Molecular Systems Biology 7; Article number 535; doi:10.1038/msb.2011.65
Citation: ***Molecular Systems Biology*** **7: 535**
© 2011 EMBO and Macmillan Publishers Limited All rights reserved 1744-4292/11
www.molecularsystemsbiology.com

molecular systems biology

REPORT

A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism—2011

Jeffrey D Orth[1], Tom M Conrad[1], Jessica Na[1], Joshua A Lerman[2], Hojung Nam[1], Adam M Feist[1] and Bernhard Ø Palsson[1,*]

[1] Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA and [2] Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA, USA
* Corresponding author. Department of Bioengineering, University of California, San Diego, 9500 Gilman Drive, Mail Code 0412, La Jolla, CA 92093-0412, USA. Tel.: +1 858 534 5668; Fax: +1 858 822 3120; E-mail: palsson@ucsd.edu

Received 5.4.11; accepted 17.8.11

The initial genome-scale reconstruction of the metabolic network of *Escherichia coli* K-12 MG1655 was assembled in 2000. It has been updated and periodically released since then based on new and curated genomic and biochemical knowledge. An update has now been built, named *i*JO1366, which accounts for 1366 genes, 2251 metabolic reactions, and 1136 unique metabolites. *i*JO1366 was (1) updated in part using a new experimental screen of 1075 gene knockout strains, illuminating cases where alternative pathways and isozymes are yet to be discovered, (2) compared with its predecessor and to experimental data sets to confirm that it continues to make accurate phenotypic predictions of growth on different substrates and for gene knockout strains, and (3) mapped to the genomes of all available sequenced *E. coli* strains, including pathogens, leading to the identification of hundreds of unannotated genes in these organisms. Like its predecessors, the *i*JO1366 reconstruction is expected to be widely deployed for studying the systems biology of *E. coli* and for metabolic engineering applications.

Molecular Systems Biology **7**: 535; published online 11 October 2011; doi:10.1038/msb.2011.65
Subject Categories: metabolic and regulatory networks; microbiology & pathogens
Keywords: constraint-based modeling; *Escherichia coli*; metabolic network reconstruction; metabolism; phenotypic screening

Introduction

A common denominator for systems biology studies of a target organism is a high-quality genome-scale metabolic network reconstruction. A network reconstruction represents a biochemically, genetically, and genomically structured knowledgebase that contains detailed information about an organism in a structured format (Thiele and Palsson, 2010). Metabolic network reconstructions contain information such as exact stoichiometry of metabolic reactions, chemical formulas and charges of metabolites, and the associations between genes, proteins, and reactions. These reconstructions form a basis for the formulation of mechanistic, and thus computable, genome-scale genotype–phenotype relationships (Palsson, 2009).

The most detailed and complete metabolic reconstruction of any organism to date is for the common laboratory strain *Escherichia coli* K-12 MG1655. The first genome-scale reconstruction of *E. coli* was *i*JE660 (Edwards and Palsson, 2000). This network was constructed through extensive searches of literature and databases to ensure correct stoichiometry and cofactor usage, and was the most extensive metabolic network reconstruction in existence at that time. An updated version of this reconstruction, *i*JR904 (Reed *et al*, 2003), had an expanded scope, including pathways for the consumption of alternate carbon sources and more specific quinone usage in the electron transport system. Hundreds of new genes and reactions were added, gene–protein–reaction associations (GPRs) were included for the first time to connect reactions with genes, and all reactions were elementally and charged balanced through the inclusion of protons. In the next update, *i*AF1260 (Feist *et al*, 2007), the scope of the network was expanded again, now including many reactions for the synthesis of cell wall components, and all metabolites were assigned to the cytoplasm, periplasm, or extracellular space. The thermodynamic properties of each reaction were calculated, and this was used to set lower bounds on predicted irreversible reactions. *i*AF1260 contained 2077 reactions, 1039 metabolites, and 1260 genes. A core model version of *i*AF1260, useful for testing and debugging new constraint-based algorithms and for educational use, has also been published (Orth *et al*, 2010a).

Here, we present an updated version of the *E. coli* metabolic network reconstruction. This new version, titled *i*JO1366, includes many newly characterized genes and reactions. Since the *i*AF1260 model was a very complete representation of the known metabolism of *E. coli*, only minor expansions in the scope of the network were made. Still, new discoveries since

2007 have made this model update necessary. Several genes were added based on the results of an experimental screen of *E. coli* knockout strains in four different media conditions. The gaps in the *i*AF1260 network were identified and characterized, and new reactions and genes were added to reduce the total number of gaps. The *i*JO1366 reconstruction can serve as a basis for metabolic network reconstructions of other *E. coli* strains and closely related organisms. We assembled preliminary reconstructions for many other *E. coli* strains, and analysis of their completeness provides insights into the metabolic networks of these organisms. *i*JO1366 is the most complete *E. coli* metabolic reconstruction to date, and like its predecessors, it will likely aid in many new discoveries (Feist and Palsson, 2008).

