PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Metabolic models and gene essentiality data reveal essential and conserved metabolism in prokaryotes

Joana C. Xavier[1,2¤], **Kiran Raosaheb Patil**[2], **Isabel Rocha**[1,3]*

1 Department of Biological Engineering, University of Minho, Campus de Gualtar, Braga, Portugal, **2** European Molecular Biology Laboratory, Heidelberg, Germany, **3** Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa (ITQB-NOVA), Oeiras, Portugal

¤ Current address: Institute of Molecular Evolution, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
* irocha@deb.uminho.pt

Abstract

Essential metabolic reactions are shaping constituents of metabolic networks, enabling viable and distinct phenotypes across diverse life forms. Here we analyse and compare modelling predictions of essential metabolic functions with experimental data and thereby identify core metabolic pathways in prokaryotes. Simulations of 15 manually curated genome-scale metabolic models were integrated with 36 large-scale gene essentiality datasets encompassing a wide variety of species of bacteria and archaea. Conservation of metabolic genes was estimated by analysing 79 representative genomes from all the branches of the prokaryotic tree of life. We find that essentiality patterns reflect phylogenetic relations both for modelling and experimental data, which correlate highly at the pathway level. Genes that are essential for several species tend to be highly conserved as opposed to non-essential genes which may be conserved or not. The tRNA-charging module is highlighted as ancestral and with high centrality in the networks, followed closely by cofactor metabolism, pointing to an early information processing system supplied by organic cofactors. The results, which point to model improvements and also indicate faults in the experimental data, should be relevant to the study of centrality in metabolic networks and ancient metabolism but also to metabolic engineering with prokaryotes.

Author summary

If we tried to list every known chemical reaction within an organism–human, plant or even bacteria–we would get quite a long and confusing read. But when this information is represented in so-called genome-scale metabolic networks, we have the means to access computationally each of those reactions and their interconnections. Some parts of the network have alternatives, while others are unique and therefore can be essential for growth. Here, we simulate growth and compare essential reactions and genes for the simplest type of unicellular species–prokaryotes–to understand which parts of their metabolism are universally essential and potentially ancestral. We show that similar patterns of essential

Citation: Xavier JC, Patil KR, Rocha I (2018) Metabolic models and gene essentiality data reveal essential and conserved metabolism in prokaryotes. PLoS Comput Biol 14(11): e1006556. https://doi.org/10.1371/journal.pcbi.1006556

Editor: Costas D. Maranas, The Pennsylvania State University, UNITED STATES

Received: January 25, 2018

Accepted: October 9, 2018

Published: November 16, 2018

Data Availability Statement: The experimental essentiality datasets used in this study are listed within the paper with the original references and are also available in accessible formats in the OGEE (http://ogee.medgenius.info/downloads/) and DEG databases (http://tubic.org/deg/). The genome-scale metabolic models used are listed in the paper with the corresponding reference from where they can be obtained in SBML format. Supplementary data including the results of simulations, translations for metabolic subsystems and the analysis of genetic conservation are provided in supplementary files.

Funding: This work was supported by grants from: the Fundação para a Ciência e a Tecnologia (http://www.fct.pt) with award number UID/BIO/04469/2013, the European Regional Development Fund (http://www.norte2020.pt) with award number NORTE-01-0145-FEDER-000004 (https://www.ceb.uminho.pt/Projects/Details/6040), Horizon 2020 (https://ec.europa.eu/programmes/horizon2020) with award number 686070 (http://dd-decaf.eu/) and COMPETE 2020 with award number POCI-01-0145-FEDER-006684 to JCX and IR and the Fundação para a Ciência e a Tecnologia (http://www.fct.pt) with award number SFRH/BD/81626/2011 to JCX. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

reactions echo phylogenetic relationships (this makes sense, as the genome provides the building plan for the enzymes that perform those reactions). Our computational predictions correlate strongly with experimental essentiality data. Finally, we show that a crucial step of protein synthesis (tRNA charging) and the synthesis and transformation of small molecules that enzymes require (cofactors) are the most essential and conserved parts of metabolism in prokaryotes. Our results are a step further in understanding the biology and evolution of prokaryotes but can also be relevant in applied studies including metabolic engineering and antibiotic design.

Introduction

Prokaryotes are the simplest contemporary life forms known, and nevertheless are characterized by an immense complexity. The debate on the features of such complexity and its breadth in the primordial life forms has been around for years [1–3], and was furthermore expanded and detailed since the advent of systems biology [4–6]. The study of essential genes has been a crucial contribution for detangling this complexity, relating some proteins with cell viability in specific conditions [7] and others with cell viability in apparently all conditions [8,9]. Genome-wide essentiality studies based on collections of targeted mutants or generated by random mutagenesis have been conducted for a number of species, aiming mainly at antibiotic design or identifying industrially relevant targets [10–15]. These data have been made available in databases such as the Online Gene Essentiality Database (OGEE) [16] and the database of essential genes (DEG) [17] but their comparative and integrative analysis, although having already provided relevant insights, is still emerging. Early work comparing genome-scale essentiality data of *Mycoplasma genitalium*, *Haemophilus influenza*, *Bacillus subtilis* and *Escherichia coli* found that it is essentiality, not expressiveness, that drives gene strand bias [18]. Later, a critical review of genome-scale essentiality datasets conducted a preliminary analysis integrating six assays corresponding to four different species [19]. Functional differences were highlighted, as the smaller number of essential genes in flavin synthesis in *B. subtilis*, a species known to have an active riboflavin salvage capability. The authors of the DEG database have also conducted a pair of integrative analyses on large-scale essentiality data. The first concluded that there are less essential genes inside than outside genomic islands, and some of those are related with virulence [20]. The second study [21] added to a previous finding based only on *E. coli* essentiality data where it was proposed that essential genes are more evolutionarily conserved than non-essential genes [22]. Luo and others used the same type of analysis, based on synonymous and non-synonymous substitution rates for 23 bacterial species, to corroborate this finding [21]. The study indicates that the most evolutionarily conserved COG categories of essential genes are carbohydrate transport and metabolism; coenzyme transport and metabolism; transcription; translation, ribosomal structure and biogenesis; lipid transport and metabolism, and replication, recombination and repair.

Genome-scale metabolic models (GSMs) are large curated repositories of metabolic data for individual species that expand possibilities of analysis of cellular physiology [23]. Apart from improving or suggesting new functional annotations by reconstructing whole pathways [24], GSMs can be used for calculating metabolic fluxes that permit the prediction of, among others, lethal phenotypes [25]. Multi-species analysis of this type of phenotype predictions with different manually curated models has been scarce [26], in part impaired by the poor knowledge basis for other species than the usual model organisms, but also by the deficient use of standards in building such models [27].

Comparative genomics is commonly used to find core essential genes for several species, and being based on the key evolutionary notion of orthology, to infer genes present in common ancestors [28]. Evolutionary parsimony indicates that genes present in a set of species have been vertically inherited from a common ancestor. Horizontal Gene Transfer (HGT) might have played a role even before the divergence of the three main domains [29], but when a gene is present in all or most species of a phylogenetic branch, the most parsimonious scenario is vertical inheritance.

In parallel with genetic comparisons, genome-scale metabolic models allow for functional comparisons at the level of metabolic capacities (reactions). Building up on this methodological advantage, in this study, 36 experimental genome-scale essentiality assays were integrated with simulation results from 15 genome-scale metabolic models to reveal common patterns of essentiality. To this analysis the screening of full genome sequences of 79 prokaryotic species was added in order to find core conserved functions in prokaryotic biology. It is expected that this knowledge on the minimal metabolic functions of prokaryotic cells can not only help untangling the fundamental complexity of cellular systems but also, by building up on the concept of orthogonalization of metabolic modules [30], here analysed in the form of metabolic subsystems (pathways), improve future engineering approaches that use this type of organisms.

Methods

Genome-scale metabolic models used in essentiality predictions

For all essentiality predictions performed in this study, 15 genome-scale metabolic models were chosen based on curation, validation, and comparability of the nomenclature of metabolites and reactions. These comprise 7 prokaryotic phyla, including one archaea. Ten of these models include more than 20% of the total of ORFs of the corresponding species. Table 1 summarizes the details on the models used, including species name, model ID, content and references.

Environmental conditions for simulations

All GSMs were collected in SBML format and then parsed to model an environmental condition corresponding to rich media: all original exchange reactions in the model were set to a maximum uptake limit of -20 mmol gDW^{-1} h^{-1} to allow for the import of all transported metabolites (including oxygen, whenever possible).

In silico, single deletion of metabolic reactions

Flux Balance Analysis (FBA) [46,47] was used to predict the essentiality of each metabolic reaction in all models. A threshold of 10% of the flux through the biomass reaction compared to the wild type was set as the limit to define an essential metabolic reaction. All modelling procedures were implemented in C++ and solved using IBM ILOG CPLEX solver. The Optflux platform [48] was used occasionally to benchmark results.

Standardizing the nomenclature of essential metabolic reactions

For comparison of the essential reactions calculated for the 15 GSMs, some nomenclature inconsistencies were resolved: standardization of suffixes used in reaction IDs, removal of unnecessary or redundant indications of reversibility and species names allocated to reactions and other redundant tags. Irrelevant and irregular characters such as dashes were filtered out of the nomenclature (the final list of standardized reaction ids for reactions essential at least once is provided in the S1 File).

Table 1. Details on models and corresponding species used in *in silico* essentiality simulations.

Phylum	Species	Model ID	# Reactions	# Metabolites	% ORFs	Reference
Firmicutes	*Bacillus subtilis*	iYO844	1020	988	21%	[31]
	Clostridium beijerinckii NCIMB 8052	iCB925	938	881	18%	[32]
	Staphylococcus aureus N315	iSB619	641	571	24%	[33]
Proteobacteria	*Escherichia coli K12*	iAF1260	2077	1039	29%	[34]
	Escherichia coli W (ATCC9637)	iCA1273	2477	1111	27%	[35]
	Helicobacter pylori 16695	iIT341	476	485	21%	[36]
	Klebsiella pneumoniae MGH 78578	iYL1228	1970	1658	24%	[37]
	Pseudomonas putida KT2440	iJN746	950	911	14%	[38]
	Salmonella typhimurium LT2	STM_v1.0	2201	1119	28%	[39]
	Shewanella oneidensis MR-1	iSO783	774	634	15%	[40]
Actinobacteria	*Mycobacterium tuberculosis H37Rv*	iNJ661	939	828	15%	[41]
Chloroflexi	*Dehalococcoides ethenogenes*	iAI549	518	549	27%	[42]
Cyanobacteria	*Synechocystis sp. PCC6803*	iJN678	863	795	21%	[43]
Thermotogales	*Thermotoga maritima MSB8*	(None)	562	503	25%	[44]
Euryarchaeota	*Methanosarcina barkeri str. Fusaro*	iAF692	476	485	14%	[45]

https://doi.org/10.1371/journal.pcbi.1006556.t001

Experimental data and subsystem mapping

Large-scale experimental data on gene essentiality were collected from two databases, OGEE [16] and DEG [17]. The content of the databases was compared and DEG was chosen for it was considerably larger, including wider and clearer annotation metadata for 36 prokaryotic datasets (Table 2). Genes were mapped to the subsystems present in the latest *Escherichia coli* genome-scale metabolic model [49] using the DEG integrated nomenclature system of gene identifiers. All essential reactions obtained after GSMs analysis were also mapped according to this updated list of subsystems, using their standardized nomenclature (see above; S2 File).

