: $P > 0.05$, $n = 8$).

Given that all auxotrophic mutants generated contained a kanamycin resistance cassette in their genome, a possible fitness cost of this marker could theoretically affect the determined epistasis values: erroneously considering the cost of the marker multiple times when calculating the fitness expected for double and triple mutants from the fitness values of single gene deletion mutants, yet just once when determining the observed fitness of double- and triple-mutants, could have resulted in an overestimation of the true epistatic effect. To assess if the kanamycin marker incurred a cost to the auxotrophic strains, the Malthusian parameter of 11 single, four double, and four triple auxotrophic strains containing the kanamycin marker was determined in coculture with the same genotypes that have been cured from the marker (Table S2) as described above. Each competition experiment was initiated by mixing ~10^5 cfus ml^{-1} of both competitors (ratio: 1:1) into 1 ml MMAB medium that contained fructose as the sole carbon source. The number of CFUs at 0 h and 24 h was determined by plating on LB agar plates that did or did not contain kanamycin (50 µg/mL) and the Malthusian parameter was calculated as described above. Finding that in these competition experiments the Malthusian parameter of none of the kanamycin-resistant mutants differed significantly from its kanamycin-sensitive counterpart (independent sample *t*-test: $P > 0.05$, $n = 10$ for all mutants, Table S2) provided no evidence for a fitness cost of this marker. The estimated minimum difference detectable by these tests [258] ranged between 0.24% and 2.8% (Table S2), which was well below the size of epistatic interactions determined (Tables S3 and S4), suggesting that a possible fitness cost of the kanamycin resistance marker used is very unlikely to have affected our results.

3.5. *Calculating epistasis*

Epistasis for multiple deletion mutants was calculated as the difference between the *observed* and *expected fitness*. *Expected fitness* was calculated by applying the multiplicative model [91–93]. Accordingly, for a genotype bearing two auxotrophy-causing mutations, the expected fitness would be the product of the observed relative fitness of the two mutations when individually present in a genotype. Epistasis was estimated as:

$$\boldsymbol{\varepsilon_{xy} = W_{xy} - W_x W_y} \tag{1}$$

$$\boldsymbol{\varepsilon_{xyz} = W_{xyz} - W_x W_y W_z} \tag{2}$$

Equation (1) shows the calculation of epistasis for double deletion mutants and equation (2) for higher order (i.e., triple deletion) mutants. W is the relative fitness, $\boldsymbol{W_{xy}}$ and $\boldsymbol{W_{xyz}}$ is the relative fitness of strains with the entire set of two or three mutations, respectively, and $\boldsymbol{W_x}$, $\boldsymbol{W_y}$, and $\boldsymbol{W_z}$ is the relative fitness of genotypes with just one deletion mutation. For higher order interactions (eq. (2), the sum of the effect of the lower order mutations was subtracted from $\boldsymbol{\varepsilon_{xyz}}$ in equation (2) to obtain the net effect of higher order epistasis as shown in (3):

$$\boldsymbol{\varepsilon'_{xyz} = \varepsilon_{xyz} - (\varepsilon_{xy} + \varepsilon_{yz} + \varepsilon_{xz})} \tag{3}$$

Error for the estimated value of $\boldsymbol{\varepsilon}$ was calculated using the method of error propagation [91,92] and epistasis was considered significant for a given combination of deletion alleles if $\boldsymbol{\varepsilon}$ was outside the error.

3.6. *Statistical analysis*

The statistical relationship between the co-occurrence of predicted auxotrophies and the distribution of the corresponding amino acids in natural environments was assessed via Kendall's rank correlation. Normal distribution of data was assessed using the Kolmogorov–Smirnov test. Homogeneity of variances was determined by applying Levene's test and variances were considered to be homogeneous when $P > 0.05$. Fitness differences between auxotrophic mutants and their wild-type competitors as well as between auxotrophic mutants in the two environments were determined with independent sample t-tests. One-way ANOVAs followed by least significant difference (LSD) post-hoc tests were employed to test if the mutants' fitness in either environment dependent on the number of mutations. Significant deviations of epistasis from zero (no epistasis) were determined by applying one sample t-tests to the values of all mutants quantified in both environments. The statistical relationship between expected fitness and epistasis was analyzed via a Pearson's product-moment correlation. The false discovery rate (FDR) procedure of Benjamini et al. (2006) was applied to correct P values after multiple testing [222]. The relationship between expected and observed fitness was analyzed using a type II regression model. Slopes of regression lines were considered to be significantly smaller than 1 when their 95% confidence intervals did not include the 45° line (i.e., the perfect correlation between expected and observed fitness, which is the null hypothesis for nonepistatic interactions). A general linear model with *fitness* as a dependent variable and *environment* as well as the presence of one of 11 mutations as *mutation 1*, *mutation 2*, and *mutation 3* as fixed factors was calculated to identify interactive effects among mutations and/or the environment. Statistical analyses were performed using the SPSS package (version 17.0, IBM, USA) and the R software [223].

4. Results

4.1. Prevalent positive co-occurrence of auxotrophies in eubacterial genomes

A recent analysis of 949 eubacterial genomes and their inferred metabolic networks suggested that biosynthetic functions for amino acids, nucleotides, and vitamins are frequently lacking in the corresponding metabolic networks, indicating that auxotrophies are prevalent in natural populations of bacteria [245]. Interestingly, by reanalyzing a more recent collection of 1,432 eubacterial genomes, 37% of all bacteria analyzed were auxotrophic for two or more metabolites. If mutations that deactivate biosynthetic functions interact epistatically (i.e. nonadditively), pairwise co-occurrence patterns of auxotrophies are expected to significantly deviate from a random distribution. Testing this prediction for a subset of 458 eubacteria that were predicted to be auxotrophic for multiple amino acids revealed for most pairwise comparisons (152 of 190) a significant positive association (FDR-corrected one-sample Wilcoxon test: $P < 0.05$, n = 8,000, Fig. 1), indicating that auxotrophies co-occur more frequently than expected by chance. A smaller fraction of 37 auxotrophy pairs co-occurred significantly

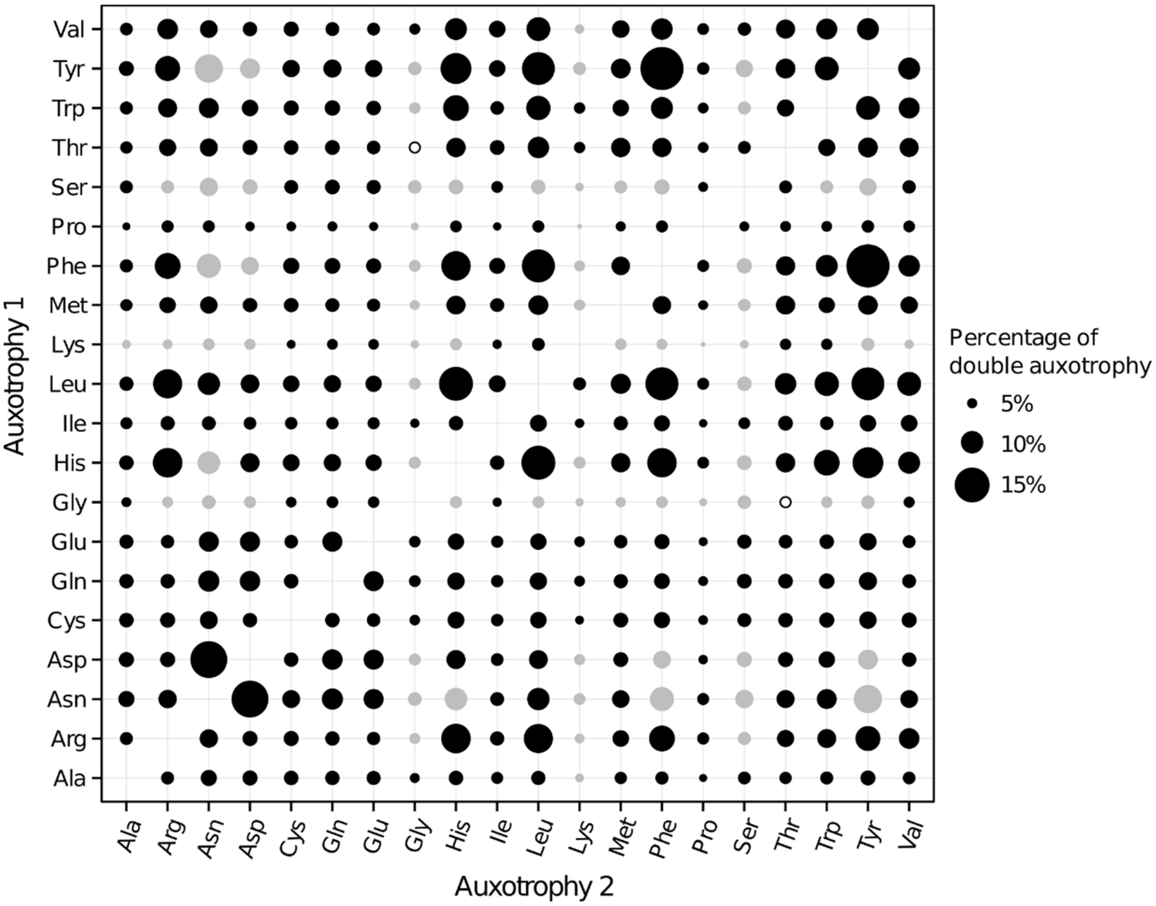


Figure 1. Predicted pairwise co-occurrence of amino acid auxotrophies in eubacterial genomes. Sizes of circles represent the proportion of genotypes (%) predicted to be simultaneously auxotroph for the two corresponding amino acids. Filled circles indicate pairs of auxotrophies that co-occurred significantly more (black) or less often (gray) than expected by chance (FDR-corrected one-sample Wilcoxon test: $P < 0.05$, $n = 8{,}000$), whereas unfilled circles depict pairs with a random co-occurrence pattern ($P > 0.05$, $n = 8{,}000$). The dataset included 1,432 eubacterial genomes that were predicted to be auxotrophic (584) for one or more of 20 different amino acids or prototrophic (848) for all amino acids.

less often than expected by chance (FDR-corrected one-sample Wilcoxon test: $P < 0.05$, $n = 8{,}000$) and only one pair (Gly-Thr) showed a distribution that was statistically undistinguishable from a random distribution (FDR-corrected one-sample Wilcoxon test: $P > 0.05$, $n = 8000$). Finding that virtually all amino acid double-auxotrophies deviate significantly in their frequency from the frequency expected by chance suggests epistatic interactions and/or environmental factors favored these combinations.

To test whether auxotrophy co-occurrences were caused by an increased propensity of certain amino acids to co-occur in natural environments, the predicted auxotrophy frequencies were correlated with quantitative measurements of relative amino acid concentrations in 69 different environments [256]. At first, the median of pairwise products of relative amino acid concentrations did not correlate significantly with the auxotrophy co-occurrence data (Kendall's rank correlation: $R_\tau = 0.04$, $P = 0.49$, $n = 120$). However, a closer look at this correlation revealed that in most environments the amino acids alanine (Ala) and glycine (Gly) were relatively abundant, while the corresponding auxotrophies were relatively rare [245]. As the two smallest proteinogenic amino acids, Ala and Gly are the metabolically cheapest to produce [61,63]. Thus, Ala and Gly auxotrophies might not be very frequent in eubacteria, because the energetic savings to lose these biosynthetic functions (i.e. the selective advantages) are relatively low. Moreover, several possible alternative biosynthetic reactions for Ala and Gly exist in prokaryotes [245], which might limit the frequency of auxotrophies. Excluding Ala and Gly from the analysis for these reasons resulted in a highly significant positive correlation between the frequency of double-auxotrophies and the pairwise abundance of amino acids in the environment (Kendall's rank correlation, $R_\tau = 0.22$, $P = 0.003$, $n = 91$, Fig. S1), which is consistent with an environmentally favored loss of metabolic genes. Taken together, the analysis of amino acid auxotrophy distributions in eubacteria suggests epistatic interactions and/or an environmentally compensated gene loss may have caused the observed co-occurrence pattern of amino acid auxotrophies.

4.2. Negative epistasis causes diminishing returns with fitness of multiply auxotrophic genotypes

How does cellular fitness scale with an increase in the number of auxotrophy-causing mutations? To address this question, one, two, or three of 11 genes that render the resulting mutant auxotrophic for amino acids were deleted from the same genetic background of *E. coli*. Altogether, 11 mutant strains bearing one (hereafter: *single mutants*), 50 mutants bearing two (hereafter: *double mutants*), and 16 strains bearing three different amino acid auxotrophy-causing mutations (hereafter: *triple mutants*) were generated (Table S1). Subsequently, the competitive fitness against prototrophic WT was determined for all 77 strains in two distinct growth environments that contained an equimolar concentration of 11 amino acids (100 µM each), yet differed in the carbon source available (i.e. either fructose or succinate). These carbon sources were chosen as they are important intermediates in the primary metabolism of most bacterial species, but derive from different points of the cells' metabolic network: fructose-6-phosphate being a core metabolite in glycolysis, while succinate is part of the tricarboxylic acid pathway.

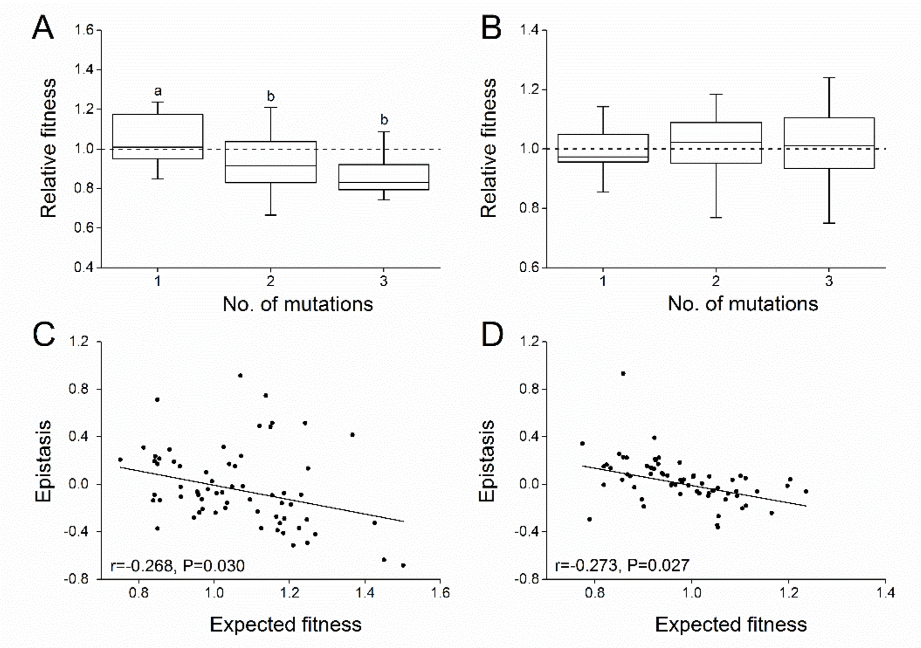


Figure 2. Change of relative fitness with increasing numbers of auxotrophy-causing mutations and relation between epistasis and expected relative fitness. (A, B) Competitive fitness of mutants bearing one, two, or three auxotrophy-causing mutations relative to prototrophic WT cells in minimal media containing either (A) fructose or (B) succinate. The dashed line represents fitness levels of the WT. Different letters indicate significant differences among deletion mutants (univariate ANOVA followed by a LSD post-hoc test: $P < 0.05$; $n = 11$ (single deletions), 50 (double deletions), and 16 (triple deletions)). Boxplots: median (horizontal lines in boxes), interquartile range (boxes, 1.5×- interquartile range (whiskers). (C, D) Relation between absolute epistasis and expected fitness determined in minimal medium containing (C) fructose and (D) succinate. Values of all double- and triple gene deletion mutants are shown and both panels include the results of a Pearson's product-moment correlation.

This analysis indicated for the fructose-containing environment that both double and triple mutants were significantly less fit than the corresponding single gene deletion mutants (one-way ANOVA followed by an LSD post-hoc test: $P < 0.05$, df = 76; Fig 2A). In contrast, the relative fitness of single-, double-, and triple-gene deletion mutants did not differ in the succinate environment (one-way ANOVA followed by an LSD post-hoc test: $P > 0.05$, df = 76; Fig 2B).

Quantitatively assessing the degree with which the effects of focal mutations deviated from expected finesses revealed on average no predominant influence of either positive or negative epistatic effects (Fig. S2). This pattern held true for both the fructose (mean epistasis: –0.05 ± 0.05, one sample *t*-test: $P = 0.26$, df = 65, Fig. S2) and the succinate environment (mean epistasis: 0.003 ± 0.03, one sample *t*-test: $P = 0.93$, df = 65, Fig. S2). However, analyzing epistatic effects for all genotypes individually

Table 1. Number of epistatic interaction identified in 50 double- and 16 triple-gene deletion mutants in both carbon (C) environments analyzed.

	Epistasis[a] in double mutants			**Epistasis[a] in triple mutants**		
C-environment	**Negative**	**Zero**	**Positive**	**Negative**	**Zero**	**Positive**
Fructose	14	27	9	6	6	4
Succinate	22	14	14	4	1	11

[a] For details on how epistatic interactions were determined please see Materials and Methods.

uncovered for the fructose environment 20 cases of significantly negative and 13 cases of significantly positive epistatic interactions (Table S3), while in the succinate environment 26 instances showed significant negative and 25 cases significant positive epistatic interactions (Table S4) (one sample t-test: $P < 0.05$, $n = 8$; Table 1). Thus, positive and negative epistatic effects cancelled each other out, thereby causing the abovementioned nonsignificant average deviation.

When the relation between expected relative fitness effects of multiple gene deletions as predicted from individual mutations and observed epistasis was scrutinized, a negative correlation (Pearson product-moment correlation: r = –0.27, P = 0.03 for fructose and r = –0.27, P = 0.03 for succinate) was observed for both carbon environments tested (Fig. 2C, D). In other words, epistatic interactions among mutations became more negative as the predicted fitness increased. Theoretically, this relationship could also be caused through a phenomenon called regression-to-the-mean [259], in which measurement error alone can cause a negative correlation between expected fitness and epistasis due to the statistical nonindependence between expected fitness and epistasis. To test whether this phenomenon could explain the observed diminishing fitness returns, a Type II regression was applied. Finding no correlation between observed and expected fitness in either carbon environment (Type II regression: $P > 0.05$, $n = 66$, Fig. S3), while both slopes were significantly smaller than 1 corroborated that the fitness of multiply auxotrophic genotypes showed a pattern of true diminishing returns.

Taken together, these experiments showed that the fitness consequences of losing conditionally essential biosynthetic genes did not increase linearly with the number of biosynthetic functions lost and that negative epistasis caused diminishing returns with mutant fitness.

4.3. Fitness consequences of auxotrophy-causing mutations depend on the available carbon source

Directly comparing the fitness levels the single gene deletion mutants achieved in both environments revealed that only the fitness of one of 11 auxotrophs tested (i.e., *ΔilvA*) was plastic with respect to the carbon source present and significantly fitter in the fructose- than in the succinate-containing environment (FDR-corrected independent samples *t*-test: $P < 0.05$, df ≥ 8; Fig. 3A). When multiply auxotrophic genotypes were also considered, a tenth (5/50) of the double mutants and two of the 16 triple mutants tested attained a significantly higher relative fitness when grown in fructose than when grown in succinate (FDR-corrected independent samples *t*-test: $P < 0.05$, df ≥ 10; Fig 3B, C). Conversely, about half of the other double and triple mutants (22/50 and 10/16, respectively) were fitter in the succinate than in the fructose-containing environment (FDR-corrected independent samples *t*-test: $P < 0.05$, df ≥ 10; Fig. 3B, C). However, the fitness of 23 double mutants (46%) and four triple mutants (25%) was unaffected by the available carbon source used (FDR-corrected independent samples *t*-test: $P < 0.05$; Fig 3B, C). Together, these results suggest that the fitness of auxotrophic mutants is highly dependent on the ambient environmental conditions.

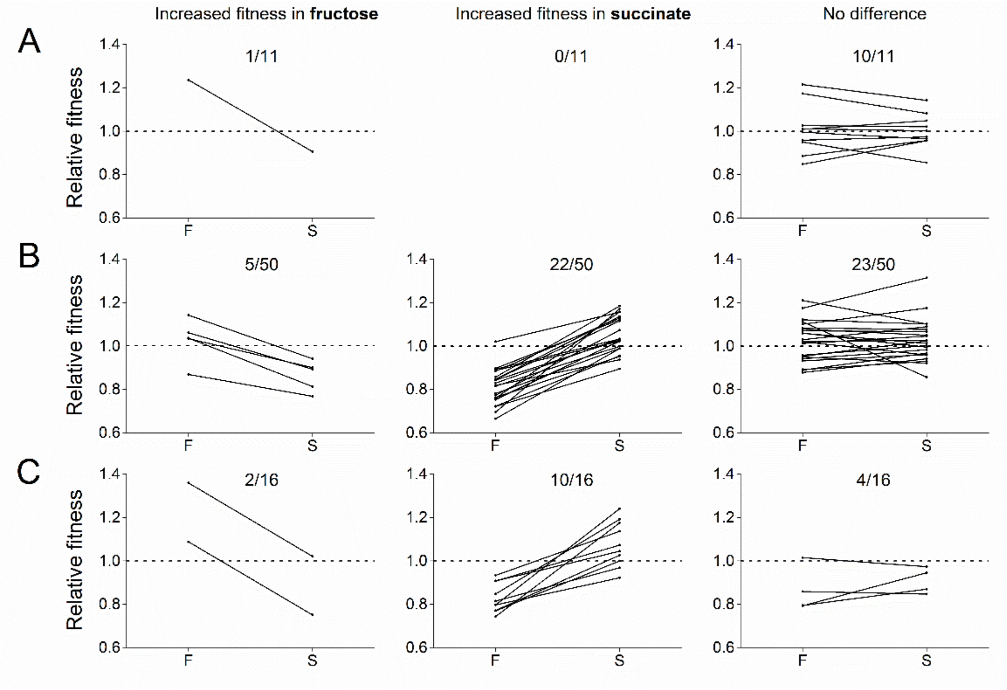


Figure 3. Reaction norms of competitive fitness of different auxotrophic genotypes against prototrophic wild-type in two different carbon environments. Each line depicts the competitive fitness of genotypes having (A) one, (B) two, or (C) three auxotrophy-causing mutations. Competition experiments against prototrophic WT (dashed line) were conducted in minimal media containing either fructose (F) or succinate (S). Differences in the mutants' relative finesses in both carbon environments were assessed using FDR-corrected independent sample *t*-tests ($P < 0.05$, df ≥ 8). Numbers above panels indicate the number of cases (left) and the total number of mutants tested (right).

4.4. Plasticity and epistasis jointly influence the fitness of multiply auxotrophic mutants

The above findings suggested that interactions among mutations (i.e. *epistasis*, G × G), interactions between mutations and the environment (G × E), and possibly also interactions of epistasis and the environment (G × G ×E) determined the fitness of defined auxotrophic mutant genotypes. Statistically evaluating the effect of these three parameters on the mutants' fitness indicated indeed a highly significant effect of epistasis (univariate ANOVA: $P < 0.0001$), G × E (univariate ANOVA: $P < 0.001$), and G × G × E (univariate ANOVA: $P < 0.0001$). Together, these findings show that the fitness associated with loosing conditionally essential biosynthetic genes is strongly affected by other metabolic mutations in the genome as well as the given nutritional environment.

5. Discussion

Knowledge on how fitness effects of a given mutation depend on other mutations present in the genome, the selective environment, or both is key to understanding adaptive processes, because the topology of the genotype–phenotype map determines the evolutionary trajectories that are accessible to organisms evolving within these fitness landscapes. Here we focus on the fitness consequences upon loss of one or more conditionally essential amino acid biosynthesis genes from bacterial genomes. Our computational analysis of 1,432 eubacterial genomes uncovered that in the vast majority of cases pairs of different auxotrophy-causing mutations co-occurred significantly more often than is expected by chance. Experimentally evaluating the fitness consequences resulting from introducing one, two, or three auxotrophy-causing mutations into the genome of *E. coli* in the presence of the required amino acids and in one of two carbon environments unravelled that (1) both positive and negative epistasis were prevalent among auxotrophy-causing genes, (2) epistasis produced diminishing returns with increasing expected genotype fitness, and (3) both the fitness of auxotrophic mutants and epistatic effects strongly depended on the carbon source available in the environment.

In our computational analysis, auxotrophy-causing genes showed a strong tendency to co-occur, which in most cases significantly exceeded what would be expected if mutations were randomly distributed (Fig. 1). Three main mechanisms may, independently or in combination, have contributed to this pattern: First, an increased co-occurrence of two amino acid auxotrophies could reflect the likelihood of the two corresponding amino acids to co-occur in the respective genotype's natural environment, thus favoring mutants that lose the corresponding biosynthesis genes. Indeed, the partial correlation observed between the co-occurrence of amino acids in nature and the co-occurrence of amino acid auxotrophies in bacterial genomes (Fig. S1) supports this scenario. Second, the probability of two amino acid biosynthesis genes to be simultaneously lost might be indicative of the fitness consequences arising upon loss of both genes in the corresponding bacterial strains. Third, amino acid biosynthesis genes that are localized in close spatial proximity on a bacterial chromosome might be simultaneously lost in large chromosomal deletion events, thus causing an increased co-occurrence of two amino acid auxotrophies.

In contrast, drift is unlikely to produce the observed co-occurrence pattern, because randomly fixing deletion alleles should rather display a frequency distribution that is not different from a random distribution. As this was the case in only two of the 190 pairwise comparisons considered, drift is unlikely to be a major determinant of the observed co-occurrence pattern.

However, statistically evaluating the relationship between the experimentally determined relative fitness or epistasis of different *E. coli* mutants with the frequencies, with which auxotrophy-causing mutations have been predicted to co-occur in eubacterial genomes (Fig. 1), did not yield significant correlations in either case (Spearman's rank correlation: $P > 0.05$). The lack of a statistical relationship between these parameters could be due to one or a combination of several of the following factors. First, bacteria frequently lose large portions of their genome. A simultaneous loss of multiple biosynthetic genes could thus explain the mismatch between the distribution of auxotrophy-causing mutations and expectations based on their epistatic interactions. Similarly, other mutations in the genome that were not considered in the present study could interact with auxotrophy-causing mutations, thus affecting the fitness of multiply-auxotrophic genotypes. Second, the effective size of bacterial populations is likely to affect the probability with which bacteria lose biosynthetic genes. Genetic drift is more effective when population sizes are small, as is the case for most endosymbiotic bacteria. Under these conditions, even nonadaptive alleles can fix in the population. Third, environmental conditions that the analyzed eubacterial strains experience in their natural environments were not evaluated in the current study, yet can affect the fitness of multiply-auxotrophic genotypes. Fourth, epistatic interactions identified for *E. coli* might not be representative for the taxonomic diversity of eubacterial genomes analyzed (Fig. 1). Fifth, in our study, exclusively structural biosynthetic genes were deleted. However, in an amino acid containing environment, natural selection might also favor mutations in regulatory elements, which could lead to the simultaneous deactivation of multiple biosynthetic pathways. Their subsequent loss from the mutants' genome would reflect regulatory relationships among groups of genes rather than epistatic interactions among multiple genes that were individually lost. Thus, future work is necessary to elucidate how auxotrophies evolve and to which extent epistatic interactions determine the mutational paths taken.

Previous work showed positive fitness effects generally accompany the loss of a conditionally essential biosynthetic gene when the focal metabolite is sufficiently available in the environment [76,209,210,245,260]. However, our study revealed that the sign and magnitude of fitness consequences can drastically change depending on the environment and the presence of additional auxotrophy-causing mutations. As such, our results strikingly matched theoretical predictions of a recent study, in which a flux-balance analysis of the metabolic network of *E. coli* identified strong effects of the carbon source used on epistatic interactions among deleted metabolic genes [88]. However, which mechanisms caused this functional relationship? One factor that could account for these observations, is that the loss of conditionally essential biosynthetic genes is likely to trigger a strong regulatory response that allows bacterial cells to survive despite amino acid starvation [261]. Most probably, these changes involve an upregulation of amino acid transporters as well as a rerouting of metabolic fluxes through multiple pathways [261] leading to a globally restructured metabolism [262]. Because such systemic changes may be specifically tailored to compensate specific shortages associated with losing certain sets of genes, these regulatory differences could explain the observed plasticity and epistatic interactions among mutations. Second, our competition experiments were performed in relatively complex nutritional environments that, besides one of two carbon sources, also contained 11 different amino acids. Thus, epistatic effects could be caused by a competitive inhibition of amino acid uptake systems [263], competition of transporters for membrane space [264], or effects resulting from alterations of cell-internal amino acid pools [265]. Future work should examine these possibilities.

Interestingly, increasing the number of metabolic auxotrophies did not result in an additive increase of fitness effects caused by individual mutations, but on average mostly showed an overall decline or neutral effect in the succinate and fructose environment, respectively (Fig. 2A, B). This observation is consistent with previous experimental works [93,113,114,253] showing that negative epistasis acts to diminish mutational effects. Finding this pattern also for auxotrophy-causing mutations suggests a common mechanism caused the beneficial fitness effects of different single gene deletion mutants. Intriguingly, a saving of protein expression costs has been previously suggested as a mechanistic cause for the fitness effects upon loss of conditionally essential biosynthetic genes from the genome of *E. coli* [245] as well as for the diminishing returns epistasis observed when four beneficial alleles were analyzed in *Methylobacterium extorquens* AM1 [113]. In any case, if epistatic interactions determine the order in which auxotrophy-causing mutations are fixed in bacterial genomes, the current work provides several testable hypotheses that could be verified in a laboratory-based evolution experiment.

Finally, a particularly strong beneficial effect upon loss of the first metabolic gene may act as a spring-loaded mechanism that facilitates the establishment of metabolic cross-feeding interactions within microbial communities [76,83,266,267] or aids the establishment of symbiotic associations between microbial symbionts and their host [268,269].

6. Conclusions and Outlook

Our study provides first empirical insights into the selective consequences bacterial genotypes face when losing multiple auxotrophy-causing mutations. Especially the observed impact the ambient environment and the number of genes lost had on the fitness of auxotrophic genotypes implies a strong context-dependency of metabolic loss-of-function mutations that needs to be taken into account when such mutations are interpreted. Observing that epistasis produced diminishing returns with increasing expected genotype fitness points to a yet unknown molecular mechanism that constrains the fitness achievable by multiply auxotrophic genotypes. Identifying this mechanism will not only shed light on what causes the strong fitness benefits conferred by auxotrophy-causing mutations, but will also help to understand the molecular links that connect different biosynthetic genes. In particular laboratory-based evolution experiments, in which bacterial populations evolve under carefully controlled environmental conditions, provide a unique opportunity to identify which genes (e.g. regulatory versus structural genes) are prime targets of natural selection during the adaptive evolution of metabolic auxotrophies. Together with the approaches used in this study, such experiments would allow to further dissect how phenotypic plasticity and epistasis interactively guide the adaptive loss of biosynthetic functions.

Acknowledgments

We thank the Experimental Ecology and Evolution group, A. de Visser as well as three anonymous referees for constructive criticism, W. Boland for support, and D. Refardt for providing us with strain *E. coli* REL 607. We are especially grateful to D. Berger, V. Cooper, and K. Flynn for advice on statistical analyses. This work was supported by the Jena School for Microbial Communication (JSMC), the Volkswagen Foundation, and the Max Planck Society. The authors declare no conflict of interest.

Data archiving

The doi for our data is 10.5061/dryad.hq701.

Chapter V

Metabolic network architecture and carbon source determine metabolite production costs

In review. Submitted to FEBS Journal on 17th September 2015

Authors

Silvio Waschina, Glen D'Souza, Christian Kost, and Christoph Kaleta

1. Abstract

Metabolism is essential to organismal life, because it provides energy and building block metabolites. Even though it is known that the biosynthesis of metabolites consumes a significant proportion of the resources available to a cell, the factors that determine their production costs remain less well understood. In this context, it is especially unclear how the nutritional environment affects the costs of metabolite production.

Here we use the amino acid metabolism of *Escherichia coli* as a model to show that the point at which a carbon source enters central metabolic pathways is a major determinant of individual metabolite production costs. Growth rates of auxotrophic genotypes, which in the presence of the required amino acid save biosynthetic costs, were compared to the growth rates that prototrophic cells achieved under the same conditions. The experimental results showed a strong concordance with computationally-estimated biosynthetic costs, which allowed us, for the first time, to systematically quantify carbon source-dependent metabolite production costs.

Thus, we demonstrate that the nutritional environment in combination with network architecture is an important but hitherto underestimated factor influencing biosynthetic costs and thus microbial growth. Our observations are highly relevant for the optimization of biotechnological processes as well as for understanding the ecology of microorganisms in their natural environments.

2. Introduction

Most bacterial species are heterotrophic and thus derive their carbon from breaking down organic compounds [59]. The structural diversity of organic compounds bacteria encounter in their natural environments is remarkable and for several species it is known that they can utilize an extremely wide range of chemically different carbon sources [14]. *Escherichia coli,* for instance, is able to utilize more than 80 compounds as sole source of energy and carbon [270]. However, since carbon sources differ drastically in terms of their energy content as well as the molecular routes how a given bacterial cell can import and degrade the corresponding chemical, bacterial growth depends decisively on the nature of the carbon source used [271]. In this context, it has been proposed that biochemical constraints in the allocation of resources may limit the growth rate of bacterial cells [60,73]. In particular, such a pattern could be caused by the distribution of fluxes through the metabolic network to provide an optimal supply of building block metabolites (i.e. amino acids, nucleotides, and lipids) and growth factors (i.e. vitamins and co-factors) for cell growth.

Assuming that the architecture of a cell's metabolic network determines fluxes through the network, flux distributions should depend on the point at which a given carbon source enters the metabolic network. Indeed, it has been shown that the entry points of a given carbon source can cause considerably higher relative fluxes through reactions closer to the entry point than fluxes of reactions more distant to the entry point of the carbon source [272]. As a consequence, locally increased fluxes could also affect the biosynthetic costs of metabolites by locally increasing the amount of substrate available to a given biosynthetic reaction relative to all other growth-related functions. Thus, differences in flux distributions caused by different carbon sources should also translate into different biosynthetic costs of metabolites.

Here we test this hypothesis by combining theoretical predictions with targeted experiments using amino acid biosynthesis of *E. coli* as a tractable model. Amino acid metabolism was chosen as a test case, because the biosynthesis of amino acids diverts an immense fraction of the total carbon source budget of a bacterial cell during growth [273]. It is therefore not surprising that bacterial species are under strong selective pressure to economize their amino acid usage [61,68]. We used a genome-scale model of the metabolic network of *E. coli* to estimate the biosynthetic cost for each of 20 proteinogenic amino acids depending on the utilized carbon source. Next, we validated these predictions by comparing the growth rates of genotypes auxotrophic for individual focal amino acids and the prototrophic wild type of *E. coli* grown on different carbon sources, while supplementing increasing concentrations of the focal amino acid to the growth environment. Under these conditions, auxotrophic genotypes increasingly saved the costs to biosynthesize the focal amino acid relative to the prototrophic wild type, and could thus invest the economized carbon source in other cell growth-related functions. By gradually relaxing the amino acid limitation for the growth of auxotrophs in this way, and comparing their maximum growth rates relative to the growth rates achieved by prototrophic wild type cells, allowed quantifying the carbon source-dependent costs to produce individual amino acids.

Both theoretical predictions and experimental results revealed strong differences in the production costs of central metabolites in bacteria depending on the point at which the utilized carbon source enters the cell's metabolic network. The observed shifts of biosynthetic costs depending on the utilized carbon source are physiologically relevant and are caused by the structure of the underlying metabolic network.

3. Results

3.1. Metabolic costs of amino acid depend on carbon source

To determine whether or not the metabolic costs to biosynthesize each of 20 proteinogenic amino acids depend on the carbon source used, amino acid production costs were computationally estimated for all 61 carbon sources, which are known to support the growth of *E. coli* K12 as sole source of carbon and energy [274] (Fig. 1). The mean costs of amino acids predicted in this way quantitatively matched previous predictions of energetic costs of amino acid biosynthesis (Pearson's product-moment correlation: $R = 0.96$, $N = 20$, $P < 0.0001$, supporting information Fig. S1, [61]). This correlation shows that the cost prediction method presented here is in line with previous estimations, but further enables to systematically assess metabolite production costs differences between various carbon substrates. To identify whether and to which extent the metabolic costs of a single amino acid were affected by the available carbon source, the metabolic cost estimated for all amino acid and 61 carbon sources were analyzed by principle component analysis (PCA) (Fig. 1C). Clustering for the correlation values of the observed amino acid costs with these three main principal components revealed three distinct groups: group 1 (blue) consisted of amino acids with precursors in glycolysis and/or the pentose phosphate pathway (i.e. Cys, Gly, His, Met, Phe, Ser, Trp, Tyr, and Ser), group 2 (green) contained pyruvate-derived amino acids (i.e. Ala, Leu, and Val), and group 3 (red) comprised amino acids with precursors from the tricarboxylic acid (TCA) cycle (i.e. Arg, Asn, Asp, Glu, Gln, Ile, Lys, Pro, and Thr) (Fig. 1A). The differences between groups reflect diverging biosynthetic costs associated to different classes of carbon sources: Amino acids of group 1 are metabolically cheaper to produce when glycolytic substrates (sugars/ sugar alcohols) are utilized as carbon source, yet more cost-intensive when only gluconeogenic substrates (e.g pyruvate, lactate, TCA-cycle intermediates) are available, and *vice versa* for the amino acids of group 2 and 3 (Tukey multiple comparisons of means: $P < 0.05$, for samples sizes see Fig. 1D). Consequently, there is a cost trade-off between the different groups of amino acids: reduced costs to produce amino acids in one group come at the expense of higher costs to synthesize amino acids of another group.