Results and discussion

Process for updating the reconstruction and its content

The updated network reconstruction of *E. coli* K-12 MG1655 began with the *i*AF1260b network (Feist *et al*, 2010), a slightly updated version of the *i*AF1260 network. In order to identify incorrect model predictions in order to improve the *E. coli* reconstruction, we experimentally determined conditional essentiality for most of the genes in the *i*AF1260b model. By comparing model predicted growth phenotypes to the measurements, errors in the reconstruction were found and several updates were made (Supplementary Table 1). For a discussion of the updates made to *i*AF1260b based on this screen, see *Experimental phenotypic screens* in the Supplementary Information. Next, literature and database searches were used to add newly characterized genes and reactions since 2007. The EcoCyc (Keseler *et al*, 2009) and KEGG (Kanehisa *et al*, 2010) databases were used extensively for this purpose. Results from the experimental screen described above also led to several model updates. After this first round of updates, the reconstruction contained 1274 genes. The network gaps (Orth and Palsson, 2010) in this version of the reconstruction were then investigated using a modified version of the GapFind algorithm (Satish Kumar *et al*, 2007). All orphan reactions (reactions without known associated genes) in the reconstruction were also identified from the model GPRs. Gaps were manually sorted into scope and knowledge gaps. Scope gaps are metabolites that are blocked in a model due to the limited scope of the network reconstruction, but have actual known producing and consuming reactions. Knowledge gaps exist because our knowledge of any metabolic network is incomplete. Targeted literature and database searches were performed for each knowledge gap to try to identify any known metabolic reactions missing from the reconstruction. We continued to add newly published metabolic information to the reconstruction during this gap-filling process, and the reconstruction was ultimately updated to *i*JO1366. All manual curation followed an established protocol (Thiele and Palsson, 2010).

*i*JO1366 represents a significant expansion of the *E. coli* reconstruction, as it contains 1366 genes, 2251 metabolic reactions, and 1136 unique metabolites. A comparison of the content of *i*JO1366 and its predecessor, *i*AF1260, is presented in Table I. Like *i*AF1260, *i*JO1366 contains a wide range of metabolic functions (Figure 1). The complete lists of reactions and metabolites in *i*JO1366 can be found in Supplementary Tables 2 and 3, with a list of all references used in Supplementary Table 4. *i*JO1366 accounts for three cellular compartments: the cytoplasm, periplasm, and extracellular space. In total, 107 new genes were added to the reconstruction, while one gene, *prpE* (b0335), was removed. Most new genes added have been characterized since *i*AF1260 was published in 2007 (Figure 1D). The fact that some references predate previous versions of the *E. coli* reconstruction does not necessarily mean that they were previously missed. Rather, as genes and reactions are often added on a pathway basis,

Table I Properties of *i*JO1366 and *i*AF1260

	*i*JO1366	*i*AF1260
	(this study)	(Feist *et al*, 2007)
Included genes	1366 (32%)[a]	1260 (29%)
Experimentally based function	1328 (97%)	1227 (97%)
Computationally predicted function	38 (3%)	33 (3%)
Unique functional proteins	1254	1148
Multigene complexes	185	167
Genes involved in complexes	483	415
Instances of isozymes[b]	380	346
Reactions	2251	2077
Metabolic reactions	1473	1387
Unique metabolic reactions[c]	1424	1339
Cytoplasmic	1272	1187
Periplasmic	193	192
Extracellular	8	8
Transport reactions	778	690
Cytoplasm to periplasm	447	390
Periplasm to extracellular	329	298
Cytoplasm to extracellular	2	2
Gene–protein–reaction associations		
Gene associated (metabolic/transport)	1382/706	1294/625
Spontaneous/diffusion reactions[d]	21/14	16/9
Total (gene associated and no association needed)	1403/720 (94%)	1310/634 (94%)
No gene association (metabolic/transport)	70/58 (6%)	77/56 (6%)
Exchange reactions	330	304
Metabolites		
Unique metabolites	1136	1039
Cytoplasmic	1039	951
Periplasmic	442	418
Extracellular	324	299

[a] Overall gene coverage based on 4325 total ORFs in *Escherichia coli* (annotation U00096.2, downloaded from ecogene.org); 2851 of these ORFs have been experimentally verified.
[b] Tabulated on a reaction basis, not including outer membrane non-specific porin transport.
[c] Reactions can occur in or between multiple compartments and metabolites can be present in more than one compartment.
[d] Diffusion reactions do not include facilitated diffusion reactions and are not included in this total if they can also be catalyzed by a gene product at a higher rate.