Analysis of genetic conservation

To analyse the conservation and infer ancestry of all the genes annotated in metabolic subsystems of GSMs, a local protein blast was performed against representative genomes of all the 35 prokaryotic phyla with at least one fully sequenced quality genome in the NCBI genome database (version June 2015). For this task, translated genomes were selected and downloaded for 53 unique species of prokaryotes for which an available GSM could be found; to these, 26 representative genomes for phyla not modelled with GSMs were added. This totalled in 79 translated genomes representing the 35 fully sequenced phyla in the current tree of life of prokaryotes (see S2 Fig, built with iTOL v. 4.2.1 [78] which can be reproduced in iTOL with the corresponding NCBI taxonomy). All of the protein-encoding genes of *E. coli* K12 (RefSeq genome NC_000913.3) were used as queries and annotated to the subsystems of the latest *E. coli* model [49]. The threshold e-value considered was 1e-10. All the procedures were implemented using the Biopython package [79].

Numerical and statistical analysis of essentiality and conservation

For assessing the conservation of essential reactions and essential genes in each metabolic subsystem, the weighted sum of essentiality (**W**) was calculated for each subsystem **m**, as:

$$W_m = \sum_{i=1}^{t} n_i * i \quad (1)$$

Table 2. Large-scale essentiality assays used in this study, number of essential genes in each and respective original reference of publication.

Species name	Number of essential genes	Reference [a]
Acinetobacter baylyi ADP1	499	[12]
Bacillus subtilis 168	271	[50]
Bacteroides fragilis 638R	547	[51]
Bacteroides thetaiotaomicron VPI-5482	325	[52]
Burkholderia pseudomallei K96243	505	[53]
Burkholderia thailandensis E264	406	[54]
Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819	228	[55]
Caulobacter crescentus	480	[56]
Escherichia coli MG1655 I	609	[10]
Escherichia coli MG1655 II	296	[57]
Francisella novicida U112	392	[11]
Haemophilus influenzae Rd KW20	642	[58]
Helicobacter pylori 26695	323	[59]
Methanococcus maripaludis S2	519	[60]
Mycobacterium tuberculosis H37Rv	614	[61]
Mycobacterium tuberculosis H37Rv II	771	[62]
Mycobacterium tuberculosis H37Rv III	687	[63]
Mycoplasma genitalium G37	381	[64]
Mycoplasma pulmonis UAB CTIP	310	[65]
Porphyromonas gingivalis ATCC 33277	463	[66]
Pseudomonas aeruginosa PAO1	117	[67]
Pseudomonas aeruginosa UCBPP-PA14	335	[68]
Salmonella enterica serovar Typhi	353	[14]
Salmonella enterica serovar Typhi Ty2	358	[69]
Salmonella enterica serovar Typhimurium SL1344	353	[69]
Salmonella enterica subsp. enterica serovar Typhimurium str. 14028S	105	[70]
Salmonella typhimurium LT2	230	[71]
Shewanella oneidensis MR-1	403	[72]
Sphingomonas wittichii RW1	535	[73]
Staphylococcus aureus N315	302	[74]
Staphylococcus aureus NCTC 8325	351	[15]
Streptococcus pneumoniae	244	[75]
Streptococcus pyogenes MGAS5448	227	[76]
Streptococcus pyogenes NZ131	241	[76]
Streptococcus sanguinis	218	[77]
Vibrio cholerae N16961	779	[13]

[a] The corresponding annotated data was obtained from the DEG database [17].

https://doi.org/10.1371/journal.pcbi.1006556.t002

$\boldsymbol{n_i}$ being the number of reactions or genes of that subsystem essential in $\boldsymbol{i}$ models or datasets, where $\mathbf{t}$ is the total of models or datasets, 15 and 36 respectively.

The score sum of experimental essentiality for each individual gene S_g was calculated as the number of times a gene was found essential (E) minus the number of times it was found not essential (N) in all experimental assays datasets:

$$S_g = E - N \tag{2}$$

All statistical analyses were performed using R (statistical software, version 3.1). Hierarchical clustering was performed using the 'pvclust' R package [80] with binary distance as the dissimilarity metric and Ward 1 method as the linkage criterion. Pvclust was also used for assessing uncertainty by calculating approximately unbiased p-values via multiscale bootstrap resampling. Both the Fisher and Kolmogorov Smirnov tests were performed in R as well with the corresponding default parameters.

Results

Single-reaction and single-gene essentiality data reflect phylogenetic patterns

To analyse the validity of the essentiality results on a large scale, the different models were clustered based on single-reaction essentiality predictions and the different datasets available on DEG [17] were clustered based on the content of essential genes (Fig 1).

In the case of the simulated essentiality (Fig 1A), strongly supported clusters (more than 80% of 1000 bootstrap replicas) are phylogenetically consistent at the level of the phylum–they cluster in a statistically significant manner with at least one sister species–with the exception of the models of *C. beijerinckii* and *P. putida*. *H. pylori* and *S. oneidensis* show up in the same cluster, but not together with the rest of the Proteobacteria, but the two high-level clusters (that exclude *M. tuberculosis* with a p-value of 98%) are not statistically supported (p-values of 58 and 62% for the left and right cluster respectively). The lower number of available exchange reactions in the models of *H. pylori*, *S. oneidensis* and *P. putida* (74, 95 and 89 respectively) compared with the models for other Proteobacteria (*K. pneumoniae*, *E. coli* K12, *S. typhimurium* and *E. coli W* with 289, 299, 305 and 310 respectively) points to a justification for these results, as less exchanges cause more reactions in the network to be essential. *C. beijerinckii*'s model is also very restricted with regards to exchange reactions, with only 19 metabolic drains available.

Regarding the experimental data (Fig 1B), there is also a pattern of clustering taxonomically related species. One well-supported phylogenetic cluster is that of several gamma and beta-proteobacteria, including *Acinetobacter baylyi*, dataset II of *E. coli* K12, three Salmonellas, one *Shewanella* and one *Francisella*. Others are the cluster of Bacteroidetes, the cluster with all three datasets of *M. tuberculosis* and the cluster of the alpha-proteobacteria, *Sphingomonas* and *Caulobacter*. The datasets of *Pseudomonas aeruginosa PAO1* and *Salmonella enterica subsp. Enterica serovar Typhimurium str*. 14028S cluster together likely due to not being saturated genome-wide gene-essentiality screens (these datasets are considerably smaller, with 117 and 105 genes respectively–see Table 2 for context). Surprisingly, Firmicutes are spread all across the tree. Also, both *E. coli* sets are very distant from each other. Although they were performed under rich media conditions, one yielded 609 essential genes and the other only 296 (Table 2). This difference is likely due to the use of different technologies to perform the large-scale assays, the first being random mutagenesis and the screening of mixed populations, and the second the screening of libraries of targeted mutants, as reviewed in [19].

Cofactor metabolism is the most essential subsystem both in simulated and experimental data

Next, all the essential reactions calculated for the 15 GSMs were mapped to the corresponding metabolic subsystem (see Methods; S1 and S2 Files). Different models show different proportions of essential reactions for each subsystem (Fig 2).

For the majority of the models, the most essential subsystem is that of cofactor and prosthetic group biosynthesis. Nearly 48% of the essential reactions in the simulations with the GSM of *E. coli* K12–54 reactions–were related with this subsystem (Fig 2). Several of the

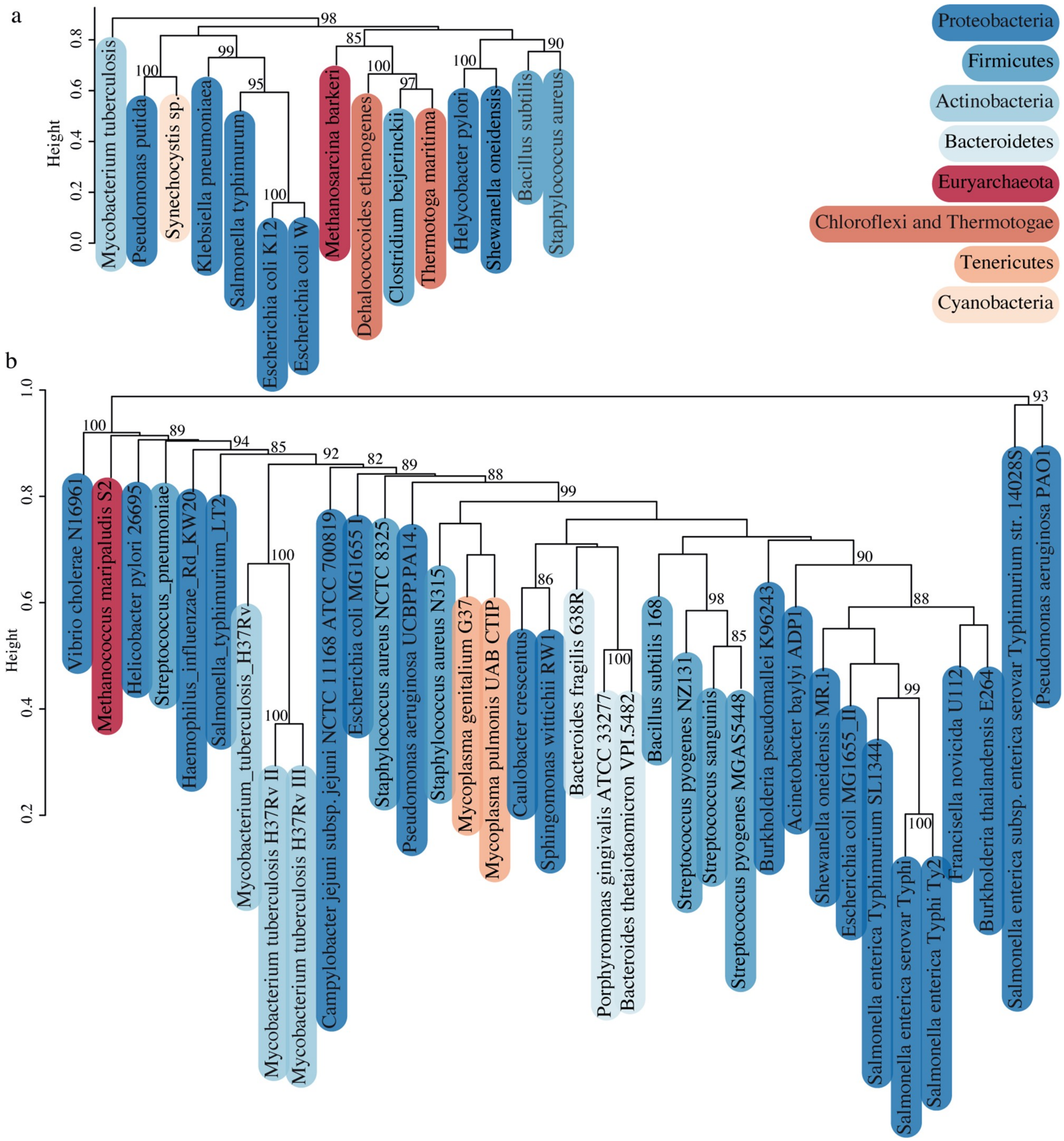

Fig 1. Clustering of (A) simulated and (B) experimental genome-scale essentialities of prokaryotes. Clusters show approximately unbiased p-values greater than 80% calculated by multiscale bootstrap re-sampling with 1000 replicas (see Methods for details). Phyla are coloured according to the taxonomical representation of the prokaryotic tree of life built with iTOL [78] (tree in S2 Fig).

https://doi.org/10.1371/journal.pcbi.1006556.g001

predicted essential reactions were confirmed to be essential steps in the biosynthesis of the active forms of cofactors that cannot be directly uptaken, e.g. dihydrofolate synthase and dihydrofolate reductase for the biosynthesis of tetrahydrofolate and derivatives [81], NAD kinase for obtaining NADP [82] and riboflavin synthase in some species [83]. For *M. tuberculosis*, *D.*

ethenogenes, *S. typhimurium and K. pneumoniae* the most represented subsystems were glycerophospholipid or membrane lipids metabolism (30.5, 25.1, 21.8 and 29.2% of all essential reactions, respectively). Discrepancies regarding results for each individual model are not only related with differences in the metabolic network but are also dependent on the formulations of the biomass equation (e.g. the biomass equation in *Klebsiella pneumoniae*'s model lacks cofactors) [84].