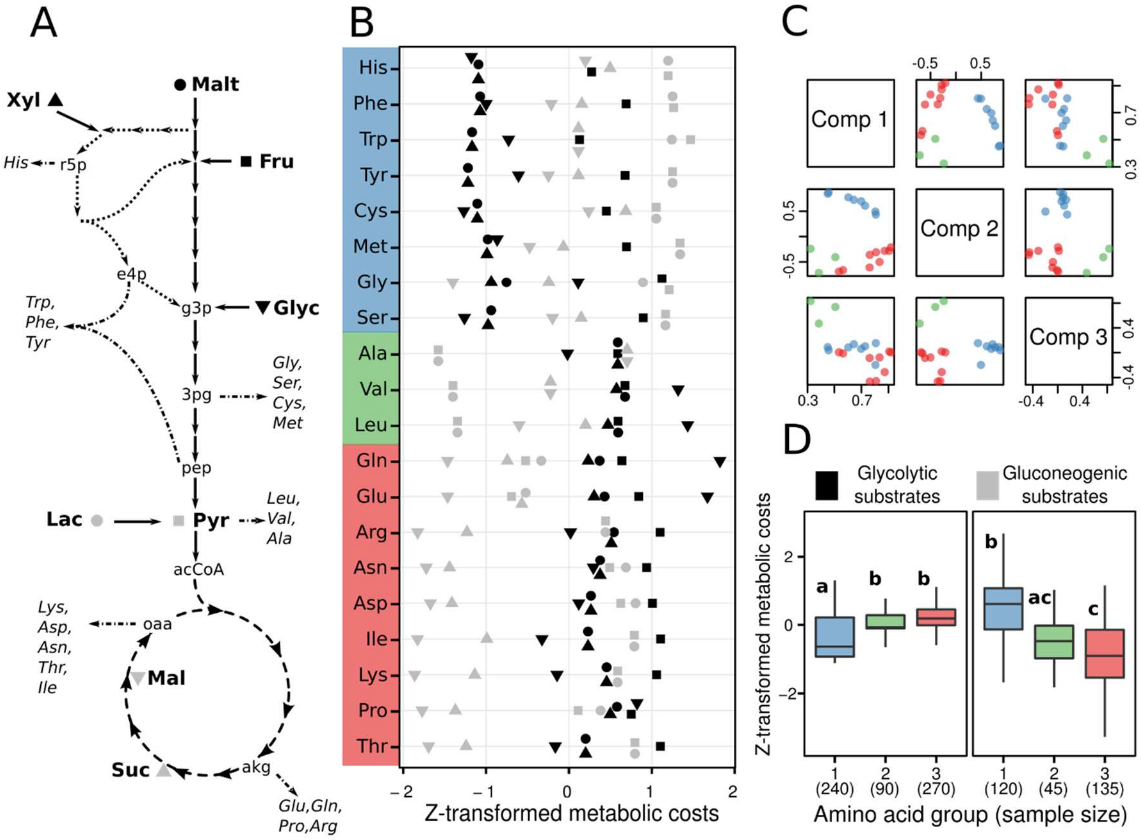


Figure 1. *In silico* estimations of carbon source- and network structure-dependent metabolic costs of proteinogenic amino acids. (A) Schematic representation of the central metabolism of *Escherichia coli* (glycolysis – solid arrows, TCA cycle – dashed arrows, pentose phosphate pathway – dotted arrows). Carbon sources used in this study are shown in boldface, amino acids in italics. (B) Estimated metabolic costs of amino acids for 8 carbon sources including 4 organic acids (grey) and 4 sugars/ sugar-alcohol (black). For a better visualization, metabolic costs of each amino acid were z-transformed (same range of values). (C) Principle component analysis (PCA) of estimated metabolic costs of amino acids based on 61 carbon sources. Shown are the correlations of the metabolic costs of each amino acid with the three main PCA components (Comp 1-3), which together explain >91% of the observed variation. Data points are colored according to k-means clustering with three centers: (group 1, blue): Cys, Gly, His, Met, Phe, Ser, Trp, and Tyr; (group 2, green): Ala, Leu, and Val; (group 3, red): Asn, Asp, Arg, Gln, Glu, Ile, Lys, Pro, and Thr. (D) Estimated (z-transformed) metabolic costs of amino acids for glycolytic- and gluconeogenic carbon sources. Amino acids are grouped according to the k-means clustering in (C). Different letters denote significant differences (Tukey multiple comparisons of means: $P < 0.05$, numbers below amino acid groups denote sample sizes).

Furthermore, the quantitative impact of different types of carbon sources on the absolute metabolic costs of amino acids varied among amino acids (Fig. 2). The highest variability of metabolic costs were observed for the leucine and glutamate, whose biosynthetic costs varied up to 13% from the mean metabolic costs based on the

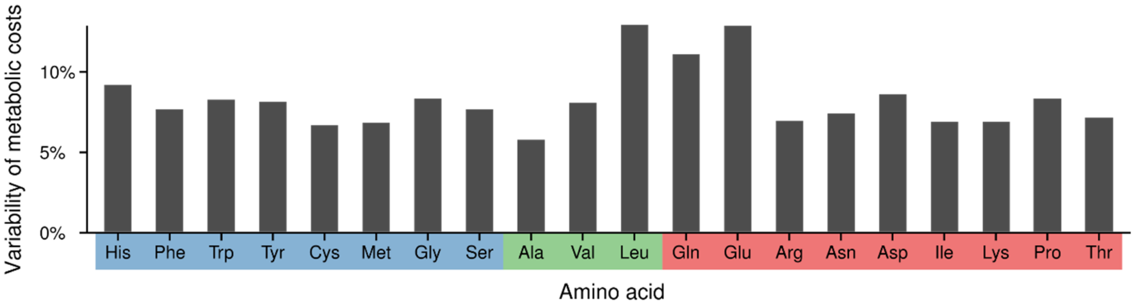


Figure 2. Variability of amino acid metabolic costs. The variability of metabolic costs was calculated as the 95% confidence interval size divided by the mean metabolic cost of the respective amino acid based on the costs estimations assuming 61 different types of carbon sources.

estimations assuming 61 different types of carbon sources. The lowest variability of 6% was observed for alanine.

3.2. The response of auxotrophs to amino acid supplementation depends on the carbon source

The maximum growth rate of all seven auxotrophs were tested under amino acid supplementation and eight different carbon source conditions. Auxotrophic strains were chosen to study the effect of amino acid supplementation and to ensure that the cells actually save the biosynthetic costs to produce the focal amino acids and use amino acids from the media. The seven auxotrophies were chosen for the experiments, because no other cellular function than the integration into proteins have been described for the respective amino acids. Other effects, than the saving of metabolic costs, of the amino acid supplementation on the growth of the *E. coli* auxotrophs could therefore be prevented.

The maximum growth rate of all seven auxotrophs increased significantly with increasing amino acid supplementation (FDR-corrected linear mixed-model fit by maximizing the restricted log-likelihood: $P < 0.05$, n=42, Fig. 3 and Fig. S2). The only exception to the otherwise consistent pattern was the case of the isoleucine auxotroph using succinate as carbon source (FDR-corrected linear mixed-model fit by maximizing the restricted log-likelihood, $P = 0.42$, $n = 42$). In contrast, the maximum growth rates of populations of prototrophic *E. coli* wild type cells did not respond significantly to increasing amino acid concentrations in 32 out of 56 amino acid-carbon source combinations analyzed (Fig. S3). In 22 cases, the maximum growth rate increased significantly with amino acid supplementation, in two cases (i.e. histidine and xylose/

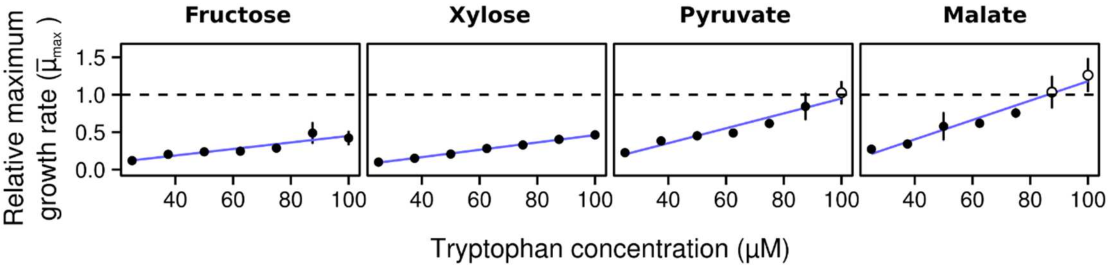


Figure 3. Carbon source-dependent growth rate response of the tryptophan auxotrophic genotype to increasing tryptophan supplementation. Shown are the mean maximum growth rates (± 95% confidence interval) of the tryptophan (Trp) auxotrophic genotype relative ($\bar{\mu}_{max}$) to the prototrophic wild type (=1, dashed line) in four carbon source regimes and seven different Trp concentrations. Filled circles indicate growth rates of the auxotrophs, which are significantly lower than the maximum growth rate of the prototrophic wild type under the same carbon source conditions without Trp supplementation (FDR-corrected Welch two sample t-tests: $P < 0.05$, $n = 6$). Unfilled circles denote no statistical difference. This figure is representative for the complete data set shown in Fig. S2.

succinate) it even decreased significantly with increasing amino acid concentrations (FDR-corrected linear mixed-model fit by maximizing the restricted log-likelihood: $P < 0.05$, $n=42$, Fig. S3).

In virtually none of the cases examined did the auxotrophic genotypes reach maximal growth levels of WT populations (Figs. 3 and S2), indicating that under the focal conditions growth of auxotrophic genotypes was mainly limited by the availability of the required amino acid. After normalizing the auxotroph's growth rate by the growth rate the prototrophic WT strain had achieved under the same carbon source condition without amino acid supplementation (in the following, $\bar{\mu}_{max}$ refers to the normalized growth rate), it became clear that the increase of the relative maximum growth rate $\bar{\mu}_{max}$ strongly depended on the carbon source provided for growth (FDR-corrected repeated measures ANOVA: all $P < 0.001$, $df_{carbon\ sources} = 7$, $df_{error} = 35$, Figs. 3 and S2). For example, the tryptophan auxotroph responded to tryptophan supplementation with an increase of 4.4 $\bar{\mu}_{max}$ (mM Trp)$^{-1}$ (mean) with fructose as carbon source, whereas with 10 $\bar{\mu}_{max}$ (mM Trp)$^{-1}$ the increase was significantly higher when utilizing pyruvate (paired t-test: $P < 0.001$, $n = 6$, Fig. 3).

Taken together, the growth-kinetic assays revealed a strong effect of the carbon source on the growth physiology of the seven amino acid auxotrophic strains tested when the availability of the required amino acid was limiting growth.

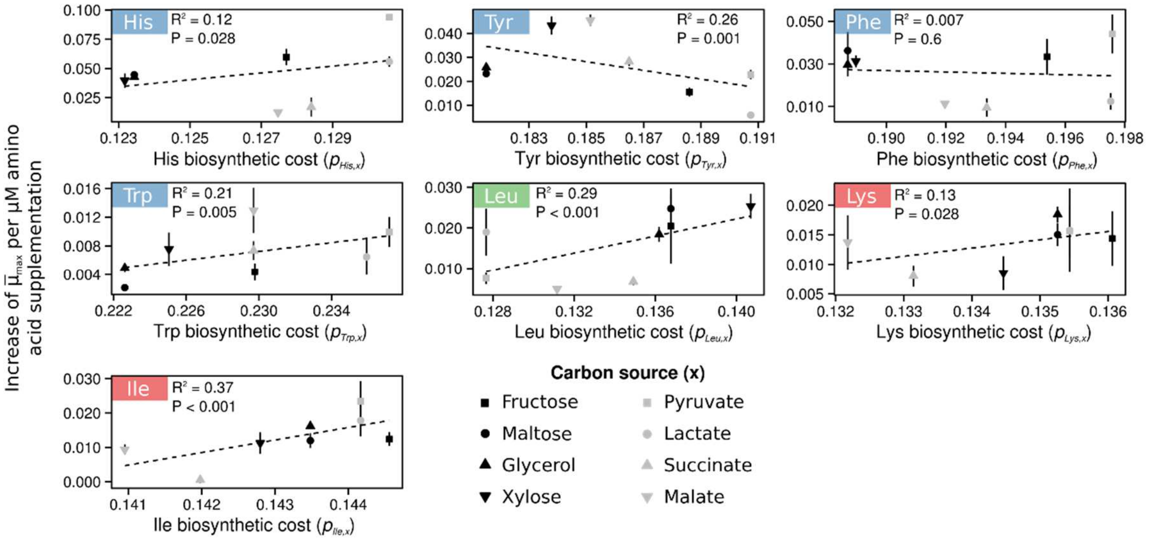


Figure 4. Correlation of predicted- and measured biosynthetic costs. The response in growth rate of auxotrophic genotypes to amino acid supplementation can be explained by the metabolic network structure. Shown are the correlations of predicted metabolic costs $\boldsymbol{p_{k,x}}$ (x-axes) for amino acid k and carbon source x and the experimentally-determined increase of the relative growth rates $\bar{\mu}_{max}$ of auxotrophs with increasing amino acid concentration (Y). Mean values ± 95% confidence intervals are shown. See Table S1 for amino acid abbreviations.

3.3. Biosynthetic costs can explain the growth rate increase upon amino acid supplementation

A significant positive correlation between estimated biosynthetic costs and growth rate increases was observed for five of the seven amino acids tested: histidine ($P < 0.05$), tryptophan ($P < 0.01$), leucine ($P < 0.001$), lysine ($P < 0.05$), and isoleucine ($P < 0.001$, FDR-corrected linear mixed-model fit by maximizing the restricted log-likelihood: n=48, Fig. 4). These five amino acids represent all three main groups identified in the above-mentioned *in silico* analysis of the costs to biosynthesize the 20 proteinogenic amino acids (Fig. 1C). In other words, the same metabolic trade-offs in the efficiencies to synthesize amino acids that were theoretically predicted (Fig. 1D) were also found experimentally (Fig. 4).

A significantly negative correlation was observed between the predicted biosynthetic costs and growth rate increases for tyrosine (FDR-corrected linear mixed-model fit by maximizing the restricted log-likelihood: $P = 0.001$, n=48, Fig. 4), while no statistical relationship between these two parameters could be detected for phenylalanine (FDR-corrected linear mixed-model fit by maximizing the restricted log-likelihood: P = 0.6, n=48, Fig. 4).

4. Discussion

Microorganisms invest a significant proportion of their available carbon resources in the biosynthesis of metabolites. The amount of carbon source a cell needs to produce individual metabolites (i.e. biosynthetic costs) can be estimated based on the organisms' genome sequence and information on the nutritional composition of the natural habitat [65]. However, natural environments can fluctuate widely in the availability of different resources [275,276] and many microorganisms are able to utilize a broad range of different carbon sources [14]. Two main questions arise from these facts: i) How are metabolite production costs affected by the nutritional environment?, and ii) How variable are these costs within an organism? Here, we tested for the first time whether the variability of biosynthetic costs within a given organism can be explained by the carbon source used. The main findings of this study are that the structure of the metabolic network determines biosynthetic costs and that these costs are variable depending on (1) the position of the precursor metabolites within the metabolic network, and (2) the point at which the carbon source enters central metabolism (Fig. 1D).

A genome-scale metabolic network of *E. coli* was employed to predict differences in amino acid production costs depending on the nutritional environment. To test the *in silico* cost estimations, growth kinetic assays of amino acid auxotrophic *E. coli* strains and the prototrophic wild type were performed for eight different carbon sources and seven amino acids in increasing concentrations. By comparing the maximum growth rates achieved by auxotrophic and prototrophic genotypes under specific conditions, it was possible to experimentally determine the biosynthetic costs, which auxotrophic genotypes saved by not having to synthetize the respective amino acid autonomously. The experimental measures derived in this way matched theoretical predictions of the carbon source-dependent biosynthetic costs for five of the seven amino acids tested: histidine, isoleucine, leucine, lysine, and tryptophan (Fig. 4). A discrepancy between cost prediction and experimental approximation was observed only for phenylalanine and tyrosine. The biosynthetic pathways for these two amino acids are closely connected: both amino acids originate from the common precursor chorismate and the biosynthetic pathways consist both of three reactions where only the second step is catalyzed by distinct enzymes [72]. Furthermore, both pathways are tightly co-regulated [277]. This interconnection of both pathways might cause additional effects besides the focal biosynthetic costs, when only one of the two amino acids is supplemented to the media.

To avoid confounding factors affecting the growth kinetics that are independent of biosynthetic costs, we focused our analysis on amino acids, which cannot be degraded and, hence, cannot be utilized as an alternative carbon source. Also, by using auxotrophic genotypes that cannot convert any other metabolite into the focal amino acid [196], it was possible to directly and precisely control the amount of the focal amino acid that was available to the cells. Taken together, our study provides, for the first time, a comparison of the growth response of *E. coli* to amino acid supplementation with a metabolic model using the flux balance analysis framework. The detected cost differences between carbon sources strongly influenced bacterial growth and thus significantly affected bacterial fitness.

To understand the evolution of a microbial metabolic network requires knowledge on the factors that determine biosynthetic costs within a given organism. The results presented in this work provide first evidence that metabolite production costs are affected by environmental factors such as the available carbon source. Most notable differences were found for the amino acids leucine, glutamate, and glutamine, whose costs varied by up to 13% between carbon sources. The observation of carbon source-dependent metabolic costs of amino acids is in line with recent findings that gene deletion mutations, which lead to the loss of biosynthetic functions, can have different fitness effects depending on which carbon source is provided for growth [278]. In addition, it has been shown that synthetically generated amino acid-, nucleotide- , and vitamin auxotrophic mutants of *E. coli* had a significant fitness advantage over their prototrophic ancestor in environments where the respective metabolite was sufficiently present – even when both strains directly competed against each other [245]. These fitness benefits are likely to be due to the biosynthetic costs, which the auxotroph save by not having to synthesize the respective metabolite [279].

Another interesting outcome of our study was, that the comparison of the biosynthetic costs of all 20 proteinogenic amino acids for 61 different carbon sources pointed to a metabolic cost trade-off between the efficiencies to produce different classes of amino acids (Fig. 1D). Thus, amino acids that are less costly to produce utilizing one specific carbon source (e.g. amino acids derived from TCA cycle intermediates) relative to another carbon source come at the expense of higher costs for other amino acids (e.g. derived from glycolysis intermediates). Biochemical trade-offs are thought to play a key role for metabolic specialization [9]. Hence, our results provide a plausible explanation for the evolution and maintenance of metabolic cross-feeding interactions where subpopulations, which specialized on preferentially performing certain metabolic functions, share the products of these functions [76]. Based on our results, metabolite cross-feeding could be especially promoted in environments, where multiple carbon sources are simultaneously present and subpopulation have specialized on utilizing distinct carbon sources. Sympatric specialization to utilize different carbon sources has been observed in laboratory evolution experiments of *E. coli* [117,119]. In a prominent example of a long term evolution experiment, in which *E. coli* was serially propagated in glucose minimal media, an adaptive mutation emerged in one population after 31,500 generations, through which the newly evolved variants acquired the ability to utilize citrate as carbon source – an abundant yet previously unused carbon source, which has been included as part of the media formulation due to its iron-chelating properties [117]. In another long-term continuous culture of *E. coli,* where glucose was provided as sole carbon source, two subpopulations evolved: one, which utilized glucose and produced acetate as a metabolic by-product and a second subpopulation, which specialized to utilize the exogenously available acetate [119]. Consequently, the utilization of different carbon sources can cause significant differences in the distribution of metabolic fluxes [47,272,280] and, as shown in this study, different biosynthetic costs. Interestingly, the amino acid biosynthetic cost differences between the two specialized subpopulations in the two above mentioned examples are highly reciprocal, because glucose is a glycolytic carbon source, whereas citrate, or acetate, respectively are gluconeogenic substrates (Fig. 1D). These differences in turn could favor the evolution of amino acid cross-feeding, where each specialized subpopulation can receive mutual benefits by saving biosynthetic costs.

Our results are not only relevant to understand adaptive processes of bacteria that are exposed to different nutritional environments, but have also implications for more applied contexts, for example the optimization of biotechnological processes where microorganism are used to produce value-added compounds such as biofuels, amino acids, or recombinant proteins. Metabolic engineering uses recombinant DNA techniques to modify the structure of metabolic networks by introducing new biosynthetic capabilities to the cell or improving the production rate of a specific molecule [281]. Another way to optimize production rates of desired metabolites is to rationally design the nutritional environment that is used as culture media [282]. Based on the presented results, it will be possible to increase the yield of a desired compound by rationally choosing a carbon source that minimizes production costs of the focal metabolite. Thus, a better understanding of the (environmental) factors that determine the production costs of desired compounds can significantly improve biotechnological production processes.

All growing cells allocate resources to different biosynthetic pathways in response to the nutritional environment. The resource costs associated with the biosynthesis of metabolites strongly affect the fitness of a species. In this study, the interplay between the chemical nature of a carbon source and its conversion into cell constituents was systematically assessed. The presented results unravel the link between a cell's nutritional environment and the architecture of its metabolic network as a key determinant of biosynthetic costs and microbial growth. As the structure of a metabolic network has evolved in response to natural selection, the here observed variability of biosynthetic costs depending on the available carbon source is indicative of the crucial role of the environmental context for the evolution of biochemical networks and the ecology of microorganisms. Future work is necessary to extent the economical concept of metabolic costs in more natural settings where multiple microbial species with diverse metabolic capabilities coexist and where several different substrates are available for cell growth.

5. Methods

5.1. *Prediction of biosynthetic costs*

The biosynthetic costs of all 20 proteinogenic amino acids for 61 different carbon sources (see Table S3), which theoretically support the growth of *E. coli*, were estimated using flux balance analysis (FBA). Biosynthetic costs $p_{k,x}$ of an amino acid k were defined as the proportion of carbon source x, which is at least required to produce 1 mmol gDW^{-1} h^{-1} of the amino acid relative to the amount of carbon source x required to form 1 mmol gDW^{-1} h^{-1} biomass. The estimation incorporates the '*dual costs of amino acids'* [61]: (1) the resources required to generate energy in form of high-energy phosphor bonds (ATP and GTP) as well as the reducing power in form of NADH, NADPH, and $FADH_2$, which is consumed by enzymes of the biosynthetic pathway, and (2) the resources required to produce precursors for amino acid synthesis.

Two optimizations were performed within the FBA-framework using a genome-scale metabolic network reconstructions of *E. coli* K12 [26]: (1) $n_{k,x}$ was defined as the minimum amount of a carbon source x (in mmol gDW^{-1} h^{-1}; DW = dry weight) to produce one unit (i.e. 1 mmol gDW^{-1} h^{-1}) of an amino acid k. $n_{k,x}$ was determined by minimizing the influx of the carbon source x and fixing an outflow reaction of the amino acid k to a flux value of 1 mmol gDW^{-1} h^{-1}. (2) m_x was defined as the minimum amount of a carbon source x required to form 1 mmol gDW^{-1} h^{-1} biomass. m_x was calculated by constraining the flux through the biomass reaction (with 53.95 GAM estimate) of the metabolic model to a value equal 1 mmol gDW^{-1} h^{-1} and by minimizing the influx of the carbon source x. The optimizations were performed within Matlab 7.14 (Mathworks, USA) using the COBRA Toolbox version 2.0.5 [283] and the TOMLAB /CPLEX version 7.9 (TOMLAB Optimization, USA) as linear programming solver. The media elements used for the genome-scale model were Ca^{2+}, Cl^-, CO_2, Co^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, H^+, H_2O, K^+, Mg^{2+}, Mn^{2+}, molybdate, Na^{2+}, NH_4^+, Ni^{2+}, O_2, phosphate, SO_4, tungstate, and Zn^{2+}.

Finally, the biosynthetic cost estimations $p_{k,x}$ for all amino acid – carbon source combinations were calculated as

$$p_{k,x} = n_{k,x}/m_x.$$

5.2. *Bacterial strains*

Amino acid auxotrophic genotypes used in this study have been generated as described previously [245] (Table S2). The auxotrophic strains were derived from the *E. coli* BW25113 strain, which is the prototrophic wild type. Mutant strains were cured of the kanamycin resistance marker by excising the kanamycin cassette from the mutant's genome using pCP20 plasmid, which harbors the FLP recombinase [257]. For unknown reasons, it was not possible to cure the tyrosine auxotroph of the kanamycin resistance. Thus, the original kanamycin resistant mutant was used for growth kinetic assays instead. However, it has been previously demonstrated that this resistance marker does not incur detectable fitness effects under non-selective (i.e. antibiotic-free) conditions [278].

5.3. *Culture conditions*

All cultures were incubated under shaking conditions at 30 °C and grown in Minimal Media for *Azospirillum brasilense* (MMAB) [215] containing K_2HPO_4 (3 g L^{-1}), NaH_2PO_4 (1 g L^{-1}), NH_4Cl (1 g L^{-1}), $MgSO_4 \cdot 7H_2O$ (0.3 g L^{-1}), KCl (0.15 g L^{-1}), $CaCl_2 \cdot 2H_2O$ (0.01 g L^{-1}), $FeSO_4 \cdot 7H_2O$ (0.0025 g L^{-1}), $Na_2MoO_4 \cdot 2H_2O$ (0.05 g L^{-1}), and using different carbon sources. The concentrations of the carbon sources were 5 g L^{-1} D-fructose, 8.86 g L^{-1} disodium succinate, 8.61 g L^{-1} potassium L-lactate, 4.42 g L^{-1} glycerol, 8.17 g L^{-1}

sodium pyruvate, 10.64 g L^{-1} disodium L-malate, 5 g L^{-1} D-maltose monohydrate or 5.06 g L^{-1} D-xylose. The concentrations of carbon sources were chosen such that – at least theoretically – the same amount of biomass could have been produced under all nutritional conditions. For this, we used the above-introduced value m_x, i.e. the minimum amount of carbon source x required to form one unit of biomass. The final concentration of carbon source x was calculated as $c_x = c_{Fru} \cdot m_x / m_{Fru}$ using 5 g L^{-1} fructose (c_{Fru} = 27.75 mM) as reference. This procedure is similar to the approach described by [131]), where concentrations were adjusted to match the number of reducible carbon atoms. However, using the genome-scale metabolic network of *E. coli* allows to take the physiological capabilities of the cell to transform a certain carbon source into biomass more precisely into account.

The eight carbon sources fructose, maltose, xylose, glycerol, pyruvate, lactate, succinate, and malate were chosen, because these substrates enter the central metabolic network of *E. coli* at different points (Fig. 1A) and the predicted biosynthetic costs of amino acids differed considerably between these carbon sources (Fig. 1B). Fructose, maltose, and glycerol are catabolized via the Embden-Meyerhof-Parnas (EMP) Pathway. Xylose is converted to the pentose phosphate pathway intermediate D-xylulose 5-phosphate. L-lactate can be oxidized to pyruvate, a central metabolite, which links the glycolysis and the tricarboxylic acid (TCA) cycle. The carbon sources succinate and L-malate are intermediates of the TCA cycle.

5.4. Growth kinetic assays

The response of seven amino acid auxotrophic *E. coli* mutants and the prototrophic wild type strain in terms of the maximum growth rate to the supplementation of the focal amino acids was quantified in growth kinetic assays. For this, seven genotypes that were auxotrophic for one of the following amino acids were selected (deleted gene in brackets): histidine (*hisD*), tyrosine (*tyrA*), phenylalanine (*pheA*), tryptophan (*trpB*), leucine (*leuB*), lysine (*lysA*), and isoleucine (*ilvA*). These amino acids were chosen based on three criteria: (1) these amino acids cannot be catabolized and utilized as carbon source by *E. coli*. Tryptophan was the exception, which can be partially degraded to pyruvate and indole (indole cannot be further degraded) [284]. (2) No other cellular functions besides protein synthesis is known for these seven amino acids. For example, *E. coli* cannot degrade methionine, but can utilize it also as a precursor for S-adenosyl-L-methionine (SAM), the major methyl group donor in the cell. (3) No other reaction is known, with which *E. coli* can transform another metabolite into the focal amino acid

[196]. The above criteria were applied to exclude unwanted confounding effects to influence the growth kinetics.

For each amino acid, six *E. coli* BW25113 wild type colonies and six colonies of the corresponding auxotrophic genotype were used to inoculate 1 ml overnight cultures (16 h) with fructose as carbon source. The media used to cultivate auxotrophic strains was supplemented with the amino acid the focal auxotroph required to grow (see Table S1 for exact amino acid concentrations). Each of these cultures was used to inoculate eight 1 ml pre-cultures (96-deep-well plates, Eppendorf, Germany), each containing one of the eight different carbon sources (i.e. fructose, maltose, xylose, glycerol, pyruvate, lactate, succinate, and malate) and the focal amino acid (see Table S1 for exact amino acid concentrations). Precultures were incubated for 26 h at 30 °C under shaking conditions (220 rpm).

To test whether the maximum growth rates of prototrophic wild type cells was sensitive to increasing amino acid concentration in the growth medium, wild type precultures were used to inoculate 50 µl cultures in 384-well plates (flat bottom and transparent, Greiner Bio-One, Kremsmünster, Austria) with an initial cell density of 10^5 colony-forming units (CFUs) mL^{-1}. The MMAB medium used for these experiments contained the same carbon source as the preculture, yet in addition one of eight different concentrations of the focal amino acid, with the lowest level corresponding to no amino acid supplementation (see Table S1 for exact amino acid concentrations). In this way, each of the 64 combinations of eight carbon sources and eight amino acid concentrations was independently replicated six times. A second 384-well plate with the exact same media layout was inoculated accordingly from the precultures of auxotrophic genotypes. Wells without amino acid supplementation were inoculated with the wild type strain as control. Growth kinetics were determined in a Tecan Infinite 200 Pro plate reader (Tecan Group, Männedorf, Switzerland) for automated kinetic measurements for 48 h at 30 °C and a 10 min kinetic cycle consisting of 7.5 minutes of orbital shaking (2 mm amplitude), 1 min waiting (no shaking), and 1.5 minutes for measuring the optical density at 600 nm (OD_{600nm}, 10 nm bandwidth) with 5 flashes.

5.5. *Statistical data analysis*

For each culture of the growth kinetic experiment, the maximum growth rate μ_{max} was determined. Since *E. coli* reaches substantially different maximum growth rates in the eight different carbon source regimes and to compare the increase of μ_{max} with increasing amino acid concentration, the μ_{max} values were normalized by the median of the maximum growth rates the wild type strain achieved under the same carbon condition without amino acid supplementation. Hereafter, we will refer to the normalized maximum growth rates as $\bar{\mu}_{max}$. The increase of the growth rate per µM of the focal amino acid (7 data points for the auxotrophic genotypes, 8 for the wild type strain) was calculated for each cognate population (i.e. populations which originated from the same clonal colony) and for each carbon source by linear regression [223,285].

For correlation analysis between increases of growth rates with either amino acid concentration or with predicted metabolic costs, a linear mixed-effects model was fitted considering the '*cognate population identity*' as random effect and the '*amino acid concentration*' or '*predicted metabolic costs*', respectively, as fixed effects. Models were fitted by maximizing the restricted log-likelihood until convergence. Conditional R^2 values of the fitted models were calculated according to [286].

Principle component analysis (PCA) and k-means clustering were performed to analyze the variance of biosynthetic costs of amino acids under various carbon sources. Only the main PCA axes, which together explained more than 90% of the observed variation, were used for k-means clustering. The algorithm by [287], for k-means clustering was applied starting with 25 random initial sets and optimization (minimizing within-group sum of squares) until convergence. P-values were corrected after multiple testing using the false discovery rate (FDR) procedure of [222]). All statistical analyses were using the *R* software (version 3.1.1) [223].

Acknowledgements

We thank the Experimental Ecology and Evolution Research Group, the Theoretical Systems Biology Research Group, Holger Merker, Gabriele Diekert, and Wilhelm Boland for constructive criticism and support.

Chapter VI

General discussion

1. Reductive evolution of metabolic networks

1.1. Metabolic causes of the fitness advantages of biosynthetic gene loss

A major outcome of this study was that biosynthetic gene deletion mutations provide the corresponding genotypes with selective advantages in environments, where the focal metabolite is available. The selective advantages may explain the prevalent reductive evolution of metabolic networks, but what causes the significant fitness advantages of auxotrophs?

Deleting different metabolic genes within the same biosynthetic pathway showed that the highest fitness benefits was gained when the genes of those enzymes were deleted, which account for the highest protein mass of all enzymes involved in the pathway (chapter III). This observation is in agreement with previous studies, which suggest that bacterial growth rate can be limited by protein costs [228–230]. Furthermore, evolution experiments with *E. coli* have shown that the increase in growth rate is often associated with the reduction of proteins, which are not, or in less abundancy, required under the specific selection environment [288,289]. These results indicate that the costs, which are associated with the production of proteins can limit the fitness of bacteria. Interestingly, 57% of the complete protein mass within an *Escherichia coli* cell are proteins involved in metabolism [290,291], which highlights the immense resources a cell needs to invest into its metabolic network.

Another factor, which provides auxotrophic genotypes with selective advantages are the resources, which are saved by auxotrophs by not transforming resources into the focal metabolite. These so-termed *metabolic-* or *biosynthetic costs* were subject of the study presented in chapter V, where it was computationally and empirically shown that these costs can also limit the growth rate of *E. coli*. Several previous studies have collected empirical evidence that these costs govern the evolution of amino acid usage in bacterial proteomes [62,61,71,65]. Additionally, due to the strong impact on the growth of bacteria, which was identified in this study (chapter V), biosynthetic costs are likely to affect also the evolution of the metabolic network structure, including the loss of biosynthetic capabilities.

Other metabolic causes of the observed fitness advantages of biosynthetic gene loss possibly include *regulatory costs* and *DNA costs*. *Regulatory costs* denote the resources, which are invested in the coordination of metabolic fluxes, e.g. as a results of expression of transcription factors or the phosphorylation of enzymes to control their catalytic activity [292]. *DNA costs* summarise the costs involved in the maintenance of the nucleotide sequence of the corresponding gene in the chromosome [293].

Presumably, all of these abovementioned factors contribute to the metabolic burden of a specific biosynthetic function of a metabolic network. However, to what extent each factor might be responsible for the selective advantages of biosynthetic gene loss in environments, in which the focal metabolite is available, remains difficult to quantify. In addition, their effect probably also varies among different microbial species and, as shown in chapter V, even for the same species but between different environmental conditions such as the available carbon source.

1.2. The pan-metabolic network and metabolic complementarity

Another significant finding of this research project was that biosynthetic functions for seemingly essential metabolites are frequently absent in the metabolic networks of a majority (i.e. 76%) of analysed eubacterial genomes. These results are in line with two previous studies, which compared the metabolic networks of 50 and 55 different *E. coli* and *Shigella*, respectively, strains [233,294]. These strains are closely related (all belong to the same family of Enterobacteriaceae) and are gut-dwelling bacteria. In contrast to their close relatedness and similar lifestyles, their metabolic capabilities are fairly diverse: Only 70% of all reactions are part of all analysed metabolic networks – termed the *core-metabolic network* [233]. This means, that 30% of reactions of the *pan-metabolic network* – the union of all reactions found in the analysed networks – are absent in at least one of the strains. Interestingly, this includes 30% of all reactions involved in amino acid metabolism, which renders some of the strains auxotrophic [233]. Taken together, the systematic comparison of bacterial metabolic networks in this and in other studies revealed that auxotrophic bacterial species are prevalent in nature and that even the biosynthetic capabilities of strains of seemingly the same species are highly diverse.

Metabolic complementarity

The observed prevalence of auxotrophies raises the questions: What is the source of essential metabolites for auxotrophic genotypes? Intuitively, the source can be explained for a range of bacteria by their nutritional-rich environments, e.g. the environments of endosymbionts or gut-dwelling bacteria. However, the auxotrophy predictions (chapter III) suggest, that also free-living bacteria are frequently auxotrophic, including bacteria living in aquatic environments, which are often poor in organic compounds [295]. In fact, auxotrophic bacterial strains have been isolated for example from freshwater lakes [296]. A plausible explanation is that essential metabolites are provided by other cells within the microbial community [83]. The Black Queen Hypothesis (see chapter I, Box 1) for example proposes, that not all essential metabolic functions need to be performed by all microbial cells within the community. Instead, the community-wide demand for the metabolic function's product may be covered by a certain fraction of cells within the community, which still performs the functions and releases its product (in part) as public good [83]. The results presented in chapter III support this hypothesis by showing that the loss of biosynthetic genes provides the auxotrophic genotypes with selective advantages and thereby may govern the adaptive evolution of metabolic function loss within microbial communities.

Nonetheless, if the loss of biosynthetic genes cause strong fitness advantages and if the metabolic demand for the focal metabolite can be covered by the production of other community members, why do most isolates have only a few auxotrophies and did not lose more biosynthetic functions? In contrast, the prediction of auxotrophies for free-living organisms showed, that most bacteria have lost the ability to produce only one or two metabolites (chapter III, Figure 1). Also other studies report only few auxotrophies per strain of free-living bacteria [233,296]. A possible explanation is the epistasis of diminishing returns in fitness of multiple biosynthetic gene deletion mutations. The results presented in chapter IV show that, in most cases, the selective advantage of biosynthetic gene loss is diminished if the mutant is already auxotrophic for another, or even two other, metabolite(s) due to prior gene deletion mutation(s) (chapter IV, Figure 2A). In other words, the highest fitness benefits are usually gained upon the first auxotrophy-causing mutations. The consequence for a community of bacteria is that the distribution of multiple auxotrophies to different sub-groups may result in a higher average fitness gain within the whole community. In line with this theory are the observations of a recent study by Garcia *et al.* (2015), which show that metabolic pathways for the biosynthesis of amino acids and vitamins are distributed across different dominant members of a freshwater microbial community [296].

Taken together, selective fitness advantages of biosynthetic gene loss and epistasis are likely to have a strong impact on the evolution of metabolic networks. More specifically, the fitness consequences of metabolic gene loss may explain also the sub-division of metabolic tasks, hence, metabolic complementarity in microbial communities and the formation of pan-metabolic networks.

1.3. *Trade-off in metabolite production costs may explain epistasis between multiple biosynthetic gene deletions*

Using a synthetic biology approach, it was shown that epistasis has a strong impact on the fitness of *E. coli* after losing multiple metabolic genes (chapter IV). Epistasis had a significant impact of the fitness of 50% percent of double- and triple gene deletion mutants with fructose as the sole carbon source and even 77% in the succinate-containing carbon environment (chapter IV, Table 1). What are the functional associations between biosynthetic gene deletion mutations that cause the observed strong epistasis?

Epistatic interactions among metabolic genes have been predicted by Joshi *et al.* (2014) for three different microorganisms, including *E. coli*, based on the structure of the underlying metabolic networks and flux balance analysis [88]. Interestingly and in line with the results presented here, the computational analysis has also predicted a strong influence of the type of carbon source, which is provided for growth, on epistasis [88]. Joshi *et al.* (2014) concluded that epistasis among metabolic genes is not necessarily due to direct interactions between the genes (e.g. the genes are targeted by the same transcription factor), but can also be explained by the metabolic fluxes through the metabolic network. Hence, the epistasis-causing functional association among metabolic genes can be the biochemical pathways that carry carbon fluxes between both corresponding reactions, e.g. if both reactions are part of the same pathway [88].