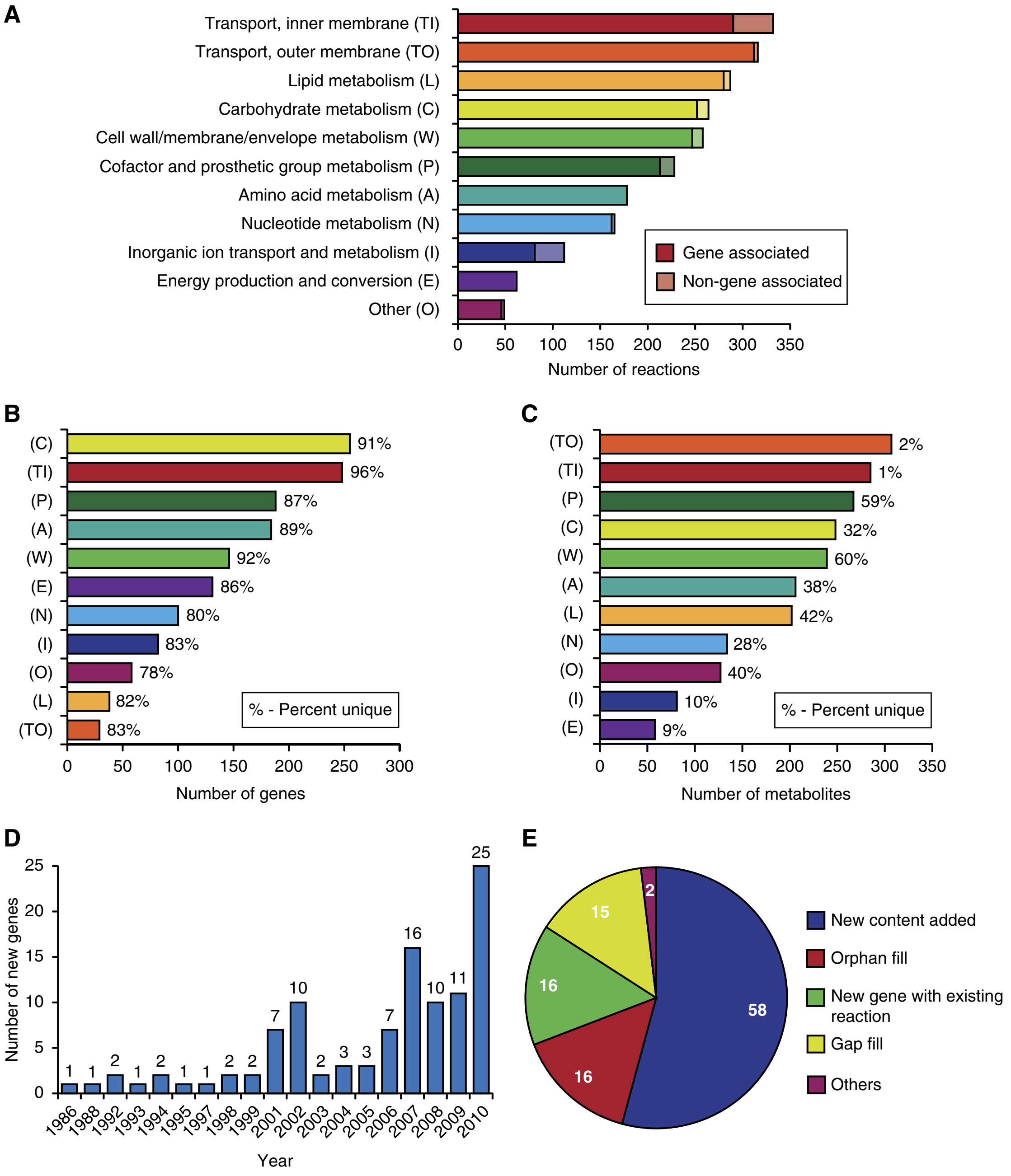

Figure 1 Properties of *i*JO1366. (**A**) The number of reactions in each of 11 functional categories. Non-gene-associated (orphan) reactions are indicated by the lighter portion at the far right of each bar. (**B**) The number of genes with associated reactions in each category. The number of genes unique to each category (i.e. associated only with reactions in one category) is given as a percentage. (**C**) The number of unique metabolites that participate in at least one reaction in each category, with the number of metabolites unique to each category indicated. (**D**) Histogram of the years in which the function of each of the 107 new genes was first unambiguously identified. (**E**) Classification of each of the 107 new genes in *i*JO1366. 'New content added' includes genes associated with new (non-gap-filling) pathways and systems in the model. 'Orphan fill' includes genes associated with orphan reactions from *i*AF1260. 'New gene with existing reaction' includes new isozymes for existing gene-associated reactions in *i*AF1260. 'Gap fill' includes genes associated with new gap-filling reactions. 'Others' includes genes that are associated both with new, non-gap-filling reactions, and with a previous orphan reaction or as a new isozyme.

complete functional pathways are typically fully elucidated over time from multiple sources. The new genes mostly add new pathways and systems to the network, but a significant number of them fill gaps and orphan reactions in existing systems (Figure 1E). A complete list of all new and removed genes, reactions, and metabolites can be found in Supplementary Table 5. The 'core' and 'wild-type' biomass reactions of *i*AF1260 have also been updated in *i*JO1366. These are reactions that drain biomass precursor compounds in experimentally determined ratios to simulate growth (Feist and Palsson, 2010). For the complete core and wild-type biomass reactions see *Updating the biomass composition and growth* *requirements* in the Supplementary Information and Supplementary Table 6. The knowledge index (number of abstracts in Medline) of the 1366 genes in the network was computed, and indicates that *i*JO1366 contains most of the best-characterized genes in *E. coli* (*Knowledge index of* i*JO1366 genes* in the Supplementary Information and Supplementary Table 7). *i*JO1366 was also compared with an automatically generated *E. coli* reconstruction from the Model SEED (Henry *et al*, 2010) (*Comparison of* i*JO1366 to the Model SEED E. coli reconstruction* in the Supplementary Information and Supplementary Table 8) and to the protein localization database EchoLocation (Horler *et al*, 2009) (*Comparison of* i*JO1366 to the EchoLocation database* in the Supplementary Information and Supplementary Table 9).

A significant number of the gaps in the *i*AF1260 network were filled during the update to *i*JO1366, and several blocked pathways were unblocked. Several different types of gaps in metabolic networks are possible. Root no-production gaps are metabolites with consuming reactions but no producing reactions. Root no-consumption gaps are metabolites with producing reactions but no consuming reactions. Downstream gaps are metabolites with producing and consuming reactions but which are unable to be produced at steady state because they are downstream of a root no-production gap. Similarly, upstream gaps are upstream of root no-consumption gaps. The final reconstruction contains 48 root no-production gaps, 63 root no-consumption gaps, 52 downstream gaps, and 69 upstream gaps. In total, 11.5% of the metabolites in *i*JO1366 are blocked under all conditions due to gaps (*Gaps and orphan reactions in the iJO1366 reconstruction* in the Supplementary Information and Supplementary Table 10). The orphan reactions in models such as *i*JO1366 can also help to identify the functions of metabolic genes. In the original *i*JR904 study (Reed *et al*, 2003), gene homology was used to predict the likely *E. coli* genes that encode the enzymes for 56 orphan reactions. Since then, 14 of these predictions have been independently confirmed to be correct (Supplementary Table 11), and these genes are now included in *i*JO1366.