To validate the predictions of essentiality of metabolic subsystems obtained with GSMs, each experimentally essential gene in DEG was annotated according to its function, using the same system used in GSM's. This system (see Methods) covered more annotations when compared with COG annotations (1363 essential metabolic genes annotated in total, when compared with a total of 906 unique metabolic COGs–S1 Fig). Moreover, this curated dataset could be directly compared to the modelling results and included some genes annotated in the "General function prediction only" COG category. After annotation, all unique essential genes were identified and the same was done for essential reactions in the models. Both the total number of unique essential reactions (modelling) or genes (experimental) varies significantly between subsystems (Fig 3). The total of reactions and genes in each subsystem (for both modelling and experimental data, respectively) also varies significantly. To test for independence of the totals of essentials from the sizes of the subsystems, we performed a Fisher's exact test, and for the majority of subsystems the null hypothesis of dependence was rejected (p-value less than 0.05).

In Fig 3 the subsystem of Cofactor and Prosthetic Groups biosynthesis appears isolated with the maximum number of unique essential enzymes both in experimental and modelling data. Three subsystems show a striking difference in the ranking between experimental and modelling data, being more represented in the latter, all related with membrane and cell wall metabolism. Several justifications can be raised for this difference. First and foremost, often in those subsystems the number of different reactions that can be encoded by the same gene is high [85], (thus there will be several essential reactions for each essential gene). For instance, in the model of *Synechocystis* there are twelve essential reactions related with fatty acid biosynthesis all encoded by the same gene, fabZ (sll1605 in the model), a gene that is essential in rich medium experimentally, but in that case, counted only once. Similarly, all of the twenty essential fatty acid synthase reactions in the model of *M. tuberculosis* are encoded by only 3 different genes: Rv1663 and Rv1662 or Rv2940c. There might also be some lack of integration in nomenclature of the reactions in these subsystems in the different GSMs that should not contribute significantly as all models follow the same nomenclature scheme, which was still manually curated for the essential reactions.

Conservation of essentiality validates predictions by GSMs with experimental data and pinpoints specific model gaps

To further explore essentiality at the level of genes and reactions, the conservation of essentiality across models and experimental datasets was analysed for each reaction and gene in each metabolic subsystem. Strikingly, no reaction was essential in all the models analysed. Three reactions annotated within aromatic amino acids metabolism (shikimate kinase, 3-phosphoshikimate 1-carboxyvinyltransferase and chorismate synthase) were essential in all models except for *K. pneumoniae.* Although there are differences in the models regarding the capacity to uptake aromatic amino acids, this does not justify the observed differences in terms of essentiality. The notable difference between the model of *K. pneumoniae* and the others is that it lacks cofactors and prosthetic groups in its biomass equation. Those three reactions correspond to the three last steps in the synthesis of chorismate, which is part of the shikimate

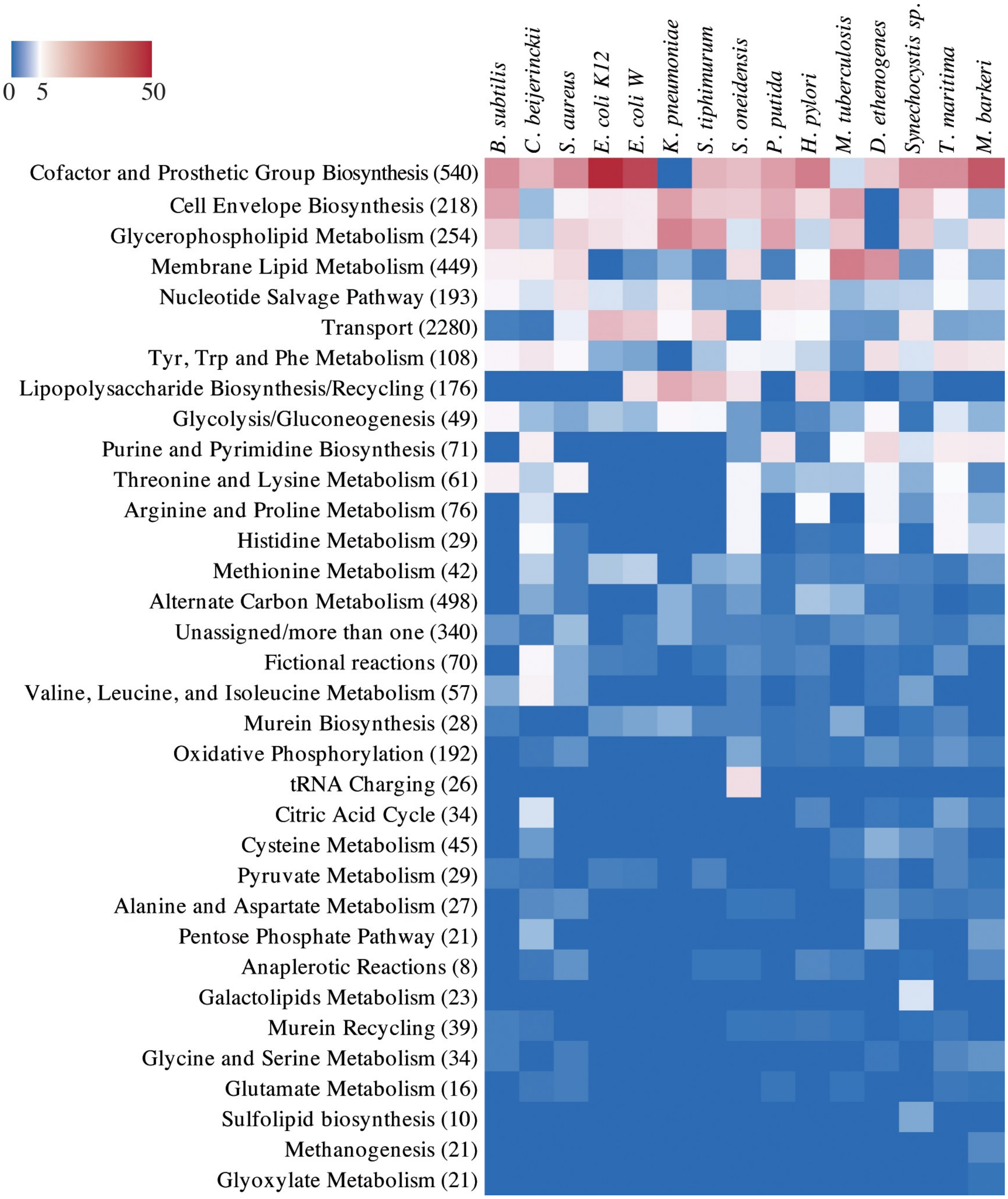

Fig 2. Essentiality for biomass production of each metabolic subsystem in 15 genome-scale manually curated metabolic models. The colour bar represents the ratio of essential reactions (in that subsystem) to the total of essential reactions in each model (0%—blue; 5%—white; 50%—red). In parenthesis next to the subsystem name is the total of reactions in that subsystem (for all models).

https://doi.org/10.1371/journal.pcbi.1006556.g002

pathway, which connects central metabolism with aromatic amino acids metabolism. However, this pathway is also the route taken to synthesize several other compounds in the cell, including quinones and folates [86], which are not present in the biomass equation of *Klebsiella*, which is likely the cause for the difference in the results as discussed in [84] and above.

Reactions essential in several models annotated within the cofactor and prosthetic group biosynthesis subsystem are dihydrofolate synthase, essential in 13 out of the 15 models, and

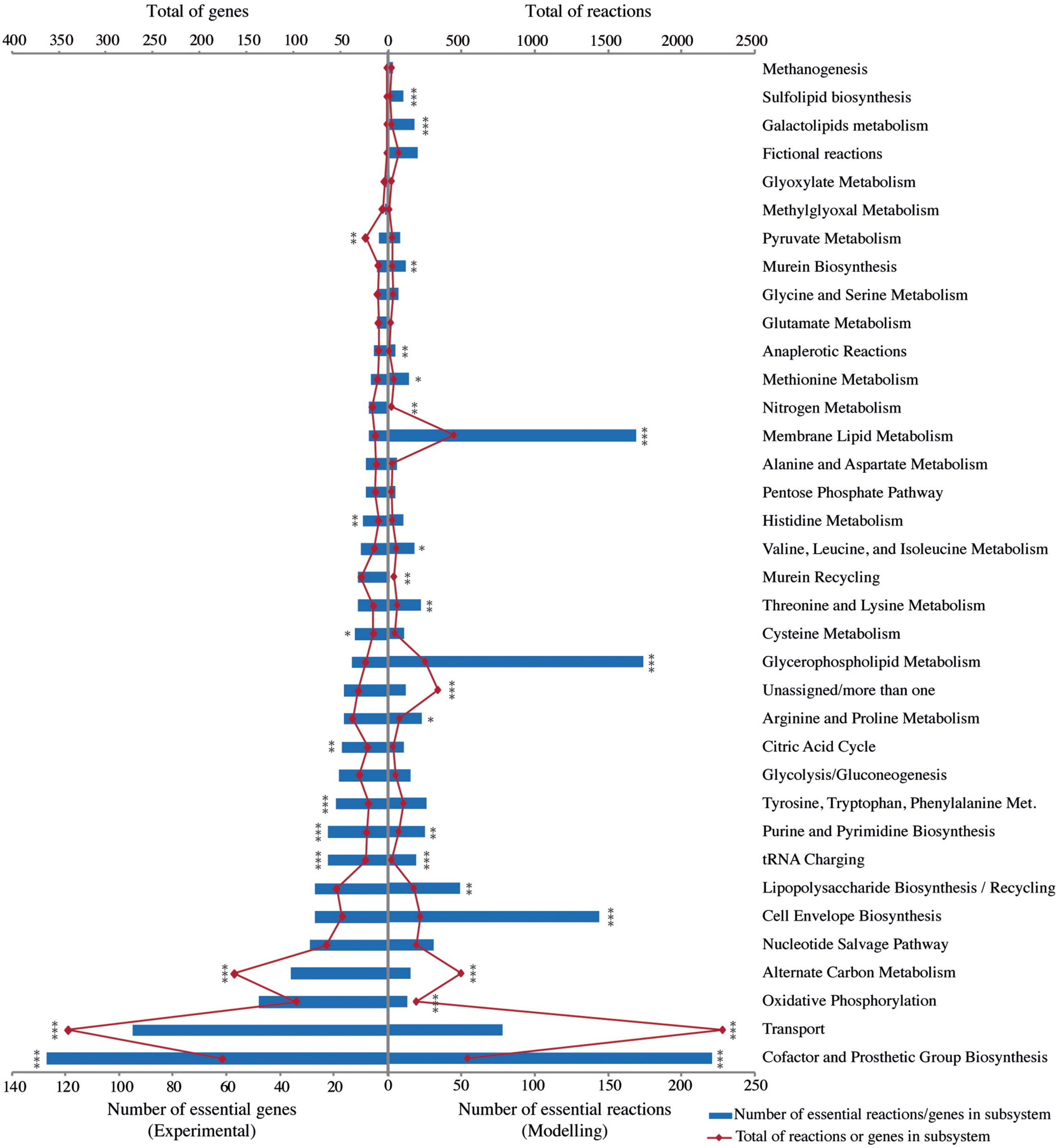

Fig 3. Number of essential reactions predicted by genome-scale metabolic models and essential genes in genome-scale experimental assays for different metabolic subsystems. Each reaction and gene were counted only once, even if present (or present and essential) in more than one model or experimental dataset. Single, double and triple asterisks indicate p-values less than 0.05, 0.01 and 0.0001, respectively, after a Fisher's exact test for count data testing independence of the number of essential vs the number of non-essential mapped genes and reactions.