However, here, epistasis was observed among metabolic genes, which encode enzymes of distinct biosynthetic pathways for different amino acids. Thus, the metabolic flux though one reaction is not further passed on to the reaction of the other biosynthetic pathway and vice versa. How can the epistatic interactions among metabolic genes of distant pathways be functionally explained? One hypothesis is that metabolic trade-offs between different biosynthetic functions are responsible for the epistasis: In chapter V, a metabolic trade-off between the costs to produce amino acids was identified. That is, amino acids, which are synthesized from precursors from glycolysis or the pentose phosphate pathway, induce lower metabolic costs if a glycolytic carbon source (e.g. fructose) is provided for growth than if a gluconeogenic carbon source (e.g. succinate) is provided (chapter V, Figure 1). In contrast, the metabolic cost differences are reciprocal between the two groups of carbon sources (glycolytic and gluconeogenic) for amino acid, whose precursor metabolites are part of the tricarboxylic acid (TCA) cycle (chapter V, Figure 1D). This trade-off among biosynthetic costs of amino acids under different carbon environments depicts possible interactions among biosynthetic pathways, which are due to the distribution of carbon fluxes through the metabolic network. For instance, the deletion of a metabolic gene involved in an amino acid biosynthetic pathway may increase the availability of the precursor metabolite which possibly also serves as precursor for another amino acid. Hence, the redistribution of fluxes upon a biosynthetic gene loss may affect the fitness consequences of a second biosynthetic gene deletion mutation.

Taken together, the strategic allocation of a limiting carbon resource to all required metabolic pathways is a functional association, even between distant reactions (in terms of the metabolic network topology), which may account for epistatic interactions among metabolic gene deletion mutations.

1.4. Resource efficiency may govern the evolution of metabolic cross-feeding through adaptive gene loss

Cooperative exchange of metabolites plays a crucial role in the ecology of many microorganisms and is, for example, essentially involved in the assembly and disassembly of bacterial biofilms [297], the degradation of organic material in microbial communities [119,298], and metabolic cross-feeding interactions [76,298]. A fundamental question in evolution of cooperative metabolic interactions is: Why does an organism exhibit a costly metabolic function to benefit another species and, at the same time, coercively depends on actions performed by the partner, instead of accomplishing all necessary metabolic functions autonomously?

In the case of obligate metabolite cross-feeding interactions, two strains exchange metabolites, which the respective other strain cannot synthesise on its own [76,84]. As a consequence, both strains save biosynthetic costs by not synthesising the metabolite that the respective partner strain is providing, but also invest resources in the biosynthesis of another metabolite to cover the demand of the partner. How can the selective advantages of the division of metabolic labour be explained? It has been proposed, that biochemical trade-offs within the metabolism of microorganisms have a strong impact on metabolic specialisation [9] and the evolution of cooperation [298,299].

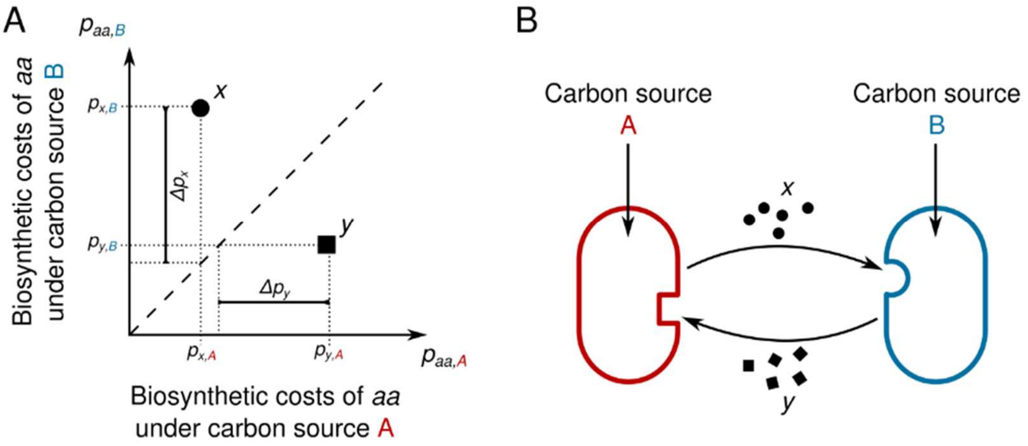


Figure 1. Carbon source(cs)-dependent trade-off between biosynthetic costs of different amino acids (aa) may promote cross-feeding interactions. (A) The plot schematically shows the biosynthetic costs $p_{aa,cs}$ (see chapter V) of two amino acids *x* and *y* under two different carbon sources A and B. The dashed line represents the isocline, at which biosynthetic costs of metabolites would be the same under both carbon sources, thus are carbon source-independent. The costs differences between the carbon sources are specified by Δp_{aa}. (B) The scheme illustrates the metabolic cross-feeding of two strains, which utilise each a different carbon source. Each strain produces the specific amino acid, which the strain produces more efficiently (lower biosynthetic costs) than the partner strain and shares the produced amino acids with their respective partner.

Here, evidence for one of such biochemical trade-offs in metabolism was found (Chapter V): Depending on the type of carbon source a bacterium utilises, the metabolic costs to produce an amino acid *x* may be low under carbon source A and high under carbon source B, whereas the costs for amino acid *y* can be high under carbon source A and low under B (Figure 1A). Assuming an environment in which multiple carbon sources, including A and B, are available and two different strains have specialised in the utilisation of carbon source A or B, respectively, the two strains will reciprocally differ in the biosynthetic costs to produce the amino acids *x* and *y* (Figure 1A). By mutually exchanging the respective amino acids, which each strain produced at lower biosynthetic costs than the other strain – i.e. the strain utilising A produces x and the B-utilising strain produces y – both strains may gain a benefit by saving limiting carbon resources (Figure 1B).

It has to be noted that such a scenario assumes that the strains utilise only one carbon source, even if multiple are present in the environment. However, it seems a common pattern in the bacterial domain that the organisms first exploit one carbon source, until that resource is (nearly) exhausted, before the cells switch to an alternative substrate [300,119,117,9].

Taken together, mutualistic interactions that are based on the exchange of metabolites between bacterial cells or between bacteria and multicellular organisms could evolve due to structure of their metabolic networks, which determine the biosynthetic costs of metabolites. Thus, the method to estimate biosynthetic costs, which was described in chapter V, could also be extended to predict the evolution of metabolic interactions between species. Therefore, the biosynthetic cost predictions could be expanded to a wide range of metabolic networks of different microbial species in order to compare the biosynthetic costs between species and their specific carbon source preferences. In this way, it is possible to forecast which metabolites are likely to be exchanged due to the highest mutual benefits both partners would gain from the specific trade [301]. Promising approaches to test these ecological predictions are to evaluate the performance of synthetically engineered consortia of different microbial strains or co-evolution experiments of co-cultures with strains, which differ in their metabolic costs to synthesise the same metabolites (see also chapter I, section 2.2).

1.5. Microbial un-culturability

Several ambitious studies aimed to extrapolate archaeal and bacterial diversity, and estimated the number of species to range in the millions [302,303]. Despite the certainty about the existence of numerous microbial species, there is a gaping lack of knowledge on the ecological characteristics of a majority of microorganisms including the metabolic niches, which the focal species in a microbial community occupy [304]. This scarce functional information of the so-called *microbial dark matter* is mainly due to our inability to cultivate the majority of microorganisms in the laboratory [303].

It has been proposed that the difficulty to mimic the physical, chemical, and biological conditions of the environment, which a microorganism is adapted to, accounts at least partly for the prevalent unculturability [305]. The analysis of hundreds of eubacterial metabolic networks in chapter III revealed that the nutritional requirements of most bacteria involve beyond energy- and bioelement resources also amino acids, nucleotides, and vitamins. These essential compounds could for example be obtained from a pool of externally supplemented and freely available metabolites in the environment. In fact, most isolated bacteria which can be cultured under laboratory conditions require rich media, which contain a wide range of organic compounds including amino acids, nucleotides, and vitamins [306]. This observation suggests that these bacteria are auxotrophic for one or more metabolites. However, missing biosynthetic capabilities (i.e. auxotrophies) are not always necessarily compensated by externally available chemicals, because it requires additionally specific transporters to import exogenously available metabolites into the cytosol across the cell membrane [307]. As discussed in the section 1.2, an alternative source for metabolites can also be neighbouring cells, which synthesise and provide the compounds.

A recent study has shown that a consortia of obligate cross-feeding auxotrophic bacteria can exchange metabolites via intercellular nanotubes [308]. These are membrane-derived structures that bacteria use for direct cell-cell connections [308]. Until now, there is no broad survey of the ubiquity of such contact-dependent mechanisms for metabolite exchange in natural microbial communities. However, several observations and evolutionary theory provide evidence that metabolite cross-feeding via physical contact is common in nature: First, auxotrophies are prevalent in free-living bacteria (chapter III) whose habitats are presumably poor in organic compounds, e.g. in marine environments. This implies that the essential metabolites are obtained from co-occurring organisms [296]. Second, if the focal compounds are obtained via nanotubes from neighbouring cells, no additional genes for transporters, which import the extracellular metabolites, are required. Third, contact-dependent metabolite exchange reduces the risk of metabolite loss due to diffusion [309], and fourth, the clustering of cells, which metabolically cooperate via cross-feeding protects cooperators of the invasion of interaction defectors. Interaction defectors are individuals, which gain benefits from the interactions (uptake of metabolites), but do not contribute to their production [308].

As a consequence, an organism which coercively relies on the acquisition of metabolites by contact-dependent mechanisms would fail to grow in pure cultures isolated from their interaction partners, no matter how nutrient-rich the cultivation media is. Thus, metabolic interdependencies and complementarities between organisms may also play a major role for the vast diversity and extensive un-culturability of bacteria [310]. In fact, there is a growing number of studies, where the growth of previous unculturable-categorised bacteria was facilitated by the cultivation of defined microbial communities consisting of the natural ecological interaction partner [296,311] or with other specific microorganism, which also were able to complement the metabolic needs of the organism of interest [310,312–314].

2. The relationship between microbial growth and metabolism

How microorganisms exploit available resources in order to grow is unquestionably tightly linked with the organism's fitness. Hence, from an evolutionary perspective, an important question to ask is: What factors limit microbial growth [73]? In 1942, Jacques Monod showed that the bacterial growth rate can be limited by the concentration of substrate, namely the carbon source, when substrate levels are low, but the growth rate does not significantly increase further with increasing substrate availability when resource levels are already high [315]. Based on these empirical results, Monod formulated following equation to describe the relationship between bacterial growth rate and substrate concentration [315]:

$$\mu = \mu_{max} \left(\frac{S}{K_S + S}\right)$$

Where μ is the specific growth rate of the bacterial population and μ_{max} is the maximum growth rate the organism can possibly reach if substrate concentration S is not limiting growth. K_S is the substrate concentration, at which the organism reaches half of its maximum growth rate μ_{max}. Note that this equation has the same form as the Michaelis-Menten-kinetics, which describes the rate of an enzymatic reaction as a function of the substrate concentration.

But what limits growth if not the nutrient availability or other external conditions? It has been shown that the substrate concentration dependency of bacterial growth can be related to the physiological ability of cells to scavenge and import resources (e.g. [316,317]). Other studies in turn report that the growth rate is not necessarily limited by the substrate transport capacity [318–320]. Alternatively, it has been proposed, that besides the availability of resources, also the strategic allocation of resources may reflect the maximal bacterial growth rate and growth yield [60,98]. In particular, the structure of the metabolic network and how the network constraints the distribution of metabolic fluxes to optimally allocate resources to biosynthetic pathways can reflect the growth rate of bacteria [46,130,131]. In chapter V it was shown that biosynthetic functions require different amounts of resources under different carbon environments to produce the same amounts of the biosynthetic product. Hence, cells need to readjust their resource allocation strategy upon a shift between chemically different substrates to form biomass in the most resource-efficient way. Any mechanism which relaxes the resource allocation problem, for example by the uptake of exogenously available amino acids (see chapter III), may lead to an increase in growth [60].

A comprehensive understanding of the relationship between growth and metabolism can help to find new methods to control microbial growth, which is essential in many practical instances, for example in medicine and biotechnology.

2.1. Inhibiting microbial growth

In cases where microorganisms are sought to be inhibited in growth or even sought to be killed, antibiotics – chemical compounds, which interfere with the reproduction and/or survival of the organism – are widely used. However, there is an increasing number of cases where bacteria evolved resistance to a wide range of available antibiotic drugs and this *antibiotic resistance crisis* [321] became one of the most pressing issues in medicine. Therefore, new antimicrobial agents and strategies to inhibit bacterial growth are pursued. To date, most antibiotics target to interrupt prokaryotic macromolecule polymerisation: DNA replication, RNA-, protein-, and cell wall synthesis [322]. Only a few antibiotics have been developed, which inhibit or damage metabolic enzymes [323]. The limited number of exploited antimicrobial metabolic targets is likely due to redundant metabolic pathways, the ability of prokaryotic pathogens to utilize a wide range of host nutrients, and too similar properties of a set of the pathogen's and host's metabolic enzymes, which might cause harmful side-effects to the host [323]. However, the reconstruction of bacterial metabolic networks (chapter II) and a subsequent structural and functional *in silico* analysis of the networks (chapter III – V) can serve as a platform to develop new antimicrobial drugs by identifying species-specific targets.

Particularly, a detailed mechanistic understanding of how a prokaryotic pathogen has metabolically adapted to the host environment can facilitate to predict enzymatic targets for a medical intervention strategy that could disrupt the integrity of the entire microbial metabolic network. Such a strategy can for example aim to disrupt the economic resource allocation (see chapter V) of a microbial cell in order reduce the available resources for cell growth.

The polyol xylitol, for example, is widely used for the prevention of caries as it inhibits the growth of *Streptococcus mutans* [324]. The growth inhibition is due to two mechanisms: First, xylitol is taken up by *S. mutans* and accumulates as xylitol-5-phosphate, which inhibits enzymes involved in glycolysis [325]. Second, the accumulated xylitol-5-phosphate is partially dephosphorylated to intracellular xylitol, which is further exported from the cell and taken up again by a phosphoenolpyruvate(PEP)-dependent phosphotransferase [325,326]. This import-export of xylitol creates a futile cycle at the expense of PEP and ultimately also depletes the energy pool of the cell [326]. This example shows, that a targeted impairment of the bacterial cell's resource economy can help to inhibit the growth of disease-causing bacteria.

2.2. Promoting microbial growth

In biotechnology, microorganisms are frequently used to produce value-added chemicals ranging from small molecules such as amino acids or vitamins to large polymer molecules, such as proteins. In these situations, an enhancement of microbial growth is usually desired. Metabolic engineering is an approach to rationally design microbial genotypes by linking growth to the overproduction of a desired compound [281,327]. Prominent strain design strategies are the knock-out of metabolic genes and the addition of new reactions or whole pathways to the metabolic network via the insertion of the corresponding enzymes-coding genes [328]. Several computational tools exist, which predict reaction knock-outs or additions to improve production yields [151,329–331]. However, these methods require a pre-defined description of the nutritional composition of the production media. The results presented in chapter IV and V clearly show that the effect of reaction knock-out mutations on the organism's growth rate, fitness, and metabolite production costs are strongly affected by the chemical nature of the provided carbon source. Specifically, this means that an organisms may need to invest a higher proportion of its available carbon resource under one carbon source than under a chemically different carbon source to produce the same amount of a specific metabolite. Hence, metabolic engineering strategies for optimal microbial strain design need to go hand in hand with a rational design of the production 'environment' to reduce resource costs and optimise production yields. This is because, reducing the carbon source requirement for the production of the desired compound may significantly reduce the economic costs. In fact, the estimated microbial biosynthetic costs of amino acids (chapter V) reflect also current market prices of these metabolites (Figure 2). The correlation is most likely due to the fact that bacteria such as *Escherichia coli* and *Corynebacterium glutamicum* are typically used to commercially produce amino acids [332] and economic costs for the growth substrate probably account to a large fraction for the market prices.

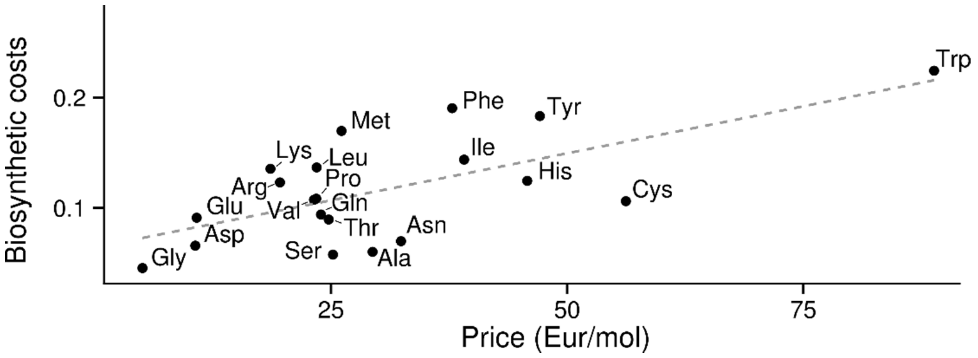


Figure 2. Correlation of amino acid catalogue prices with estimated biosynthetic costs in *E. coli*. Prices for amino acids in Euro per mole were calculated based on the purchase prices from the online catalogue of Carl Roth GmbH + Co. KG, Karlsruhe (*http://www.carlroth.com*; retrieved on 7th September 2015). The biosynthetic costs estimates correspond to the predictions from chapter V for glucose as carbon source. Pearson's product-moment correlation, $r = 0.66$, $P = 0.002$, $n = 20$.

In summary, it is beyond dispute, that metabolism is tightly interwoven with cell and population growth [126]. However, due to the complexity of metabolic networks and their regulation, is has been difficult to elucidate how different structural and functional properties of metabolic networks contribute to the microbial growth kinetics. The results on the interplay between different carbon sources and loss-of-biosynthetic-function mutations presented in this thesis (chapter III – V) revealed that resource-efficiency is a major criteria for the metabolic adaptation of bacteria to improve their ability to grow.

3. The evolution of autocatalytic reaction cycles in metabolic networks

In chapter II, the concept of autocatalytic metabolites was introduced. These compounds are part of the metabolic network and are required for their own biosynthesis. Interestingly, the number of autocatalytic metabolites within metabolic networks is usually small and the compounds, which have the autocatalytic property are highly conserved (typically ATP, NAD, and coenzyme A) across different species; including multicellular organisms [30]. This fact has been used to unravel inconsistencies in genome-scale metabolic network reconstructions (chapter II).

The strong conservation of the set of autocatalytic cycles within metabolic networks raises the question about their evolution and whether natural selection has favoured the small number and the identity of autocatalytic compounds within metabolic networks [31]. Interestingly, the most common autocatalytic metabolites within metabolic networks (i.e. ATP, NAD, and coenzyme A) are involved in many enzymatic reactions within the networks – so-called *metabolic hubs* – even though most metabolites are consumed or produced only by a few enzymes [29]. This connectivity distribution, which follows the power law, deviated significantly from random networks [29] and the evolutionary processes, which led to this connectivity distribution are highly debated (see e.g. [96]). Jeong *et al.* (2000), for example, argued that this network feature might be adaptive as it increases (compared to random reaction networks) the network's robustness against mutations, which lead to the removal of reactions [29]. In contrast, Pfeiffer *et al.* (2005) showed using computer simulations of network evolution that the high connectivity of a few metabolites could indirectly be the results of selection for growth rate in the early evolution of metabolic networks [333].

The relation between autocatalytic cycles and the connectivity of metabolic hubs remains elusive and needs further theoretical and experimental analysis; especially to elucidate whether or not natural selection has favoured these traits. However, both network features, autocatalytic sets and metabolic hubs, likely emerged during the early evolution of metabolic networks [32,333,334], which could explain the strong conservation across all domains of life.

4. Concluding remarks

During the past decades, awareness of the tremendous bacterial genetic and metabolic diversity has grown [13]. Especially in ecological, medical, and biotechnological contexts there is a growing need to understand the factors that govern bacterial adaptation to various environments and thus cause this multiplicity. The ability to exploit available resources and to use them for cell growth and reproduction has thereby a strong impact on the adaptive evolution of bacteria [335]. The metabolic network has central role for the utilisation of resources and consists of hundreds to thousands of different enzymatic reactions.

In this thesis, I employed data mining approaches to compare the metabolic capabilities of more than 900 bacterial metabolic networks and flux balance analysis to address the question: How do metabolic networks evolve? More specifically, I focussed on the evolutionary factors that may govern the reductive evolution of metabolic networks. In short, the results presented here (i) indicate that the loss of metabolic capabilities to produce central metabolites such as amino acids and vitamins is much more common among bacteria than initially assumed, (ii) provide evidence that the prevalent loss of biosynthetic capabilities in bacteria can be explained by selective advantages gained by auxotrophic genotypes, (iii) provide a first report on carbon source-dependent epistatic effects between mutations, which disrupt different biosynthetic pathways in *E. coli*, (iv) revealed the structure of the metabolic network as major determinant of metabolite production costs, and (v) provide a new computational method that helps to find and correct flaws in genome-scale metabolic network reconstructions by identifying autocatalytic cycles within the network.

Due to the intrinsic role of metabolism for the evolutionary fate of a bacterial species, adaptive evolution of metabolic networks may explain complex microbiological phenomena, whose evolutionary origin is so far unclear. The results presented in this thesis give plausible explanations for some of these phenomena including the unculturability of a majority or know species, the tremendous genetic and metabolic diversity of bacteria, the evolution of cooperative interactions between microorganisms, the plasticity of bacterial genomes, and metabolic complementarity of organisms within natural microbial communities.

Acknowledgements

First of all, I would like to thank Christoph Kaleta and Christian Kost. You guided this project with impressive scientific knowledge and creativity. You introduced me into a fascinating world of science and taught me how to address a biological question and how to generate new ones. Thank you for everything!

Thank you endlessly Holger Merker and Samay Pande! When I first started to do real laboratory work as a basically pure theoretical bioinformatician, you helped me tirelessly and showed me how to conduct meaningful experiments. I am so grateful that I had the chance to learn from you two ingenious scientists.

Glen D'Souza, Thank you for being the best collaborator and a real friend! I learned so much from you about evolution, microbiology and Indian cuisine. I will never forget how much you have contributed to this project.

Dear Prof. Gabriele Diekert, the committee meetings were always very helpful and have opened new directions for my project. Thank you very much for your ideas and support!

Dear Prof. Wilhelm Boland and Prof. Stefan Schuster, Thank you for your support and constructive criticism.

Shraddha Shitut, Anne-Kathrin Dietel, Daniel Preußger, Lisa Freund, Sandra Scholz, Ariane Zander, Muhammad Atigur Rahman, Ghada Mohamed Yousif Abdelsalam, Tina Schiklang, Samay Pande, Holger Merker, and Glen D'Souza. You are the best colleagues I can imagine! Thank you for all the discussions we have had; they helped me immensely. But there are many more things I thank you for: cake, institute terrace parties, lab gatherings, crazy ideas, water fights, sweets, Indian sweets; You are the best! Shraddha, Thank you also for all the nonsense!

Johannes Zimmerman, Juliana Gebauer, Jan Ewald, Peter Großmann, Martin Kötzing, Martin Bartl, and Sascha Schäuble; the view was only the second best thing about our

offices in the tower. I enjoyed working with you every minute. Thank you for all the interesting discussions we have had. And for all the coffee.

Hea Reung ‘Sindy’ Park, Juliane Fischer, Hella Schmidt, Katharina Eick and Alessio Garrone. Thank you for all the fun time we had, even during busy PhD-work-loaded times. And thank you for your patience with my tardiness.

Prof. Martin Kaltenpoth, Hassan Salem, Aileen Berasategui, Sailendharan Sudakaran, Laura Victoria Flórez Patino (I didn’t know your name is that long), Tobias Engl, Peter Biedermann, Shantanu Shukla, Thomas Ogao Onchuru, Eric Tang, and Taras Nechitaylo; Thank you for all the great joint group events! They were really productive (some of the ideas produced there are part of this thesis) and great fun. I hope that there will be anther round of *Bang!* soon.

Marie Vasse; So glad I know you. Thank you for our discussions, for the books, music and the places you have shown me.

Danke, Angelika Berg, für all deine Hilfe im Labor. Ohne dich hätte vieles nicht funktioniert.

Sabine Mentzel, Grit Winnefeld, Katja Präfke, Kathrin Schowtka, Carsten Thoms, Jan Büllesbach und Rebecca Schmückling; jedes Mal kam ich mit einem Problem zu Euch, und jedes Mal bin ich mit einer Lösung gegangen. Vielen lieben Dank für Eure Hilfe und Unterstützung!

Teresa Lehnert und Christian Dreßler. Ihr beide wart von Anfang an dabei und habt mir geholfen auch mal wieder auf andere Gedanken zu kommen. Danke dafür und für die wunderbare Auszeit, die wir fast jeden Freitag zusammen hatten.

Michael Seiler. Du hast dich immer dafür interessiert, woran ich arbeite und hast immer wieder großartige Ideen. Danke auch, dass du mir das Jonglieren beigebracht hast.

Christina Lange, vielen lieben Dank für deine Hilfe! Und deinen Schreibtisch.

Liebe Mutti, lieber Papa, lieber Daniel, eure grenzenlose Unterstützung und euer Zuspruch haben es erst ermöglicht, dass ich diese Arbeit schreiben konnte oder ich es mir überhaupt vorstellen konnte sie anzufangen. Ich bin euch unendlich dankbar. Ich liebe euch!

References

1 Kirschner M & Gerhart J (1998) Evolvability. *Proc. Natl. Acad. Sci.* **95**, 8420–8427.

2 Orzack S & Sober E (eds.) (2001) *Adaptationism and optimality* Cambridge University Press, Cambridge ; New York.

3 del Giorgio PA & Cole JJ (1998) Bacterial Growth Efficiency in Natural Aquatic Systems. *Annu. Rev. Ecol. Syst.* **29**, 503–541.

4 Roller BR & Schmidt TM (2015) The physiology and ecological implications of efficient growth. *ISME J.* **9**, 1481–1487.

5 Dunham MJ (2007) Synthetic ecology: a model system for cooperation. *Proc. Natl. Acad. Sci.* **104**, 1741–1742.

6 Feist AM, Herrgård MJ, Thiele I, Reed JL & Palsson BØ (2008) Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* **7**, 129–143.

7 Wagner A (2012) Metabolic Networks and Their Evolution. In *Evolutionary Systems Biology* (Soyer OS, ed), pp. 29–52. Springer New York.

8 Chubukov V, Gerosa L, Kochanowski K & Sauer U (2014) Coordination of microbial metabolism. *Nat. Rev. Microbiol.* **12**, 327–340.

9 Johnson DR, Goldschmidt F, Lilja EE & Ackermann M (2012) Metabolic specialization and the assembly of microbial communities. *ISME J.* **6**, 1985–1991.

10 Mazumdar V, Amar S & Segrè D (2013) Metabolic proximity in the order of colonization of a microbial community. *PloS One* **8**, e77617.

11 Waters CM & Bassler BL (2005) QUORUM SENSING: Cell-to-Cell Communication in Bacteria. *Annu. Rev. Cell Dev. Biol.* **21**, 319–346.

12 Carlson RP (2007) Metabolic systems cost-benefit analysis for interpreting network structure and regulation. *Bioinformatics* **23**, 1258–1264.

13 Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol* **3**, 1–0003.

14 Gottschalk G (2009) *Bacterial metabolism*, 2. ed Springer, New York.

15 Crawford IP (1989) Evolution of a biosynthetic pathway: the tryptophan paradigm. *Annu. Rev. Microbiol.* **43**, 567–600.

16 Radwanski ER (1995) Tryptophan Biosynthesis and Metabolism: Biochemical and Molecular Genetics. *Plant Cell* **7**, 921–934.

17 Matias Rodrigues JF & Wagner A (2009) Evolutionary Plasticity and Innovations in Complex Metabolic Reaction Networks. *PLoS Comput. Biol.* **5**, e1000613.

18 Samal A, Rodrigues JFM, Jost J, Martin OC & Wagner A (2010) Genotype networks in metabolic reaction spaces. *BMC Syst. Biol.* **4**, 30.

19 Heinemann M & Sauer U (2010) Systems biology of microbial metabolism. *Curr. Opin. Microbiol.* **13**, 337–343.

20 Notebaart RA, Szappanos B, Kintses B, Pál F, Györkei Á, Bogos B, Lázár V, Spohn R, Csörgő B, Wagner A, Ruppin E, Pál C & Papp B (2014) Network-level architecture and the evolutionary potential of underground metabolism. *Proc. Natl. Acad. Sci.* **111**, 11762–11767.

21 de la Cruz F & Davies J (2000) Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* **8**, 128–133.

22 Pál C, Papp B & Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* **37**, 1372–1375.

23 Toft C & Andersson SGE (2010) Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat. Rev. Genet.* **11**, 465–475.

24 Thomas GH, Zucker J, Macdonald SJ, Sorokin A, Goryanin I & Douglas AE (2009) A fragile metabolic network adapted for cooperation in the symbiotic bacterium Buchnera aphidicola. *BMC Syst. Biol.* **3**, 24.

25 González-Domenech CM, Belda E, Patiño-Navarrete R, Moya A, Peretó J & Latorre A (2012) Metabolic stasis in an ancient symbiosis: genome-scale metabolic networks from two Blattabacterium cuenoti strains, primary endosymbionts of cockroaches. *BMC Microbiol.* **12**, S5.

26 Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM & Palsson BØ (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol. Syst. Biol.* **7**, 535–535.

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

28 Drouin G, Godin J-R & Pagé B (2011) The Genetics of Vitamin C Loss in Vertebrates. *Curr. Genomics* **12**, 371–378.

29 Jeong H, Tombor B, Albert R, Oltvai ZN & Barabási AL (2000) The large-scale organization of metabolic networks. *Nature* **407**, 651–654.

30 Kun Á, Papp B & Szathmáry E (2008) Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks. *Genome Biol* **9**, 51.

31 Filisetti A, Villani M, Damiani C, Graudenzi A, Roli A, Hordijk W & Serra R (2014) On RAF sets and autocatalytic cycles in random reaction networks. In *Advances in Artificial Life and Evolutionary Computation* pp. 113–126. Springer.

32 Hordijk W, Hein J & Steel M (2010) Autocatalytic Sets and the Origin of Life. *Entropy* **12**, 1733–1742.

33 Fenchel T (2002) *The Origin and Early Evolution of Life* Oxford University Press.

34 Tanay A, Regev A & Shamir R (2005) Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast. *Proc. Natl. Acad. Sci.* **102**, 7203–7208.

35 Samal A & Jain S (2008) The regulatory network of E. coli metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response. *BMC Syst. Biol.* **2**, 21.

36 Wagner A (2013) *Robustness and Evolvability in Living Systems.* Princeton University Press.

37 Hua Q, Yang C, Baba T, Mori H & Shimizu K (2003) Responses of the Central Metabolism in Escherichia coli to Phosphoglucose Isomerase and Glucose-6-Phosphate Dehydrogenase Knockouts. *J. Bacteriol.* **185**, 7053–7067.

38 Kabir M & Shimizu K (2003) Gene expression patterns for metabolic pathway in pgi knockout Escherichia coli with and without phb genes based on RT-PCR. *J. Biotechnol.* **105**, 11–31.

39 Charusanti P, Conrad TM, Knight EM, Venkataraman K, Fong NL, Xie B, Gao Y & Palsson BØ (2010) Genetic Basis of Growth Adaptation of Escherichia coli after Deletion of pgi, a Major Metabolic Gene. *PLoS Genet.* **6**, e1001186.

40 Harcombe WR, Delaney NF, Leiby N, Klitgord N & Marx CJ (2013) The Ability of Flux Balance Analysis to Predict Evolution of Central Metabolism Scales with the Initial Distance to the Optimum. *PLoS Comput. Biol.* **9**, e1003091.

41 Elena SF & Lenski RE (2003) Microbial genetics: Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.* **4**, 457–469.

42 Philippe N, Crozat E, Lenski RE & Schneider D (2007) Evolution of global regulatory networks during a long-term experiment withEscherichia coli. *BioEssays* **29**, 846–860.

43 Goodarzi H, Bennett BD, Amini S, Reaves ML, Hottes AK, Rabinowitz JD & Tavazoie S (2010) Regulatory and metabolic rewiring during laboratory evolution of ethanol tolerance in E. coli. *Mol. Syst. Biol.* **6**.

44 Lind PA, Farr AD & Rainey PB (2015) Experimental evolution reveals hidden diversity in evolutionary pathways. *eLife* **4**, e07074.

45 Conrad TM, Lewis NE & Palsson BØ (2014) Microbial laboratory evolution in the era of genome-scale science. *Mol. Syst. Biol.* **7**, 509–509.

46 Varma A & Palsson BØ (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Appl. Environ. Microbiol.* **60**, 3724–3731.

47 Hua Q, Joyce AR, Palsson BØ & Fong SS (2007) Metabolic characterization of Escherichia coli strains adapted to growth on lactate. *Appl. Environ. Microbiol.* **73**, 4639–4647.

48 Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI & Koonin EV (2014) Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol.* **12**, 66.

49 Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108**, 583–586.

50 Yus E, Maier T, Michalodimitrakis K, van Noort V, Yamada T, Chen W-H, Wodke JAH, Guell M, Martinez S, Bourgeois R, Kuhner S, Raineri E, Letunic I, Kalinina OV, Rode M, Herrmann R, Gutierrez-Gallego R, Russell RB, Gavin A-C, Bork P & Serrano L (2009) Impact of Genome Reduction on Bacterial Metabolism and Its Regulation. *Science* **326**, 1263–1268.

51 Wagner A (2000) Robustness against mutations in genetic networks of yeast. *Nat. Genet.* **24**, 355–361.

52 Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD & Broadbelt LJ (2005) Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**, 1603–1609.

53 Masel J (2011) Genetic drift. *Curr. Biol.* **21**, R837–R838.

54 Ochman H & Moran NA (2001) Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292**, 1096–1099.

55 Moran NA & Plague GR (2004) Genornic changes following host restriction in bacteria. *Curr. Opin. Genet. Dev.* **14**, 627–633.

56 McCutcheon JP & Moran NA (2007) Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proc. Natl. Acad. Sci.* **104**, 19392–19397.

57 Nilsson AI, Koskiniemi S, Eriksson S, Kugelberg E, Hinton JCD & Andersson DI (2005) Bacterial genome size reduction by experimental evolution. *Proc. Natl. Acad. Sci.* **102**, 12112–12116.

58 Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M & others (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**, 1242–1245.

59 Reddy TBK, Thomas AD, Stamatis D, Bertsch J, Isbandi M, Jansson J, Mallajosyula J, Pagani I, Lobos EA & Kyrpides NC (2014) The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.*, gku950.

60 Goelzer A & Fromion V (2011) Bacterial growth rate reflects a bottleneck in resource allocation. *Biochim. Biophys. Acta BBA - Gen. Subj.* **1810**, 978–988.

61 Akashi H & Gojobori T (2002) Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci.* **99**, 3695–3700.

62 Craig CL & Weber RS (1998) Selection costs of amino acid substitutions in ColE 1 and colIa gene clusters harbored by Escherichia coli. *Mol. Biol. Evol.* **15**, 774–776.

63 Kaleta C, Schäuble S, Rinas U & Schuster S (2013) Metabolic costs of amino acid and protein production in *Escherichia coli. Biotechnol. J.* **8**, 1105–1114.

64 Seligmann H (2003) Cost-minimization of amino acid usage. *J. Mol. Evol.* **56**, 151–161.

65 Swire J (2007) Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. *J. Mol. Evol.* **64**, 558–571.

66 Hibbing ME, Fuqua C, Parsek MR & Peterson SB (2010) Bacterial competition: surviving and thriving in the microbial jungle. *Nat. Rev. Microbiol.* **8**, 15–25.

67 Sajitz-Hermstein M (2014) On costs and benefits of individual reactions in biochemical networks. *Dissertation*. University of Potsdam.

68 Richmond RC (1970) Non-Darwinian Evolution: A Critique. *Nature* **225**, 1025–1028.

69 King JL & Jukes TH (1969) Non-Darwinian evolution. *Science* **164**, 788–798.

70 Dayhoff MO (1979) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation.

71 Heizer EM (2006) Amino Acid Cost and Codon-Usage Biases in 6 Prokaryotic Genomes: A Whole-Genome Analysis. *Mol. Biol. Evol.* **23**, 1670–1680.

72 Pittard J & Yang J (2008) Biosynthesis of the Aromatic Amino Acids. *EcoSal Plus*.

73 Koch AL (1997) Microbial physiology and ecology of slow growth. *Microbiol. Mol. Biol. Rev.* **61**, 305–318.

74 Phelan VV, Liu W-T, Pogliano K & Dorrestein PC (2012) Microbial metabolic exchange—the chemotype-to-phenotype link. *Nat. Chem. Biol.* **8**, 26–35.

75 Foster KR & Bell T (2012) Competition, Not Cooperation, Dominates Interactions among Culturable Microbial Species. *Curr. Biol.* **22**, 1845–1850.

76 Pande S, Merker H, Bohl K, Reichelt M, Schuster S, de Figueiredo LF, Kaleta C & Kost C (2014) Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. *ISME J.* **8**, 953–962.

77 Rosenzweig RF, Sharp RR, Treves DS & Adams J (1994) Microbial evolution in a simple unstructured environment: genetic differentiation in Escherichia coli. *Genetics* **137**, 903–917.

78 Doebeli M (2002) A model for the evolutionary dynamics of cross-feeding polymorphisms in microorganisms. *Popul. Ecol.* **44**, 59–70.

79 Lukjancenko O, Wassenaar TM & Ussery DW (2010) Comparison of 61 Sequenced Escherichia coli Genomes. *Microb. Ecol.* **60**, 708–720.

80 Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL & Mori H (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008.

81 Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, Mori H, Lesely SA, Palsson BØ & Agarwalla S (2006) Experimental and computational assessment of conditionally essential genes in Escherichia coli. *J. Bacteriol.* **188**, 8259–8271.

82 Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui M El, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguenec C, Lescat M, Mangenot S, Martinez-Jehanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Saint Ruf C, Schneider D, Tourret J, Vacherie B, Vallenet D, Medigue C, Rocha

EPC & Denamur E (2009) Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. *Plos Genet.* **5**, e1000344.

83 Morris JJ, Lenski RE & Zinser ER (2012) The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio* **3**, e00036–12.

84 Wintermute EH & Silver PA (2010) Emergent cooperation in microbial metabolism. *Mol. Syst. Biol.* **6**.

85 Little AEF, Robinson CJ, Peterson SB, Raffa KF & Handelsman J (2008) Rules of Engagement: Interspecies Interactions that Regulate Microbial Communities. *Annu. Rev. Microbiol.* **62**, 375–401.