Prediction of metabolic phenotypes

Flux balance analysis (FBA) (Orth *et al*, 2010b) can be used with a constraint-based model to predict metabolic flux distributions, growth rates, substrate uptake rates, and product secretion rates. The *i*AF1260 model and its predecessors were already very accurate at making phenotypic predictions such as growth rates and central metabolic flux distributions, so improved predictive capabilities in these areas were not expected with *i*JO1366. Instead, the value of the updated model is in its ability to predict phenotypes under a wider range of conditions than its predecessors.

To demonstrate the utility of the *i*JO1366 model in making these phenotypic predictions, we generated two large-scale sets of model phenotype predictions. First, the growth phenotypes of *E. coli* on all possible carbon, nitrogen, phosphorus, and sulfur sources were predicted. The numbers of growth-supporting substrates are summarized in Table II, and the full results of this screen given in Supplementary Table 12 and discussed in *Prediction of all growth-supporting carbon, nitrogen, phosphorus, and sulfur sources* in the Supplementary Information. We also performed a screen of model predicted growth phenotypes for all possible single gene knockout strains. Growth phenotypes were predicted on both glucose and glycerol minimal media, and the results were compared with experimental data sets (Table III; Supplementary Table 13). Not unexpectedly, *i*JO1366 is slightly less accurate at predicting overall gene essentiality than *i*AF1260. This difference is due to the fact that the 107 new genes added to this model version are from less well-studied systems and pathways than the existing genes in *i*AF1260, as discussed more thoroughly in *Prediction of gene essentiality* in the Supplementary Information.

Table II Growth supporting carbon, nitrogen, phosphorus, and sulfur sources

Source	*i*JO1366		*i*AF1260	
	Potential substrates	Growth supporting	Potential substrates	Growth supporting
Carbon	285	180	262	174
Nitrogen	178	94	163	78
Phosphorus	64	49	63	49
Sulfur	28	11	25	11

Table III Gene essentiality predictions on glucose and glycerol minimal media

	Experimental	
	Essential	Non-essential
	Growth on glucose	
Computational		
Essential	168 (12.3%)	39 (2.8%)
Non-essential	80 (5.9%)	1079 (79.0%)
	Growth on glycerol	
Computational		
Essential	161 (11.8%)	45 (3.3%)
Non-essential	87 (6.4%)	1073 (78.5%)

Mapping *i*JO1366 to closely related strains

Although *i*JO1366 is a model of *E. coli* K-12 MG1655, gene homology mapping can be used to create models of other *E. coli* and *Shigella* strains. To date, the metabolic reconstruction of *E. coli* W (Archer *et al*, 2011) is the only published reconstruction for *E. coli* strains other than K-12. The *i*JO1366 reconstruction should prove useful for the study of other recently sequenced *E. coli* strains. A previous analysis of multiple *E. coli* genomes showed that there is a moderate level of variability with respect to metabolic gene content within the species (Vieira *et al*, 2011). This analysis went one step beyond genetic conservation, also investigating the conservation of network topology, but it stopped short of computing the capacity to carry flux through specific growth-supporting pathways.

While it is known that equivalent function is not guaranteed by gene homology, it is still one of the most commonly used and effective methods of genome annotation (Frazer *et al*,