https://doi.org/10.1371/journal.pcbi.1006556.g003

dihydrofolate reductase and NAD kinase, both essential in 12 models. These are related with the biosynthesis of folates and the phosphorylation of NAD to produce NADP. Two reactions involved in the salvage pathways of nucleotides–the biosynthesis of GDP and dTTP—were also essential for 14 models. Three reactions related with the biosynthesis of cell wall components are essential in all models except for *B. subtilis* and *D. ethenogenes*—glucosamine-1-phosphate N-acetyltransferase, phosphoglucosamine mutase and UDP-N-acetylglucosamine 1-carboxyvinyltransferase. Acetyl-CoA carboxylase, related with membrane lipid metabolism, is essential in 12 of the 15 models. One reaction not assigned to any subsystem, the HCO_3^- equilibration reaction, was essential in 11 of all 15 models.

To overview the relationship between modelling and experimental results, the conservation of essentiality for each set of results was compared. The inset plot in Fig 4 shows the high correlation obtained between the weighted essentiality for each subsystem between simulated and experimental data. However, there are some differences to be noted. Firstly, regarding the tRNA charging subsystem, it is modelled in only one GSM (*S. oneidensis*, Fig 2). This causes this category to appear much more evidently as the second most conserved essential subsystem in experimental data, in contrast with the low result in the simulations of GSMs. It is expected that future GSMs will include this subsystem, but for the sake of comparison of these results, it was excluded from the correlation.

Interestingly, regarding the three highly essential reactions in modelling results related with chorismate biosynthesis, these were not found as significantly essential in experimental data. It is known that in minimal media the knock-out of chorismate synthase (aroC) in *E. coli* impairs growth [87]. The non-essentiality of this enzyme in the rich media analysed here indicates that there must be a compound in the media compensating for its absence. It has been shown that, when provided with p-aminobenzoic acid (PABA), para-hydroxybenzoic acid (PHBA) or a combination of a precursor from PABA with a non-biological catalyst, the growth of *E. coli* aroC mutant in M9 minimal medium can be rescued [88]. PABA and its derivatives cannot be uptaken in the genome-scale model of *Escherichia coli* or any other of the working set. Transporters for these compounds or others that might compensate in rich media for lethal phenotypes in minimal media remain to be integrated in the genome-scale metabolic models and further explored.

Again, the subsystem of cofactor and prosthetic groups metabolism has the highest number of reactions appearing as essential in more datasets in experimental data, in accordance with modelling data. Dihydrofolate synthase and reductase (highest ranking in modelling) are essential in 11 and 25 experimental datasets, respectively, indicating that either several organisms can overcome the lack of both enzymes by intermediate pathways not yet modelled, or that the experimental assays have produced false negatives (see the differences between the sizes of the two experimental assays for *E. coli* in Table 2 discussed above; a special case, dihydrofolate synthase, is further explored in the Discussion). NAD kinase appears as highly essential, also in accordance with simulations, in 24 datasets. Several other genes encoding for enzymes essential for the biosynthesis of cofactors are highly essential for cell viability experimentally, even in rich media (eg. nadE for NAD; coaD and coaE for coenzyme A; hemC for tetrapyrroles; dxr for isoprenoids). Cell envelope biosynthesis genes follow as the third most conserved essential functional module, in accordance with the modelling results as well.

Core conserved metabolic subsystems

Based on the premises of evolutionary parsimony and orthology [28], we proceeded to the analysis at a large scale of the conservation of metabolic genes in the prokaryotic tree of life to

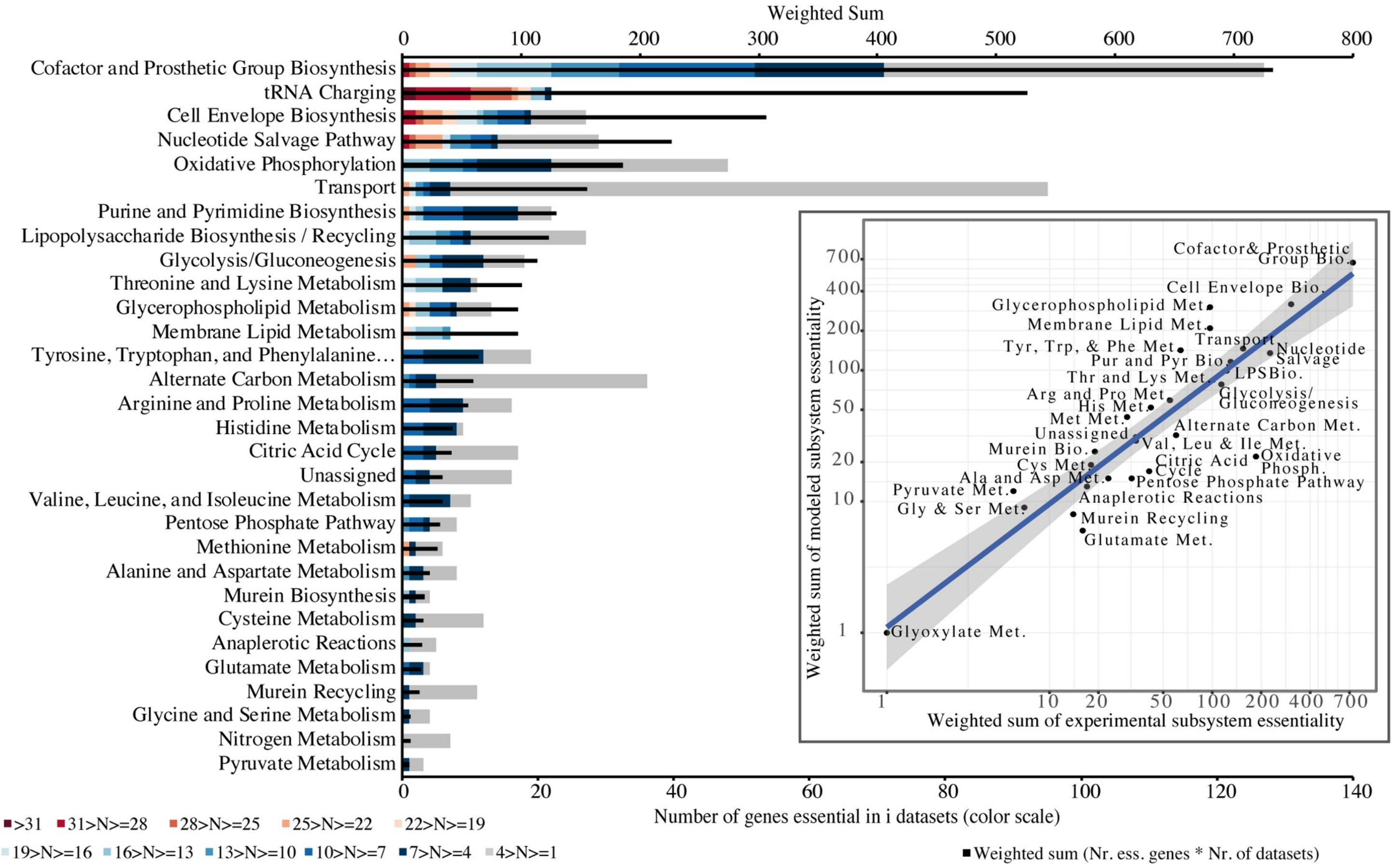

Fig 4. Conservation of essentiality of metabolic subsystems in 36 large-scale gene essentiality datasets and correlation with modelling predictions (inset). Red indicates highest conservation (genes essential in more than 31 datasets) and grey the least (essential in less than 4 datasets); black bar: weighted sum of essential genes given the number of datasets in which they are essential (see Methods). Inset plot: Correlation between modelling and experimental genome-scale essentiality data at subsystem level (adjusted R^2 of 0.821, Pearson correlation coefficient of 0.909 with p-value 9.65e-14); axes are represented in log scale for visualization purposes only.

https://doi.org/10.1371/journal.pcbi.1006556.g004

infer potential ancestral metabolic functions. Seventy-nine genomes were assayed, representing all the known prokaryotic phyla with a fully sequenced genome (see Methods for details). A phylogenetic tree with these 79 species is available in S2 Fig. All annotated metabolic genes of *E. coli* K12 were used as queries to search the set of genomes for conserved metabolic genes and respective functions. The results on conservation of metabolic genes are summarized in Fig 5.