86 Maharjan RP, Ferenci T, Reeves PR, Li Y, Liu B & Wang L (2012) The multiplicity of divergence mechanisms in a single evolving population. *Genome Biol.* **13**, R41.

87 Phillips PC (2008) Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9**, 855–867.

88 Jagdishchandra Joshi C & Prasad A (2014) Epistatic interactions among metabolic genes depend upon environmental conditions. *Mol. Biosyst.* **10**, 2578.

89 He X, Qian W, Wang Z, Li Y & Zhang J (2010) Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. *Nat. Genet.* **42**, 272–276.

90 Snitkin ES & Segrè D (2011) Epistatic Interaction Maps Relative to Multiple Metabolic Phenotypes. *PLoS Genet.* **7**, e1001294.

91 Trindade S, Sousa A, Xavier KB, Dionisio F, Ferreira MG & Gordo I (2009) Positive Epistasis Drives the Acquisition of Multidrug Resistance. *Plos Genet.* **5**, e1000578.

92 da Silva J, Coetzer M, Nedellec R, Pastore C & Mosier DE (2010) Fitness Epistasis and Constraints on Adaptation in a Human Immunodeficiency Virus Type 1 Protein Region. *Genetics* **185**, 293–303.

93 Flynn KM, Cooper TF, Moore FB-G & Cooper VS (2013) The Environment Affects Epistatic Interactions to Alter the Topology of an Empirical Fitness Landscape. *Plos Genet.* **9**, e1003426.

94 Weinreich DM, Watson RA, Chao L & Harrison R (2005) Perspective: sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* **59**, 1165–1174.

95 Bataillon T, Zhang T & Kassen R (2011) Cost of Adaptation and Fitness Effects of Beneficial Mutations in Pseudomonas fluorescens. *Genetics* **189**, 939–949.

96 Papp B, Teusink B & Notebaart RA (2009) A critical view of metabolic network adaptations. *HFSP J.* **3**, 24–35.

97 Simpson GG (1944) *Tempo and mode in evolution.* Columbia University Press.

98 Bosdriesz E (2015) Darwin's invisible hand: optimality principles in cellular resource allocation. .

99 Zaslaver A, Mayo AE, Rosenberg R, Bashkin P, Sberro H, Tsalyuk M, Surette MG & Alon U (2004) Just-in-time transcription program in metabolic pathways. *Nat. Genet.* **36**, 486–491.

100 Bartl M, Kötzing M, Schuster S, Li P & Kaleta C (2013) Dynamic optimization identifies optimal programmes for pathway regulation in prokaryotes. *Nat. Commun.* **4**.

101 Ewald J, Kötzing M, Bartl M & Kaleta C (2015) Footprints of Optimal Protein Assembly Strategies in the Operonic Structure of Prokaryotes. *Metabolites* **5**, 252–269.

102 Schuster S, Dandekar T & Fell DA (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.* **17**, 53–60.

103 Schuetz R, Kuepfer L & Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Mol. Syst. Biol.* **3**.

104 Schuster S, Pfeiffer T & Fell DA (2008) Is maximization of molar yield in metabolic networks favoured by evolution? *J. Theor. Biol.* **252**, 497–504.

105 Molenaar D, van Berlo R, de Ridder D & Teusink B (2009) Shifts in growth strategies reflect tradeoffs in cellular economics. *Mol. Syst. Biol.* **5**.

106 Kimura M (1991) Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. *Proc. Natl. Acad. Sci.* **88**, 5969–5973.

107 Baumler DJ, Ma B, Reed JL & Perna NT (2013) Inferring ancient metabolism using ancestral core metabolic models of enterobacteria. *BMC Syst. Biol.* **7**, 46.

108 Forst CV, Flamm C, Hofacker IL & Stadler PF (2006) Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. *BMC Bioinformatics* **7**, 67.

109 Pál C, Papp B, Lercher MJ, Csermely P, Oliver SG & Hurst LD (2006) Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**, 667–670.

110 Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M & Tanabe M (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–D205.

111 Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Weerasinghe D, Zhang P & Karp PD (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **42**, D459–D471.

112 Kazamia E, Aldridge DC & Smith AG (2012) Synthetic ecology – A way forward for sustainable algal biofuel production? *J. Biotechnol.* **162**, 163–169.

113 Chou H-H, Chiu H-C, Delaney NF, Segre D & Marx CJ (2011) Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. *Science* **332**, 1190–1192.

114 Khan AI, Dinh DM, Schneider D, Lenski RE & Cooper TF (2011) Negative Epistasis Between Beneficial Mutations in an Evolving Bacterial Population. *Science* **332**, 1193–1196.

115 Shou W, Ram S & Vilar JM (2007) Synthetic cooperation in engineered yeast populations. *Proc. Natl. Acad. Sci.* **104**, 1877–1882.

116 Willey JM, Sherwood L, Woolverton CJ & Prescott LM (2008) *Microbiology* McGraw-Hill Higher Education, New York.

117 Blount ZD, Borland CZ & Lenski RE (2008) Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proc. Natl. Acad. Sci.* **105**, 7899–7906.

118 Ibarra RU, Edwards JS & Palsson BØ (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420**, 186–189.

119 Treves DS, Manning S & Adams J (1998) Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of Escherichia coli. *Mol. Biol. Evol.* **15**, 789–797.

120 Satterwhite RS & Cooper TF (2015) Constraints on adaptation of *Escherichia coli* to mixed-resource environments increase over time. *Evolution* **69**, 2067–2078.

121 Doebeli M & Dieckmann U (2000) Evolutionary branching and sympatric speciation caused by different types of ecological interactions. *Am. Nat.* **156**, S77–S101.

122 Friesen ML, Saxer G, Travisano M & Doebeli M (2004) Experimental evidence for sympatric ecological diversification due to frequency-dependent competition in Escherichia coli. *Evolution* **58**, 245–260.

123 Bachmann H, Fischlechner M, Rabbers I, Barfa N, Santos FB dos, Molenaar D & Teusink B (2013) Availability of public goods shapes the evolution of competing metabolic strategies. *Proc. Natl. Acad. Sci.* **110**, 14302–14307.

124 Kitano H (2002) Systems Biology: A Brief Overview. *Science* **295**, 1662–1664.

125 Price ND, Papin JA, Schilling CH & Palsson BØ (2003) Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol.* **21**, 162–169.

126 Stelling J, Klamt S, Bettenbrock K, Schuster S & Gilles ED (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420**, 190–193.

127 Schuster S & Hilgetag C (1994) On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.* **02**, 165–182.

128 Covert MW, Xiao N, Chen TJ & Karr JR (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. *Bioinformatics* **24**, 2044–2050.

129 Antoniewicz MR (2013) Dynamic metabolic flux analysis — tools for probing transient states of metabolic networks. *Curr. Opin. Biotechnol.* **24**, 973–978.

130 Segre D, Vitkup D & Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci.* **99**, 15112–15117.

131 Adadi R, Volkmer B, Milo R, Heinemann M & Shlomi T (2012) Prediction of Microbial Growth Rate versus Biomass Yield by a Metabolic Network with Kinetic Parameters. *PLoS Comput. Biol.* **8**, e1002575.

132 Lee JW, Na D, Park JM, Lee J, Choi S & Lee SY (2012) Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat. Chem. Biol.* **8**, 536–546.

133 Jung YK, Kim TY, Park SJ & Lee SY (2010) Metabolic engineering of *Escherichia coli* for the production of polylactic acid and its copolymers. *Biotechnol. Bioeng.* **105**, 161–171.

134 Lee KH, Park JH, Kim TY, Kim HU & Lee SY (2007) Systems metabolic engineering of Escherichia coli for L-threonine production. *Mol. Syst. Biol.* **3**.

135 Park JH, Lee KH, Kim TY & Lee SY (2007) Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *Proc. Natl. Acad. Sci.* **104**, 7797–7802.

136 Qian Z-G, Xia X-X & Lee SY (2009) Metabolic engineering of *Escherichia coli* for the production of putrescine, a four carbon diamine. *Biotechnol. Bioeng.* **104**, 651–662.

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

138 Teusink B, van Enckevort FHJ, Francke C, Wiersma A, Wegkamp A, Smid EJ & Siezen RJ (2005) In Silico Reconstruction of the Metabolic Pathways of Lactobacillus plantarum: Comparing Predictions of Nutrient Requirements with Those from Growth Experiments. *Appl. Environ. Microbiol.* **71**, 7253–7262.

139 Smid EJ, Molenaar D, Hugenholtz J, Vos WM de & Teusink B (2005) Functional ingredient production: application of global metabolic models. *Curr. Opin. Biotechnol.* **16**, 190–197.

140 Shen Y, Liu J, Estiu G, Isin B, Ahn Y-Y, Lee D-S, Barabasi A-L, Kapatral V, Wiest O & Oltvai ZN (2010) Blueprint for antimicrobial hit discovery targeting metabolic networks. *Proc. Natl. Acad. Sci.* **107**, 1082–1087.

141 Orth JD, Thiele I & Palsson BØ (2010) What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–248.

142 Feist AM & Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat. Biotechnol.* **26**, 659–667.

143 Schuster S, Marhl M & Höfer T (2002) Modelling of simple and complex calcium oscillations. *Eur. J. Biochem.* **269**, 1333–1355.

144 Chakrabarti A, Miskovic L, Soh KC & Hatzimanikatis V (2013) Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints. *Biotechnol. J.* **8**, 1043–1057.

145 Smallbone K, Simeonidis E, Swainston N & Mendes P (2010) Towards a genome-scale kinetic model of cellular metabolism. *BMC Syst. Biol.* **4**, 6.

146 Trinh CT, Wlaschin A & Srienc F (2009) Elementary Mode Analysis: A Useful Metabolic Pathway Analysis Tool for Characterizing Cellular Metabolism. *Appl. Microbiol. Biotechnol.* **81**, 813–826.

147 Acuña V, Marchetti-Spaccamela A, Sagot M-F & Stougie L (2010) A note on the complexity of finding and enumerating elementary modes. *Biosystems* **99**, 210–214.

148 Behre J, de Figueiredo L, Schuster S & Kaleta C (2012) Detecting Structural Invariants in Biological Reaction Networks. In *Bacterial Molecular Networks* (van Helden J, Toussaint A, & Thieffry D, eds), pp. 377–407. Springer New York.

149 Kaleta C, de Figueiredo LF, Behre J & Schuster S (2009) EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks. In *Lecture Notes in Informatics-Proceedings* pp. 179–189.

150 Figueiredo LF de, Podhorski A, Rubio A, Kaleta C, Beasley JE, Schuster S & Planes FJ (2009) Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics* **25**, 3158–3165.

151 Bohl K, de Figueiredo LF, Hädicke O, Klamt S, Kost C, Schuster S & Kaleta C (2010) CASOP GS: Computing Intervention Strategies Targeted at Production Improvement in Genome-scale Metabolic Networks. In *GCB* pp. 71–80.

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

153 Wessely F, Bartl M, Guthke R, Li P, Schuster S & Kaleta C (2011) Optimal regulatory strategies for metabolic pathways in Escherichia coli depending on protein costs. *Mol. Syst. Biol.* **7**, 515.

154 Kaleta C, de Figueiredo LF, Werner S, Guthke R, Ristow M & Schuster S (2011) In Silico Evidence for Gluconeogenesis from Fatty Acids in Humans. *Plos Comput. Biol.* **7**, e1002116.

155 Ruppin E, Papin JA, de Figueiredo LF & Schuster S (2010) Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Curr. Opin. Biotechnol.* **21**, 502–510.

156 Lewis NE, Nagarajan H & Palsson BØ (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* **10**, 291–305.

157 Wang Q, Chen X, Yang Y & Zhao X (2006) Genome-scale in silico aided metabolic analysis and flux comparisons of Escherichia coli to improve succinate production. *Appl. Microbiol. Biotechnol.* **73**, 887–894.

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

159 Barua D, Kim J & Reed JL (2010) An Automated Phenotype-Driven Approach (GeneForce) for Refining Metabolic and Regulatory Models. *Plos Comput. Biol.* **6**, e1000970.

160 Jerby L, Shlomi T & Ruppin E (2010) Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol. Syst. Biol.* **6**, 401.

161 Li P, Dada JO, Jameson D, Spasic I, Swainston N, Carroll K, Dunn W, Khan F, Malys N, Messiha HL, Simeonidis E, Weichart D, Winder C, Wishart J, Broomhead DS, Goble CA, Gaskell SJ, Kell DB, Westerhoff HV, Mendes P & Paton NW (2010) Systematic integration of experimental data and models in systems biology. *BMC Bioinformatics* **11**, 582.

162 Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B & Stevens RL (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28**, 977–U22.

163 Orth JD & Palsson BØ (2010) Systematizing the Generation of Missing Metabolic Knowledge. *Biotechnol. Bioeng.* **107**, 403–412.

164 Kumar VS, Dasika MS & Maranas CD (2007) Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* **8**, 212.

165 Latendresse M, Krummenacker M, Trupp M & Karp PD (2012) Construction and completion of flux balance models from pathway databases. *Bioinformatics* **28**, 388–396.

166 Ponce-de-Leon M, Montero F & Pereto J (2013) Solving gap metabolites and blocked reactions in genome-scale models: application to the metabolic network of Blattabacterium cuenoti. *Bmc Syst. Biol.* **7**, 114.

167 Eigen M & Schuster P (1977) The Hypercycle - Principle of Natural Self-Organization. *Naturwissenschaften* **64**, 541–565.

168 Handorf T, Ebenhoh O & Heinrich R (2005) Expanding metabolic networks: Scopes of compounds, robustness, and evolution. *J. Mol. Evol.* **61**, 498–512.

169 de Figueiredo LF, Schuster S, Kaleta C & Fell DA (2008) Can sugars be produced from fatty acids? A test case for pathway analysis tools. *Bioinformatics* **24**, 2615–2621.

170 Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M & Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**, D354–D357.

171 Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Paley S, Popescu L, Pujar A, Shearer AG, Zhang P & Karp PD (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **38**, D473–D479.

172 Zhang P, Dreher K, Karthikeyan A, Chi A, Pujar A, Caspi R, Karp P, Kirkup V, Latendresse M, Lee C, Mueller LA, Muller R & Rhee SY (2010) Creation of a Genome-Wide Metabolic Pathway Database for Populus trichocarpa Using a New Approach for Reconstruction and Curation of Metabolic Pathways for Plants. *Plant Physiol.* **153**, 1479–1491.

173 Dal'Molin CG de O, Quek L-E, Palfreyman RW, Brumbley SM & Nielsen LK (2010) AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis. *Plant Physiol.* **152**, 579–589.

174 Chang RL, Ghamsari L, Manichaikul A, Hom EFY, Balaji S, Fu W, Shen Y, Hao T, Palsson BØ, Salehi-Ashtiani K & Papin JA (2011) Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Mol. Syst. Biol.* **7**, 518.

175 Mo ML, Palsson BØ & Herrgard MJ (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *Bmc Syst. Biol.* **3**, 37.

176 Dal'Molin CG de O, Quek L-E, Palfreyman RW, Brumbley SM & Nielsen LK (2010) C4GEM, a Genome-Scale Metabolic Model to Study C-4 Plant Metabolism. *Plant Physiol.* **154**, 1871–1885.

177 Vass M, Allen N, Shaffer CA, Ramakrishnan N, Watson LT & Tyson JJ (2004) The JigCell Model Builder and Run Manager. *Bioinformatics* **20**, 3680–3681.

178 Feist AM & Palsson BØ (2010) The biomass objective function. *Curr. Opin. Microbiol.* **13**, 344–349.

179 Turner WL, Waller JC, Vanderbeld B & Snedden WA (2004) Cloning and characterization of two NAD kinases from arabidopsis. Identification of a calmodulin binding isoform. *Plant Physiol.* **135**, 1243–1255.

180 Waller JC, Dhanoa PK, Schumann U, Mullen RT & Snedden WA (2010) Subcellular and tissue localization of NAD kinases from Arabidopsis: compartmentalization of de novo NADP biosynthesis. *Planta* **231**, 305–317.

181 Lawhorn BG, Mehl RA & Begley TP (2004) Biosynthesis of the thiamin pyrimidine: the reconstitution of a remarkable rearrangement reaction. *Org. Biomol. Chem.* **2**, 2538–2546.

182 Swoboda JG, Campbell J, Meredith TC & Walker S (2010) Wall Teichoic Acid Function, Biosynthesis, and Inhibition. *Chembiochem* **11**, 35–45.

183 Inaoka T & Ochi K (2012) Undecaprenyl Pyrophosphate Involvement in Susceptibility of Bacillus subtilis to Rare Earth Elements. *J. Bacteriol.* **194**, 5632–5637.

184 Moulin M, Nguyen GTDT, Scaife MA, Smith AG & Fitzpatrick TB (2013) Analysis of Chlamydomonas thiamin metabolism in vivo reveals riboswitch plasticity. *Proc. Natl. Acad. Sci.* **110**, 14622–14627.

185 Bouvier F, Linka N, Isner J-C, Mutterer J, Weber APM & Camara B (2006) Arabidopsis SAMT1 defines a plastid transporter regulating plastid biogenesis and plant development. *Plant Cell* **18**, 3088–3105.

186 Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Marechal-Drouard L, Marshall WF, Qu L-H, Nelson DR, Sanderfoot AA, Spalding MH, Kapitonov VV, Ren Q, Ferris P, Lindquist E, Shapiro H, Lucas SM, Grimwood J, Schmutz J, Cardol P, Cerutti H, Chanfreau G, Chen C-L, Cognat V, Croft MT, Dent R, Dutcher S, Fernandez E, Fukuzawa H, Gonzalez-Ballester D, Gonzalez-Halphen D, Hallmann A, Hanikenne M, Hippler M, Inwood W, Jabbari K, Kalanon M, Kuras R, Lefebvre PA, Lemaire SD, Lobanov AV, Lohr M, Manuell A, Meier I, Mets L, Mittag M, Mittelmeier T, Moroney JV, Moseley J, Napoli C, Nedelcu AM, Niyogi K, Novoselov SV, Paulsen IT, Pazour G, Purton S, Ral J-P, Riano-Pachon DM, Riekhof W, Rymarquis L, Schroda M, Stern D, Umen J, Willows R, Wilson N, Zimmer SL, Allmer J, Balk J, Bisova K, Chen C-J, Elias M, Gendler K, Hauser C, Lamb MR, Ledford H, Long JC, Minagawa J, Page MD, Pan J, Pootakham W, Roje S, Rose A, Stahlberg E, Terauchi AM, Yang P, Ball S, Bowler C, Dieckmann CL, Gladyshev VN, Green P, Jorgensen R, Mayfield S, Mueller-Roeber B, Rajamani S, Sayre RT, Brokstein P, Dubchak I, Goodstein D, Hornick L, Huang YW, Jhaveri J, Luo Y, Martinez D, Ngau WCA, Otillar B, Poliakov A, Porter A, Szajkowski L, Werner G, Zhou K, Grigoriev IV, Rokhsar DS & Grossman AR (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **318**, 245–251.

187 Grabinska K & Palamarczyk G (2002) Dolichol biosynthesis in the yeast Saccharomyces cerevisiae: an insight into the regulatory role of farnesyl diphosphate synthase. *Fems Yeast Res.* **2**, 259–265.

188 Belanger FC, Leustek T, Chu BY & Kriz AL (1995) Evidence for the thiamine biosynthetic pathway in higher-plant plastids and its developmental regulation. *Plant Mol. Biol.* **29**, 809–821.

189 Alexandrov NN, Brover VV, Freidin S, Troukhan ME, Tatarinova TV, Zhang H, Swaller TJ, Lu Y-P, Bouck J, Flavell RB & Feldmann KA (2009) Insights into corn genes derived from large-scale cDNA sequencing. *Plant Mol. Biol.* **69**, 179–194.

190 Hanson AD & Gregory JF (2011) Folate Biosynthesis, Turnover, and Transport in Plants. In *Annual Review of Plant Biology, Vol 62* (Merchant SS, Briggs WR, & Ort D, eds), pp. 105–125.

191 Hesse H, Kreft O, Maimann S, Zeh M & Hoefgen R (2004) Current understanding of the regulation of methionine biosynthesis in plants. *J. Exp. Bot.* **55**, 1799–1808.

192 Eckermann C, Eichel J & Schroder J (2000) Plant methionine synthase: New insights into properties and expression. *Biol. Chem.* **381**, 695–703.

193 Eichel J, Gonzalez J, Hotze M, Matthews R & Schroder J (1995) Vitamin-b-12-independent methionine synthase from a higher-plant (catharanthus-roseus) - molecular characterization, regulation, heterologous expression, and enzyme properties. *Eur. J. Biochem.* **230**, 1053–1058.

194 Ravanel S, Gakiere B, Job D & Douce R (1998) The specific features of methionine biosynthesis and metabolism in plants. *Proc. Natl. Acad. Sci.* **95**, 7805–7812.

195 Vollmer W & Holtje JV (2004) The architecture of the murein (peptidoglycan) in gram-negative bacteria: Vertical scaffold or horizontal layer(s)? *J. Bacteriol.* **186**, 5978–5987.

196 Bertels F, Merker H & Kost C (2012) Design and characterization of auxotrophy-based amino acid biosensors. *PloS One* **7**, e41349.

197 Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Ros IMY, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou LW, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R & Fraser CM (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." *Proc. Natl. Acad. Sci.* **102**, 13950–13955.

198 Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V & Ravel J (2008) The pangenome structure of Escherichia coli: Comparative genomic analysis of E-coli commensal and pathogenic isolates. *J. Bacteriol.* **190**, 6881–6893.

199 Karberg KA, Olsen GJ & Davis JJ (2011) Similarity of genes horizontally acquired by Escherichia coli and Salmonella enterica is evidence of a supraspecies pangenome. *Proc. Natl. Acad. Sci.* **108**, 20154–20159.

200 van de Guchte M, Penaud S, Grimaldi C, Barbe V, Bryson K, Nicolas P, Robert C, Oztas S, Mangenot S, Couloux A & others (2006) The complete genome sequence of Lactobacillus bulgaricus reveals extensive and ongoing reductive evolution. *Proc. Natl. Acad. Sci.* **103**, 9274–9279.

201 Maughan H, Callicotte V, Hancock A, Birky CW, Nicholson WL & Masel J (2006) The population genetics of phenotypic deterioration in experimental populations of Bacillus subtilis. *Evolution* **60**, 686–695.

202 Shigenobu S, Watanabe H, Hattori M, Sakaki Y & Ishikawa H (2000) Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp APS. *Nature* **407**, 81–86.

203 Andersson SG. & Kurland CG (1998) Reductive evolution of resident genomes. *Trends Microbiol.* **6**, 263–268.

204 Kuo C-H, Moran NA & Ochman H (2009) The consequences of genetic drift for bacterial genome complexity. *Genome Res.* **19**, 1450–1454.

205 Moran NA, McCutcheon JP & Nakabachi A (2008) Genomics and Evolution of Heritable Bacterial Symbionts. In *Annual Review of Genetics* pp. 165–190.

206 Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE & Kim JF (2009) Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* **461**, 1243–U74.

207 Koskiniemi S, Sun S, Berg OG & Andersson DI (2012) Selection-Driven Gene Loss in Bacteria. *Plos Genet.* **8**, e1002787.

208 Lee M-C & Marx CJ (2012) Repeated, Selection-Driven Genome Reduction of Accessory Genes in Experimental Populations. *Plos Genet.* **8**, e1002651.

209 Zamenhof S & Eichhorn HH (1967) Study of Microbial Evolution through Loss of Biosynthetic Functions: Establishment of “Defective” Mutants. *Nature* **216**, 456–458.

210 Dykhuizen D (1978) Selection for Tryptophan Auxotrophs of Escherichia coli in Glucose-Limited Chemostats as a Test of the Energy Conservation Hypothesis of Evolution. *Evolution* **32**, 125.

211 Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Pujar A, Shearer AG, Travers M, Weerasinghe D, Zhang P & Karp PD (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **40**, D742–D753.

212 Vallenet D, Engelen S, Mornico D, Cruveiller S, Fleury L, Lajus A, Rouy Z, Roche D, Salvignol G, Scarpelli C & Medigue C (2009) MicroScope: a platform for microbial genome annotation and comparative genomics. *Database* **2009**, bap021–bap021.

213 Gomez-Valero L, Rocha EPC, Latorre A & Silva FJ (2007) Reconstructing the ancestor of Mycobacterium leprae: The dynamics of gene loss and genome reduction. *Genome Res.* **17**, 1178–1185.

214 Pagani I, Liolios K, Jansson J, Chen I-MA, Smirnova T, Nosrat B, Markowitz VM & Kyrpides NC (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **40**, D571–D579.

215 Vanstockem M, Michiels K, Vanderleyden J & Van Gool AP (1987) Transposon mutagenesis of Azospirillum brasilense and Azospirillum lipoferum: physical analysis of Tn5 and Tn5-Mob insertion mutants. *Appl. Environ. Microbiol.* **53**, 410–415.

216 Ogata H, Goto S, Sato K, Fujibuchi W, Bono H & Kanehisa M (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27**, 29–34.

217 Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muniz-Rascado L, Bonavides-Martinez C, Paley S, Krummenacker M, Altman T, Kaipa P, Spaulding A, Pacheco J, Latendresse M, Fulcher C, Sarker M, Shearer AG, Mackie A, Paulsen I, Gunsalus RP & Karp PD (2011) EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Res.* **39**, D583–D590.

218 Thomason LC, Costantino N & Court DL (2007) E. coli genome manipulation by P1 transduction. *Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al* **Chapter 1**, Unit 1.17.

219 Lenski R, Rose M, Simpson S & Tadler S (1991) Long-term experimental evolution in Escherichia coli .1. Adaptation and divergence during 2,000 generations. *Am. Nat.* **138**, 1315–1341.

220 de Berardinis V, Vallenet D, Castelli V, Besnard M, Pinet A, Cruaud C, Samair S, Lechaplais C, Gyapay G, Richez C, Durot M, Kreimeyer A, Le Fevre F, Schaechter V, Pezo V, Doering V, Scarpelli C, Medigue C, Cohen GN, Marliere P, Salanoubat M & Weissenbach J (2008) A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. *Mol. Syst. Biol.* **4**, 174.

221 Tepper N & Shlomi T (2011) Computational Design of Auxotrophy-Dependent Microbial Biosensors for Combinatorial Metabolic Engineering Experiments. *Plos One* **6**, e16274.

222 Benjamini Y, Krieger AM & Yekutieli D (2006) Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507.

223 R Core Team (2014) *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria.

224 Vyass S & Maas W (1963) Feedback inhibition of acetylglutamate synthetase by arginine in Escherichia coli. *Arch. Biochem. Biophys.* **100**, 542–546.

225 Pabst M, Kuhn J & Somervil R (1973) Feedback regulation in anthranilate aggregate from wild-type and mutant strains of Escherichia coli. *J. Biol. Chem.* **248**, 901–914.

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

227 McCutcheon JP & Moran NA (2010) Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. *Genome Biol. Evol.* **2**, 708–718.

228 Dekel E & Alon U (2005) Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**, 588–592.

229 Scott M, Gunderson CW, Mateescu EM, Zhang Z & Hwa T (2010) Interdependence of Cell Growth and Gene Expression: Origins and Consequences. *Science* **330**, 1099–1102.

230 Shachrai I, Zaslaver A, Alon U & Dekel E (2010) Cost of Unneeded Proteins in E. coli Is Reduced after Several Generations in Exponential Growth. *Mol. Cell* **38**, 758–767.

231 Flamholz A, Noor E, Bar-Even A, Liebermeister W & Milo R (2013) Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proc. Natl. Acad. Sci.* **110**, 10039–10044.

232 Hottes AK, Freddolino PL, Khare A, Donnell ZN, Liu JC & Tavazoie S (2013) Bacterial Adaptation through Loss of Function. *Plos Genet.* **9**, e1003617.

233 Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, Feist AM & Palsson BØ (2013) Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl. Acad. Sci.* **110**, 20338–20343.

234 Munster U (1993) Concentrations and fluxes of organic-carbon substrates in the aquatic environment. *Antonie Van Leeuwenhoek Int. J. Gen. Mol. Microbiol.* **63**, 243–274.

235 Ihssen J, Grasselli E, Bassin C, Francois P, Piffaretti J-C, Koester W, Schrenzel J & Egli T (2007) Comparative genomic hybridization and physiological characterization of environmental isolates indicate that significant (eco-)physiological properties are highly conserved in the species Escherichia coli. *Microbiol.-Sgm* **153**, 2052–2066.

236 Boles BR, Thoendel M & Singh PK (2004) Self-generated diversity produces “insurance effects” in biofilm communities. *Proc. Natl. Acad. Sci.* **101**, 16630–16635.

237 Giraud A, Matic I, Tenaillon O, Clara A, Radman M, Fons M & Taddei F (2001) Costs and benefits of high mutation rates: Adaptive evolution of bacteria in the mouse gut. *Science* **291**, 2606–2608.

238 Ensminger AW, Yassin Y, Miron A & Isberg RR (2012) Experimental Evolution of Legionella pneumophila in Mouse Macrophages Leads to Strains with Altered Determinants of Environmental Survival. *Plos Pathog.* **8**, e1002731.

239 Ellers J, Kiers ET, Currie CR, McDonald BR & Visser B (2012) Ecological interactions drive evolutionary loss of traits. *Ecol. Lett.* **15**, 1071–1082.

240 Fraser C, Gocayne J, White O, Adams M, Clayton R, Fleischmann R, Bult C, Kerlavage A, Sutton G, Kelley J, Fritchmann J, Weidman J, Small K, Sandusky M, Fuhrmann J, Nguyen D, Utterback T, Saudek D, Phillips C, Merrick J, Tomb J, Dougherty B, Bott K, Hu P, Lucier T, Peterson S, Smith H, Hutchison C &

Venter J (1995) The minimal gene complement of mycoplasma-genitalium. *Science* **270**, 397–403.

241 McCutcheon JP & von Dohlen CD (2011) An Interdependent Metabolic Patchwork in the Nested Symbiosis of Mealybugs. *Curr. Biol.* **21**, 1366–1372.

242 Bernstein HC, Paulson SD & Carlson RP (2012) Synthetic Escherichia coli consortia engineered for syntrophy demonstrate enhanced biomass productivity. *J. Biotechnol.* **157**, 159–166.

243 Amann R, Ludwig W & Schleifer K (1995) Phylogenetic identification and in-situ detection of individual microbial-cells without cultivation. *Microbiol. Rev.* **59**, 143–169.

244 Hugenholtz P, Goebel BM & Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180**, 4765–4774.

245 D'Souza G, Waschina S, Pande S, Bohl K, Kaleta C & Kost C (2014) Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria. *Evolution* **68**, 2559–2570.

246 Mee MT, Collins JJ, Church GM & Wang HH (2014) Syntrophic exchange in synthetic microbial communities. *Proc. Natl. Acad. Sci.* **111**, E2149–E2156.

247 Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, Pavlov A, Pavlova N, Karamychev V, Polouchine N, Shakhova V, Grigoriev I, Lou Y, Rohksar D, Lucas S, Huang K, Goodstein DM, Hawkins T, Plengvidhya V, Welker D, Hughes J, Goh Y, Benson A, Baldwin K, Lee J-H, Diaz-Muniz I, Dosti B, Smeianov V, Wechter W, Barabote R, Lorca G, Altermann E, Barrangou R, Ganesan B, Xie Y, Rawsthorne H, Tamir D, Parker C, Breidt F, Broadbent J, Hutkins R, O'Sullivan D, Steele J, Unlu G, Saier M, Klaenhammer T, Richardson P, Kozyavkin S, Weimer B & Mills D (2006) Comparative genomics of the lactic acid bacteria. *Proc. Natl. Acad. Sci.* **103**, 15611–15616.

248 Richards VP, Palmer SR, Bitar PDP, Qin X, Weinstock GM, Highlander SK, Town CD, Burne RA & Stanhope MJ (2014) Phylogenomics and the Dynamic Genome Evolution of the Genus Streptococcus. *Genome Biol. Evol.* **6**, 741–753.

249 Luo H, Cusros M, Hughes AL & Moran MA (2013) Evolution of Divergent Life History Strategies in Marine Alphaproteobacteria. *Mbio* **4**, e00373–13.

250 Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, Gonzalez JM, Luo H, Wright JJ, Landry ZC, Hanson NW, Thompson BP, Poulton NJ, Schwientek P, Acinas SG, Giovannoni SJ, Moran MA, Hallam SJ, Cavicchioli R, Woyke T & Stepanauskas R (2013) Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl. Acad. Sci.* **110**, 11463–11468.

251 Almaas E, Kovacs B, Vicsek T, Oltvai ZN & Barabasi AL (2004) Global organization of metabolic fluxes in the bacterium Escherichia coli. *Nature* **427**, 839–843.

252 Weinreich DM, Delaney NF, DePristo MA & Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114.

253 Schenk MF, Szendro IG, Salverda MLM, Krug J & de Visser JAGM (2013) Patterns of Epistasis between Beneficial Mutations in an Antibiotic Resistance Gene. *Mol. Biol. Evol.* **30**, 1779–1787.

254 Remold SK & Lenski RE (2004) Pervasive joint influence of epistasis and plasticity on mutational effects in Escherichia coli. *Nat. Genet.* **36**, 423–426.

255 Verhelst N, Hatzinger R & Mair P (2007) The Rasch Sampler. *J. Stat. Softw.* **20**, 1–14.

256 Moura A, Savageau MA & Alves R (2013) Relative Amino Acid Composition Signatures of Organisms and Environments. *Plos One* **8**, e77319.

257 Datsenko KA & Wanner BL (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc. Natl. Acad. Sci.* **97**, 6640–6645.

258 Zar JH (1999) *Biostatistical Analysis*, 4th ed. Prentice Hall.

259 Berger D & Postma E (2014) Biased Estimates of Diminishing-Returns Epistasis? Empirical Evidence Revisited. *Genetics* **198**, 1417–+.

260 Kim W & Levy SB (2008) Increased fitness of Pseudomonas fluorescens Pf0-1 leucine auxotrophs in soil. *Appl. Environ. Microbiol.* **74**, 3644–3651.

261 Traxler MF, Summers SM, Nguyen H-T, Zacharia VM, Hightower GA, Smith JT & Conway T (2008) The global, ppGpp-mediated stringent response to amino acid starvation in Escherichia coli. *Mol. Microbiol.* **68**, 1128–1148.

262 Paul BJ, Berkmen MB & Gourse RL (2005) DksA potentiates direct activation of amino acid promoters by ppGpp. *Proc. Natl. Acad. Sci.* **102**, 7823–7828.

263 Braun PR, Al-Younes H, Gussmann J, Klein J, Schneider E & Meyer TF (2008) Competitive inhibition of amino acid uptake suppresses chlamydial growth: Involvement of the chlamydial amino acid transporter BrnQ. *J. Bacteriol.* **190**, 1822–1830.

264 Zhuang K, Vemuri GN & Mahadevan R (2011) Economics of membrane occupancy and respiro-fermentation. *Mol. Syst. Biol.* **7**, 500.

265 Leavitt R & Umbarger H (1962) Isoleucine and valine metabolism in Escherichia coli .XI. Valine inhibition of growth of Escherichia coli strain K-12. *J. Bacteriol.* **83**, 624–&.

266 Harcombe W (2010) Novel cooperation experimentally evolved between species. *Evolution* **64**, 2166–2172.

267 Hosoda K, Suzuki S, Yamauchi Y, Shiroguchi Y, Kashiwagi A, Ono N, Mori K & Yomo T (2011) Cooperative Adaptation to Establishment of a Synthetic Bacterial Mutualism. *Plos One* **6**, e17105.

268 Douglas AE (1998) Nutritional interactions in insect-microbial symbioses: Aphids and their symbiotic bacteria Buchnera. *Annu. Rev. Entomol.* **43**, 17–37.

269 Moran NA (2007) Symbiosis as an adaptive process and source of phenotypic complexity. *Proc. Natl. Acad. Sci.* **104**, 8627–8633.

270 Yoon SH, Han M-J, Jeong H, Lee CH, Xia X-X, Lee D-H, Shim JH, Lee SY, Oh TK & Kim JF (2012) Comparative multi-omics systems analysis of Escherichia coli strains B and K-12. *Genome Biol.* **13**, R37.

271 Marr AG (1991) Growth rate of Escherichia coli. *Microbiol. Rev.* **55**, 316–333.

272 Zhao J & Shimizu K (2003) Metabolic flux analysis of Escherichia coli K12 grown on 13 C-labeled acetate and glucose using GC-MS and powerful flux calculation method. *J. Biotechnol.* **101**, 101–117.

273 Neidhardt FC, Ingraham JL & Schaechter M (1990) *Physiology of the Bacterial Cell: A Molecular Approach* Sinauer Associates Inc, Sunderland, Mass.

274 Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C, Fulcher C, Huerta AM, Kothari A, Krummenacker M, Latendresse M, Muniz-Rascado L, Ong Q, Paley S, Schroder I, Shearer AG, Subhraveti P, Travers M, Weerasinghe D, Weiss V, Collado-Vides J, Gunsalus RP, Paulsen I & Karp PD (2013) EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* **41**, D605–D612.

275 Savageau MA (1983) Escherichia coli Habitats, Cell Types, and Molecular Mechanisms of Gene Control. *Am. Nat.* **122**, pp. 732–744.

276 van Elsas JD, Semenov AV, Costa R & Trevors JT (2011) Survival of Escherichia coli in the environment: fundamental and public health aspects. *ISME J.* **5**, 173–183.

277 Im SWK, Davidson H & Pittard J (1971) Phenylalanine and tyrosine biosynthesis in Escherichia coli K-12: mutants derepressed for 3-deoxy-D-arabinoheptulosonic acid 7-phosphate synthetase (phe), 3-deoxy-D-

arabinoheptulosonic acid 7-phosphate synthetase (tyr), chorismate mutase T-prephenate dehydrogenase, and transaminase A. *J. Bacteriol.* **108**, 400–409.

278 D'Souza G, Waschina S, Kaleta C & Kost C (2015) Plasticity and epistasis strongly affect bacterial fitness after losing multiple metabolic genes. *Evolution* **69**, 1244–1254.

279 Peebo K, Valgepea K, Maser A, Nahku R, Adamberg K & Vilu R (2015) Proteome reallocation in Escherichia coli with increasing specific growth rate. *Mol Bio Syst* **11**, 1184–1193.

280 Sabarly V, Aubron C, Glodt J, Balliau T, Langella O, Rigal O, Bourgais A, Picard B, de Vienne D, Denamur E, Bouvet O & Dillmann C (2015) Interactions between genotype and environment drive the metabolic phenotype within *Escherichia coli* isolates: Metabolic diversity within *Escherichia coli* isolates. *Environ. Microbiol.*

281 Bailey JE (1991) Toward a science of metabolic engineering. *Science* **252**, 1668–1675.

282 Masurekar PS (2008) Nutritional and engineering aspects of microbial process development. In *Natural Compounds as Drugs Volume I* pp. 291, 293–328.