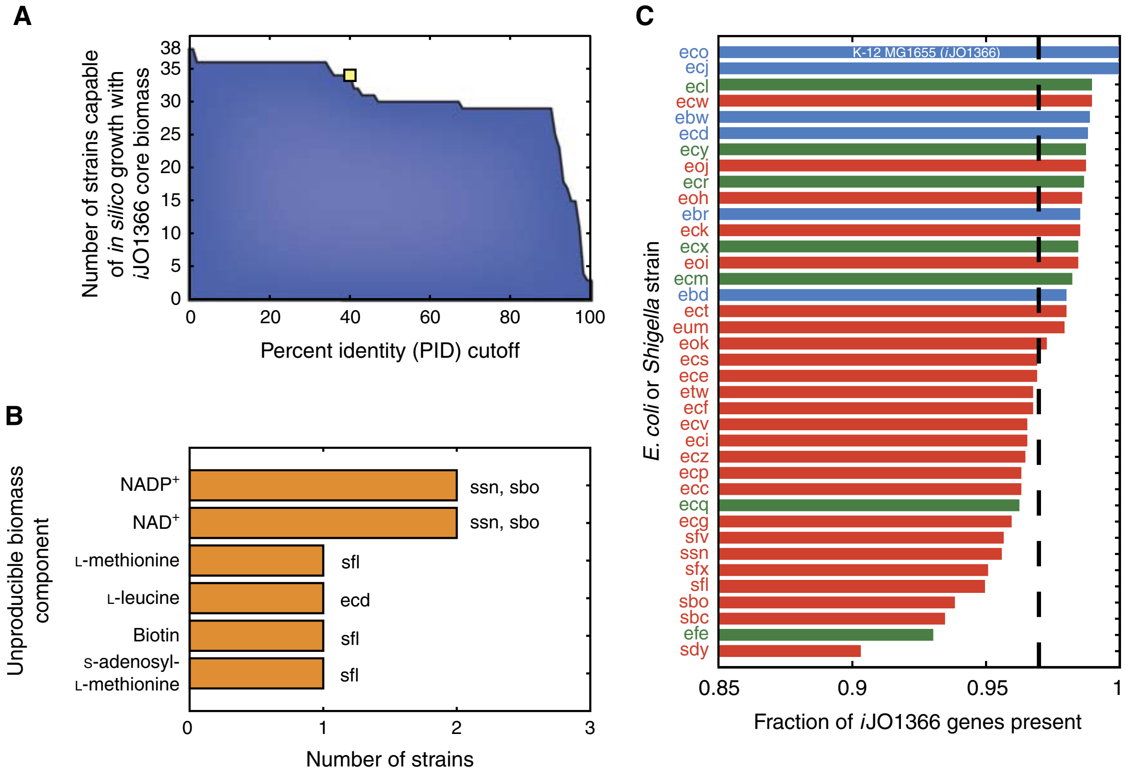

Figure 2 Results of the mapping between the *i*JO1366 reconstruction and all 38 available *E. coli* and *Shigella* strains. (**A**) The number of strain models capable of producing all components of the *i*JO1366 core biomass reaction at different PID cutoffs. The PID of 40% used in parts (**B**, **C**) of this figure is indicated with a yellow square. At a PID of 40%, only four strains are incapable of complete biomass synthesis. (**B**) Biomass components that cannot be produced in one or more models with a PID of 40%. The strains are indicated by their KEGG organism code. (**C**) The fraction of *i*JO1366 genes present in all 38 strains at a PID of 40%. Strains are listed by their KEGG organism code. Laboratory strains are colored blue, commensal and environmental strains are in green, and pathogens are in red. A dashed line indicates the average fraction of genes present (97%).

2003). In addition, the combination of sequence homology with constraint-based analysis of metabolic networks can be used to determine the most likely metabolic gene content of an organism by generating models that match known biology. Using this approach, we predicted with FBA whether metabolic models based on *i*JO1366 for 38 *E. coli* and *Shigella* strains are capable of producing biomass on glucose minimal media at various conservation thresholds (Figure 2A). At each percent identity (PID) cutoff, the set of genes that are absent from *i*JO1366 was determined using Smith–Waterman alignments. The set of missing genes was then used to impose constraints on metabolic reactions utilizing the *i*JO1366 GPRs. As expected, the results show that all strains were capable of producing biomass with no conservation requirement, but many strains lost this ability as the requirement was made more stringent. At a PID of 40%, only four strains were incapable of producing biomass. This PID was used for further analysis, and was justified through network-based analysis of auxotrophies (Figure 2B and *Mapping* i*JO1366 to closely related strains* in the Supplementary Information). Hundreds of unannotated genes in various *E. coli* and *Shigella* strains were identified through comparisons to *i*JO1366 genes, and are listed in Supplementary Table 14.

By analyzing the genetic content of 38 *E. coli* and *Shigella* genomes, 1006 metabolic genes in *i*JO1366 were found to be common to all genomes. The average genome in this analysis contains ∼97% of the genes in *i*JO1366 (Figure 2C). The mapping procedure described here led to the creation of 38 strain-specific models. It is important to note that these models are not yet on the same level as typical 'draft' metabolic models, as new metabolic functions (functions beyond those occurring in *E. coli* K-12 MG1655) have not yet been considered. Because only the 1366 metabolic genes of *i*JO1366 were considered for this mapping, the common set of genes (74%) appears larger than in the previous study by Vieira *et al*, which considered a set of 1545 reactions. This difference is also partly due to the fact that Vieira *et al* added orphan reactions to their *E. coli* networks and compared metabolic content on a reaction basis, rather than on a gene basis.

Conclusions

Although the metabolism of *E. coli* has been the subject of active research for decades, the new content added to the *i*JO1366 reconstruction is a result of the new discoveries that continue to be made. In fact, most of the newly characterized genes in *i*JO1366 were actually characterized in 2010 (Figure 1D). Although this is a small sample size, it appears that the pace of new discoveries is not slowing as more of the metabolic network is uncovered. There are still numerous uncharacterized *E. coli* K-12 MG1655 genes, many of which are predicted to have metabolic functions (Riley *et al*, 2006). As more discoveries are made in the future, additional updates to the *E. coli* network reconstruction will need to be made. Future increases in the scope of the *E. coli* metabolic network reconstruction will likely include the integration of this network with reconstructions of other cellular systems such as transcription and translation (Thiele *et al*, 2009) or transcriptional regulation (Covert *et al*, 2004). *i*JO1366 is the most advanced and comprehensive metabolic reconstruction of any microorganism to date, and can thus continue to serve as a basis for the metabolic reconstruction of other bacteria. Based on the success of its predecessors, we expect that *i*JO1366 will be an important tool in microbial systems biology for years to come.