The metabolic subsystem with more prevalent genes is Transport, followed by the tRNA charging subsystem with 18 nearly universal aminoacyl-tRNA synthetases. It should be noted though that nearly all of the 41 transport genes conserved in all 79 genomes correspond to ABC transporters (S1 Table) with a ubiquitous ATP-binding domain. Two genes involved in oxidative phosphorylation were also conserved in all genomes analysed: atpA and atpD (ATP synthase subunits alpha and beta, respectively). In the subsystem of cofactors and prosthetic group biosynthesis, glutX and sufC were also conserved in all genomes analysed. It should be noted that glutX corresponds to a tRNA charging protein, a glutamyl-tRNA synthetase involved in the biosynthesis of heme, that should have a double annotation; sufC is an atypical cytoplasmic ABC/ATPase, required for the assembly of iron-sulphur clusters [89]. Although the vast majority of genes found conserved in all the genomes analysed correspond to ABC

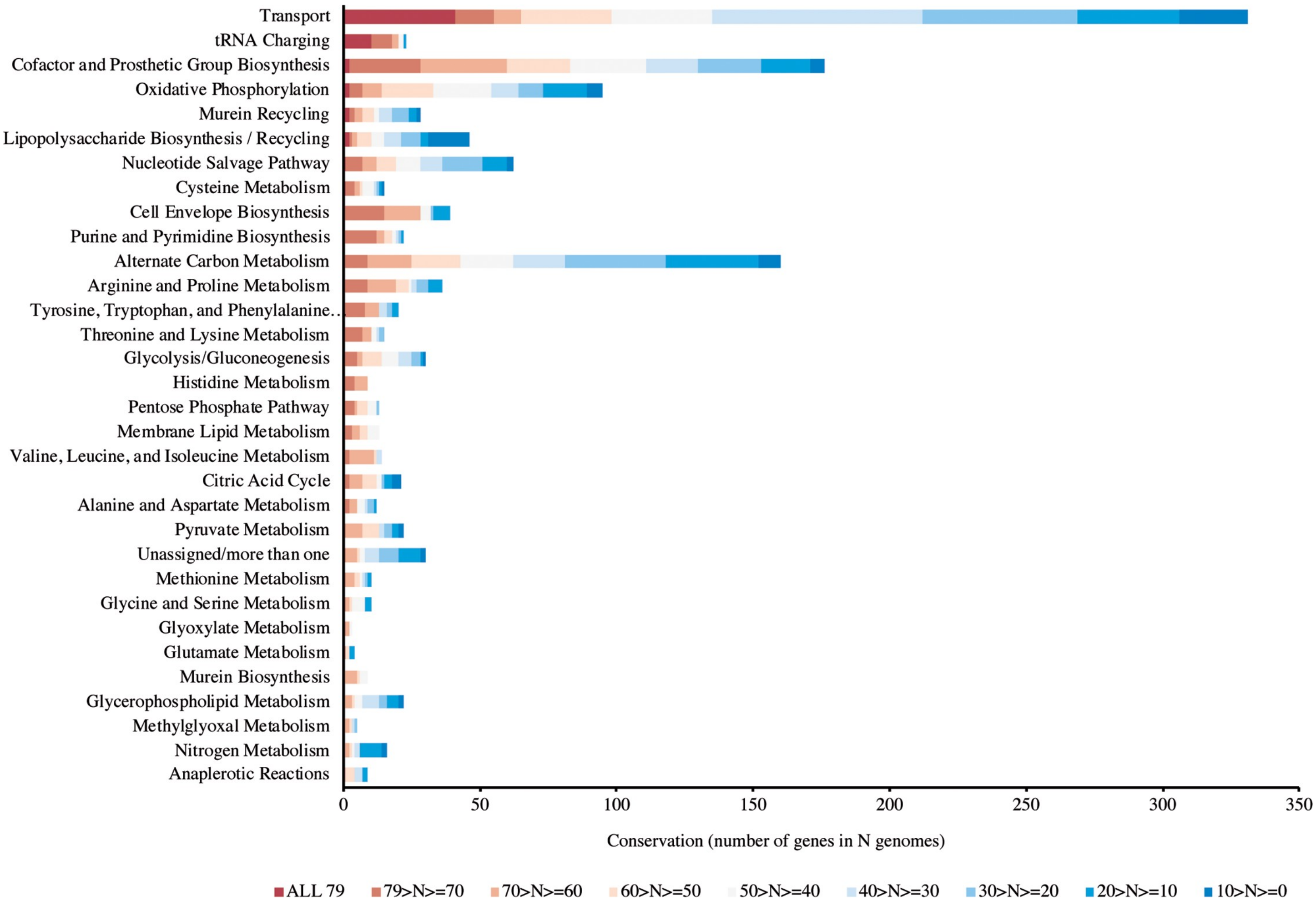

Fig 5. Conservation of metabolic subsystems in genomes of all prokaryotic phyla with at least one fully sequenced genome. Dark red indicates the highest conservation (genes that are present in all 79 genomes accessed) and dark blue the least (present in less than 10 genomes).

https://doi.org/10.1371/journal.pcbi.1006556.g005

ubiquitous domains, the high conservation (between 70 and 79 genomes) of 214 other genes is still prominent. Twenty-eight of those are classified in the subsystem of cofactor and prosthetic group biosynthesis genes (Fig 5, S2 Table).

Common essential genes are rarer and prone to be highly conserved, contrarily to common non-essential genes

On a first look, there is no correlation between essentiality (Fig 4) and conservation (Fig 5) at the individual gene level. The same is even more evident in the case of the highly conserved ABC domains in transporters (S1 Table), with the majority (20/33) not being essential in any dataset in DEG. This substantiates the fact that highly conserved genes are not necessarily highly essential, likely due to genetic redundancy (see Discussion). The only subsystem for which the correlation is positive and significant (Pearson coefficient 0.95, p-value 1.07e-06) is Membrane Lipid Metabolism. To assess the unbiased relationship between essentiality and conservation at the individual gene level, we compared data on conservation with the experimental data on essentiality. For this purpose, we calculated a sum score for each gene's experimental essentiality and compared it with its conservation (Methods, Fig 6). This allows for measuring the level of evidence of essentiality for each gene individually with precision.

In the bottom panel of Fig 6, the vast majority of genes lie on the bottom area of the plot (sum of essentiality below zero), meaning that they show up in more datasets as non-essential than as essential. However, on the top side of the same plot, where the genes with a positive sum of essentiality lie, the clear majority are shifted to the right–meaning that they are highly conserved. There are some interesting outliers, such as glyS–glycine-tRNA ligase–highlighted in Fig 6. This corresponds to one instance in which the monophyly rule is violated: *E. coli's* type is common for most bacteria, but another type is common to some other bacteria, archaea and eukarya [90,91]. Another interesting case is holA, also highlighted, a DNA polymerase delta subunit that is also divergent in its evolutionary history [50,92] but which has been poorly studied. Upon splitting the genes into two sets–those with a positive sum of essentiality and those with a null or negative sum, two very different distributions are obtained, shown in the top panel of Fig 6. A Kolmogorov-Smirnov test states on the independence of both distributions–essential genes are much more likely to be conserved, whereas non-essential genes can or not be highly conserved. The full data on conservation and essentiality sums can be found in S3 File.

Discussion

The integration done here was, to our knowledge, the first of the kind for a wide variety of phyla of the bacteria and archaea domains, encompassing experimental phenotypic data, results of large-scale computational simulations and sequence data. The experimental genome-scale essentiality data reveal that approximately 25% of prokaryotic essential genes encode for unknown or general functions (categories S and R in S1 Fig), which is a strong warning on the need for experimental studies on the phenotype of these essential proteins for prokaryotic physiology. Moreover, the organisms for which genome-wide essentiality data are available are relatively scarce. While more experimental data are not available, computational models can be valuable tools aiding in the task of decoding prokaryotic metabolism. It is well known that GSMs are limited by the quality of the genome annotations, the formulation of the biomass equation and the pre-defined environmental conditions and other modelling artefacts [27,84], of which the impact in our results we expand below. Here we tried to reduce the impact of these limitations by basing the choice of the models on a large survey of high-quality manually curated models [84] from which 15 balanced, validated, comparable models were chosen, that at the same time represented a wide phylogenetic diversity (Table 1). The analysis of common patterns of essentiality filtered out the unique essential reactions that might represent specific errors related with individual models. In this manner, we can find core and common features in the overlap of all models. Using GSMs is particularly interesting in this sense, as they allow us to step one level above that of genomes and all their redundancy in the form of isozymes, duplications and confusing phylogenetic events as lateral gene transfers and gene losses. Manually curated GSMs are not mere reflections of genomes–they include thorough revisions of the network and addition of necessary reactions that are encoded by unknown genes or spontaneous chemical transformations. These are also assigned subsystems and counted in our results. Last, as recognized by other authors, the predictive power of comparative analysis can be significantly enhanced by using it within the functional context of pathways and subsystems [19].

The GSMs prediction of which metabolic subsystem has more genes that are commonly essential in multiple species–Cofactor and Prosthetic Group Biosynthesis–was accurate (Fig 4). The exception of the experimentally highly essential tRNA-charging functionality that was

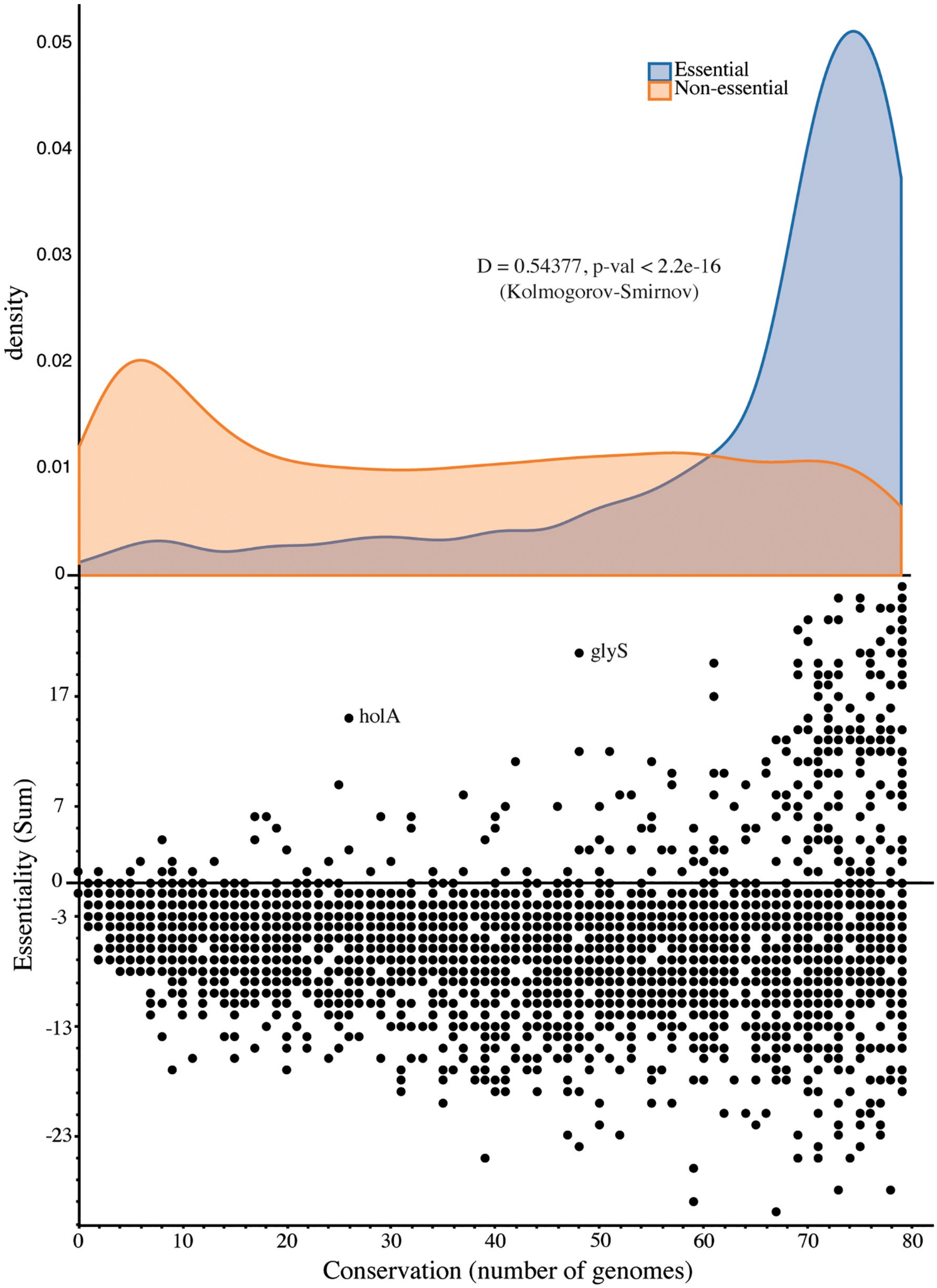

Fig 6. Conservation and essentiality of protein-encoding genes. Conservation is calculated as the number of genomes where a gene is present and essentiality as the number of datasets where a gene is essential minus the datasets where it is non-essential. The full list of genes is given in S3 File. In the bottom panel, all data is plotted, with conservation on the x axis and essentiality on the y axis. The top panel depicts the density distribution of the conservation of genes (on the shared x axis) with a sum of essentiality larger than 0 –

blue–and that of the genes with a sum of essentiality equal or smaller than 0 –orange. The results of a Kolmogorov–Smirnov test for the independence of both distributions are shown, D for the value of the test statistic and the p-value of the test.