283 Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, Kang J, Hyduke DR & Palsson BØ (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* **6**, 1290–1307.

284 Snell EE (1975) Tryptophanase: Structure, Catalytic Activities, and Mechanism of Action. In *Advances in Enzymology and Related Areas of Molecular Biology* (Meister A, ed), pp. 287–333. John Wiley & Sons, Inc., Hoboken, NJ, USA.

285 Bates D, Mächler M, Bolker B & Walker S (2014) Fitting linear mixed-effects models using lme4. *ArXiv Prepr. ArXiv14065823*.

286 Nakagawa S & Schielzeth H (2013) A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* **4**, 133–142.

287 Hartigan JA & Wong MA (1979) Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **28**, 100.

288 Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D, Weitz KK, Eils R, König R, Smith RD & Palsson BØ (2010) Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **6**.

289 Pelosi L, Kühn L, Guetta D, Garin J, Geiselmann J, Lenski RE & Schneider D (2006) Parallel Changes in Global Protein Profiles During Long-Term Experimental Evolution in Escherichia coli. *Genetics* **173**, 1851–1869.

290 Lu P, Vogel C, Wang R, Yao X & Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25**, 117–124.

291 Liebermeister W, Noor E, Flamholz A, Davidi D, Bernhardt J & Milo R (2014) Visual account of protein investment in cellular functions. *Proc. Natl. Acad. Sci.* **111**, 8488–8493.

292 Krebs EG & Beavo JA (1979) Phosphorylation-dephosphorylation of enzymes. *Annu. Rev. Biochem.* **48**, 923–959.

293 Nielsen KM, Ray JL & Johnsen PJ (2009) Horizontal Gene Transfer: Uptake of Extracellular DNA by Bacteria. In *Encyclopedia of Microbiology (Third Edition)* (Schaechter M, ed), pp. 587–596. Academic Press, Oxford.

294 Vieira G, Sabarly V, Bourguignon P-Y, Durot M, Fèvre FL, Mornico D, Vallenet D, Bouvet O, Denamur E, Schachter V & Médigue C (2011) Core and Panmetabolism in Escherichia coli. *J. Bacteriol.* **193**, 1461–1472.

295 Alexander M (1971) Biochemical ecology of microorganisms. *Annu. Rev. Microbiol.* **25**, 361–392.

296 Garcia SL, Buck M, McMahon KD, Grossart H-P, Eiler A & Warnecke F (2015) Auxotrophy and intra-population complementary in the “interactome” of a cultivated freshwater model community. *Mol. Ecol.* **24**, 4449–4459.

297 Kolodkin-Gal I, Romero D, Cao S, Clardy J, Kolter R & Losick R (2010) D-Amino Acids Trigger Biofilm Disassembly. *Science* **328**, 627–629.

298 Pfeiffer T & Bonhoeffer S (2004) Evolution of Cross-Feeding in Microbial Populations. *Am. Nat.* **163**, E126–E135.

299 Pfeiffer T, Schuster S & Bonhoeffer S (2001) Cooperation and Competition in the Evolution of ATP-Producing Pathways. *Science* **292**, 504–507.

300 Monod J (1942) Research on the Glowth of Bacterial Cultures. *Hermann Cie Paris*, 211 pp.

301 Tasoff J, Mee MT & Wang HH (2015) An Economic Framework of Microbial Trade. *PLOS ONE* **10**, e0132907.

302 Pedrós-Alió C (2006) Marine microbial diversity: can it be determined? *Trends Microbiol.* **14**, 257–263.

303 Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P & Woyke T (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437.

304 Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, Szeto E, Platt D, Hugenholtz P, Relman DA & others (2007) Dissecting biological “dark matter” with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl. Acad. Sci.* **104**, 11889–11894.

305 Barer MR & Harwood CR (1999) Bacterial Viability and Culturability. In *Advances in Microbial Physiology* (Poole RK, ed), pp. 93–137. Academic Press.

306 Atlas RM (2004) *Handbook of Microbiological Media, Third Edition* CRC Press.

307 Kadner RJ (1977) Transport and utilization of D-methionine and other methionine sources in Escherichia coli. *J. Bacteriol.* **129**, 207–216.

308 Pande S, Shitut S, Freund L, Westermann M, Bertels F, Colesie C, Bischofs IB & Kost C (2014) Metabolic cross-feeding via intercellular nanotubes among bacteria. *Nat. Commun.* **6**, 6238.

309 Boyer M & Wisniewski-Dyé F (2009) Cell–cell signalling in bacteria: not simply a matter of quorum: Cell–cell signalling in bacteria. *FEMS Microbiol. Ecol.* **70**, 1–19.

310 Morris JJ (2015) Black Queen evolution: the role of leakiness in structuring microbial communities. *Trends Genet.* **31**, 475–482.

311 Kaeberlein T, Lewis K & Epstein SS (2002) Isolating “Uncultivable” Microorganisms in Pure Culture in a Simulated Natural Environment. *Science* **296**, 1127–1129.

312 D’Onofrio A, Crawford JM, Stewart EJ, Witt K, Gavrish E, Epstein S, Clardy J & Lewis K (2010) Siderophores from Neighboring Organisms Promote the Growth of Uncultured Bacteria. *Chem. Biol.* **17**, 254–264.

313 Morris JJ, Kirkegaard R, Szul MJ, Johnson ZI & Zinser ER (2008) Facilitation of Robust Growth of Prochlorococcus Colonies and Dilute Liquid Cultures by “Helper” Heterotrophic Bacteria. *Appl. Environ. Microbiol.* **74**, 4530–4534.

314 Vartoukian SR, Palmer RM & Wade WG (2010) Cultivation of a Synergistetes strain representing a previously uncultivated lineage. *Environ. Microbiol.* **12**, 916–928.

315 Monod J (1942) Recherches sur la croissance des cultures bacteriennes. *Hermann Cie Paris.*

316 Von Meyenburg K (1971) Transport-Limited Growth Rates in a Mutant of Escherichia coli. *J. Bacteriol.* **107**, 878–888.

317 Death A, Notley L & Ferenci T (1993) Derepression of LamB protein facilitates outer membrane permeation of carbohydrates into Escherichia coli under conditions of nutrient stress. *J. Bacteriol.* **175**, 1475–1483.

318 Neijssel OM, Hueting S & Tempest DW (1977) Glucose transport capacity is not the rate-limiting step in the growth of some wild-type strains of *Escherichia coli* and *Klebsiella aerogenes* in chemostat culture. *FEMS Microbiol. Lett.* **2**, 1–3.

319 Rutgers M, Balk PA & van Dam K (1989) Effect of concentration of substrates and products on the growth of Klebsiella pneumoniae in chemostat cultures. *Biochim. Biophys. Acta* **977**, 142–149.

320 Ferenci T (1999) `Growth of bacterial cultures' 50 years on: towards an uncertainty principle instead of constants in bacterial growth kinetics. *Res. Microbiol.* **150**, 431–438.

321 Neu HC (1992) The crisis in antibiotic resistance. *Science* **257**, 1064–1073.

322 Walsh C (2003) *Antibiotics* American Society of Microbiology.

323 Becker D, Selbach M, Rollenhagen C, Ballmaier M, Meyer TF, Mann M & Bumann D (2006) Robust Salmonella metabolism limits possibilities for new antimicrobials. *Nature* **440**, 303–307.

324 Maguire A & Rugg-Gunn AJ (2003) Xylitol and caries prevention — is it a magic bullet? *Br. Dent. J.* **194**, 429–436.

325 Trahan L (1995) Xylitol: a review of its action on mutans streptococci and dental plaque--its clinical significance. *Int. Dent. J.* **45**, 77–92.

326 Hausman SZ, Thompson J & London J (1984) Futile xylitol cycle in Lactobacillus casei. *J. Bacteriol.* **160**, 211–215.

327 Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD & Palsson BØ (2005) In silico design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol. Bioeng.* **91**, 643–648.

328 Stephanopoulos G (1999) Metabolic Fluxes and Metabolic Engineering. *Metab. Eng.* **1**, 1–11.

329 Burgard AP, Pharkya P & Maranas CD (2003) Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84**, 647–657.

330 Hädicke O & Klamt S (2010) CASOP: A Computational Approach for Strain Optimization aiming at high Productivity. *J. Biotechnol.* **147**, 88–101.

331 Choon YW, Mohamad MS, Deris S, Illias RM, Chong CK & Chai LE (2014) A hybrid of bees algorithm and flux balance analysis with OptKnock as a platform for in silico optimization of microbial strains. *Bioprocess Biosyst. Eng.* **37**, 521–532.

332 Wendisch VF, Bott M & Eikmanns BJ (2006) Metabolic engineering of Escherichia coli and Corynebacterium glutamicum for biotechnological production of organic acids and amino acids. *Curr. Opin. Microbiol.* **9**, 268–274.

333 Pfeiffer T, Soyer OS & Bonhoeffer S (2005) The Evolution of Connectivity in Metabolic Networks. *PLoS Biol* **3**, e228.

334 Sousa FL, Hordijk W, Steel M & Martin WF (2015) Autocatalytic sets in E. coli metabolism. *J. Syst. Chem.* **6**.

335 Goel A, Wortel MT, Molenaar D & Teusink B (2012) Metabolic shifts: a fitness perspective for microbial cell factories. *Biotechnol. Lett.* **34**, 2147–2160.

Author contributions to manuscripts

Author abbreviations: Silvio Waschina (SW), Glen D'Souza (GD), Christoph Kaleta (CKa), Christian Kost (CKo), Ralf Schmidt (RS), Daniela Boettger-Schmidt (DBS), Samay Pande (SP), and Katrin Bohl (KB)

Chapter II – Computing autocatalytic sets to unravel inconsistencies in metabolic network reconstructions	
Conceived the project	CKa 100%
Conceived *in silico* analysis	RS 70%, CKa 15%, SW 15%
Designed *in-silico* analyses	RS 50%, CKa 25%, SW 25%
Implemented and performed in-silico analyses	RS 80%, SW 20%
Analysed in-silico predictions	RS 50%, SW 40%, DBS 10%
Wrote manuscript	RS 55%, SW 30%, DBS 5%, CKa 5%, CKo 5%

Chapter III – Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria	
Conceived the project	CKo 100%
Designed the experiments	GD 50%, CKo 50%
Performed all experiments	GD 100%
Constructed *A. baylyi* strains	SP 100%
Analysed the data	GD 56%, SW 40%, CKo 4%
Conceived *in silico* analysis	CKo 100%
Designed *in-silico* analyses	SW 20%, CKa 70%, CKo 10%
Performed in-silico analyses	SW 95%, KB 5%

Chapter IV – Plasticity and epistasis strongly affect bacterial fitness after losing multiple metabolic genes	
Conceived the project	GD 50%, CKo 50%
Designed the experiments	GD 70%, CKo 25%, SW 5%
Performed all experiments	GD 100%
Analysed experimental data	GD 95%, CKo 5%
Conceived *in silico* analysis	GD 40%, CKo 40%, SW 20%
Designed *in-silico* analyses	SW 75%, CKa 25%
Implemented and performed in-silico analyses	SW 100%
Analysed in-silico predictions	SW 90%, CKa 10%
Wrote manuscript	GD 70%, SW 15%, CKo% 12.5%, CKa 2.5%

Chapter V – Metabolic network architecture and carbon source determine metabolite production costs	
Conceived the project	SW 60%, CKo 20%, CKa 20%
Designed the experiments	SW 90%, GD 10%
Performed the experiments	SW 100%
Analysed experimental data	SW 100%
Designed auxotrophic strains	GD 100%
Conceived *in silico* analysis	SW 75%, CKa 20%, CKo 5%
Designed *in-silico* analyses	SW 100%
Implemented and performed in-silico analyses	SW 100%
Analysed in-silico predictions	SW 76%, GD 8%, CKo 8%, CKa 8%
Wrote manuscript	SW 85%, GD 5%, CKo 5%, CKa 5%

Supporting information for chapter II

Computing autocatalytic sets to unravel inconsistencies in metabolic network reconstructions

Instructions to run ASBIG

The method is available as source code (additionally, the file including the initial seed set for the test model is provided). To successfully run the method, the JigCell SBML parser (available on SBML Parser) has to be included in the JAVA Build Path.

For testing purposes, we recommend to use the model ARAGEM, which can be downloaded from the original publication (doi: http://dx.doi.org/10.1104/pp.109.148817).

Initial seed set

The initial seed set contains three different parts: (1) the elements of a minimal medium, (2) additional autocatalytic compounds according to Kun *et al.* and (3) metabolites to circumvent artifacts of the modeling procedure.

Minimal medium:

H_2O, O_2, HO_4P, NH^{3-}/H_4N, Glucose, Mn^{2-},Zn^{2-}, SO_4^{2-}, Cu^{2+}, Ca^{2+}, H^+, Cl^-, Co^{2+}, K^+, NO_3^-, Ni^{2+}, Mg^{2+}, Na^+, Fe^{2+}, H_2MoO_4, Fe^{3+}

Known/assumed autocatalytic metabolites:

ATP, NAD, CoA

Additional metabolites:

Apo-ACP, Holo-[carboxylase] (Biotin-Protein), Thioredoxin, Dihydrolipoamide, Cbl (Cob (I)alamin) (Vitamin B12s) (only as external metabolite)
In a few models three special pseudo-compounds are implemented:
DNA replication, RNA transcription, protein biosynthesis

Reference:

Á. Kun, B. Papp, and E. Szathmáry, ‘Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks’, *Genome Biol*, vol. 9, no. 3, p. 51, 2008

Table S1. Frequency of compounds identified in 190 automatically reconstructed metabolic networks.

Compound	**Identified in x% of the models**
peptidoglycan polymer (n-1 subunits)	68%
spermidine	47%
tRNA-Glu	26%
Thiamine	24%
alanylhistidine	22%
glycyl-L-asparagine	21%
glycerol teichoic acid (n=45)	20%
L-methionine	18%
glycyl-L-tyrosine	17%
L-tryptophan	12%
6,7-dimethyl-8-(1-D-ribityl)lumazine	12%
glycyl-L-phenylalanine	12%
adenosylcobinamide-GDP	12%
glycyl-L-cysteine	11%
5-formyltetrahydrofolate	11%
N-L-alanyl-L-threonine	11%
N-glycyl-L-methionine	11%
glycyl-L-leucine	11%
L-arginine	10%
L-valine	10%
riboflavin	9%
L-2-lysophosphatidylethanolamine	8%
inosine 5‘-triphosphate	7%
L-isoleucine	7%
meso-2,6-diaminoheptanedioate	7%
dihydropteroate	7%
(R)-S-Lactoylglutathione	7%
alpha-ribazole-5‘-phosphate	7%
L-alanyl- L-glutamine	7%

Table S2. Examined *E. coli* amino acid auxotrophs.

Deleted gene	**Affected reaction**	**Auxotrophy caused**	**Add-on metabolite identified**
argH	arginosuccinate lyase	L-arginine	L-arginine
hisD	histidinol dehydrogenase	L-histidine	L-histidine
ilvA	L-threonine deaminase	L-isoleucine	2-oxobutanoate
leuB	3-isopropylmalate dehydrogenase	L-leucine	3-carboxy-4-methyl-2-oxopentanoate
lysA	diaminopimelate decarboxylase	L-lysine	fructoselysine
metB	O-succinylhomoserine lyase	L-methionine	L-methionine
pheA	chorismate mutase	L-phenylalanine	prephenate
proC	pyrroline-5-carboxylate reductase	L-proline	L-prolinylglycine
thrC	threonine synthase	L-threonine	L-threonine O-3-phosphate
trpC	indole-3-glycerol-phosphate synthase	L-tryptophan	N-methyltryptophan

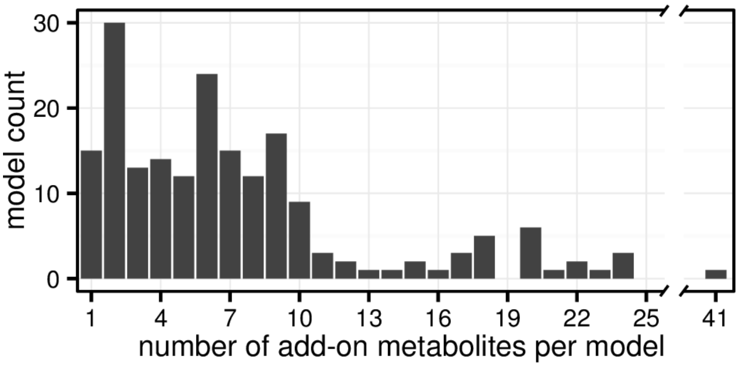


Figure S1. Histogram of identified add-on metabolites using ASBIG per automatically reconstructed metabolic network from the Model SEED repository. A total number of 190 networks were investigated.

Supporting information for chapter III

Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria

Supporting methods

Computation of protein and DNA sequence biosynthetic costs

Biosynthetic costs were estimated as the amount of carbon source that is required to produce (1) the amount of a certain protein per cell and (2) the DNA sequence of a certain gene. We used flux balance analysis within the Cobra toolbox v2.0 (Schellenberger et al. 2011) in a genome-scale metabolic network of *Escherichia coli* K12 (Orth et al. 2011). For each protein, the artificial reactions for protein synthesis,

$$\textbf{(1)}\ l\left(\sum_{\mathrm{a}_j\in \mathrm{AA}} n_j^a\ \mathrm{a}_j + (q * m_p)\mathrm{ATP}\right) \rightarrow l * m_p\ \ \mathrm{H_2O} + q\ \mathrm{ADP} +\ q\ \mathrm{Phosphate} +\ q\ \mathrm{H}^+$$

or for the synthesis of the corresponding DNA sequence

$$\textbf{(2)}\ k * \sum_{\mathrm{d}_j\in NA} n_j^d\ \mathrm{d}_j \rightarrow k * m_{DNA}\ \mathrm{Pyrophosphate}$$

have been included into the model. AA is the set of all 20 proteinogenic amino acids, NA is the set of the four desoxynucleoside triphosphates dATP, dCTP, dGTP, and dTTP. n^a_j represents the number of occurrences of amino acid a_j in the amino acid sequence of the protein and n^d_j the number of occurrences of the desoxynucleoside triphosphate d_j in the DNA sequence of the gene. m_p is the length of the amino acid sequence of the protein and m_{DNA} the length of the corresponding DNA sequence. The abundance of the protein per cell has been incorporated in the calculations by the parameter *l* and the number of DNA sequence copies by parameter k. Protein abundance data were taken from Wessely *et al.* (2011) and a maximum of 6.54 DNA sequence copies k=6.54 were assumed for all sequences, which corresponds to the maximal chromosomal copy-number of a single-locus gene near the origin of replication in an *E. coli* cell at 2.5 doublings per hour (Klumpp *et al.* 2009). The parameter q represents the ATP requirement per amino acid residue during the polymerization process of translation. A previously reported value of q=4.2 was used (Kaleta *et al.* 2013).

The lower bound for the flux of these reactions was set to a value of 1. The consumption of fructose as sole carbon source was minimized by linear programming to determine the minimal amount of fructose required to produce the corresponding proteins and DNA sequence. DNA and protein sequences were retrieved from the EcoCyc database (Keseler *et al.* 2013).

Construction of auxotrophic strains of Acinetobacter baylyi

Linear constructs of the kanamycin cassette with 5'-overhangs homologous to the insertion site were produced by PCR. To this end, plasmid pKD4 Datsenko and Wanner 2000 DNA was used as a template to amplify the kanamycin resistance cassette. Upstream and downstream regions homologous to *argH, hisD, leuB*, and *trpB* were amplified using primers with a 5'-extension that was complementary to the primers used to amplify the kanamycin cassette (Table S4). The three resulting products were combined by PCR to finally obtain the kanamycin cassette fused to the upstream and downstream homologous overhangs. Natural competence of *A. baylyi* was utilized to transform the linear fragments into the WT strain. Transformation was done by diluting 20 µl of a 16 h old culture grown in LB medium. This diluted culture was again incubated at 30 °C with shaking. 50 µl PCR mix containing the deletion cassette was added to this culture and again incubated at 30 °C with shaking for 2 h. Finally, the whole culture volume was concentrated to 100 µl and plated on LB agar plates containing kanamycin and incubated at 30 °C for colonies to appear.

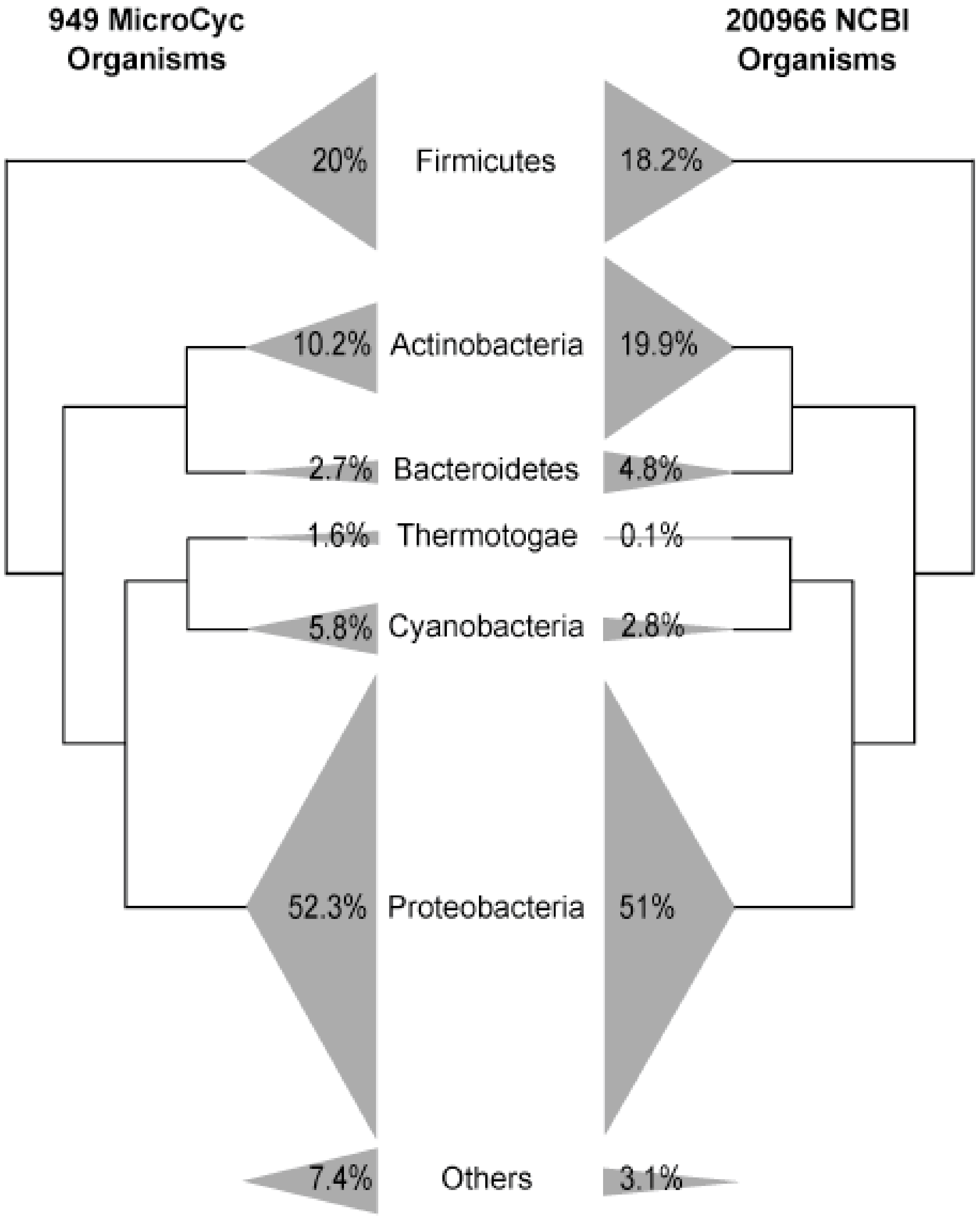


Figure S2. Taxonomic distribution of eubacterial strains used for *in silico* prediction of auxotrophies. Triangle size indicates the proportion of the phylum in the sample of 949 bacterial species from the MicroCyc database, which were used for auxotrophy prediction left cladogram and, for comparison, the proportion of each phylum in the National Center for Biotechnology Information NCBI taxonomy database of all phylum-classified Eubacteria (status: March 2013). Phylogeny adapted from Ciccarelli *et al.* (2006).

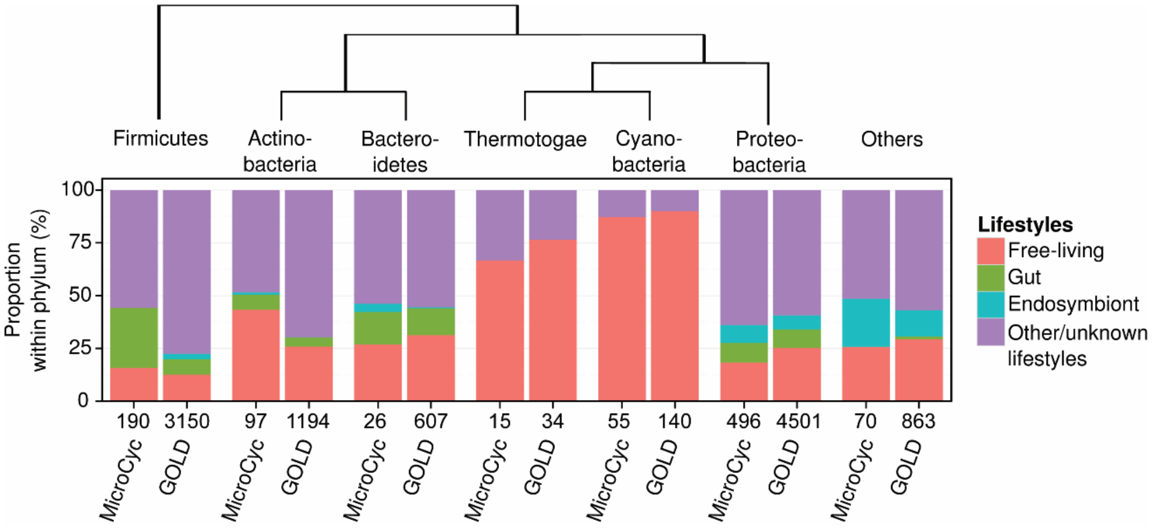
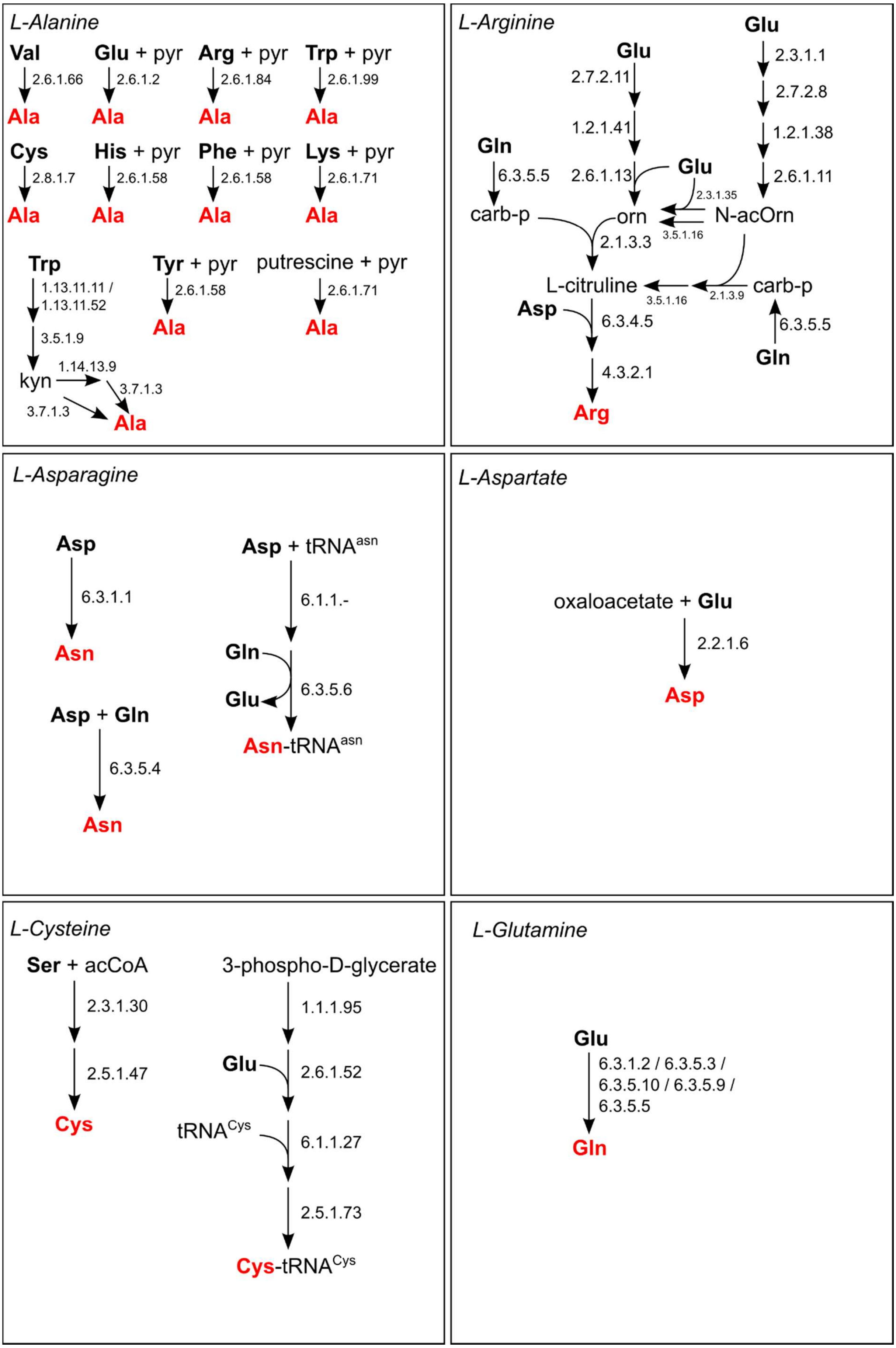
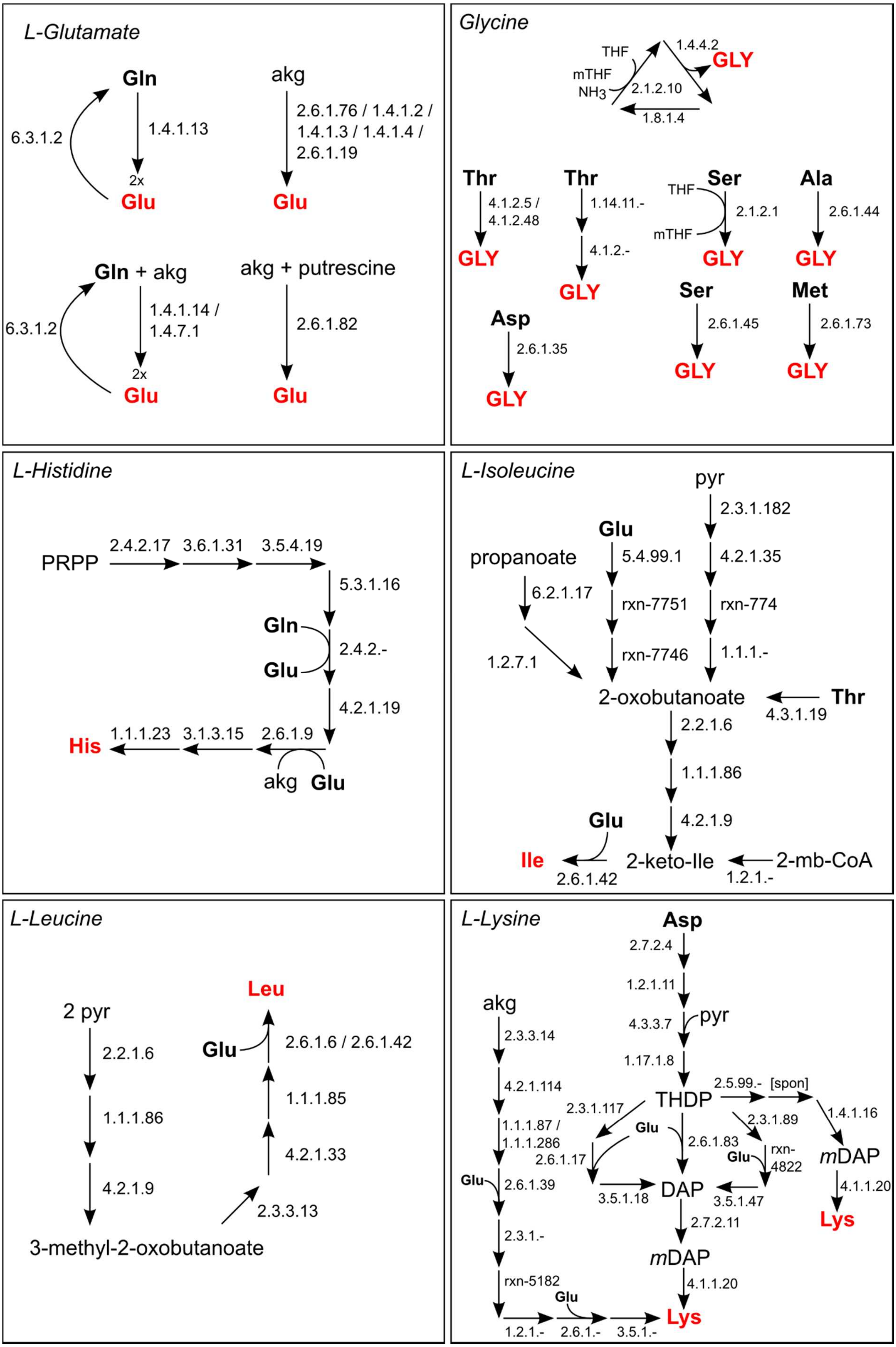
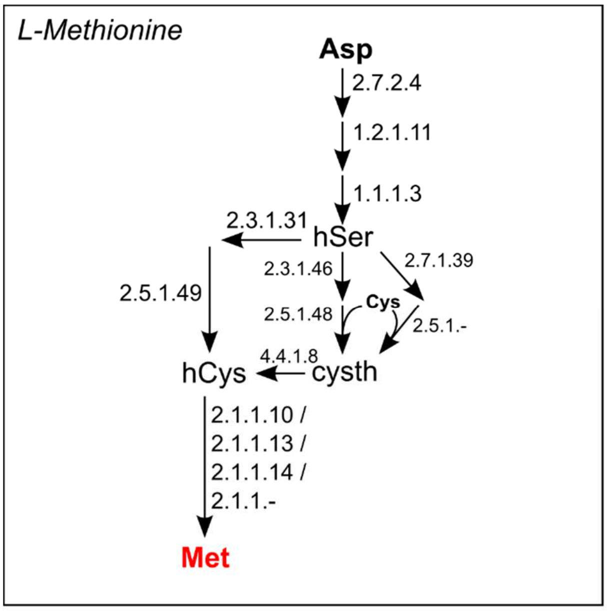
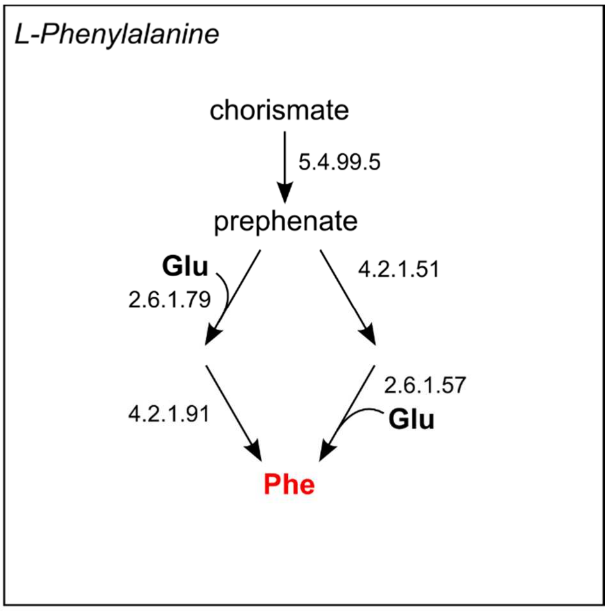


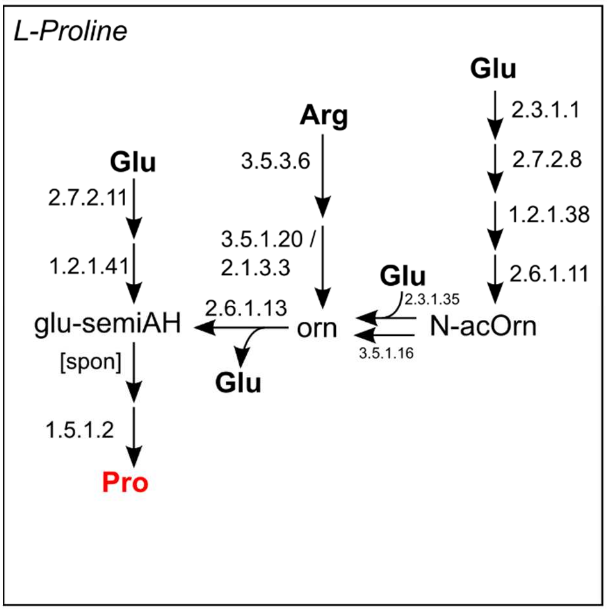
Figure S4. Phylogenetic distribution of free-living, gut-inhabiting, and endosymbiotic bacteria within the MicroCyc database (i.e. 949 organisms; Vallenet et al. 2009) and the Genomes OnLine Database GOLD; 10,489 organisms; Pagani et al. 2012). Only organisms for which the whole genome sequence as well as its lifestyle as listed in the Genomes OnLine Database were known were included. Numbers below bars indicate the number of organisms within the corresponding phylum and database. Phylogeny adapted from Ciccarelli et al. 2006.