Materials and methods

Experimental phenotypic screens

The *i*AF1260 *E. coli* K-12 MG1655 metabolic reconstruction was used to make computational predictions. The parent strain of the Keio Collection, BW25113, is derived from K-12 MG1655 and is missing several genes that are present in K-12 MG1655: *araBAD*, *rhaBAD*, and *lacZ*. Therefore, flux through the associated reactions without isozymes (ARAI, RBK_L1, RMPA, LYXI, RMI, RMK, and LACZ) was constrained by setting the upper and lower flux bounds of these reactions to zero. For aerobic growth, oxygen uptake was allowed by setting the lower bound of the oxygen exchange reaction to $-18.5\,\mathrm{mmol\,gDW^{-1}\,h^{-1}}$. Anaerobic growth was modeled by setting the lower bound of this reaction to zero. For growth on glucose, the lower bound of the glucose exchange reaction was set to $-8\,\mathrm{mmol\,gDW^{-1}\,h^{-1}}$. For growth on L-lactate and succinate, the lower bounds were set to $-16\,\mathrm{mmol\,gDW^{-1}\,h^{-1}}$. These are arbitrary bounds based on typical substrate uptake rates. After setting the bounds for each condition, the predicted effect of the deletion of each gene in *i*AF1260 for each condition was computed using the singleGeneDeletion COBRA Toolbox function (Becker *et al*, 2007), which uses GPRs to constrain the appropriate reactions to carry zero flux and then predicts maximum growth using FBA. A gene was considered essential for the simulated condition if deletion of the gene reduced optimal growth rate to $<5\%$ of the wild-type strain as computed by FBA. When a strict threshold of zero growth was used to determine essentiality instead, only one additional essential gene was predicted: *lldD* (b3605) on lactate minimal medium.

Knockout strains were taken from the Keio Collection (Baba *et al*, 2006; Yamamoto *et al*, 2009) (supplied by Open Biosystems), a genome-scale collection of *E. coli* K-12 gene knockouts. Stocks of knockout mutants were streaked onto LB agar with kanamycin ($50\,\mu\mathrm{g\,ml^{-1}}$) from odd-numbered Keio Collection plates only. Two single colonies of the Keio knockout strain were inoculated into 96-well plates containing 200 µl of LB media with kanamycin ($50\,\mu\mathrm{g\,ml^{-1}}$) and incubated overnight at 37 °C without shaking. Plates were then centrifuged and pelleted cells were washed twice with 200 µl of $1\times$ M9 salts per well. Disposable replicator pins were used to transfer cells from the preculture plate to four new plates, two containing glucose M9 minimal medium, one containing lactate M9 minimal medium, and one containing succinate M9 minimal medium. There was 200 µl of minimal medium per well, and the minimal media also contained $50\,\mu\mathrm{g\,ml^{-1}}$ kanamycin. The 96-well plates were covered with Aeraseal breathable film (Sigma) to minimize cross-well contamination. For aerobic conditions, the plates were incubated at 37 °C without shaking in a sterile cabinet. For anaerobic conditions, the plates were incubated at 37 °C in an anaerobic chamber ([O$_2$] <50 p.p.m.). After 48 h, absorbance at 600 nm of each well was determined using a VERSAmax microplate reader (Molecular Devices, Sunnyvale, CA).

A well was considered to have no growth or slow growth if its absorbance was less than a cutoff value specific to each of the four conditions. The cutoff value was determined by visual inspection of a histogram of absorbance values for all wells measured from a given condition. ‘Normal’ growers are supposed to result in measurements lying in the roughly Gaussian-shaped distribution that makes up most of the data, while slow- and non-growers are supposed to result in measurements falling outside the upper 0.95 area of the Gaussian distribution. The cutoffs were OD$_{600}$=0.26 for glucose aerobic plates, OD$_{600}$=0.21 for glucose anaerobic plates, OD$_{600}$=0.21 for lactate aerobic plates, and OD$_{600}$=0.20 for succinate aerobic plates. If both colonies of a gene knockout were determined to have normal growth, then the gene was considered a true positive (TP) if the model had predicted growth for the knockout, or a tentative false negative (TFN) if the model indicated the knockout should not have grown or previous experiments in glucose MOPS minimal medium had indicated the gene was essential (Baba *et al*, 2006). Similarly, if both colonies of a gene knockout were determined to have slow/no growth, then the gene was considered a true negative (TN) if the model predicted the same outcome, or a tentative false positive (TFP) otherwise. A gene was considered inconclusive (INC) if for any condition only one colony showed slow/no growth.

A second round of screening was performed for TFN, TFP, and INC gene knockouts. TFP and INC gene knockouts were rescreened with two colonies each; TFN gene knockouts were rescreened with four colonies and their pellets were washed four times before transfer to minimal media instead of twice to ensure that growth was not due to contamination from trace amounts of rich media from the precultures. For TFP and INC gene knockouts, a gene was considered essential for a condition if at least one colony showed slow/no growth in the condition in both the first and second round screens. If an essential gene was predicted by the model to be non-essential in the experimental condition, then the gene was concluded to be a genuine false positive (FP) for that condition. For TFN gene knockouts, a gene was considered a genuine false negative (FN) if two or more colonies demonstrated normal growth in the secondary screen.

Metabolic network reconstruction procedure

The *i*JO1366 reconstruction was assembled by updating the *i*AF1260b *E. coli* metabolic reconstruction (Feist *et al*, 2010), an updated version of the *i*AF1260 reconstruction (Feist *et al*, 2007). *i*AF1260b contains six additional reactions that were not in *i*AF1260 (ALAt2rpp, ASPt2rpp, CITt3pp, DHORDfum, GLYt2rpp, and MALt3pp). A 96-step procedure for metabolic network reconstruction was recently published (Thiele and Palsson, 2010), and the appropriate steps were followed when adding new genes, reactions, and metabolites to form *i*JO1366. The reconstruction was assembled using the SimPheny (Genomatica Inc., San Diego, CA) software platform. All new metabolites were checked against public databases (KEGG, PubChem) for correct structure and charge at a pH of 7.2. New reactions were mass and charge balanced and reversibility was assigned based on experimental studies, thermodynamic information, or the heuristic rules in the standard reconstruction protocol (Thiele and Palsson, 2010). Reactions were associated with genes and functional proteins to form GPRs. The *i*JO1366 model was exported from SimPheny as an SBML file and the COBRA Toolbox (Becker *et al*, 2007), a Matlab (The MathWorks Inc., Natick, MA) Toolbox, was used for additional model testing. The Tomlab (Tomlab Optimization Inc., Seattle, WA) CPLEX linear programming solver was used for all optimization procedures.