https://doi.org/10.1371/journal.pcbi.1006556.g006

not reflected in the simulations is due to the hindrance of just one model including this subsystem [40] but it should be fixed if all the models represent appropriately this subsystem in the future. However, the analysis done here, by integrating experimental data with several different models still allowed us to identify this subsystem as highly essential and conserved (Fig 4 and Fig 5). The problem of the unstandardized biomass composition, evidenced by the GSM of *K. pneumoniae* not predicting any essential reaction involved in cofactor and prosthetic group biosynthesis due to the fact that none of those compounds is present in the biomass equation (Fig 2) is relevant to the results, a subject that we have already addressed in a recent publication [84]. Due to the incompleteness of the networks, it was not possible to complete the biomass equations with the missing cofactors without an impractical manual editing and curation of most models. However, considering the results obtained here, this incompleteness could readily be identified (Fig 2) and did not impair the prediction of an overwhelming majority of essential reactions related with the subsystem of cofactor and prosthetic group biosynthesis (Fig 4)–the overlap of all other models and the experimental data reveals the conserved essentiality of this subsystem.

The comparison of the modelling and experimental results can help raise specific hypotheses and directions for more detailed investigation, as discussed above for the essentiality of chorismate synthase in GSMs. In the analysed experimental datasets obtained in rich media, chorismate synthase is not essential, but is has been shown that in minimal media in *E. coli* the knock-out of its gene impairs growth [87] that can be restored with p-aminobenzoic acid (PABA) or derivatives [88]. The transporters for these compounds should be added to the models, and we expect that a curated analysis of experimental studies of auxotrophies in the literature can point several more additions to GSMs that will improve predictions at the gene-level. Moreover, these results point also to the necessity of performing more often essentiality experiments in defined media. On the other direction, genome-scale models can also indicate improvements for experimental assays. At the moment of writing of this manuscript, a new genome-wide screen for *Bacillus subtilis* was published [93], where folC (dihydrofolate synthase) was found to be essential, confirming our modelling results and contradicting the previous experimental results for *B. subtilis* that were used here [50].

The analysis of conservation of metabolic genes here was the first performed using a manually curated annotation system for metabolic pathways and subsystems, with the most complete genome-scale metabolic model of a prokaryote to date [49]. Regarding inferences on ancestry, there are some limitations to our approach. We chose to use a single e-value threshold in a local BLAST–this might be a lax threshold, but at the same time it allows us to recover potential very ancient homologs, tracing back all the way to the Last Universal Common Ancestor (LUCA), and not to incur in debates about in and out-paralogs. Moreover, we used only a sample of prokaryotic genomes– 79 –although we made sure to include representatives of all sequenced phyla to date (S2 Fig), and we took a stringent threshold to indicate and not affirm ancestry (at least 70 out of 79 genomes). Looking at the conservation of subsystems (Fig 5) also allows us to overcome the phylogenetic distribution bias. On another note, looking only for universal genes as markers of ancestry can be a limited approach, due to the phenomena of gene loss and lateral gene transfer–ideally, phylogenetic trees should be built for all genes. A recent study used an innovative and large-scale approach to infer on the genome of LUCA, building all trees for protein families based on 1981 prokaryotic genomes [94]. Interestingly, although using a completely alternative approach, the study also concluded on tRNA charging and cofactor metabolism as being ancient subsystems. These findings corroborate that the genes identified here as present in all genomes of all representative phyla are most likely genes present in the last common ancestor [28]. Overall, our results of high conservation of the tRNA charging system, Transport and Oxidative Phosphorylation point to a last common ancestor metabolic network of prokaryotes where most of the nutrients were uptaked with nonspecific transporters at the expense of ATP and in which tRNA charging was already present. The results also suggest that the catalytic role of cofactors and prosthetic groups was a coin highly sought for in early prebiotic systems still maintained today, as this is the most conserved metabolic subsystem after transport and tRNA charging. It is highly likely that genes encoding for enzymes aiding in cofactor biosynthesis were selected for early in primordial evolution, as was suggested elsewhere for the origin of anabolic pathways in prebiotic systems [95].

This work expanded considerably on previous related studies regarding the relationship between gene conservation and essentiality in width and depth. The demonstration that essential genes are more evolutionary conserved that non-essential [22], corroborated later with more datasets [21], used the ratio of non-synonymous substitutions to synonymous substitutions in the genomes to estimate conservation (Ka/Ks). Here, 36 experimentally essential datasets were used, that included one Archaea (Table 2). The conservation was analysed by looking at the presence of each gene in 79 genomes that were manually selected to represent all the phyla with one fully-sequenced genome in the prokaryotic tree of life. Because each gene might be essential in some datasets in DEG, non-essential in others and not assayed in yet others, instead of analysing essential genes separately from non-essential as in the two aforementioned studies, we used a measure of essentiality for each gene (sum of essentiality) that takes into account the datasets where it shows up as essential and those where it is non-essential (Fig 6). The results show that genes with a positive sum of essentiality (more datasets showing essential than non-essential) are much scarcer than those with a negative sum; however, it is much more likely that they are highly conserved. For genes with a negative sum of essentiality, there is no tendency for high or low conservation, with a uniform distribution of these genes for all the values of conservation (corroborating results by Fang et al. [92]). We also expanded on previous studies [21,22] by integrating *in silico* simulations and functional assessment of the data, with the conclusion that with the exception of the tRNA charging subsystem, the majority of highly conserved genes related with transport and cofactor biosynthesis are not highly essential (Fig 6, S1 and S2 Tables). These two subsystems show low single-gene essentiality most likely due to metabolic redundancy caused by known alternative metabolic routes (for which multiple knock-outs ought to be performed to test for subsystem-level essentiality) complemented with enzymatic activities not yet known (supported by the percentage of genes with general function prediction only S1 Fig) that might also include promiscuous enzymes [96]. The remarkable redundancy of metabolic networks is reflected in the resilience and robustness of prokaryotic life for the billions of years that it has inhabited Earth.

Supporting information

S1 Fig. COG functional categories for prokaryotic essential genes in DEG. COG metabolic functional categories are less detailed than those used in the annotation of metabolic models: both the “Energy production and Conversion” and “Amino acid transport and metabolism” functional categories encompass several of those that are detailed within GSMs, except for the “Transport” category lumped in GSMs, and distributed for each of the major biomolecules in the COG system. Misleading COG annotations included thiO, an essential gene for the biosynthesis of thiamine diphosphate annotated in category E (Amino acid transport and

metabolism) and csd, a cysteine desulfurase, essential in 7 datasets and involved in the formation of Fe-S clusters, also annotated in category E.
(PDF)

S2 Fig. Phylogenetic reference tree for species used in the analysis of conservation of essential genes (based on NCBI taxonomy).
(PDF)

S1 Table. Ubiquitous transporter genes. The list includes genes conserved in all 79 prokaryotic genomes analysed. Essentiality is given as the number of datasets in DEG (out of 36) in which each gene is essential. The description is that of the corresponding annotated ORF in the genome of *E. coli* K12.
(PDF)

S2 Table. Highly conserved cofactor biosynthesis genes in prokaryotic genomes. Essentiality is given as the number of datasets in DEG (out of 36) in which each gene is essential. Conservation is given as the percentage of the 79 genomes where a significant homolog for this gene was found. The description is that of the corresponding annotated ORF in the genome of *E. coli* K12. Biosynthesized cofactors were manually retrieved from the detailed information available in the Metacyc database.
(PDF)

S1 File. Essential reactions in metabolic models. Presence-absence matrix of essential reactions in the 15 GSMs used.
(XLSX)

S2 File. Translation of metabolic subsystems of GSMs to a standardized nomenclature. Lists of original subsystem classifications and corresponding translation for all models and all reactions.
(XLSX)

S3 File. Essentiality and conservation scores. List of genes with corresponding scores of essentiality and conservation.
(XLSX)

Acknowledgments

We thank Fábio Ramalho and Sergej Andrejev for crucial assistance with bioinformatics tools, in particular with the Biopython package and FBA simulations, respectively.

Author Contributions

Conceptualization: Joana C. Xavier, Kiran Raosaheb Patil, Isabel Rocha.

Data curation: Joana C. Xavier.

Formal analysis: Joana C. Xavier.

Funding acquisition: Isabel Rocha.

Investigation: Joana C. Xavier.

Methodology: Joana C. Xavier, Kiran Raosaheb Patil, Isabel Rocha.

Project administration: Isabel Rocha.

Resources: Kiran Raosaheb Patil, Isabel Rocha.

Supervision: Kiran Raosaheb Patil, Isabel Rocha.

Validation: Joana C. Xavier.

Visualization: Joana C. Xavier.

Writing – original draft: Joana C. Xavier.

Writing – review & editing: Kiran Raosaheb Patil, Isabel Rocha.