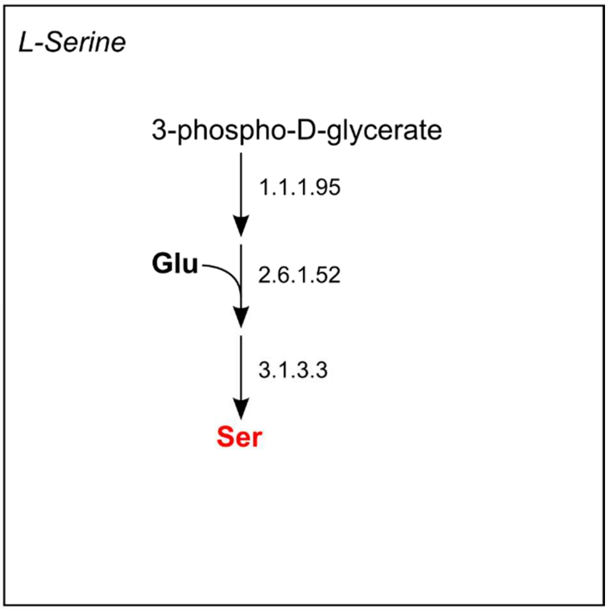

L-Alanine
Val
2.6.1.66
Ala
Glu + pyr
2.6.1.2
Ala
Arg + pyr
2.6.1.84
Ala
Trp + pyr
2.6.1.99
Ala
Cys
2.8.1.7
Ala
His + pyr
2.6.1.58
Ala
Phe + pyr
2.6.1.58
Ala
Lys + pyr
2.6.1.71
Ala
Trp
1.13.11.11 /
1.13.11.52
3.5.1.9
kyn
1.14.13.9
3.7.1.3
3.7.1.3
Ala
Tyr + pyr
2.6.1.58
Ala
putrescine + pyr
2.6.1.71
Ala
L-Arginine
Glu
2.7.2.11
1.2.1.41
2.6.1.13
Glu
2.3.1.35
Glu
2.3.1.1
2.7.2.8
1.2.1.38
2.6.1.11
N-acOrn
3.5.1.16
orn
Gln
6.3.5.5
carb-p
2.1.3.3
L-citruline
3.5.1.16
2.1.3.9
carb-p
6.3.5.5
Gln
Asp
6.3.4.5
4.3.2.1
Arg
L-Asparagine
Asp
6.3.1.1
Asn
Asp + Gln
6.3.5.4
Asn
Asp + tRNA^{asn}
6.1.1.-
Gln
Glu
6.3.5.6
Asn-tRNA^{asn}
L-Aspartate
oxaloacetate + Glu
2.2.1.6
Asp
L-Cysteine
Ser + acCoA
2.3.1.30
2.5.1.47
Cys
3-phospho-D-glycerate
1.1.1.95
Glu
2.6.1.52
tRNA^{Cys}
6.1.1.27
2.5.1.73
Cys-tRNA^{Cys}
L-Glutamine
Glu
6.3.1.2 / 6.3.5.3 /
6.3.5.10 / 6.3.5.9 /
6.3.5.5
Gln

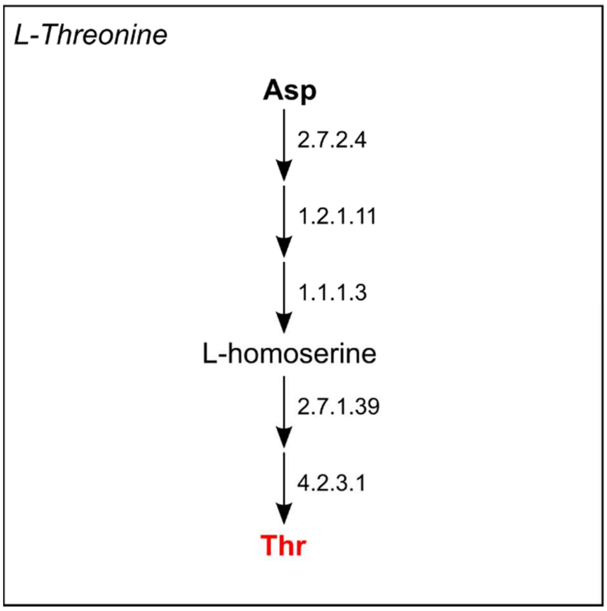

L-Glutamate
Gln
1.4.1.13
2x
Glu
6.3.1.2
akg
2.6.1.76 / 1.4.1.2 / 1.4.1.3 / 1.4.1.4 / 2.6.1.19
Glu
Gln + akg
1.4.1.14 / 1.4.7.1
2x
Glu
6.3.1.2
akg + putrescine
2.6.1.82
Glu
Glycine
THF
mTHF
NH3
1.4.4.2
GLY
2.1.2.10
1.8.1.4
Thr
4.1.2.5 / 4.1.2.48
GLY
Thr
1.14.11.-
4.1.2.-
GLY
THF
Ser
mTHF
2.1.2.1
GLY
Ala
2.6.1.44
GLY
Asp
2.6.1.35
GLY
Ser
2.6.1.45
GLY
Met
2.6.1.73
GLY
L-Histidine
PRPP
2.4.2.17
3.6.1.31
3.5.4.19
5.3.1.16
Gln
2.4.2.-
Glu
4.2.1.19
His
1.1.1.23
3.1.3.15
2.6.1.9
akg
Glu
L-Isoleucine
pyr
2.3.1.182
Glu
propanoate
5.4.99.1
4.2.1.35
6.2.1.17
rxn-7751
rxn-774
1.2.7.1
rxn-7746
1.1.1.-
2-oxobutanoate
Thr
4.3.1.19
2.2.1.6
1.1.1.86
Glu
4.2.1.9
Ile
2.6.1.42
2-keto-Ile
1.2.1.-
2-mb-CoA
L-Leucine
2 pyr
2.2.1.6
1.1.1.86
4.2.1.9
3-methyl-2-oxobutanoate
Leu
Glu
2.6.1.6 / 2.6.1.42
1.1.1.85
4.2.1.33
2.3.3.13
L-Lysine
Asp
2.7.2.4
1.2.1.11
4.3.3.7
pyr
1.17.1.8
akg
2.3.3.14
4.2.1.114
THDP
2.5.99.-
[spon]
1.4.1.16
2.3.1.117
2.3.1.89
1.1.1.87 / 1.1.1.286
Glu
2.6.1.83
rxn-4822
Glu
mDAP
2.6.1.17
Glu
2.6.1.39
3.5.1.18
DAP
3.5.1.47
4.1.1.20
Lys
2.7.2.11
2.3.1.-
mDAP
rxn-5182
4.1.1.20
Glu
Lys
1.2.1.-
2.6.1.-
3.5.1.-

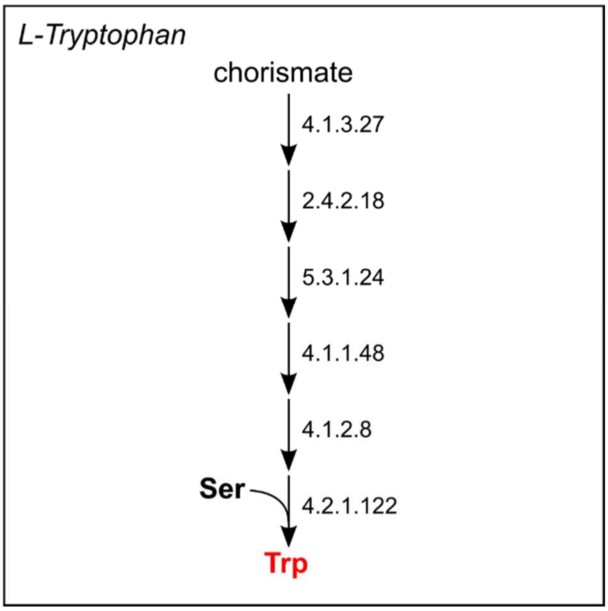

L-Methionine
Asp
2.7.2.4
1.2.1.11
1.1.1.3
2.3.1.31
hSer
2.3.1.46
2.7.1.39
2.5.1.49
2.5.1.48
Cys
2.5.1.-
4.4.1.8
hCys
cysth
2.1.1.10 /
2.1.1.13 /
2.1.1.14 /
2.1.1.-
Met

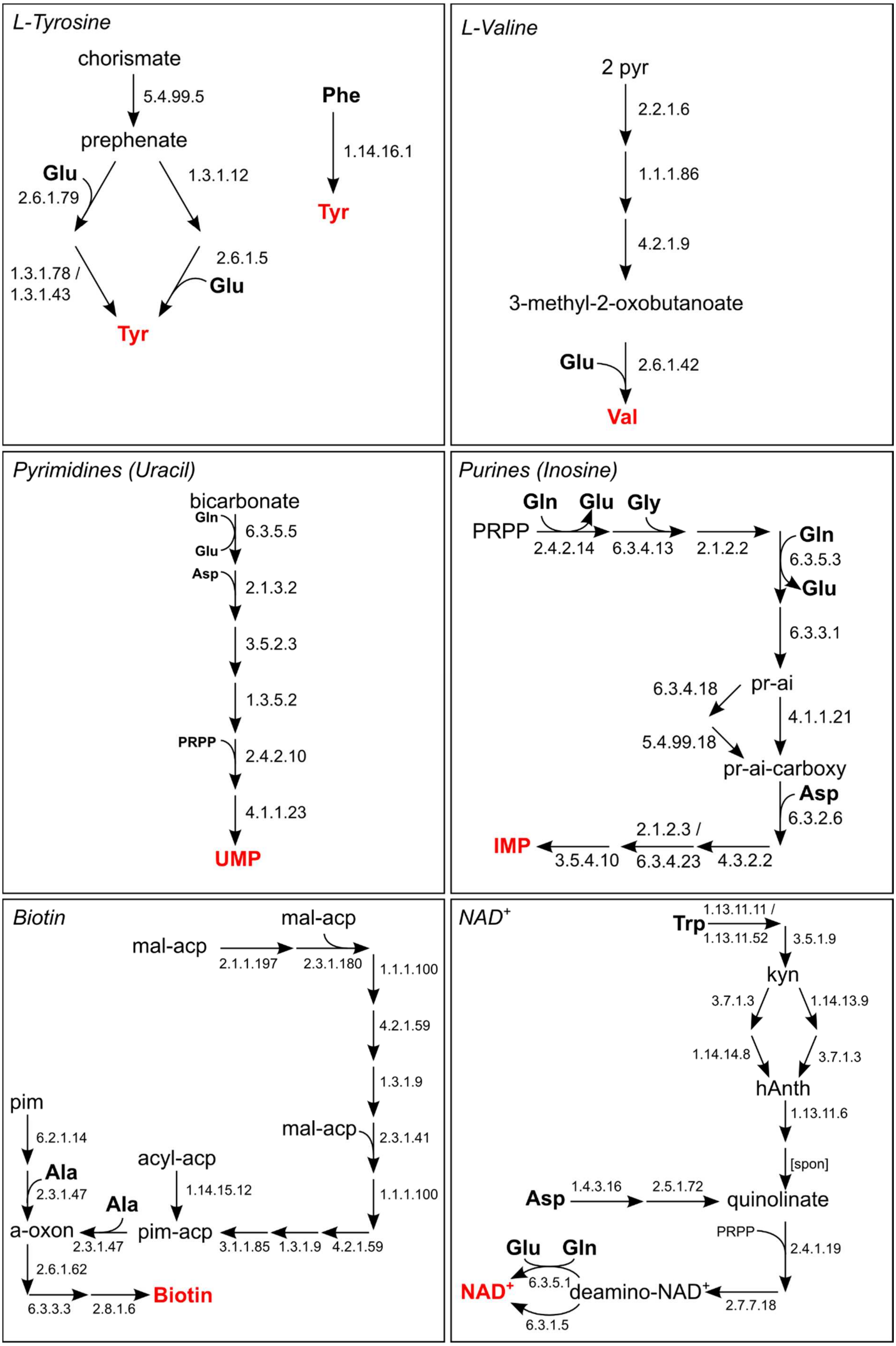

L-Phenylalanine
chorismate
5.4.99.5
prephenate
Glu
2.6.1.79
4.2.1.51
4.2.1.91
2.6.1.57
Glu
Phe


L-Proline
Glu
2.3.1.1
Arg
2.7.2.8
Glu
3.5.3.6
2.7.2.11
1.2.1.38
3.5.1.20 /
2.1.3.3
1.2.1.41
Glu
2.6.1.11
2.3.1.35
2.6.1.13
glu-semiAH
orn
N-acOrn
3.5.1.16
[spon]
Glu
1.5.1.2
Pro


L-Serine
3-phospho-D-glycerate
1.1.1.95
Glu
2.6.1.52
3.1.3.3
Ser


L-Threonine
Asp
2.7.2.4
1.2.1.11
1.1.1.3
L-homoserine
2.7.1.39
4.2.3.1
Thr


L-Tryptophan
chorismate
4.1.3.27
2.4.2.18
5.3.1.24
4.1.1.48
4.1.2.8
Ser
4.2.1.122
Trp


L-Tyrosine
chorismate
5.4.99.5
prephenate
Glu
2.6.1.79
1.3.1.12
1.3.1.78 /
1.3.1.43
2.6.1.5
Glu
Tyr
Phe
1.14.16.1
Tyr
L-Valine
2 pyr
2.2.1.6
1.1.1.86
4.2.1.9
3-methyl-2-oxobutanoate
Glu
2.6.1.42
Val
Pyrimidines (Uracil)
bicarbonate
Gln
6.3.5.5
Glu
Asp
2.1.3.2
3.5.2.3
1.3.5.2
PRPP
2.4.2.10
4.1.1.23
UMP
Purines (Inosine)
Gln
Glu
Gly
PRPP
2.4.2.14
6.3.4.13
2.1.2.2
Gln
6.3.5.3
Glu
6.3.3.1
pr-ai
6.3.4.18
4.1.1.21
5.4.99.18
pr-ai-carboxy
Asp
6.3.2.6
2.1.2.3 /
6.3.4.23
4.3.2.2
3.5.4.10
IMP
Biotin
mal-acp
2.1.1.197
mal-acp
2.3.1.180
1.1.1.100
4.2.1.59
1.3.1.9
mal-acp
2.3.1.41
1.1.1.100
4.2.1.59
1.3.1.9
3.1.1.85
pim-acp
acyl-acp
1.14.15.12
Ala
2.3.1.47
a-oxon
pim
6.2.1.14
Ala
2.3.1.47
2.6.1.62
6.3.3.3
2.8.1.6
Biotin
NAD$^+$
Trp
1.13.11.11 /
1.13.11.52
3.5.1.9
kyn
3.7.1.3
1.14.13.9
1.14.14.8
3.7.1.3
hAnth
1.13.11.6
[spon]
Asp
1.4.3.16
2.5.1.72
quinolinate
PRPP
2.4.1.19
Glu
Gln
6.3.5.1
NAD$^+$
deamino-NAD$^+$
2.7.7.18
6.3.1.5

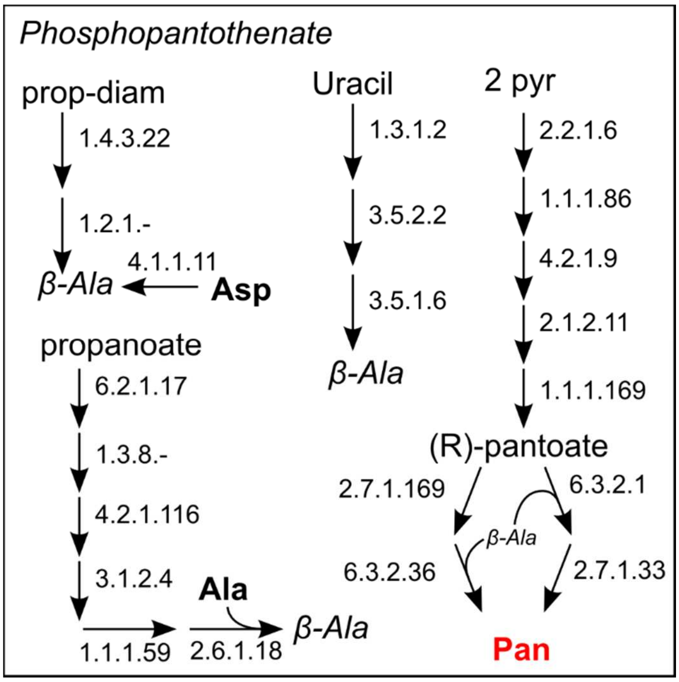


Figure S1. Metabolic pathways that were considered for the prediction of auxotrophies. Pathways including EC numbers were collected from the MetaCyc database (Caspi *et al.* 2012). Target compounds of each metabolic route are written in red. Metabolites written in black bold type are indicating dependencies on other biosynthetic pathways. All reactions are named by the corresponding EC number or the MetaCyc reaction ID if no EC number is assigned to the reaction in MetaCyc. UMP uridine monophosphate is the precursor for cytosine and IMP inosine monophosphate is the precursor for guanosine. Abbreviations: pyr: pyruvate, acCoA: acetyl CoA, tRNA^{Cys}: uncharged tRNA for L-cysteine, Cys-tRNA^{Cys}: L-cysteine-charged tRNA for L-cysteine, 2-keto-Ile: 2-keto-isoleucine, 2-mb-CoA: 2-methylbutanoyl-CoA, carb-p: carbamyl-phosphate, N-acOrn: N-acetyl-L-ornithine, orn: L-ornithine, akg: α-keto-glutarat, glu-semiAH: L-glutamate-5-semialdehyd, THDP: S-2,3,4,5-tetrahydrodipicolinate, DAP: L,L-diaminopimelate, mDAP: meso-diaminopimelate, kyn: L-kynurenine, cysth: L-cystathionine, hSer: L-homoserine, hCys: L-homocysteine, THF: tetrahydrofolate, mTHF: 5,10-methylenetetrahydrofolate, acp: acyl carrier protein, mal-acp: a malonyl acp, acyl-acp: a long chain acyl-acp, pim: pimelate, pim-acp: pimelyl-acp, a-oxon: 8-amino-7-oxononanoate, hAnth: 3-hydroxyanthranilate, pr-ai: 5-amino-1-5-phospho-β-D-ribosylimidazole, pr-ai-carboxy: 5-amino-1-5-phospho-D-ribosylimidazole-4-carboxylate, prop-diam: propane-1,3-diamine, [spon]: spontaneous reaction.

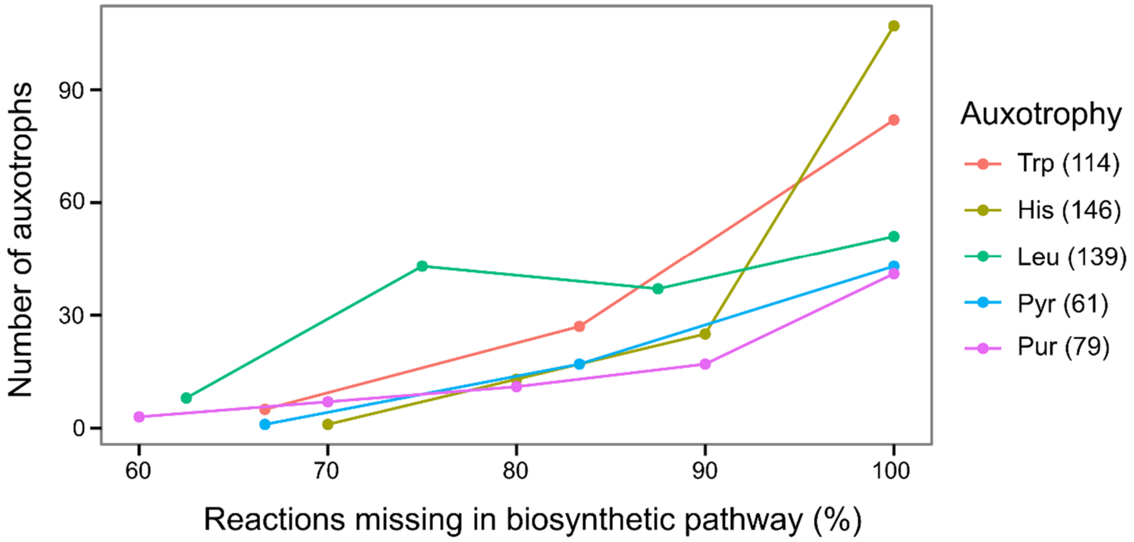


Figure S3: Incompleteness of the biosynthetic pathways forming tryptophan, histidine, leucine, pyrimidine, and purine within all Eubacteria predicted to be auxotrophic for these metabolites. The numbers behind the auxotrophy indicate the total number of strains, which are predicted to be auxotrophic for the corresponding compound. Predictions are based on analyses of the MicroCyc database (i.e. 949 organisms; Vallenet *et al*. 2009). These five pathways were chosen, because they are the longest linear pathways in the data set (see Fig. S1).

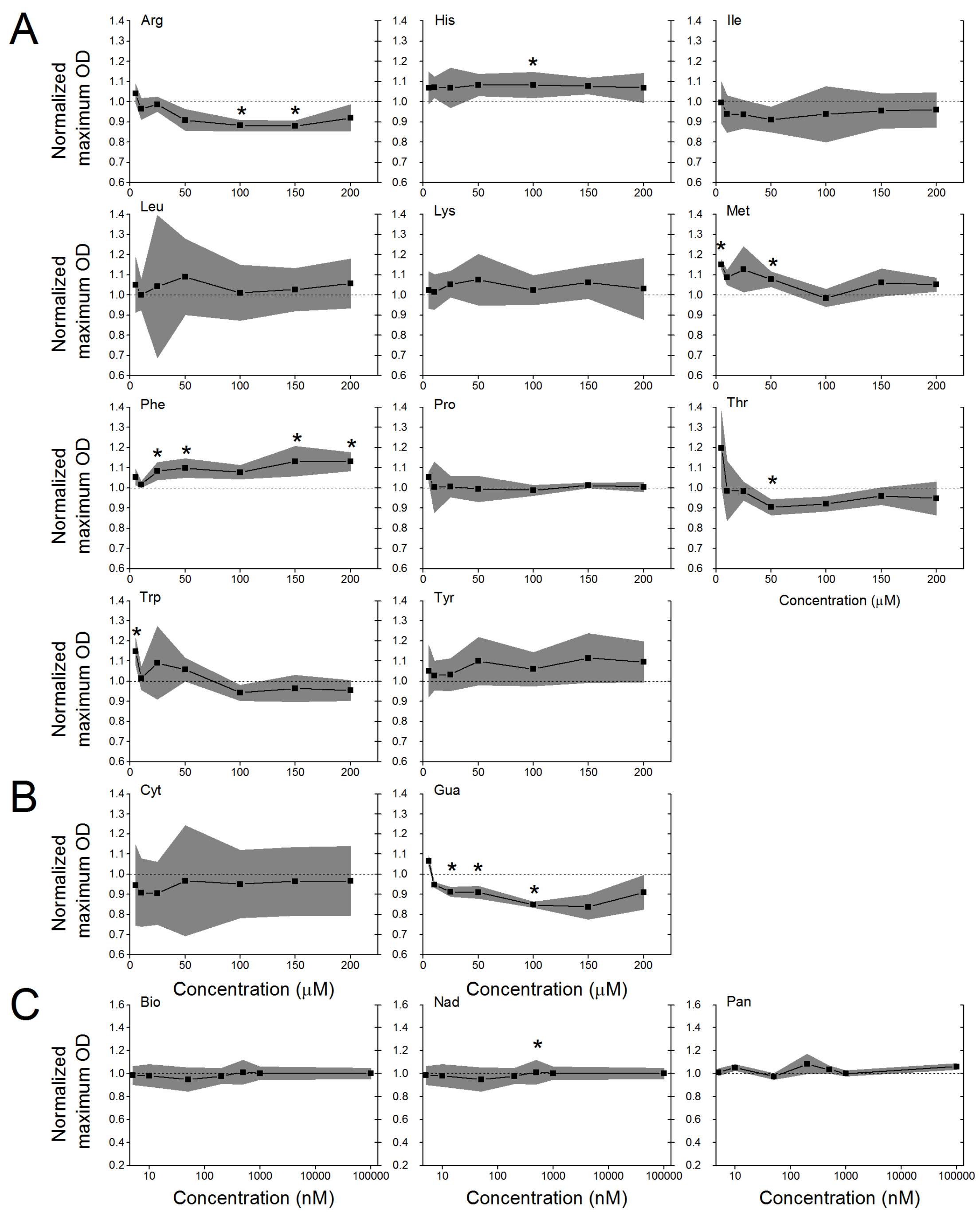


Figure S5: Growth response of *Escherichia coli* WT to increasing concentrations of the focal metabolites. Growth within 24 h was determined as optical density at 600 nm and is displayed as growth in minimal medium that contained a particular metabolite at a certain concentration relative to its growth in pure minimal medium. Each plot shows the concentration-dependent normalized growth response of WT in the presence of (**A**) an amino acid, (**B**) a nucleobase, or (**C**) a vitamin. All values are medians of four replicates and the grey-shaded area delimits the 95% confidence intervals. Asterisks mark significant differences from the growth of the WT in the absence of the focal compound (i.e. dashed line; FDR-corrected independent sample t-tests: *P<0.05, **P<0.01, and ***P<0.001, n=4). See Table S1 for abbreviations of metabolite names.

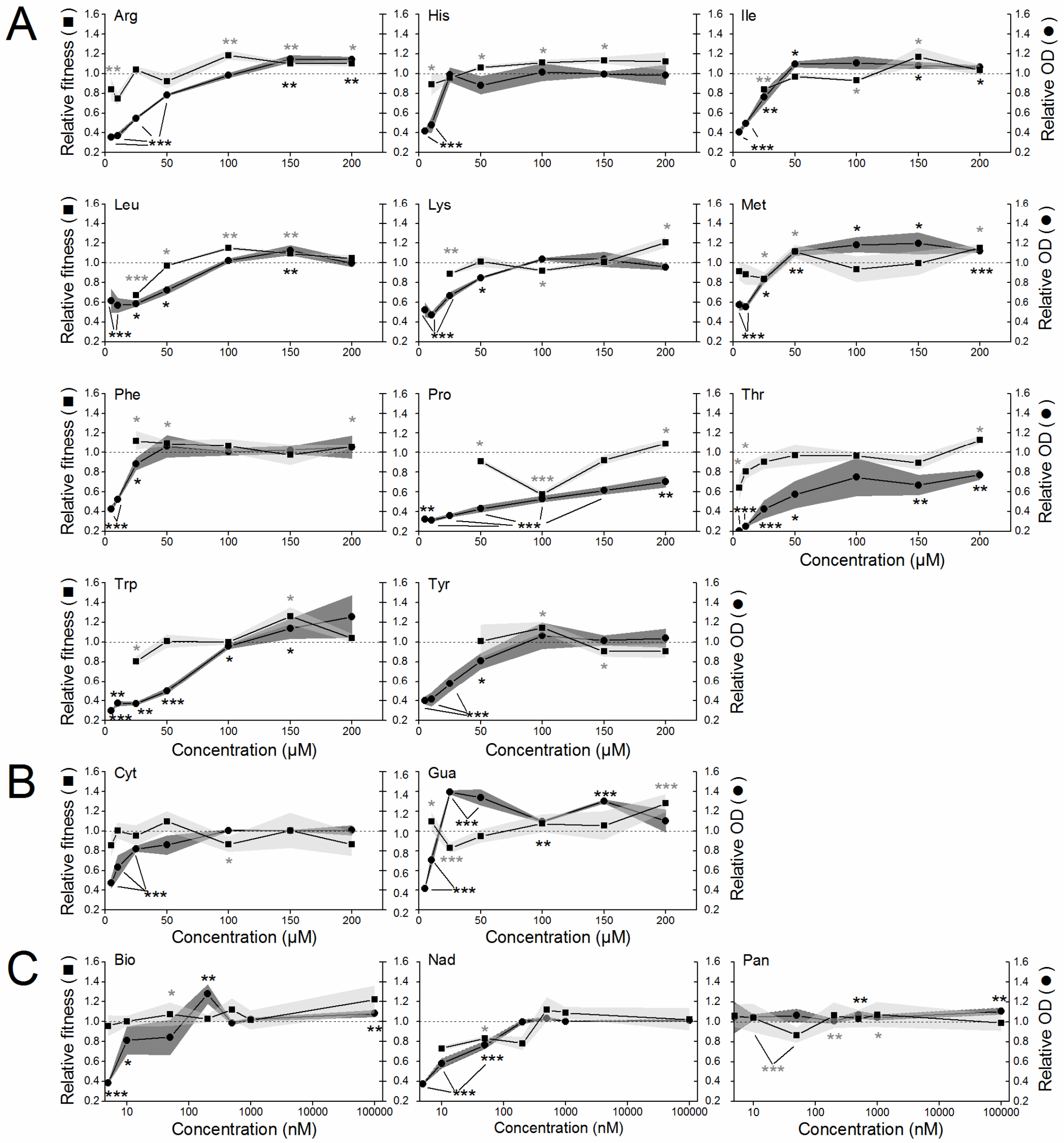


Figure S6. Productivity and competitive fitness of *Escherichia coli* auxotrophs relative to WT in increasing concentrations of the focal metabolites. Productivity (i.e. OD) and fitness of auxotrophic mutants within 24 h was determined relative to WT in both mono- circles or coculture squares using minimal medium that has been supplemented with (**A**) an amino acid, (**B**) a nucleobase, or (**C**) a vitamin in increasing concentrations. Relative OD of monocultures was determined as the ratio of the auxotroph's and the WT's optical densities measured at 600 nm and the relative fitness of cocultures is expressed as the ratio of their Malthusian parameters. Medians of four replicates are displayed. The dark and light grey regions mark the 95% confidence intervals for mono- and cocultures, respectively. Black and light grey asterisks mark significant differences of auxotrophs to WT levels (i.e. dashed line in mono- and cocultures, respectively monocultures: FDR-corrected independent sample t-tests, cocultures: FDR-corrected paired sample t-tests: *P<0.05, **P<0.01, and ***P<0.001; n=4). See Table S1 for abbreviations of metabolite names.

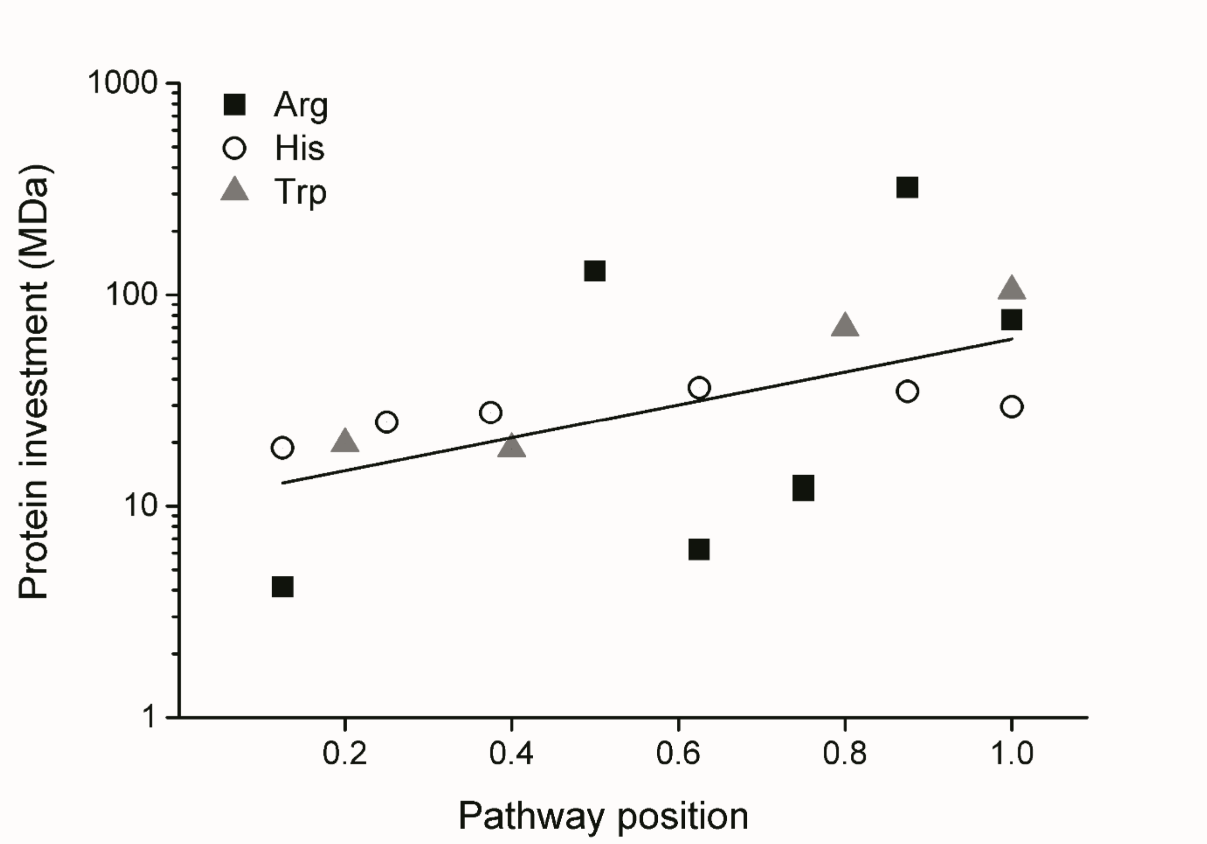


Figure S7. Relationship between the amount of protein invested by *Escherichia coli* into a certain biosynthetic step and the position of the gene within the biosynthetic pathways of arginine (Arg), histidine (His), and tryptophan (Trp). Protein investment in Mega Dalton is the mass of the individual protein multiplied with the abundance of protein copies per cell. Data was obtained from Wessely *et al.* (2011). Pathway position is the normalised localisation of each gene between the start (0.1) and the end (1.0) of the pathway. The line is the linear fit line between both variables.

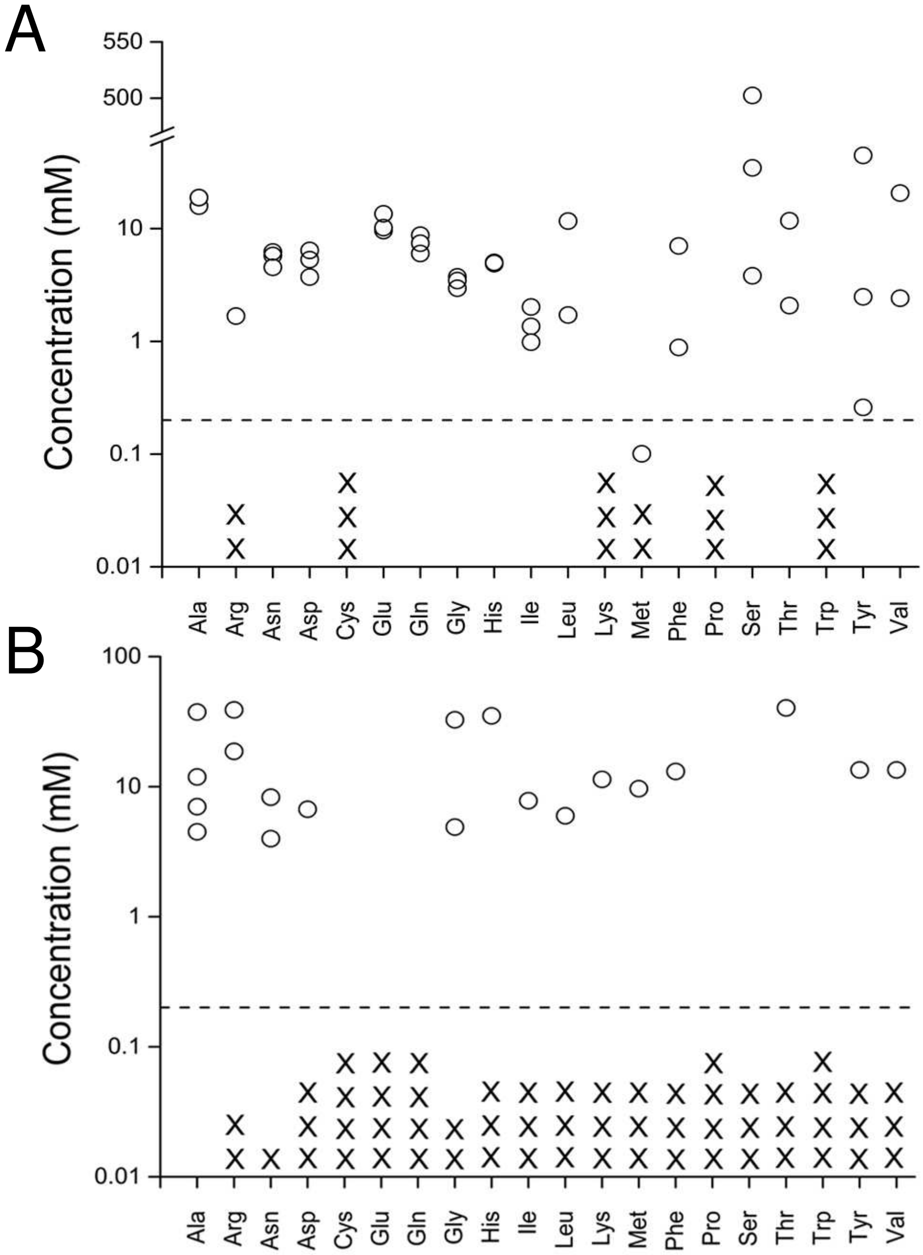


Figure S8. Amino acid concentrations in natural habitats of bacteria. Concentrations of individual amino acids found in (**A**) three different soil samples mM kg^{-1} soil (Werdin-Pfisterer *et al.* 2012) and (**B**) the gut of four different termite species mM gut^{-1} (Fujita and Abe 2002). Each circle indicates the amount of amino acid quantified in either a single soil sample or termite species. The dashed line represents the upper limit of amino acid concentrations i.e. 200 µM used in this study to determine the fitness of auxotrophic mutants. Crosses (X) signify instances in which the corresponding amino acid was not detected.

Table S1. Overview over the different auxotrophies analysed and the abbreviations used.

			Auxotrophy analysed in	
Class	Metabolite	Abbreviation	*Escherichia coli*	*Acinetobacter baylyi*
Amino acid	Alanine	Ala	●	
	Arginine	Arg	●	●
	Asparagine	Asn	●	
	Aspartic acid	Asp	●	
	Cysteine	Cys	●	
	Glutamine	Gln	●	
	Glutamic acid	Glu	●	
	Glycine	Gly	●	
	Histidine	His	●	●
	Isoleucine	Ile	●	
	Leucine	Leu	●	●
	Lysine	Lys	●	
	Methionine	Met	●	
	Phenylalanine	Phe	●	
	Proline	Pro	●	
	Serine	Ser	●	
	Threonine	Thr	●	
	Tryptophan	Trp	●	●
	Tyrosine	Tyr	●	
	Valine	Val	●	
Nucleobase	Cytosine	Cyt	●	
	Guanine	Gua	●	
Vitamin	Biotin	Bio	●	
	Nicotinamide adenine dinucleotide	Nad	●	
	Pantothenate	Pan	●	

Table S2. Strains used in this study. Abbreviations: ara$^{+/-}$ = ability to use arabinose as a C-source absent/ present, AT = auxotroph, WT = wild type.