The GapFind MILP algorithm (Satish Kumar *et al*, 2007) was encoded in the COBRA Toolbox and used to identify all blocked metabolites in the *i*AF1260 and *i*JO1366 models. This algorithm was modified from the published version. Specifically, an option was included to change the mass balance constraint $\sum_j S_{ij}v_j \geqslant 0$ to $\sum_j S_{ij}v_j = 0$, allowing for metabolites without consuming reactions to be identified as gaps. For each GapFind run, the lower bounds of all exchange reactions were set to $-1000\,\mathrm{mmol\,gDW^{-1}\,h^{-1}}$ and the upper bounds of all model reactions were set to $10^9\,\mathrm{mmol\,gDW^{-1}\,h^{-1}}$. The GapFind algorithm was then run twice, once with each mass balance constraint option. Root no-production and no-consumption metabolites were identified from the model stoichiometric matrix (**S**) by searching for rows containing only negative or positive coefficients, respectively. Downstream no-production and upstream no-consumption gaps were identified by removing the root gaps from the GapFind outputs. The root gaps of each downstream gap were identified through targeted computational experiments in which metabolite source reactions were added to the network to restore connectivity. Orphan reactions were identified as all reactions without associated GPRs.

The core and wild-type biomass reactions were modified from the *i*AF1260 biomass reactions. Biotin was added with a coefficient based on a published biotin concentration of 250 molecules per cell (Delli-Bovi *et al*, 2010), while the related cofactor lipoate was added with a similar number of molecules per cell assumed. Iron–sulfur clusters were added with coefficients based on the predictions that 5% of all *E. coli* proteins contain these clusters (Fontecave, 2006), and that the majority (90%) of these clusters are of the [4Fe–4S] type. The coefficient of Fe^{2+} was decreased to account for Fe used in iron–sulfur clusters. Molybdenum cofactors were added with coefficients based on the measurement that the inorganic ion content of *E. coli* is 0.80% Mo (Cvetkovic *et al*, 2010), and the fact that the majority of this Mo is in the bis-molybdopterin guanine dinucleotide form (85%). The coefficients of Cu, Mn, Zn, Ni, and Co were also adjusted based on recent *in vivo* measurements (Cvetkovic *et al*, 2010). All other biomass components remain the same as in *i*AF1260. Growth-associated and non-growth-associated maintenances values were recalculated based on recent *E. coli* K-12 MG1655 chemostat data for growth on glucose minimal media (Taymaz-Nikerel *et al*, 2010). The slope and intercept of this experimental data were identified by linear regression. For these calculations, the electron transport system NADH dehydrogenase reactions NADH16pp (*nuo*) and NADH5 (*ndh*) were constrained to carry identical fluxes by replacing these reactions with an equivalent 'flux split' reaction, in order to constrain the model to a realistic P/O ratio of 1.375 (Noguchi *et al*, 2004). The NGAM of 3.15 mmol ATP $gDW^{-1} h^{-1}$ was identified by FBA as the maximum amount of ATP produced at a glucose uptake rate of 0.17 mmol $gDW^{-1} h^{-1}$, the intercept of the experimental data. The GAM of 53.95 mmol ATP gDW^{-1} was identified by FBA as the value that would give the correct experimentally determined slope of 10.83 mmol glucose $gDW^{-1} h^{-1}$/(μ) h^{-1} when using the core biomass reaction.

The *i*JO1366 model is available in SBML format at BioModels (accession: MODEL1108160000) and as Supplementary Model 1.

Comparison of *i*JO1366 to the Model SEED *E. coli* reconstruction

The *E. coli* K-12 MG1655 model Seed83333.1 V20.21 was downloaded from the Model SEED database (Henry *et al*, 2010) in SBML format. The set of 1139 genes in this model was compared by gene ID (b-number) to the 1366 genes in *i*JO1366 to identify common genes and the unique genes in each model. The genes in the Model SEED model that were not in *i*JO1366 were then investigated one at a time. EcoCyc was used to identify gene functions, and several genes with verified metabolic functions were then added to *i*JO1366.

Comparison of *i*JO1366 to the EchoLocation database

Predicted and experimentally determined protein location data were obtained for all *E. coli* K-12 genes from the EchoLocation database (Horler *et al*, 2009). The cellular locations of the proteins associated with the 1366 model genes from this database were then compared, one at a time, to the compartments of the metabolites in the reactions associated with each gene in *i*JO1366. For each location in the EchoLocation database, a Boolean rule was written to determine if the location is consistent with the associated model metabolites. For example, 'Cytoplasmic' proteins are consistent with genes only associated with cytoplasmic metabolites, and are inconsistent with genes associated with any periplasmic or extracellular metabolites. See Supplementary Table 9 for the full list of Boolean rules. The genes whose locations were inconsistent with model metabolites were then investigated individually, except for 'Cytoplasmic' and 'Periplasmic' genes with both cytoplasmic and periplasmic metabolites in the model, because there were too many of these inconsistencies to investigate manually. Literature and database information was used to determine whether the EchoLocation database, the *i*JO1366 reconstruction, or both were correct.