References

1. Kauffman S. At Home in the Universe: The Search for the Laws of Self-Organization and Complexity. Oxford University Press; 1995.
2. Rasmussen S, Bedau MA, Chen L, Deamer D, Krakauer DC, Packard NH, et al., editors. Protocells. 1st ed. London: The MIT Press; 2008.
3. Schuster P. How does complexity arise in evolution:Nature's recipe for mastering scarcity, abundance, and unpredictability. Complexity. 1996; 2: 22–30.
4. Kim KM, Caetano-Anollés G. The proteomic complexity and rise of the primordial ancestor of diversified life. BMC Evol Biol. 2011; 11: 140. https://doi.org/10.1186/1471-2148-11-140 PMID: 21612591
5. Oltvai ZN, Barabási A-L. Systems biology. Life's complexity pyramid. Science. 2002; 298: 763–764. https://doi.org/10.1126/science.1078563 PMID: 12399572
6. Peretó J. Out of fuzzy chemistry: from prebiotic chemistry to metabolic networks. Chem Soc Rev. 2012; 41: 5394. https://doi.org/10.1039/c2cs35054h PMID: 22508108
7. Skouloubris S, Thiberge JM, Labigne A, De Reuse H. The Helicobacter pylori UreI protein is not involved in urease activity but is essential for bacterial survival in vivo. Infect Immun. 1998; 66: 4517–21. PMID: 9712811
8. Fayet O, Ziegelhoffer T, Georgopoulos C. The groES and groEL heat shock gene products of Escherichia coli are essential for bacterial growth at all temperatures. J Bacteriol. 1989; 171: 1379–85. PMID: 2563997
9. Wu J, Ohta N, Zhao J-L, Newton A. A novel bacterial tyrosine kinase essential for cell division and differentiation. Proc Natl Acad Sci. 1999; 96: 13068–13073. https://doi.org/10.1073/pnas.96.23.13068 PMID: 10557274
10. Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, et al. Experimental Determination and System Level Analysis of Essential Genes in Escherichia coli MG1655. J Bacteriol. 2003; 185: 5673–5684. https://doi.org/10.1128/JB.185.19.5673-5684.2003 PMID: 13129938
11. Gallagher LA, Ramage E, Jacobs MA, Kaul R, Brittnacher M, Manoil C. A comprehensive transposon mutant library of Francisella novicida, a bioweapon surrogate. Proc Natl Acad Sci U S A. 2007; 104: 1009–14. https://doi.org/10.1073/pnas.0606713104 PMID: 17215359
12. de Berardinis V, Vallenet D, Castelli V, Besnard M, Pinet A, Cruaud C, et al. A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. Mol Syst Biol. 2008; 4: 174. https://doi.org/10.1038/msb.2008.10 PMID: 18319726
13. Cameron DE, Urbach JM, Mekalanos JJ. A defined transposon mutant library and its use in identifying motility genes in Vibrio cholerae. Proc Natl Acad Sci U S A. 2008; 105: 8736–41. https://doi.org/10.1073/pnas.0803281105 PMID: 18574146
14. Langridge GC, Phan M-D, Turner DJ, Perkins TT, Parts L, Haase J, et al. Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res. 2009; 19: 2308–16. https://doi.org/10.1101/gr.097097.109 PMID: 19826075
15. Chaudhuri RR, Allen AG, Owen PJ, Shalom G, Stone K, Harrison M, et al. Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). BMC Genomics. 2009; 10: 291. https://doi.org/10.1186/1471-2164-10-291 PMID: 19570206
16. Chen W-H, Minguez P, Lercher MJ, Bork P. OGEE: an online gene essentiality database. Nucleic Acids Res. 2012; 40: D901–6. https://doi.org/10.1093/nar/gkr986 PMID: 22075992
17. Luo H, Lin Y, Gao F, Zhang C-TT, Zhang R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic Acids Res. 2014; 42: 574–580. https://doi.org/10.1093/nar/gkt1131 PMID: 24243843
18. Rocha EPC, Danchin A. Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res. 2003; 31: 6570–6577. https://doi.org/10.1093/nar/gkg859 PMID: 14602916

19. Gerdes S, Edwards R, Kubal M, Fonstein M, Stevens R, Osterman A. Essential genes on metabolic maps. Curr Opin Biotechnol. 2006; 17: 448–56. https://doi.org/10.1016/j.copbio.2006.08.006 PMID: 16978855

20. Zhang X, Peng C, Zhang G, Gao F. Comparative analysis of essential genes in prokaryotic genomic islands. Sci Rep. Nature Publishing Group; 2015; 5: 12561. https://doi.org/10.1038/srep12561 PMID: 26223387

21. Luo H, Gao F, Lin Y. Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. Sci Rep. Nature Publishing Group; 2015; 5: 13210. https://doi.org/10.1038/srep13210 PMID: 26272053

22. Jordan IK, Rogozin IB, Wolf YI, Koonin E V. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 2002; 12: 962–8. https://doi.org/10.1101/gr.87702 PMID: 12045149

23. Bordbar A, Monk JM, King ZA, Palsson BO. Constraint-based models predict metabolic and associated cellular functions. Nat Rev Genet. Nature Research; 2014; 15: 107–20. https://doi.org/10.1038/nrg3643 PMID: 24430943

24. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014; 42: D206–14. https://doi.org/10.1093/nar/gkt1226 PMID: 24293654

25. Edwards JS, Palsson BO. The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. Proc Natl Acad Sci. 2000; 97: 5528–5533. https://doi.org/10.1073/pnas.97.10.5528 PMID: 10805808

26. Oberhardt MA, Palsson BØ, Papin JA. Applications of genome-scale metabolic reconstructions. Mol Syst Biol. Department of Biomedical Engineering, University of Virginia, Health System, Charlottesville, VA, USA.; 2009; 5. https://doi.org/10.1038/msb.2009.77 PMID: 19888215

27. Ebrahim A, Almaas E, Bauer E, Bordbar A, Burgard AP, Chang RL, et al. Do genome-scale models need exact solvers or clearer standards? Mol Syst Biol. EMBO Press; 2015; 11: 831–831. https://doi.org/10.15252/msb.20156157 PMID: 26467284

28. Koonin E V. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol. 2003; 1: 127–36. https://doi.org/10.1038/nrmicro751 PMID: 15035042

29. Fournier GP, Andam CP, Gogarten JP. Ancient horizontal gene transfer and the last common ancestors. BMC Evol Biol. 2015; 15: 70. https://doi.org/10.1186/s12862-015-0350-0 PMID: 25897759

30. Mampel J, Buescher JM, Meurer G, Eck J. Coping with complexity in metabolic engineering. Trends Biotechnol. 2013; 31: 52–60. https://doi.org/10.1016/j.tibtech.2012.10.010 PMID: 23183303

31. Oh Y-K, Palsson BO, Park SM, Schilling CH, Mahadevan R. Genome-scale Reconstruction of Metabolic Network in Bacillus subtilis Based on High-throughput Phenotyping and Gene Essentiality Data. J Biol Chem. 2007; 282: 28791–28799. https://doi.org/10.1074/jbc.M703759200 PMID: 17573341

32. Milne CB, Eddy JA, Raju R, Ardekani S, Kim P-J, Senger RS, et al. Metabolic network reconstruction and genome-scale model of butanol-producing strain Clostridium beijerinckii NCIMB 8052. BMC Syst Biol. 2011; 5: 130. https://doi.org/10.1186/1752-0509-5-130 PMID: 21846360

33. Becker S a, Palsson BØ. Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. BMC Microbiol. 2005; 5: 8. https://doi.org/10.1186/1471-2180-5-8 PMID: 15752426

34. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, et al. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol. 2007; 3: 121. https://doi.org/10.1038/msb4100155 PMID: 17593909

35. Archer CT, Kim JF, Jeong H, Park JH, Vickers CE, Lee SY, et al. The genome sequence of E. coli W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of E. coli. BMC Genomics. BioMed Central Ltd; 2011; 12: 9. https://doi.org/10.1186/1471-2164-12-9 PMID: 21208457

36. Thiele I, Vo TD, Price ND, Palsson BØ. Expanded Metabolic Reconstruction of Helicobacter pylori (i IT341 GSM / GPR): an In Silico Genome-Scale Characterization of Single- and Double-Deletion Mutants. J Bacteriol. 2005; 187: 5818–5830. https://doi.org/10.1128/JB.187.16.5818-5830.2005 PMID: 16077130

37. Liao YC, Huang TW, Chen FC, Charusanti P, Hong JSJ, Chang HY, et al. An experimentally validated genome-scale metabolic reconstruction of Klebsiella pneumoniae MGH 78578, iYL1228. J Bacteriol. 2011; 193: 1710–1717. https://doi.org/10.1128/JB.01218-10 PMID: 21296962

38. Nogales J, Palsson BØ, Thiele I. A genome-scale metabolic reconstruction of Pseudomonas putida KT2440: iJN746 as a cell factory. BMC Syst Biol. 2008; 2: 79. https://doi.org/10.1186/1752-0509-2-79 PMID: 18793442

39. Thiele I, Hyduke DR, Steeb B, Fankam G, Allen DK, Bazzani S, et al. A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. BMC Syst Biol. BioMed Central Ltd; 2011; 5: 8. https://doi.org/10.1186/1752-0509-5-8 PMID: 21244678

40. Pinchuk GE, Hill EA, Geydebrekht O V, De Ingeniis J, Zhang X, Osterman A, et al. Constraint-based model of Shewanella oneidensis MR-1 metabolism: a tool for data analysis and hypothesis generation. PLoS Comput Biol. 2010; 6: e1000822. https://doi.org/10.1371/journal.pcbi.1000822 PMID: 20589080

41. Jamshidi N, Palsson BØ. Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. BMC Syst Biol. 2007; 1: 26. https://doi.org/10.1186/1752-0509-1-26 PMID: 17555602

42. Ahsanul Islam M, Edwards E a., Mahadevan R. Characterizing the metabolism of Dehalococcoides with a constraint-based model. PLoS Comput Biol. 2010; 6. https://doi.org/10.1371/journal.pcbi.1000887 PMID: 20811585

43. Nogales J, Gudmundsson S, Knight EM, Palsson BØ, Thiele I. Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis. Proc Natl Acad Sci U S A. 2012; 109: 2678–83. https://doi.org/10.1073/pnas.1117907109 PMID: 22308420

44. Zhang Y, Thiele I, Weekes D, Li Z, Jaroszewski L, Ginalski K, et al. Three-dimensional structural view of the central metabolic network of Thermotoga maritima. Science. 2009; 325: 1544–1549. https://doi.org/10.1126/science.1174671 PMID: 19762644

45. Feist AM, Scholten JCM, Palsson BØ, Brockman FJ, Ideker T. Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. Mol Syst Biol. 2006; 2. https://doi.org/10.1038/msb4100046 PMID: 16738551

46. Savinell JM, Palsson BØ. Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. J Theor Biol. 1992; 154: 421–54. https://doi.org/10.1016/S0022-5193(05)80161-4 PMID: 1593896

47. Varma A, Palsson BØ. Metabolic Capabilities of Escherichia coli: I. Synthesis of Biosynthetic Precursors and Cofactors. J Theor Biol. 1993; https://doi.org/10.1006/jtbi.1993.1202

48. Rocha I, Maia P, Evangelista P, Vilaça P, Soares S, Pinto JP, et al. OptFlux: an open-source software platform for in silico metabolic engineering. BMC Syst Biol. 2010; 4: 45. https://doi.org/10.1186/1752-0509-4-45 PMID: 20403172

49. Orth JD, Conrad TM, Na J, Lerman J a, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. Mol Syst Biol. 2011; 7: 1–9. https://doi.org/10.1038/msb.2011.65 PMID: 21988831

50. Kobayashi K, Ehrlich SD, Albertini a, Amati G, Andersen KK, Arnaud M, et al. Essential Bacillus subtilis genes. Proc Natl Acad Sci U S A. 2003; 100: 4678–83. https://doi.org/10.1073/pnas.0730515100 PMID: 12682299

51. Veeranagouda Y, Husain F, Tenorio EL, Wexler HM. Identification of genes required for the survival of B. fragilis using massive parallel sequencing of a saturated transposon mutant library. BMC Genomics. 2014; 15: 429. https://doi.org/10.1186/1471-2164-15-429 PMID: 24899126

52. Goodman AL, McNulty NP, Zhao Y, Leip D, Mitra RD, Lozupone C a., et al. Identifying Genetic Determinants Needed to Establish a Human Gut Symbiont in Its Habitat. Cell Host Microbe. 2009; 6: 279–289. https://doi.org/10.1016/j.chom.2009.08.003 PMID: 19748469

53. Moule MG, Hemsley CM, Seet Q, Guerra-Assuncao JA, Lim J, Sarkar-Tyson M, et al. Genome-wide saturation mutagenesis of Burkholderia pseudomallei K96243 predicts essential genes and novel targets for antimicrobial development. MBio. 2014; 5: e00926–13. https://doi.org/10.1128/mBio.00926-13 PMID: 24520057