Strain	**Genotype**	**Phenotype**	**Reference**
Escherichia coli BW25113 ara$^-$	F-, *ΔaraD-araB567*, *ΔlacZ4787*::rrnB-3, *λ*$^-$, *rph-1*, *ΔrhaD-rhaB568*, *hsdR514*	WT Red	Baba *et al.* 2006
Escherichia coli BW25113 ara$^+$	F-, *ΔaraD-araB567*, *ΔlacZ4787*::rrnB-3, *λ*$^-$, *rph-1*, *ΔrhaD-rhaB568*, *hsdR514, araA*	WT White	This study
Δ*argH* ara$^-$	WT ara$^-$, Δ*argH*::kan^R	AT	This study
Δ*hisD* ara$^-$	WT ara$^-$, Δ*hisD*::kan^R	AT	This study
Δ*ilvA* ara$^-$	WT ara$^-$, Δ*ilvA*::kan^R	AT	This study
Δ*leuB* ara$^-$	WT ara$^-$, Δ*leuB*::kan^R	AT	This study
Δ*lysA* ara$^-$	WT ara$^-$, Δ*lysA*::kan^R	AT	This study
Δ*metA* ara$^-$	WT ara$^-$, Δ*metA*::kan^R	AT	This study
Δ*pheA* ara$^-$	WT ara$^-$, Δ*pheA*::kan^R	AT	This study
Δ*proC* ara$^-$	WT ara$^-$, Δ*proC*::kan^R	AT	This study
Δ*thrC* ara$^-$	WT ara$^-$, Δ*thrC*::kan^R	AT	This study
Δ*trpB* ara$^-$	WT ara$^-$, Δ*trpB*::kan^R	AT	This study
Δ*tyrA* ara$^-$	WT ara$^-$, Δ*tyrA*::kan^R	AT	This study
Δ*pyrF* ara$^-$	WT ara$^-$, Δ*pyrF*::kan^R	AT	This study
Δ*guaB* ara$^-$	WT ara$^-$, Δ*guaB*::kan^R	AT	This study
Δ*bioF* ara$^-$	WT ara$^-$, Δ*bioH*::kan^R	AT	This study
Δ*nadA* ara$^-$	WT ara$^-$, Δ*nadA*::kan^R	AT	This study
Δ*panC* ara$^-$	WT ara$^-$, Δ*panC*::kan^R	AT	This study
Δ*argH* ara$^+$	WT ara$^+$, Δ*argH*::kan^R	AT	This study
Δ*hisD* ara$^+$	WT ara$^+$, Δ*hisD*::kan^R	AT	This study
Δ*ilvA* ara$^+$	WT ara$^+$, Δ*ilvA*::kan^R	AT	This study
Δ*leuB* ara$^+$	WT ara$^+$, Δ*leuB*::kan^R	AT	This study
Δ*lysA* ara$^+$	WT ara$^+$, Δ*lysA*::kan^R	AT	This study
Δ*metA* ara$^+$	WT ara$^+$, Δ*metA*::kan^R	AT	This study
Δ*pheA* ara$^+$	WT ara$^+$, Δ*pheA*::kan^R	AT	This study
Δ*proC* ara$^+$	WT ara$^+$, Δ*proC*::kan^R	AT	This study
Δ*thrC* ara$^+$	WT ara$^+$, Δ*thrC*::kan^R	AT	This study
Δ*trpB* ara$^+$	WT ara$^+$, Δ*trpB*::kan^R	AT	This study
Δ*tyrA* ara$^+$	WT ara$^+$, Δ*tyrA*::kan^R	AT	This study
Δ*pyrF* ara$^+$	WT ara$^+$, Δ*pyrF*::kan^R	AT	This study
Δ*guaB* ara$^+$	WT ara$^+$, Δ*guaB*::kan^R	AT	This study
Δ*bioH* ara$^+$	WT ara$^+$, Δ*bioH*::kan^R	AT	This study
Δ*nadA* ara$^+$	WT ara$^+$, Δ*nadA*::kan^R	AT	This study
Δ*panC* ara$^+$	WT ara$^+$, Δ*panC*::kan^R	AT	This study
Δ*argA* ara$^-$	WT ara$^-$, Δ*argA*::kan^R	AT	This study
Δ*argB* ara$^-$	WT ara$^-$, Δ*argB*::kan^R	AT	This study
Δ*argC* ara$^-$	WT ara$^-$, Δ*argC*::kan^R	AT	This study

Δ*argE* ara$^-$	WT ara$^-$, Δ*argE*::kan^R	AT	This study
Δ*argG* ara$^-$	WT ara$^-$, Δ*argG*::kan^R	AT	This study
Δ*argA* ara$^+$	WT ara$^+$, Δ*argA*::kan^R	AT	This study
Δ*argB* ara$^+$	WT ara$^+$, Δ*argB*::kan^R	AT	This study
Δ*argC* ara$^+$	WT ara$^+$, Δ*argC*::kan^R	AT	This study
Δ*argE* ara$^+$	WT ara$^+$, Δ*argE*::kan^R	AT	This study
Δ*argG* ara$^+$	WT ara$^+$, Δ*argG*::kan^R	AT	This study
Δ*trpA* ara$^-$	WT ara$^-$, Δ*trpA*::kan^R	AT	This study
Δ*trpD* ara$^-$	WT ara$^-$, Δ*trpD*::kan^R	AT	This study
Δ*trpE* ara$^-$	WT ara$^-$, Δ*trpE*::kan^R	AT	This study
Δ*trpA* ara$^+$	WT ara$^+$, Δ*trpA*::kan^R	AT	This study
Δ*trpD* ara$^+$	WT ara$^+$, Δ*trpD*::kan^R	AT	This study
Δt*rpE* ara$^+$	WT ara$^+$, Δ*trpE*::kan^R	AT	This study
Δ*hisA* ara$^-$	WT ara$^-$, Δ*hisA*::kan^R	AT	This study
Δ*hisB* ara$^-$	WT ara$^-$, Δ*hisB*::kan^R	AT	This study
Δ*hisC* ara$^-$	WT ara$^-$, Δ*hisC*::kan^R	AT	This study
Δ*hisA* ara$^+$	WT ara$^+$, Δ*hisA*::kan^R	AT	This study
Δ*hisB* ara$^+$	WT ara$^+$, Δ*hisB*::kan^R	AT	This study
Δ*hisC* ara$^+$	WT ara$^+$, Δ*hisC*::kan^R	AT	This study
Δ*argH*	WT, Δ*argH*::kan^R	AT	Baba *et al.* 2006
Δ*hisD*	WT, Δ*hisD*::kan^R	AT	Baba *et al.* 2006
Δ*ilvA*	WT, Δ*ilvA*::kan^R	AT	Baba *et al.* 2006
Δ*leuB*	WT, Δ*leuB*::kan^R	AT	Baba *et al.* 2006
Δ*lysA*	WT, Δ*lysA*::kan^R	AT	Baba *et al.* 2006
Δ*metA*	WT, Δ*metA*::kan^R	AT	Baba *et al.* 2006
Δ*pheA*	WT, Δ*pheA*::kan^R	AT	Baba *et al.* 2006
Δ*proC*	WT, Δ*proC*::kan^R	AT	Baba *et al.* 2006
Δ*thrC*	WT, Δ*thrC*::kan^R	AT	Baba *et al.* 2006
Δ*trpB*	WT, Δ*trpB*::kan^R	AT	Baba *et al.* 2006
Δ*tyrA*	WT , Δ*tyrA*::kan^R	AT	Baba *et al.* 2006
Δ*pyrF*	WT , Δ*pyrF*::kan^R	AT	Baba *et al.* 2006
Δ*guaB*	WT, Δ*guaB*::kan^R	AT	Baba *et al.* 2006
Δ*bioF*	WT, Δ*bioH*::kan^R	AT	Baba *et al.* 2006
Δ*nadA*	WT, Δ*nadA*::kan^R	AT	Baba *et al.* 2006
Δ*panC*	WT, Δ*panC*::kan^R	AT	Baba *et al.* 2006
Δ*argA*	WT ,Δ*argA*::kan^R	AT	Baba *et al.* 2006
Δ*argB*	WT ,Δ*argB*::kan^R	AT	Baba *et al.* 2006
Δ*argC*	WT ,Δ*argC*::kan^R	AT	Baba *et al.* 2006
Δ*argE*	WT, Δ*argE*::kan^R	AT	Baba *et al.* 2006
Δ*argG*	WT, Δ*argG*::kan^R	AT	Baba *et al.* 2006
Δ*trpA*	WT, Δ*trpA*::kan^R	AT	Baba *et al.* 2006

Δ*trpD*	WT, Δ*trpD*::kan^R	AT	Baba *et al.* 2006
Δ*trpE*	WT, Δ*trpE*::kan^R	AT	Baba *et al.* 2006
Δ*hisA*	WT, Δ*hisA*::kan^R	AT	Baba *et al.* 2006
Δ*hisB*	WT, Δ*hisB*::kan^R	AT	Baba *et al.* 2006
Δ*hisC*	WT, Δ*hisC*::kan^R	AT	Baba *et al.* 2006
REL 606	F-,*tsx-467Am*, *araA* *92D*, *lon*, *rpsL227* strR, *hsdR*, *[mal*+*]*LamS		Studier *et al.* 2009
REL 607	F-, *tsx-467*Am, *araA* 92G, lon-, *rpsL*227 strR, hsdR-, [mal+]LamS		Lenski *et al.* 1991
Acinetobacter baylyi ADP1		WT	Vaneechoutte *et al.* 2006
A. baylyi Δ*argH*	WT, Δ*argH*::kan^R	AT	This study
A. baylyi Δ*hisD*	WT, Δ*hisD*::kan^R	AT	This study
A. baylyi Δl*euB*	WT, Δ*leuB*::kan^R	AT	This study
A. baylyi Δ*trpB*	WT, Δ*trpB*::kan^R	AT	This study

Table S3. Comparison of biosynthetic costs for DNA sequence and the corresponding protein. Cost are given as fructose molecules that are at least needed to produce the DNA sequence of the gene or the amount of the protein. NA=no data available.

Gene	**Cost of DNA Sequence 10^4 fructose molecules**	**Cost of Protein 10^4 fructose molecules**
argA	1.783502731	3.439609569
argB	1.088531764	NA
argC	1.373221287	NA
argE	1.555713634	5.507498182
argG	1.796792481	273.9335755
argH	1.8338215	64.23576185
hisA	1.042527654	NA
hisB	1.451117531	NA
hisC	1.452515028	30.78864344
hisD	1.739683547	24.86644536
trpA	1.126263614	61.31053206
trpB	1.607108823	89.72362679
trpD	2.103003962	16.01084019
trpE	2.061666192	15.98466297

Table S4. Primers used for the construction of *Acinetobacter baylyi* auxotrophs. UF = upstream forward, UR = Upstream reverse, DF = downstream forward, DR = downstream reverse.

Gene	Primer	Sequence 5'-3'
Kanamycin resistance cassette	UF	TGTAGGCTGGAGCTGCTTC
	UR	CATATGAATATCCTCCTTA
argH	UF	GAGGTCTGGGTTGAGGTTGG
	UR	GAAGCAGCTCCAGCCTACATAACGCTGCATTTGCAC
hisD	UF	TATGCAAGCCTTGGTGAGCA
	UR	GAAGCAGCTCCAGCCTACACAGCCTCTTCCACTTGA
leuB	UF	CCGTTTACAGGGCTCAGTGT
	UR	GAAGCAGCTCCAGCCTACATCACCCAATCCTGTCAC
trpB	UF	AACCACACACGCTTTTGCAG
	UR	GAAGCAGCTCCAGCCTACAGCTGATCCACATTGGACT
argH	DF	TAAGGAGGATATTCATATGTGCTTCTGGTTTCCAGC
	DR	GGATTTTGCGCCATTCCCTG
hisD	DF	TAAGGAGGATATTCATATGGTAACTGCTCTACGGGG
	DR	ATGCGTCTGCCTGATCTACC
leuB	DF	TAAGGAGGATATTCATATGTTGCCCGAACACCGATC
	DR	CGTTCACGAATCCATGCAAGT
trpB	DF	TAAGGAGGATATTCATATGACGTGATGTGGAAATGG
	DR	AGTTGGGGCTGGATGTCTTG

LITERATURE CITED

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

Caspi, R., T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, et al. 2012. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 40:D742-D753.

Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283-1287.

Datsenko, K. A. and B. L. Wanner. 2000. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc. Natl. Acad. Sci. USA 97:6640-6645.

Fujita, A. I. and T. Abe. 2002. Amino acid concentration and distribution of lysozyme and protease activities in the guts of higher termites. Physiol. Entomol. 27:76-78.

Kaleta, C., S. Schäuble, U. Rinas, and S. Schuster. 2013. Metabolic costs of amino acid and protein production in *Escherichia coli*. Biotechnol. J. 8:1105-1114.

Keseler, I. M., A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martínez, C. Fulcher, A. M. Huerta, A. Kothari, M. Krummenacker, et al. 2013. EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res. 41:D605-D612.

Klumpp, S., Z. Zhang, and T. Hwa. 2009. Growth rate-dependent global effects on gene expression in bacteria. Cell 139:1366-1375.

Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler. 1991. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. Am. Nat. 138:1315-1341.

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

Pagani, I., K. Liolios, J. Jansson, I. M. A. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz, and N. C. Kyrpides. 2012. The Genomes OnLine Database GOLD v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 40:D571-D579.

Schellenberger, J., R. Que, R. M. T. Fleming, I. Thiele, J. D. Orth, A. M. Feist, D. C. Zielinski, A. Bordbar, N. E. Lewis, S. Rahmanian, et al. 2011. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat. Protocols 6:1290-1307.

Studier, F. W., P. Daegelen, R. E. Lenski, S. Maslov, and J. F. Kim. 2009. Understanding the differences between genome sequences of *Escherichia coli* B strains REL606 and BL21DE3 and comparison of the *E. coli* B and K-12 genomes. J. Mol. Biol. 94:653-680.

Vallenet, D., S. Engelen, D. Mornico, S. Cruveiller, L. Fleury, A. Lajus, Z. Rouy, D. Roche, G. Salvignol, C. Scarpelli, et al. 2009. MicroScope: a platform for microbial genome annotation and comparative genomics. J. Biol. Datab. Curation doi: 10.1093/database/bap021.

Vaneechoutte, M., D. M. Young, L. N. Ornston, T. De Baere, A. Nemec, T. Van Der Reijden, E. Carr, I. Tjernberg, and L. Dijkshoorn. 2006. Naturally transformable *Acinetobacter* sp. strain ADP1 belongs to the newly described species *Acinetobacter baylyi*. Appl. Environ. Microbiol. 72:932-936.

Werdin-Pfisterer, N. R., K. Kielland, and R. D. Boone. 2012. Buried organic horizons represent amino acid reservoirs in boreal forest soils. Soil Biol. Biochem. 55:122-131.

Wessely, F., M. Bartl, R. Guthke, P. Li, S. Schuster, and C. Kaleta. 2011. Optimal regulatory strategies for metabolic pathways in *Escherichia coli* depending on protein costs. Mol. Syst. Biol. 7:515.

Supporting information for chapter IV

Plasticity and epistasis strongly affect bacterial fitness after losing multiple metabolic genes

SUPPORTING METHODS

Adjustment of fructose and succinate concentrations

We used Flux-Balance-Analysis and a genome-scale metabolic model of *E. coli* (Orth *et al.* 2011) to calculate how many mole of a carbon source are needed to produce '1 mole of biomass'. We refer to this value using q_x for a carbon source *x*. In detail, q_x was calculated by constraining the flux through the biomass reaction (growth associated maintenance (GAM) estimate: 53.95) of the model to a value equal 1 mmol x gDW^{-1} x h^{-1} and by minimizing the influx of the carbon source *x*. The optimization was performed within Matlab 7.14 (Mathworks) with the COBRA Toolbox version 2.0.5 (Schellenberger et al. 2011) and the TOMLAB v7.9 as linear programming solver. The final concentration of carbon source x was calculated as

$$c_x = c_{Fru} \cdot q_x / q_{Fru}$$

using 5 g l^{-1} fructose (c_{Fru} = 27.75 mM) as reference. The corresponding concentration of disodium succinate was 8.86 g l^{-1} (c_{Suc} = 54.68 mM).

This procedure is similar to the approach used by Adadi *et al.* (2012), where concentrations were adjusted to match the number of reducible carbon atoms. Using the genome-scale metabolic network of *E. coli* also takes the physiological capabilities of the cell to transform a certain carbon source into biomass into account.

SUPPORTING FIGURES

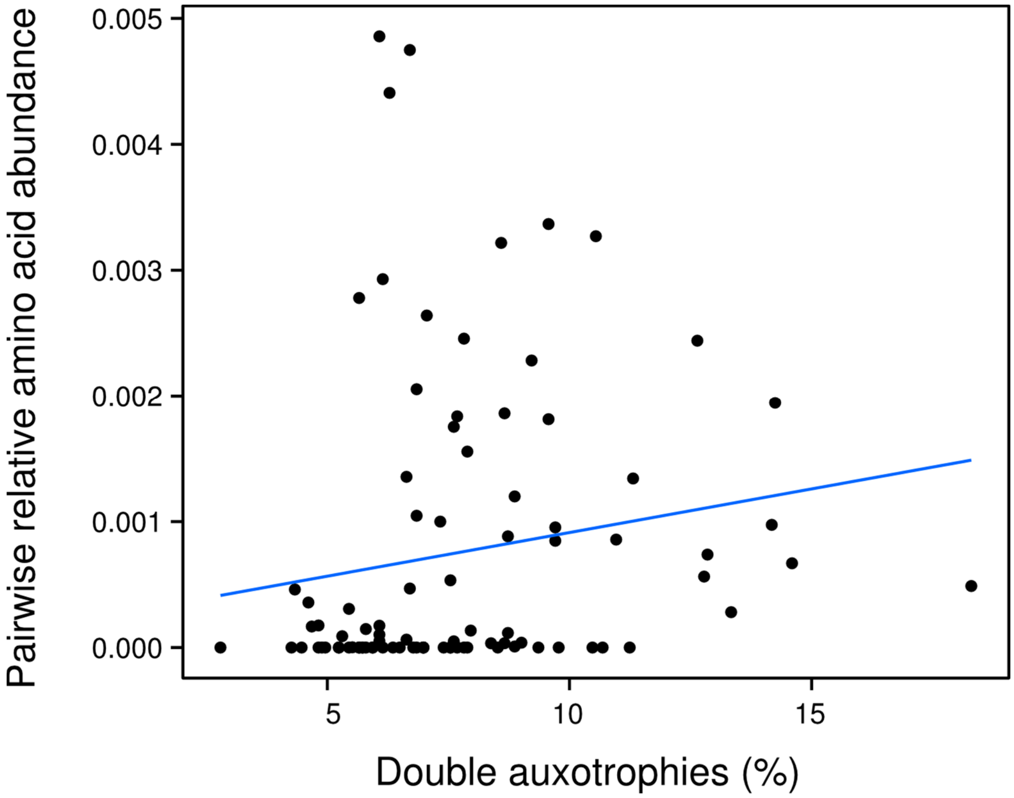


Figure S1. Correlation of the frequency of double auxotrophies among 1,432 eubacteria and the median of pairwise products of amino acid abundances in 69 natural environments (Moura *et al.* 2013). Kendall's rank correlation: $R_\tau = 0.22$, $P = 0.003$, $n = 91$.

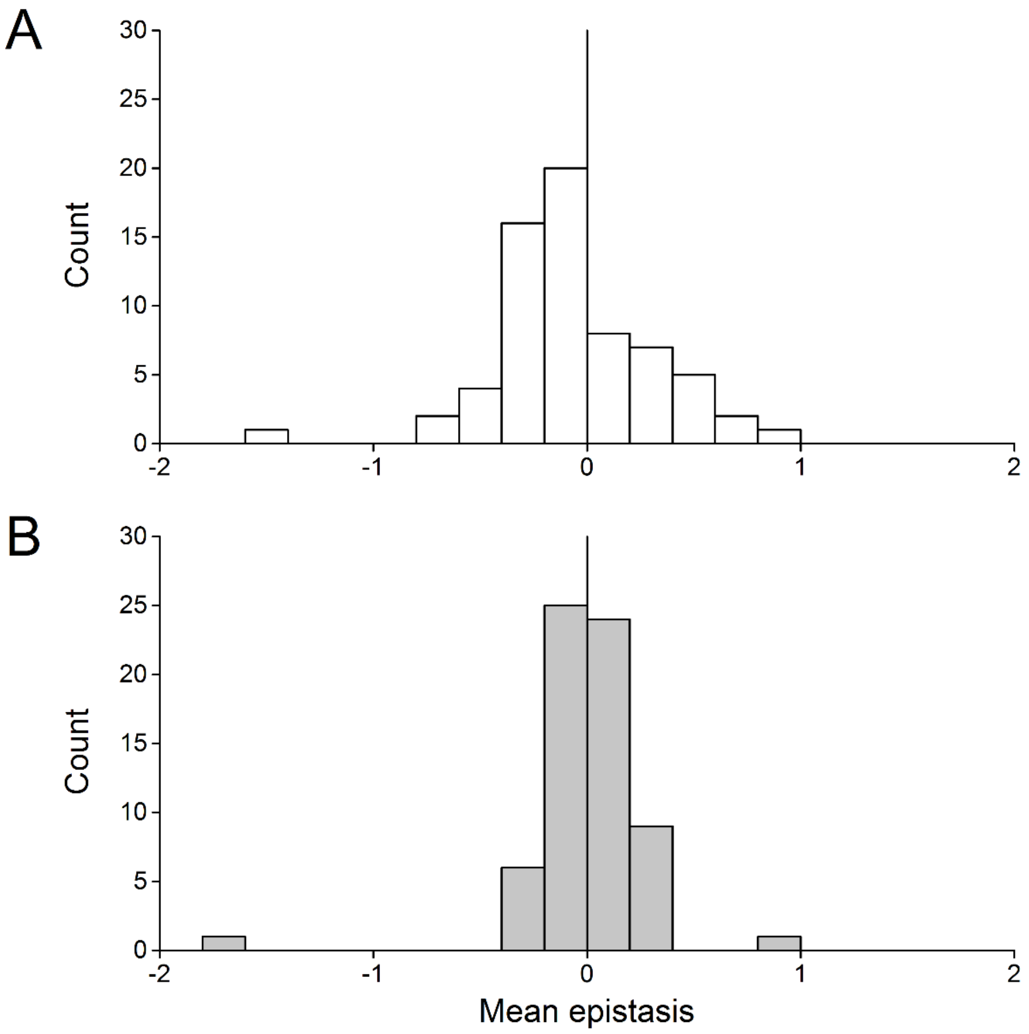


Figure S2. Frequency distribution of epistatic effects for 55 double- and 16 triple gene deletion mutants as determined in (A) the fructose- and (B) the succinate-containing environment.

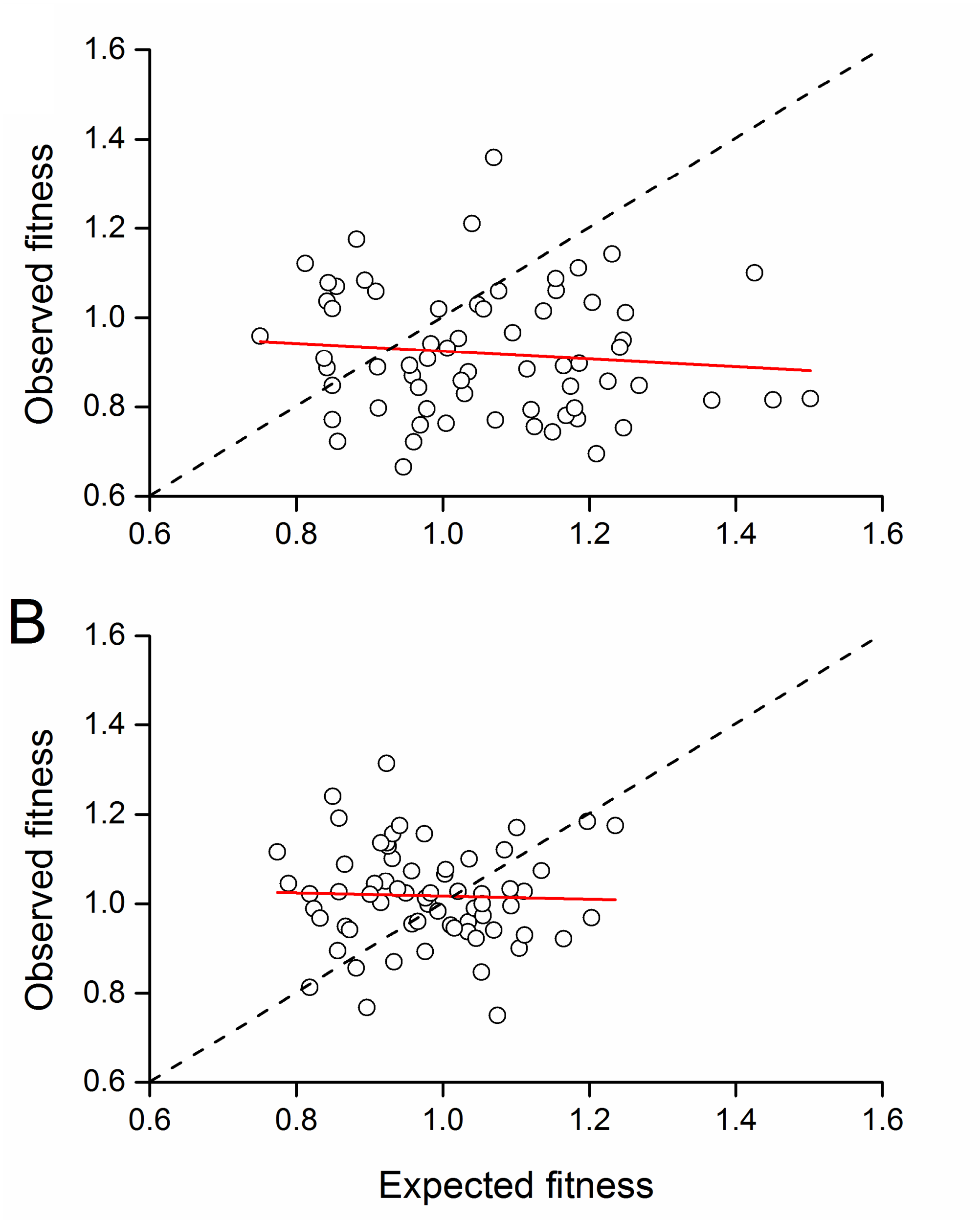


Figure S3. Type II Standard Major Axis (SMA) regression of observed and expected fitness as determined in (A) the fructose- and (B) the succinate-containing environment. The solid red line represents the regression ($P > 0.05$, $n = 66$), while the dotted black line indicates the null model assuming no epistasis.

SUPPORTING TABLES

Table S1. Strains used in this study. Abbreviations: ara$^{+/-}$ = ability to use arabinose as a carbon source present/ absent, AT = auxotroph, WT = wild type.

Strain	**Genotype**	**Phenotype**	**Reference**
Escherichia coli BW25113 ara^{-}	F-, *ΔaraD-araB567*, *ΔlacZ4787*::rrnB-3, *λ*$^{-}$, *rph-1*, *ΔrhaD-rhaB568*, *hsdR514*	WT Red	Baba et al. 2006
Escherichia coli BW25113 ara^{+}	F-, *ΔaraD-araB567*, *ΔlacZ4787*::rrnB-3, *λ*$^{-}$, *rph-1*, *ΔrhaD-rhaB568*, *hsdR514, araA*	WT White	D'Souza et al. 2014
Δ*argH*	WT ara^{-}, Δ*argH*::kan^{R}	AT	D'Souza et al. 2014
Δ*hisD*	WT ara^{-}, Δ*hisD*::kan^{R}	AT	D'Souza et al. 2014
Δ*ilvA*	WT ara^{-}, Δ*ilvA*::kan^{R}	AT	D'Souza et al. 2014
Δ*leuB*	WT ara^{-}, Δ*leuB*::kan^{R}	AT	D'Souza et al. 2014
Δ*lysA*	WT ara^{-}, Δ*lysA*::kan^{R}	AT	D'Souza et al. 2014
Δ*metA*	WT ara^{-}, Δ*metA*::kan^{R}	AT	D'Souza et al. 2014
Δ*pheA*	WT ara^{-}, Δ*pheA*::kan^{R}	AT	D'Souza et al. 2014
Δ*proC*	WT ara^{-}, Δ*proC*::kan^{R}	AT	D'Souza et al. 2014
Δ*thrC*	WT ara^{-}, Δ*thrC*::kan^{R}	AT	D'Souza et al. 2014
Δ*trpB*	WT ara^{-}, Δ*trpB*::kan^{R}	AT	D'Souza et al. 2014
Δ*tyrA*	WT ara^{-}, Δ*tyrA*::kan^{R}	AT	D'Souza et al. 2014
Δ*argH* Δ*ilvA*	WT ara^{-}, Δ*argH,* Δ*ilvA*::kan^{R}	AT	This study
Δ*argH* Δ*leuB*	WT ara^{-}, Δ*argH,* Δ*leuB*::kan^{R}	AT	This study
Δ*argH* Δ*lysA*	WT ara^{-}, Δ*argH,* Δ*lysA*::kan^{R}	AT	This study
Δ*metA* Δ*argH*	WT ara^{-}, Δ*metA,* Δ*argH*::kan^{R}	AT	This study
Δ*argH* Δ*pheA*	WT ara^{-}, Δ*argH,* Δ*pheA*::kan^{R}	AT	This study
Δ*proC* Δ*argH*	WT ara^{-}, Δ*proC,* Δ*argH*::kan^{R}	AT	This study
Δ*argH* Δ*thrC*	WT ara^{-}, Δ*argH,* Δ*thrC*::kan^{R}	AT	This study
Δ*argH* Δ*trpB*	WT ara^{-}, Δ*argH,* Δ*trpB*::kan^{R}	AT	This study
Δ*argH* Δ*tyrA*	WT ara^{-}, Δ*argH,* Δ*tyrA*::kan^{R}	AT	This study
Δ*ilvA* Δ*hisD*	WT ara^{-}, Δ*ilvA,* Δ*hisD*::kan^{R}	AT	This study
Δ*leuB* Δ*hisD*	WT ara^{-}, Δ*leuB,* Δ*hisD*::kan^{R}	AT	This study
Δ*lysA* Δ*hisD*	WT ara^{-}, Δ*lysA,* Δ*hisD*::kan^{R}	AT	This study
Δ*metA* Δ*hisD*	WT ara^{-}, Δ*metA,* Δ*hisD*::kan^{R}	AT	This study
Δ*hisD* Δ*pheA*	WT ara^{-}, Δ*hisD,* Δ*pheA*::kan^{R}	AT	This study
Δ*hisD* Δ*proC*	WT ara^{-}, Δ*hisD,* Δ*proC*::kan^{R}	AT	This study
Δ*hisD* Δ*thrC*	WT ara^{-}, Δ*hisD,* Δ*thrC*::kan^{R}	AT	This study
Δ*hisD* Δ*trpB*	WT ara^{-}, Δ*hisD,* Δ*trpB*::kan^{R}	AT	This study
Δ*hisD* Δ*tyrA*	WT ara^{-}, Δ*hisD,* Δ*tyrA*::kan^{R}	AT	This study
Δ*ilvA* Δ*leuB*	WT ara^{-}, Δ*ilvA,* Δ*leuB*::kan^{R}	AT	This study
Δ*ilvA* Δ*lysA*	WT ara^{-}, Δ*ilvA,* Δ*lysA*::kan^{R}	AT	This study
Δ*ilvA* Δ*metA*	WT ara^{-}, Δ*ilvA,* Δ*metA*::kan^{R}	AT	This study
Δ*ilvA* Δ*pheA*	WT ara^{-}, Δ*ilvA,* Δ*pheA*::kan^{R}	AT	This study

Δ*ilvA* Δ*proC*	WT ara$^-$, Δ*ilvA,* Δ*proC*::kan^R	AT	This study
Δ*ilvA* Δ*thrC*	WT ara$^-$, Δ*ilvA,* Δ*thrC*::kan^R	AT	This study
Δ*trpB* Δ*ilvA*	WT ara$^-$, Δ*trpB,* Δ*ilvA*::kan^R	AT	This study
Δ*ilvA* Δ*tyrA*	WT ara$^-$, Δ*ilvA,* Δ*tyrA*::kan^R	AT	This study
Δ*leuB* Δ*lysA*	WT ara$^-$, Δ*leuB,* Δ*lysA*::kan^R	AT	This study
Δ*metA* Δ*leuB*	WT ara$^-$, Δ*metA,* Δ*leuB*::kan^R	AT	This study
Δ*pheA* Δ*leuB*	WT ara$^-$, Δ*pheA,* Δ*leuB*::kan^R	AT	This study
Δ*proC* Δ*leuB*	WT ara$^-$, Δ*proC,* Δ*leuB*::kan^R	AT	This study
Δ*thrC* Δ*leuB*	WT ara$^-$, Δ*thrC,* Δ*leuB*::kan^R	AT	This study
Δ*leuB* Δ*trpB*	WT ara$^-$, Δ*leuB,* Δ*trpB*::kan^R	AT	This study
Δ*lysA* Δ*metA*	WT ara$^-$, Δ*lysA,* Δ*metA*::kan^R	AT	This study
Δ*lysA* Δ*pheA*	WT ara$^-$, Δ*lysA,* Δ*pheA* ::kan^R	AT	This study
Δ*lysA* Δ*proC*	WT ara$^-$, Δ*lysA,* Δ*proC*::kan^R	AT	This study
Δ*thrC* Δ*lysA*	WT ara$^-$, Δ*thrC,* Δ*lysA*::kan^R	AT	This study
Δ*lysA* Δ*trpB*	WT ara$^-$, Δ*lysA,* Δ*trpB*::kan^R	AT	This study
Δ*metA* Δ*pheA*	WT ara$^-$, Δ*metA,* Δ*pheA*::kan^R	AT	This study
Δ*proC* Δ*metA*	WT ara$^-$, Δ*proC,* Δ*metA*::kan^R	AT	This study
Δ*metA* Δ*thrC*	WT ara$^-$, Δ*metA,* Δ*thrC*::kan^R	AT	This study
Δ*metA* Δ*trpB*	WT ara$^-$, Δ*metA,* Δ*trpB*::kan^R	AT	This study
Δ*pheA* Δ*proC*	WT ara$^-$, Δ*pheA,* Δ*proC*::kan^R	AT	This study
Δ*pheA* Δ*thrC*	WT ara$^-$, Δ*pheA,* Δ*thrC*::kan^R	AT	This study
Δ*pheA* Δ*trpB*	WT ara$^-$, Δ*pheA,* Δ*trpB*::kan^R	AT	This study
Δ*pheA* Δ*tyrA*	WT ara$^-$, Δ*pheA,* Δ*tyrA*::kan^R	AT	This study
Δ*proC* Δ*thrC*	WT ara$^-$, Δ*proC,* Δ*thrC*::kan^R	AT	This study
Δ*trpB* Δ*proC*	WT ara$^-$, Δ*trpB* , Δ*proC*::kan^R	AT	This study
Δ*thrC* Δ*trpB*	WT ara$^-$, Δ*thrC,* Δ*trpB*::kan^R	AT	This study
Δ*thrC* Δ*tyrA*	WT ara$^-$, Δ*thrC,* Δ*tyrA*::kan^R	AT	This study
Δ*trpB* Δ*tyrA*	WT ara$^-$, Δ*trpB,* Δ*tyrA*::kan^R	AT	This study
Δ*trpB* Δ*pheA* Δ*tyrA*	WT ara$^-$, Δ*trpB,* Δ*pheA,* Δ*tyrA*::kan^R	AT	This study
Δ*lysA* Δ*metA* Δ*argH*	WT ara$^-$, Δ*lysA,* Δ*metA,* Δ*argH*::kan^R	AT	This study
Δ*lysA* Δ*metA* Δ*thrC*	WT ara$^-$, Δ *lysA* Δ*metA* Δ*thrC*::kan^R	AT	This study
Δ*trpB* Δ*pheA* Δ*metA*	WT ara$^-$, Δ*trpB,* Δ*pheA,* Δ*metA*::kan^R	AT	This study
Δ*trpB* Δ*pheA* Δ*leuB*	WT ara$^-$, Δ*trpB,* Δ*pheA,* Δ*leuB*::kan^R	AT	This study
Δ*metA* Δ*thrC* Δ*argH*	WT ara$^-$, Δ*metA,* Δ*thrC,* Δ*argH*::kan^R	AT	This study
Δ*metA* Δ*thrC* Δ*hisD*	WT ara$^-$, Δ*metA,* Δ*thrC,* Δ*hisD*::kan^R	AT	This study
Δ*thrC* Δ*lysA* Δ*hisD*	WT ara$^-$, Δ*thrC,* Δ*lysA,* Δ*hisD*::kan^R	AT	This study
Δ*trpB* Δ*pheA* Δ*hisD*	WT ara$^-$, Δ*trpB,* Δ*pheA,* Δ*hisD*::kan^R	AT	This study
Δ*proC* Δ*thrC* Δ*ilvA*	WT ara$^-$, Δ*proC,* Δ*thrC,* Δ*ilvA*::kan^R	AT	This study
Δ*trpB* Δ*leuB* Δ*thrC*	WT ara$^-$, Δ*trpB,* Δ*leuB,* Δ*thrC*::kan^R	AT	This study
Δ*metA* Δ*argH* Δ*hisD*	WT ara$^-$, Δ*metA,* Δ*argH,* Δ*hisD*::kan^R	AT	This study
Δ*proC* Δ*lysA* Δ*hisD*	WT ara$^-$, Δ*proC,* Δ*lysA,* Δ*hisD*::kan^R	AT	This study

Δ*proC* Δ*lysA* Δ*tyrA*	WT ara$^-$, Δ*proC*, Δ*lysA*, Δ*tyrA*::kan^R	AT	This study
Δ*ilvA* Δ*thrC* Δ*trpB*	WT ara$^-$, Δ*ilvA*, Δ*thrC*, Δ*trpB*::kan^R	AT	This study
Δ*trpB* Δ*pheA* Δ*thrC*	WT ara$^-$, Δ*trpB*, Δ*pheA*, Δ*thrC*::kan^R	AT	This study
Δ*argH*::kan^S	WT ara$^-$, Δ*argH*::kan^S	AT	This study
Δ*hisD*::kan^S	WT ara$^-$, Δ*hisD*::kan^S	AT	This study
Δ*ilvA*::kan^S	WT ara$^-$, Δ*ilvA*::kan^S	AT	This study
Δ*leuB*::kan^S	WT ara$^-$, Δ*leuB*::kan^S	AT	This study
Δ*lysA*::kan^S	WT ara$^-$, Δ*lysA*::kan^S	AT	This study
Δ*metA*::kan^S	WT ara$^-$, Δ*metA*::kan^S	AT	This study
Δ*pheA*::kan^S	WT ara$^-$, Δ*pheA*::kan^S	AT	This study
Δ*proC*::kan^S	WT ara$^-$, Δ*proC*::kan^S	AT	This study
Δ*thrC*::kan^S	WT ara$^-$, Δ*thrC*::kan^S	AT	This study
Δ*trpB*::kan^S	WT ara$^-$, Δ*trpB*::kan^S	AT	This study
Δ*ilvA* Δ*leuB*::kan^S	WT ara$^-$, Δ*ilvA*, Δ*leuB*::kan^S	AT	This study
Δ*ilvA* Δ*thrC*::kan^S	WT ara$^-$, Δ*ilvA*, Δ*thrC*::kan^S	AT	This study
Δ*thrC* Δ*trpB*::kan^S	WT ara$^-$, Δ*thrC*, Δ*trpB*::kan^S	AT	This study
Δ*thrC* Δ*trpB*::kan^S	WT ara$^-$, Δ*thrC*, Δ*trpB*::kan^S	AT	This study
Δ*ilvA* Δ*thrC* Δ*trpB*::kan^S	WT ara$^-$, Δ*ilvA*, Δ*thrC*, Δ*trpB*::kan^S	AT	This study
Δ*lysA* Δ*metA* Δ*thrC*::kan^S	WT ara$^-$, Δ*lysA*, Δ*metA*, Δ*thrC*::kan^S	AT	This study
Δ*trpB* Δ*pheA* Δ*thrC*::kan^S	WT ara$^-$, Δ*trpB*, Δ*pheA*, Δ*thrC*::kan^S	AT	This study
Δ*trpB* Δ*leuB* Δ*thrC*::kan^S	WT ara$^-$, Δ*trpB*, Δ*leuB*, Δ*thrC*::kan^S	AT	This study
Δ*argH*	WT, Δ*argH*::kan^R	AT	Baba *et al.* 2006
Δ*hisD*	WT, Δ*hisD*::kan^R	AT	Baba *et al.* 2006
Δ*ilvA*	WT, Δ*ilvA*::kan^R	AT	Baba *et al.* 2006
Δ*leuB*	WT, Δ*leuB*::kan^R	AT	Baba *et al.* 2006
Δ*lysA*	WT, Δ*lysA*::kan^R	AT	Baba *et al.* 2006
Δ*metA*	WT, Δ*metA*::kan^R	AT	Baba *et al.* 2006
Δ*pheA*	WT, Δ*pheA*::kan^R	AT	Baba *et al.* 2006
Δ*proC*	WT, Δ*proC*::kan^R	AT	Baba *et al.* 2006
Δ*thrC*	WT, Δ*thrC*::kan^R	AT	Baba *et al.* 2006
Δ*trpB*	WT, Δ*trpB*::kan^R	AT	Baba *et al.* 2006
Δ*tyrA*	WT , Δ*tyrA*::kan^R	AT	Baba *et al.* 2006

Table S2. Fitness cost of the kanamycin resistance marker. Mean Malthusian parameter (± 95% confidence interval (CI)) of kanamycin resistant (kan^{R}) and sensitive (kan^{S}) auxotrophic mutants was determined by coculturing both competitors in the fructose-containing environment for 24 h. Each comparison has been replicated 10 times. P values of independent sample t-tests are given. % MDD = minimum detectable difference calculated as described (Zar 1999).

Genotype	Malthusian parameter	± 95% CI	P value	% MDD
Δ*argH*::kan^{R}	5.59	0.24	0.95	2.80
Δ*argH*::kan^{S}	5.60	0.20		
Δ*hisD*::kan^{R}	4.64	0.10	0.10	0.24
Δ*hisD*::kan^{S}	4.76	0.09		
Δ*ilvA*::kan^{R}	5.93	0.30	0.73	2.24
Δ*ilvA*::kan^{S}	6.01	0.28		
Δ*leuB*::kan^{R}	5.48	0.13	0.31	0.92
Δ*leuB*::kan^{S}	5.55	0.11		
Δ*lysA*::kan^{R}	5.10	0.17	0.31	0.85
Δ*lysA*::kan^{S}	5.01	0.05		
Δ*metA*::kan^{R}	5.24	0.07	0.66	1.84
Δ*metA*::kan^{S}	5.27	0.07		
Δ*pheA*::kan^{R}	5.09	0.10	0.62	1.67
Δ*pheA*::kan^{S}	5.05	0.09		
Δ*proC*::kan^{R}	4.89	0.11	0.05	0.13
Δ*proC*::kan^{S}	4.75	0.06		
Δ*thrC*::kan^{R}	5.24	0.05	0.82	2.27
Δ*thrC*::kan^{S}	5.23	0.06		
Δ*trpB*::kan^{R}	6.29	0.07	0.65	2.15
Δ*trpB*::kan^{S}	6.27	0.06		
Δ*ilvA* Δ*leuB*::kan^{R}	4.82	0.24	0.91	2.31
Δ*ilvA* Δ*leuB*::kan^{S}	4.83	0.18		
Δ*ilvA* Δ*thrC*::kan^{R}	5.38	0.67	0.33	0.93
Δ*ilvA* Δ*thrC*::kan^{S}	4.88	0.73		
Δ*lysA* Δ*metA*::kan^{R}	5.04	0.07	0.67	1.77
Δ*lysA* Δ*metA*::kan^{S}	5.02	0.07		
Δ*thrC* Δ*trpB*::kan^{R}	4.92	0.58	0.42	1.07
Δ*thrC* Δ*trpB*::kan^{S}	4.59	0.52		
Δ*ilvA* Δ*thrC* Δ*trpB*::kan^{R}	6.19	0.19	0.41	1.35
Δ*ilvA* Δ*thrC* Δ*trpB*::kan^{S}	6.29	0.11		
Δ*lysA* Δ*metA* Δ*thrC*::kan^{R}	5.66	0.09	0.91	2.72
Δ*lysA* Δ*metA* Δ*thrC*::kan^{S}	5.67	0.14		
Δ*trpB* Δ*pheA* Δ*thrC*::kan^{R}	5.87	0.07	0.68	2.09
Δ*trpB* Δ*pheA* Δ*thrC*::kan^{S}	5.89	0.08		
Δ*trpB* Δ*leuB* Δ*thrC*::kan^{R}	5.78	0.07	0.49	1.51
Δ*trpB* Δ*leuB* Δ*thrC*::kan^{S}	5.83	0.12		

Table S3. Relative fitness and epistatic interactions among auxotrophy-causing mutations in the fructose-containing environment. Mean fitness of each mutant genotype relative to wild type (± 95% confidence interval (CI)) was calculated from 8 replicates. Epistasis was estimated by comparing estimated and observed fitness values using a multiplicative model. Instances of significant epistasis are depicted in bold. NA = not applicable.

Genotype	Relative fitness	± 95% CI	Epistasis
Δ*argH*	0.84	0.08	NA
Δ*hisD*	0.95	0.05	NA
Δ*ilvA*	1.23	0.13	NA
Δ*leuB*	1.02	0.08	NA
Δ*lysA*	0.88	0.05	NA
Δ*metA*	1.21	0.07	NA
Δ*pheA*	1.17	0.10	NA
Δ*proC*	1.00	0.04	NA
Δ*thrC*	0.99	0.07	NA
Δ*trpB*	0.95	0.08	NA
Δ*tyrA*	1.01	0.04	NA
Δ*argH* Δ*ilvA*	1.02	0.03	-0.01
Δ*argH* Δ*leuB*	1.02	0.03	-0.03
ΔargH ΔlysA	**0.95**	**0.06**	**0.20**
ΔmetA ΔargH	**0.82**	**0.07**	**-0.20**
Δ*argH* Δ*pheA*	1.02	0.04	0.02
ΔproC ΔargH	**1.07**	**0.03**	**0.21**
ΔargH ΔthrC	**1.07**	**0.11**	**0.23**
ΔargH ΔtrpB	**1.12**	**0.06**	**0.30**
ΔargH ΔtyrA	**0.72**	**0.09**	**-0.13**
ΔilvA ΔhisD	**0.84**	**0.05**	**-0.32**
Δ*leuB* Δ*hisD*	0.88	0.09	0.04
ΔlysA ΔhisD	**1.03**	**0.05**	**0.19**
Δ*metA* Δ*hisD*	1.06	0.05	-0.09
ΔhisD ΔpheA	**0.88**	**0.06**	**-0.23**
Δ*hisD* Δ*proC*	0.86	0.03	-0.08
ΔhisD ΔthrC	**0.66**	**0.05**	**-0.28**
Δ*hisD* Δ*trpB*	0.88	0.08	-0.02
ΔhisD ΔtyrA	**0.72**	**0.03**	**-0.23**
ΔilvA ΔleuB	**0.84**	**0.03**	**-0.42**
ΔilvA ΔlysA	**0.96**	**0.11**	**-0.12**
ΔilvA ΔmetA	**0.81**	**0.08**	**-0.68**
ΔilvA ΔpheA	**0.81**	**0.06**	**-0.63**
ΔilvA ΔproC	**0.75**	**0.06**	**-0.49**
Δ*ilvA* Δ*thrC*	1.14	0.09	-0.08
Δ*trpB* Δ*ilvA*	1.11	0.22	-0.07
ΔilvA ΔtyrA	**1.01**	**0.07**	**-0.23**
ΔleuB ΔlysA	**1.05**	**0.16**	**0.15**
ΔmetA ΔleuB	**0.94**	**0.17**	**-0.29**
ΔpheA ΔleuB	**1.03**	**0.05**	**-0.17**
ΔproC ΔleuB	**0.87**	**0.06**	**-0.15**

Δ*thrC* Δ*leuB*	0.95	0.06	-0.06
Δ*leuB* Δ*trpB*	0.94	0.05	-0.04
Δ*lysA* Δ*metA*	1.06	0.04	-0.01
Δ*lysA* Δ*pheA*	**1.21**	**0.12**	**0.17**
Δ*lysA* Δ*proC*	**1.08**	**0.07**	**0.19**
Δ*thrC* Δ*lysA*	**1.17**	**0.08**	**0.29**
Δ*lysA* Δ*trpB*	**1.02**	**0.08**	**0.17**
Δ*metA* Δ*pheA*	**1.10**	**0.09**	**-0.32**
Δ*proC* Δ*metA*	**0.85**	**0.02**	**-0.36**
Δ*metA* Δ*thrC*	**0.69**	**0.04**	**-0.51**
Δ*metA* Δ*trpB*	**0.89**	**0.12**	**-0.27**
Δ*pheA* Δ*proC*	**0.77**	**0.06**	**-0.41**
Δ*pheA* Δ*thrC*	**0.78**	**0.10**	**-0.38**
Δ*pheA* Δ*trpB*	**0.75**	**0.09**	**-0.37**
Δ*pheA* Δ*tyrA*	**0.89**	**0.08**	**-0.28**
Δ*proC* Δ*thrC*	**0.76**	**0.08**	**-0.24**
Δ*trpB* Δ*proC*	**0.84**	**0.05**	**-0.12**
Δ*thrC* Δ*trpB*	0.89	0.10	-0.06
Δ*thrC* Δ*tyrA*	0.93	0.10	-0.07
Δ*trpB* Δ*tyrA*	**0.75**	**0.07**	**-0.21**
Δ*trpB* Δ*pheA* Δ*tyrA*	**1.01**	**0.13**	**0.74**
Δ*lysA* Δ*metA* Δ*argH*	**0.77**	**0.07**	**0.23**
Δ*lysA* Δ*metA* Δ*thrC*	**0.79**	**0.04**	**-0.10**
Δ*trpB* Δ*pheA* Δ*metA*	**0.81**	**0.04**	**0.41**
Δ*trpB* Δ*pheA* Δ*leuB*	**1.08**	**0.12**	**0.51**
Δ*metA* Δ*thrC* Δ*argH*	**0.85**	**0.10**	**0.31**
Δ*metA* Δ*thrC* Δ*hisD*	**0.74**	**0.04**	**0.48**
Δ*thrC* Δ*lysA* Δ*hisD*	**0.90**	**0.06**	**-0.13**
Δ*trpB* Δ*pheA* Δ*hisD*	**0.79**	**0.11**	**0.49**
Δ*proC* Δ*thrC* Δ*ilvA*	**1.35**	**0.12**	**0.91**
Δ*trpB* Δ*leuB* Δ*thrC*	**0.93**	**0.07**	**0.51**
Δ*metA* Δ*argH* Δ*hisD*	**0.90**	**0.08**	**0.10**
Δ*proC* Δ*lysA* Δ*hisD*	**0.79**	**0.07**	**-1.48**
Δ*proC* Δ*lysA* Δ*tyrA*	**0.77**	**0.07**	**-0.37**
Δ*ilvA* Δ*thrC* Δ*trpB*	**0.84**	**0.08**	**0.71**
Δ*trpB* Δ*pheA* Δ*thrC*	**0.79**	**0.07**	**-0.16**

Table S4. Relative fitness and epistatic interactions among auxotrophy-causing mutations in the succinate-containing environment. Mean fitness of each mutant genotype relative to wild type (± 95% confidence interval (CI)) was calculated from 8 replicates. Epistasis was estimated by comparing estimated and observed fitness values using a multiplicative model. Instances of significant epistasis are depicted in bold. NA = not applicable.

Genotype	Relative fitness	± 95% CI	Epistasis
Δ*argH*	0.95	0.15	NA
Δ*hisD*	0.85	0.07	NA
Δ*ilvA*	0.90	0.03	NA
Δ*leuB*	1.02	0.08	NA
Δ*lysA*	0.95	0.06	NA
Δ*metA*	1.14	0.05	NA
Δ*pheA*	1.08	0.05	NA
Δ*proC*	1.04	0.06	NA
Δ*thrC*	0.96	0.05	NA
Δ*trpB*	0.97	0.03	NA
Δ*tyrA*	1.00	0.03	NA
Δ*argH* Δ*ilvA*	**1.08**	**0.06**	**0.22**
Δ*argH* Δ*leuB*	1.02	0.08	-0.08
Δ*argH* Δ*lysA*	1.00	0.05	0.08
Δ*metA* Δ*argH*	**1.03**	**0.07**	**-0.05**
Δ*argH* Δ*pheA*	**0.95**	**0.03**	**-0.07**
Δ*proC* Δ*argH*	1.06	0.07	0.06
Δ*argH* Δ*thrC*	**1.05**	**0.05**	**0.12**
Δ*argH* Δ*trpB*	**1.10**	**0.10**	**0.17**
Δ*argH* Δ*tyrA*	0.95	0.04	-0.00
Δ*ilvA* Δ*hisD*	**1.11**	**0.05**	**0.34**
Δ*leuB* Δ*hisD*	**1.02**	**0.03**	**0.20**
Δ*lysA* Δ*hisD*	0.81	0.07	-0.00
Δ*metA* Δ*hisD*	**0.89**	**0.07**	**-0.08**
Δ*hisD* Δ*pheA*	**1.12**	**0.03**	**0.20**
Δ*hisD* Δ*proC*	**0.76**	**0.06**	**-0.12**
Δ*hisD* Δ*thrC*	**0.98**	**0.06**	**0.16**
Δ*hisD* Δ*trpB*	**0.96**	**0.04**	**0.13**
Δ*hisD* Δ*tyrA*	0.89	0.05	0.03
Δ*ilvA* Δ*leuB*	**1.13**	**0.04**	**0.21**
Δ*ilvA* Δ*lysA*	0.94	0.08	0.08
Δ*ilvA* Δ*metA*	**0.93**	**0.03**	**-0.09**
Δ*ilvA* Δ*pheA*	0.99	0.05	0.01
Δ*ilvA* Δ*proC*	1.02	0.05	0.07
Δ*ilvA* Δ*thrC*	0.94	0.06	0.06
Δ*trpB* Δ*ilvA*	0.85	0.04	-0.02
Δ*ilvA* Δ*tyrA*	**1.04**	**0.11**	**0.13**
Δ*leuB* Δ*lysA*	1.01	0.05	0.03
Δ*metA* Δ*leuB*	**0.92**	**0.07**	**-0.24**
Δ*pheA* Δ*leuB*	**0.90**	**0.03**	**-0.20**
Δ*proC* Δ*leuB*	**0.94**	**0.08**	**-0.12**

ΔthrC ΔleuB	1.02	0.04	0.04
ΔleuB ΔtrpB	**0.98**	**0.03**	**-0.01**
ΔlysA ΔmetA	**0.99**	**0.11**	**-0.09**
ΔlysA ΔpheA	1.10	0.07	0.06
ΔlysA ΔproC	1.07	0.07	0.07
ΔthrC ΔlysA	**1.31**	**0.13**	**0.39**
ΔlysA ΔtrpB	**1.15**	**0.04**	**0.22**
ΔmetA ΔpheA	**1.17**	**0.07**	**-0.06**
ΔproC ΔmetA	1.18	0.05	-0.01
ΔmetA ΔthrC	1.17	0.10	0.07
ΔmetA ΔtrpB	**0.93**	**0.06**	**-0.18**
ΔpheA ΔproC	**1.07**	**0.18**	**-0.06**
ΔpheA ΔthrC	0.98	0.08	-0.05
ΔpheA ΔtrpB	1.02	0.10	-0.03
ΔpheA ΔtyrA	1.12	0.17	0.03
ΔproC ΔthrC	**0.95**	**0.09**	**-0.05**
ΔtrpB ΔproC	1.02	0.04	0.00
ΔthrC ΔtrpB	**1.03**	**0.05**	**0.09**
ΔthrC ΔtyrA	0.96	0.04	-0.00
ΔtrpB ΔtyrA	**1.15**	**0.13**	**0.18**
ΔtrpB ΔpheA ΔtyrA	**0.97**	**0.02**	**-0.26**
ΔlysA ΔmetA ΔargH	**1.00**	**0.04**	**-0.36**
ΔlysA ΔmetA ΔthrC	0.92	0.02	-0.05
ΔtrpB ΔpheA ΔmetA	0.96	0.08	0.04
ΔtrpB ΔpheA ΔleuB	0.75	0.11	-0.07
ΔmetA ΔthrC ΔargH	**0.84**	**0.05**	**-0.34**
ΔmetA ΔthrC ΔhisD	1.17	0.14	0.08
ΔthrC ΔlysA ΔhisD	**1.04**	**0.11**	**-0.29**
ΔtrpB ΔpheA ΔhisD	0.94	0.17	-0.07
ΔproC ΔthrC ΔilvA	**1.02**	**0.06**	**-0.18**
ΔtrpB ΔleuB ΔthrC	**1.13**	**0.05**	**0.13**
ΔmetA ΔargH ΔhisD	1.07	0.06	-0.00
ΔproC ΔlysA ΔhisD	**0.87**	**0.04**	**-1.66**
ΔproC ΔlysA ΔtyrA	**1.02**	**0.07**	**0.22**
ΔilvA ΔthrC ΔtrpB	**1.19**	**0.08**	**0.93**
ΔtrpB ΔpheA ΔthrC	**1.24**	**0.06**	**0.25**

SUPPORTING REFERENCES

Adadi, R., B. Volkmer, R. Milo, M. Heinemann, and T. Shlomi. 2012. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. PLoS Comp. Biol. 8:e1002575.

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

D'Souza, G., S. Waschina, S. Pande, K. Bohl, C. Kaleta, and C. Kost. 2014. Less is more: Selective advantages can explain the prevalent loss of biosynthetic genes in bacteria. Evolution 68-9: 2559–2570.

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

Schellenberger, J., R. Que, R. M. Fleming, I. Thiele, J. D. Orth, A. M. Feist, D. C. Zielinski, A. Bordbar, N. E. Lewis, and S. Rahmanian. 2011. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0. Nat. Protoc. 6:1290-1307.

Zar, J. H. 1999. Biostatistical analysis. 4th ed. Prentice-Hall, Inc., New Jersey, USA.

Supporting information for chapter V

Metabolic network architecture and carbon source determine metabolite production costs

Table S1. Final amino acid concentrations (µM) in the media used for the precultures of auxotrophs and for the growth kinetic assays.

Amino acid	**Auxotroph precultures**	**Growth kinetic assays**							
		Lvl 1	Lvl 2	Lvl 3	Lvl 4	Lvl 5	Lvl 6	Lvl 7	Lvl 8
His	15	0	2.5	3.75	5	6.25	7.5	8.75	10
Tyr	30	0	5	7.5	10	12.5	15	17.5	20
Phe	30	0	5	7.5	10	12.5	15	17.5	20
Trp	150	0	25	37.5	50	62.5	75	87.5	100
Leu	60	0	10	15	20	25	30	35	40
Lys	60	0	10	15	20	25	30	35	40
Ile	45	0	7.5	11.25	15	18.75	22.5	26.25	30

Table S2. Strains used in this study.

Strain	Genotype	Phenotype	Reference
Escherichia coli **BW25113 ara$^-$**	F-, *ΔaraD-araB567*, *ΔlacZ4787*::rrnB-3, λ^-, *rph-1*, *ΔrhaD-rhaB568*, *hsdR514*	WT	Baba *et al.* (2006)
Δ*hisD*	WT ara$^-$, Δ*hisD*::kan^R	AT	D'Souza et al. (2014)
Δ*pheA*	WT ara$^-$, Δ*pheA*::kan^R	AT	D'Souza et al. (2014)
Δ*tyrA*	WT ara$^-$, Δ*tyrA*::kan^R	AT	D'Souza et al. (2014)
Δ*trpB*	WT ara$^-$, Δ*trpB*::kan^R	AT	D'Souza et al. (2014)
Δ*leuB*	WT ara$^-$, Δ*leuB*::kan^R	AT	D'Souza et al. (2014)
Δ*lysA*	WT ara$^-$, Δ*lysA*::kan^R	AT	D'Souza et al. (2014)
Δ*ilvA*	WT ara$^-$, Δ*ilvA*::kan^R	AT	D'Souza et al. (2014)
Δ*hisD*	WT ara$^-$, Δ*hisD*	AT, kan^S	This study
Δ*pheA*	WT ara$^-$, Δ*pheA*	AT, kan^S	This study
Δ*trpB*	WT ara$^-$, Δ*trpB*	AT, kan^S	This study
Δ*leuB*	WT ara$^-$, Δ*leuB*	AT, kan^S	This study
Δ*lysA*	WT ara$^-$, Δ*lysA*	AT, kan^S	This study
Δ*ilvA*	WT ara$^-$, Δ*ilvA*	AT, kan^S	This study

Abbreviations: ara$^-$ – inability to use arabinose as carbon source, WT – wild type, AT – auxotroph, kan^S – kanamycine sensitive.

References

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.

D'Souza G, Waschina S, Pande S, Bohl K, Kaleta C, Kost C: Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria. Evolution 2014, **68**:2559–2570.

Table S3. Carbon sources considered for biosynthetic cost estimation.

D-fructose, L-lactate, succinate, L-malate, α-ketoglutarate, D-galactose, maltose, D-glucsoe, pyruvate, acetate, L-arabinose, N-acetyl-D-glucosamine, D-glucarate, L-aspartate, D-alanine, threhalose, D-mannose, D-sorbitol, glycerol, L-fucose, D-glucuronate, D-gluconate, glycerol 3-phosphate, D-xylose, D-mannitol, L-glutamate, D-glucose 6-phosphate, D-malate, D-ribose, L-rhamnose, melibiose, thymidine, L-asparagine, octadecenoate, fumarate, butyrate, phenylacetaldehyde, 5-dehydro-D-gluconate, acetoacetate, adenosine, L-alanine, D-allose, D-fructose 6-phosphate, D-galactarate, galactitol, D-galacturonate, D-glucosamine, deoxyadenosine, dihydroxyacetone, L-glutamine, inosine, (S)-Propane-1,2-diol, L-tartrate, lactose, maltotriose, N-acetyl-D-mannosamine, N-acetylneuraminate, propionate, uridine, D-glucose 1-phosphate, and L-lyxose.

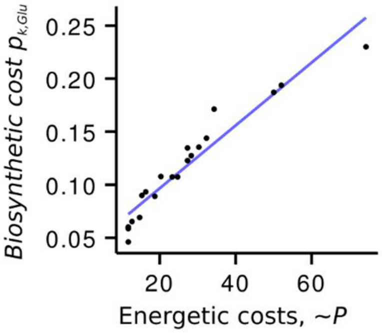


Figure S1. Biosynthetic cost estimations from this study are in line with previously reported estimations (Akashi and Gojobori 2002).

Reference

Akashi H, Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc Natl Acad Sci USA*. 2002, **99**:3695–3700.

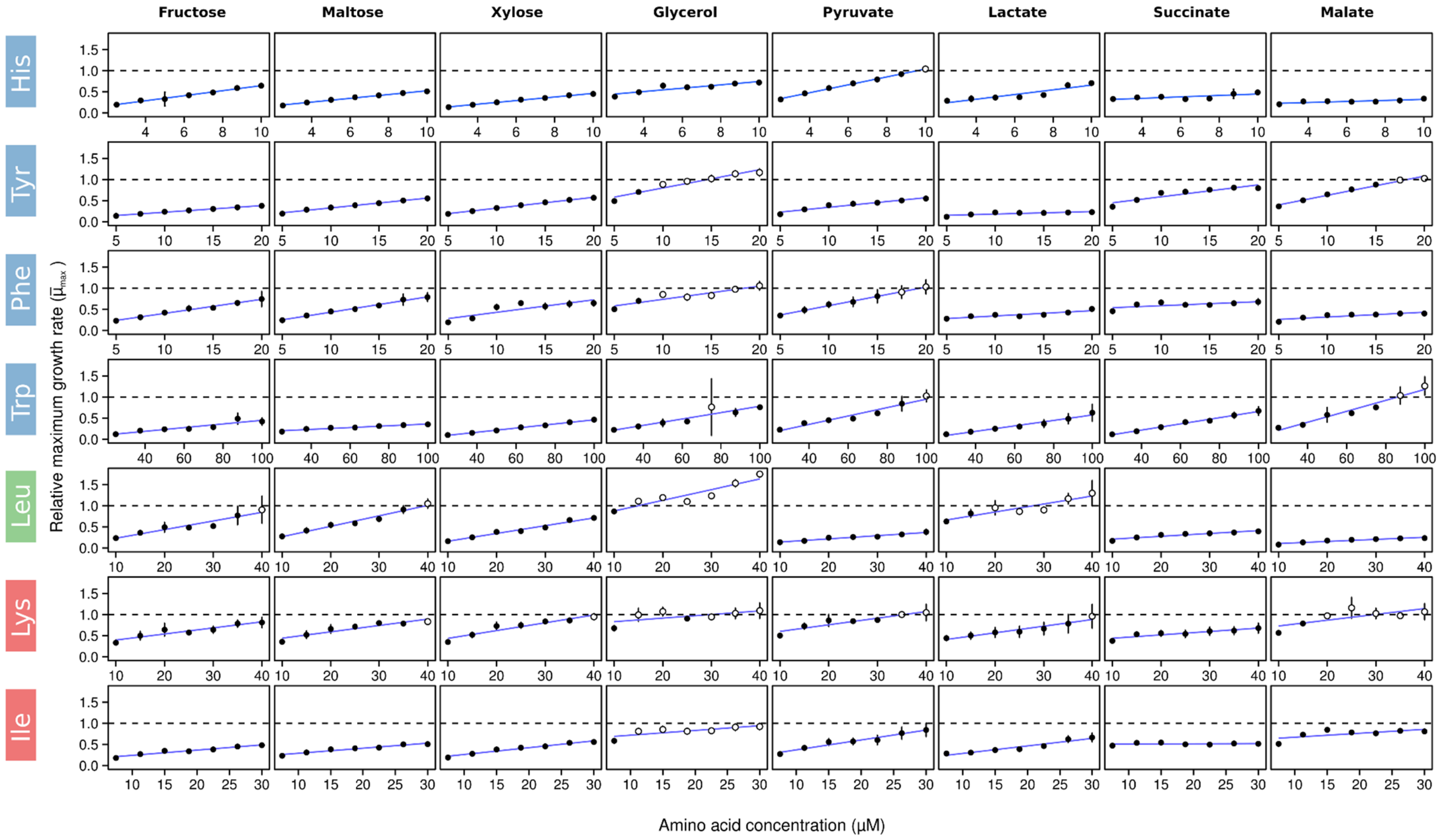


Figure S2. Maximum growth rates of auxotrophs under various carbon sources and amino acid concentrations relative to the maximum growth rate level of the wild type growing under the same carbon source and without amino acid supplementation (=1, dashed line). Error bars indicate the 95% confidence intervals. Filled circles denote the growth rates of the auxotrophs which are significantly lower than the WT strain growth rate (FDR-corrected Welch two sample t-tests, $P < 0.05$, $n = 6$), empty circles indicate no significant difference.

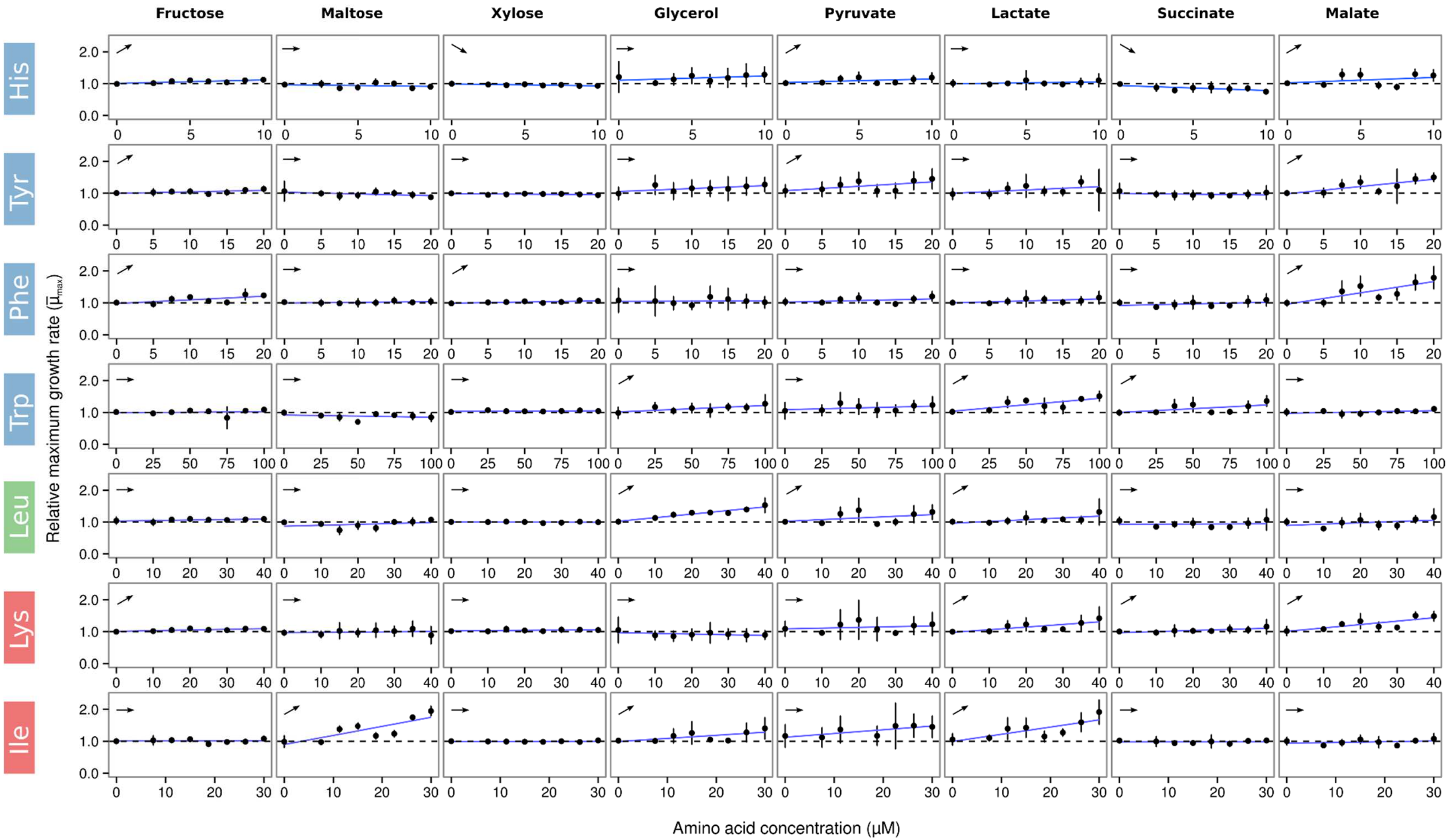


Figure S3. Maximum growth rates of the *E. coli* wild type strain under various carbon sources and amino acid concentrations relative to the maximum growth rate level of the wild type growing under the same carbon source and without amino acid supplementation (=1, dashed line). Error bars indicate the 95% confidence intervals. Arrows indicate significant correlation (up- or down arrows) or no significant correlation (horizontal arrows) of the two axes (FDR corrected linear mixed-model fit by maximizing the restricted log-likelihood, n=48).

Curriculum Vitae

Silvio Waschina

Geboren am 9. Dezember 1986 in Jena

Käthe-Kollwitz-Str. 4
07743 Jena
Silvio.waschina@jsmc.info

Ausbildung

1997 – 2005	**Abitur** am Adolf-Reichwein-Gymnasium in Jena Leistungsfächer: Mathematik und Biologie
2006 – 2012	**Bioinformatik Studium** an der Friedrich-Schiller-Universität Jena mit Diplomabschluss; Diplomarbeitsthema: *Theoretical design and experimental verification of amino acid overproducing strains of Escherichia coli using CASOP GS*
2009 – 2010	**Auslandsstudienjahr** an der Aarhus University in Dänemark
Juli 2013 – jetzt	**Dissertation** in den beiden Forschungsgruppen von Christoph Kaleta (theoretische Systembiologie an der FSU Jena) und Christian Kost (experimentelle Ökologie und Evolution am MPI für chemische Ökologie) Thema: *Evolutionary systems biology of bacterial metabolic adaptation.*

Arbeitserfahrung

2005 - 2006	**Zivildienst** an der Kinderklinik Jena
November 2008 – Juli 2009	**Studentische Hilfskraft** am MPI für Ökonomik in der Abteilung für strategische Interaktionen geleitet von Prof. Dr. Werner Güth.
Jan. – Mai 2013	**Forschungspraktikum** am King Mongkut's Institute of Technology Latkrabang in Bangkok, Thailand. Forschungsthema: *Analysis of the transcriptional regulation of hydrogenase genes in Anabaena sp.*

Publikationen

(* geteilte Erstautorenschaften)

D'Souza G, **Waschina S**, Pande S, Bohl K, Kaleta C and Kost C (2014). *Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria.* Evolution, 68: 2559-2570

Schmidt R*, **Waschina S***, Boettger-Schmidt D, Kost C and Kaleta C (2015). *Computing autocatalytic sets to unravel inconsistencies in metabolic network reconstructions.* Bioinformatics, 31: 373-381

D'Souza G*, **Waschina S***, Kaleta C and Kost C (2015). *Plasticity and epistasis strongly affect bacterial fitness after losing multiple metabolic genes.* Evolution, 69, 1244-1254

Waschina S, D'Souza G, Kost C and Kaleta C (in review). *Metabolic network architecture and carbon source determine metabolite production costs.* Submitted to FEBS Journal

Vorträge

Carbon source-dependent metabolic costs of amino acid biosynthesis in Escherichia coli. Understanding Microbial Communities; Function, Structure and Dynamics. 27th – 31st Oktober 2014 am Newton Institute in Cambridge, Großbritannien.

Differences in metabolic profiles promote cross-feeding in microbial communities. 19th European Meeting of PhD Students in Evolutionary Biology (EMPSEB), 3. – 7. September 2013 in Exeter, Großbritannien

Differences in metabolic profiles promote cross-feeding in microbial communities. 3rd International Student Conference on Microbial Communication, 5. -8. November 2012 in Jena

Carbon source-dependent metabolic costs of amino acid biosynthesis in Escherichia coli. Jena School for Microbial Communication (JSMC) Symposium, 2.-3. September 2014 in Jena

Posterpräsentationen

(* präsentierender Autor)

Gebauer J*, Gentsch C, Brandes S, **Waschina S**, Ristow S, Schuster S, Schäuble S, Kaleta C. *Reconstruction of a genome-scale metabolic model of Caenorhabditis elegans*. Conference on Constraint-Based Reconstruction and Analysis (COBRA) 2015 in Heidelberg

Waschina S*, D'Souza G, Kaleta C, Kost C. *The adaptive potential of metabolic heterogeneity in communities of Escherichia coli.* Economy of a Cell: Resource Allocation, Trade-Offs and Efficiency in Living Systems. 23. – 27. Juni 2014 am "International Centre for Theoretical Physics (ICTP)" in Trieste, Italien

Waschina S*, D'Souza G*, Kaleta C, Kost C. *Selective benefits can explain the prevalent loss of biosynthetic genes in bacteria.* Scientific Advisory Board Meeting. 14. – 16. Mai 2014 am MPI für chemische Ökologie in Jena.

D'Souza G*, **Waschina S**, Kaleta C, Kost C. *Lose to gain: Selective advantages can explain the prevalent loss of biosynthetic genes in bacteria.* 566th WE-Heraeus-Seminar on Mechanisms, Strategies, and Evolution of Microbial Systems. 15. – 19. Juni 2014 in Bad Honnef

Waschina S*, Kost C, Kaleta C. *Theoretical design and experimental verification of amino acid overproducing strains of Escherichia coli using CASOP GS.* 2nd Conference on Constraint-Based Reconstruction and Analysis. 7. – 10. Oktober 2012 in Kopenhagen, Dänemark. Poster

Waschina S*, Kost C, Kaleta C. *Theoretical design and experimental verification of amino acid overproducing strains of Escherichia coli using CASOP GS.* German Conference on Bioinformatics (GCB), 19. – 22. September 2012 in Jena

Eigenständigkeitserklärung

Entsprechend der geltenden, mir bekannten Promotionsordnung der Biolgisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität Jena erkläre ich, dass ich die vorliegende Dissertation eigenständig angefertigt und alle von mir benutzten Hilfsmittel und Quellen angegeben habe. Personen, die mich bei der Auswahl und Auswertung des Materials sowie bei der Fertigstellung der Manuskripte unterstützt haben, sind am Beginn eines jeden Kapitels genannt. Es wurde weder die Hilfe eines Promotionsberaters in Anspruch genommen, noch haben Dritte für Arbeiten, welche im Zusammenhang mit dem Inhalt der vorliegenden Dissertation stehen, geldwerte Leistungen erhalten. Die vorgelegte Dissertation wurde außerdem weder als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung noch als Dissertation an einer anderen Hochschule eingereicht.

Jena, den 17. September 2015

Silvio Waschina