54. Baugh L, Gallagher LA, Patrapuvich R, Clifton MC, Gardberg AS, Edwards TE, et al. Combining functional and structural genomics to sample the essential Burkholderia structome. PLoS One. 2013; 8: e53851. https://doi.org/10.1371/journal.pone.0053851 PMID: 23382856

55. Metris A, Reuter M, Gaskin DJH, Baranyi J, van Vliet AHM. In vivo and in silico determination of essential genes of Campylobacter jejuni. BMC Genomics. 2011; 12: 535. https://doi.org/10.1186/1471-2164-12-535 PMID: 22044676

56. Christen B, Abeliuk E, Collier JM, Kalogeraki VS, Passarelli B, Coller JA, et al. The essential genome of a bacterium. Mol Syst Biol. 2011; 7: 528. https://doi.org/10.1038/msb.2011.58 PMID: 21878915

57. Baba T, Ara T, Hasegawa M. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006; 2. https://doi.org/10.1038/msb4100050 PMID: 16738554

58. Akerley BJ, Rubin EJ, Novick VL, Amaya K, Judson N, Mekalanos JJ. A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. Proc Natl Acad Sci U S A. 2002; 99: 966–71. https://doi.org/10.1073/pnas.012602299 PMID: 11805338

59. Salama NR, Shepherd B, Falkow S. Global transposon mutagenesis and essential gene analysis of Helicobacter pylori. J Bacteriol. 2004; 186: 7926–35. https://doi.org/10.1128/JB.186.23.7926-7935.2004 PMID: 15547264

60. Sarmiento F, Mrazek J, Whitman WB. Genome-scale analysis of gene function in the hydrogenotrophic methanogenic archaeon Methanococcus maripaludis. Proc Natl Acad Sci U S A. 2013; https://doi.org/10.1073/pnas.1220225110 PMID: 23487778

61. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol. 2003; 48: 77–84. https://doi.org/10.1046/j.1365-2958.2003.03425.x PMID: 12657046

62. Griffin JE, Gawronski JD, Dejesus MA, Ioerger TR, Akerley BJ, Sassetti CM. High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. PLoS Pathog. 2011; 7: e1002251. https://doi.org/10.1371/journal.ppat.1002251 PMID: 21980284

63. Zhang YJ, Ioerger TR, Huttenhower C, Long JE, Sassetti CM, Sacchettini JC, et al. Global assessment of genomic regions required for growth in Mycobacterium tuberculosis. PLoS Pathog. 2012; 8: e1002946. https://doi.org/10.1371/journal.ppat.1002946 PMID: 23028335

64. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, et al. Essential genes of a minimal bacterium. Proc Natl Acad Sci U S A. 2006; 103: 425–430. https://doi.org/10.1073/pnas.0510013103 PMID: 16407165

65. French CT, Lao P, Loraine AE, Matthews BT, Yu H, Dybvig K. Large-scale transposon mutagenesis of Mycoplasma pulmonis. Mol Microbiol. 2008; 69: 67–76. https://doi.org/10.1111/j.1365-2958.2008.06262.x PMID: 18452587

66. Klein B a, Tenorio EL, Lazinski DW, Camilli A, Duncan MJ, Hu LT. Identification of essential genes of the periodontal pathogen Porphyromonas gingivalis. BMC Genomics. 2012. https://doi.org/10.1186/1471-2164-13-578 PMID: 23114059

67. Gallagher LA, Shendure J, Manoil C. Genome-scale identification of resistance functions in Pseudomonas aeruginosa using Tn-seq. MBio. 2011; 2: e00315–10. https://doi.org/10.1128/mBio.00315-10 PMID: 21253457

68. Liberati NT, Urbach JM, Miyata S, Lee DG, Drenkard E, Wu G, et al. An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. Proc Natl Acad Sci U S A. 2006; 103: 2833–8. https://doi.org/10.1073/pnas.0511100103 PMID: 16477005

69. Barquist L, Langridge GC, Turner DJ, Phan M-D, Turner AK, Bateman A, et al. A comparison of dense transposon insertion libraries in the Salmonella serovars Typhi and Typhimurium. Nucleic Acids Res. 2013; 41: 4549–64. https://doi.org/10.1093/nar/gkt148 PMID: 23470992

70. Khatiwara A, Jiang T, Sung S-S, Dawoud T, Kim JN, Bhattacharya D, et al. Genome scanning for conditionally essential genes in Salmonella enterica Serotype Typhimurium. Appl Environ Microbiol. 2012; 78: 3098–107. https://doi.org/10.1128/AEM.06865-11 PMID: 22367088

71. Knuth K, Niesalla H, Hueck CJ, Fuchs TM. Large-scale identification of essential Salmonella genes by trapping lethal insertions. Mol Microbiol. 2004; 51: 1729–44. PMID: 15009898

72. Deutschbauer A, Price MN, Wetmore KM, Shao W, Baumohl JK, Xu Z, et al. Evidence-based annotation of gene function in Shewanella oneidensis MR-1 using genome-wide fitness profiling across 121 conditions. PLoS Genet. 2011; 7: e1002385. https://doi.org/10.1371/journal.pgen.1002385 PMID: 22125499

73. Roggo C, Coronado E, Moreno-Forero SK, Harshman K, Weber J, Van der Meer JR. Genome-wide transposon insertion scanning of environmental survival functions in the polycyclic aromatic hydrocarbon degrading bacterium Sphingomonas wittichiiRW1. Environ Microbiol. 2013; 15: 2681–2695. https://doi.org/10.1111/1462-2920.12125 PMID: 23601288

74. Ji Y, Zhang B, Van SF, Horn Warren P, Woodnutt G, et al. Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. Science. 2001; 293: 2266–9. https://doi.org/10.1126/science.1063566 PMID: 11567142

75. Thanassi JA, Hartman-Neumann SL, Dougherty TJ, Dougherty BA, Pucci MJ. Identification of 113 conserved essential genes using a high-throughput gene disruption system in Streptococcus pneumoniae. Nucleic Acids Res. 2002; 30: 3152–3162. https://doi.org/10.1093/nar/gkf418 PMID: 12136097

76. Le Breton Y, Belew AT, Valdes KM, Islam E, Curry P, Tettelin H, et al. Essential genes in the core genome of the human pathogen Streptococcus pyogenes. Sci Rep. 2015; 5: 9838. https://doi.org/10.1038/srep09838 PMID: 25996237

77. Xu P, Ge X, Chen L, Wang X, Dou Y, Xu JZ, et al. Genome-wide essential gene identification in Streptococcus sanguinis. Sci Rep. 2011; 1: 1–9. https://doi.org/10.1038/srep00001

78. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. Oxford University Press; 2016; 44: W242–W245. https://doi.org/10.1093/nar/gkw290 PMID: 27095192

79. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25: 1422–3. https://doi.org/10.1093/bioinformatics/btp163 PMID: 19304878

80. Suzuki R, Shimodaira H. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006; 22: 1540–1542. https://doi.org/10.1093/bioinformatics/btl117 PMID: 16595560

81. Bermingham A, Derrick JP. The folic acid biosynthesis pathway in bacteria: evaluation of potential for antibacterial drug discovery. BioEssays. 2002; 24: 637–648. https://doi.org/10.1002/bies.10114 PMID: 12111724

82. Magni G, Orsomando G, Raffaelli N. Structural and functional properties of NAD kinase, a key enzyme in NADP biosynthesis. Mini Rev Med Chem. 2006; 6: 739–746. https://doi.org/10.2174/138955706777698688 PMID: 16842123

83. Saeed-Kothe A, Yang W, Mills SD. Use of the riboflavin synthase gene (ribC) as a model for development of an essential gene disruption and complementation system for Haemophilus influenzae. Appl Environ Microbiol. American Society for Microbiology; 2004; 70: 4136–43. https://doi.org/10.1128/AEM.70.7.4136-4143.2004 PMID: 15240293

84. Xavier JCJC, Patil KR, Rocha I. Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes. Metab Eng. 2017; 39: 200–208. https://doi.org/10.1016/j.ymben.2016.12.002 PMID: 27939572

85. Rock CO, Jackowski S. Forty years of bacterial fatty acid synthesis. Biochem Biophys Res Commun. 2002; 292: 1155–1166. https://doi.org/10.1006/bbrc.2001.2022 PMID: 11969206

86. Coggins JR, Abell C, Evans LB, Frederickson M, Robinson D a, Roszak a W, et al. Experiences with the shikimate-pathway enzymes as targets for rational drug design. Biochem Soc Trans. 2003; 31: 548–552. https://doi.org/10.1042/ PMID: 12773154

87. Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, et al. Experimental and computational assessment of conditionally essential genes in Escherichia coli. J Bacteriol. 2006; 188: 8259–71. https://doi.org/10.1128/JB.00740-06 PMID: 17012394

88. Lee Y, Umeano A, Balskus EP. Rescuing auxotrophic microorganisms with nonenzymatic chemistry. Angew Chem Int Ed Engl. 2013; 52: 11800–3. https://doi.org/10.1002/anie.201307033 PMID: 24115592

89. Nachin L. SufC: an unorthodox cytoplasmic ABC/ATPase required for [Fe-S] biogenesis under oxidative stress. EMBO J. 2003; 22: 427–437. https://doi.org/10.1093/emboj/cdg061 PMID: 12554644

90. Mazauric M-H, Reinbolt J, Lorber B, Ebel C, Keith G, Giege R, et al. An Example of Non-Conservation of Oligomeric Structure in Prokaryotic Aminoacyl-tRNA Synthetases. Biochemical and Structural Properties of Glycyl-tRNA Synthetase from Thermus thermophilus. Eur J Biochem. 1996; 241: 814–826. https://doi.org/10.1111/j.1432-1033.1996.00814.x PMID: 8944770

91. Woese CR, Olsen GJ, Ibba M, Soll D. Aminoacyl-tRNA Synthetases, the Genetic Code, and the Evolutionary Process. Microbiol Mol Biol Rev. 2000; 64: 202–236. https://doi.org/10.1128/MMBR.64.1.202–236.2000 PMID: 10704480

92. Fang G, Rocha E, Danchin A. How Essential Are Nonessential Genes? Mol Biol Evol. Oxford University Press; 2005; 22: 2147–2156. https://doi.org/10.1093/molbev/msi211 PMID: 16014871

93. Koo B-M, Kritikos G, Farelli JD, Todor H, Tong K, Kimsey H, et al. Construction and Analysis of Two Genome-Scale Deletion Libraries for Bacillus subtilis. Cell Syst. ASM Press; 2017; 4: 291–305.e7. https://doi.org/10.1016/j.cels.2016.12.013 PMID: 28189581

94. Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, et al. The physiology and habitat of the last universal common ancestor. Nat Microbiol. Nature Publishing Group; 2016; 1: 16116. https://doi.org/10.1038/nmicrobiol.2016.116 PMID: 27562259

95. Fani R, Fondi M. Origin and evolution of metabolic pathways. Phys Life Rev. Elsevier B.V.; 2009; 6: 23–52. https://doi.org/10.1016/j.plrev.2008.12.003 PMID: 20416849

96. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. Multicopy suppression underpins metabolic evolvability. Mol Biol Evol. 2007; 24: 2716–22. https://doi.org/10.1093/molbev/msm204 PMID: 17884825