Constraint-based modeling

The *i*JO1366 model, constructed in SimPheny, was exported as an SBML file and used to perform simulations and constraint-based analyses using the COBRA Toolbox and Tomlab CPLEX linear programming solver. The constraint-based model consists of an **S** matrix with 1805 rows and 2583 columns, where 1805 is the number of distinct metabolites (in all three compartments) and 2583 is the number of reactions including exchange and biomass reactions. Each of the reactions has an upper and lower bound on the flux it can carry. Reversible reactions have an upper bound of 1000 mmol $gDW^{-1} h^{-1}$ and a lower bound of -1000 mmol $gDW^{-1} h^{-1}$, making them practically unconstrained, while irreversible reactions have a lower bound of zero.

By default, the core biomass reaction is set as the objective to be maximized. Certain reactions are by default constrained to carry zero flux to avoid unrealistic behaviors. These reactions are CAT, DHPTDNR, DHPTDNRN, FHL (formate hydrogen lyase), SPODM, SPODMpp, SUCASPtpp, SUCFUMtpp, SUCMALtpp, and SUCTARTtpp. CAT, SPODM, and SPODMpp are hydrogen peroxide producing and consuming reactions that can carry flux in unrealistic energy generating loops. DHPTDNR and DHPTDNRN form a closed loop that can carry an arbitrarily high flux. The succinate antiporters SUCASPtpp, SUCFUMtpp, SUCMALtpp, and SUCTARTtpp can form unrealistic flux loops with other transporters for aspartate, fumarate, malate, and tartrate. The genes encoding FHL are known to be active under anaerobic conditions, but this reaction is constrained to zero to avoid unrealistic aerobic hydrogen production. The NGAM constraint is imposed by a lower bound of 3.15 mmol $gDW^{-1} h^{-1}$ on the reaction ATPM. The exchange reactions that allow for extracellular metabolites to pass in and out of the system are defined such that a positive flux indicates flow out. All exchange reactions have a lower bound of zero except for glucose (-10 mmol $gDW^{-1} h^{-1}$), the vitamin B_{12} precursor cob(I)alamin (-0.01 mmol $gDW^{-1} h^{-1}$), and oxygen and all inorganic ions required by the biomass reaction (-1000 mmol $gDW^{-1} h^{-1}$). The default lower bound on glucose uptake is based on typical glucose uptake rates. Because only a very small amount of B_{12} is required for growth, the lower bound on cob(I)alamin uptake is arbitrary and never actually constraining in practice. The *i*JO1366 computational model also includes drain reactions for six cytoplasmic metabolites without known consuming reactions that must be drained from the system to allow simulation of steady-state cell growth. These metabolites are *p*-cresol, 5′-deoxyribose, aminoacetaldehyde, S-adenosyl-4-methylthio-2-oxobutanoate, (2R,4S)-2-methyl-2,3,3,4-tetrahydroxytetrahydrofuran, and oxamate.

Prediction of all growth-supporting carbon, nitrogen, phosphorus, and sulfur sources

The possible growth-supporting carbon, nitrogen, phosphorus, and sulfur sources of *E. coli* were identified using FBA. First, all exchange reactions for extracellular metabolites containing the four elements were identified from the metabolite formulas. Every extracellular compound containing carbon was considered a potential carbon source, for example. Next, to determine possible growth-supporting carbon sources, the lower bound of the glucose exchange reaction was constrained to zero. Then the lower bound of each carbon exchange reaction was set, one at a time, to -10 mmol $gDW^{-1} h^{-1}$ (a typical uptake rate for growth-supporting substrates), and growth was maximized by FBA using the core biomass reaction. The target substrate was considered growth supporting if the predicted growth rate was above zero. While identifying carbon sources, the default nitrogen, phosphorus, and sulfur sources were ammonium (nh4), inorganic phosphate (pi), and inorganic sulfate (so4), respectively. Prediction of growth-supporting sources of these other three elements was performed in the same manner as growth on carbon, with glucose as the default carbon source.

Prediction of gene essentiality

To simulate the effects of gene knockouts, the *i*JO1366 model with its default constraints and core biomass reaction objective was modified to match the genotype of *E. coli* BW25113 (see Experimental phenotypic screens). For growth on glucose, the lower bound of the glucose exchange reaction was set to $-10\,\mathrm{mmol\,gDW^{-1}\,h^{-1}}$. For growth on glycerol, the lower bound of the glucose exchange reaction was set to zero while the lower bound of the glycerol exchange reaction was set to $-10\,\mathrm{mmol\,gDW^{-1}\,h^{-1}}$. All 1366 genes in the model were knocked out one a time and growth was simulated by FBA using the singleGeneDeletion COBRA Toolbox function. Gene knockout strains with a growth rate above zero were considered non-essential. Experimental gene essentiality data for growth on glucose (Baba *et al*, 2006) and glycerol (Joyce *et al*, 2006) was then obtained and adjusted based on an updated analysis of the Keio Collection strains (Yamamoto *et al*, 2009). The newly identified essential genes were added to the lists of essential genes under both conditions, while the genes whose essentiality was identified as uncertain were not changed from their original designations.

Mapping *i*JO1366 to closely related strains

The protein sequences of all available *E. coli* and *Shigella* strains (38 strains total) were downloaded from KEGG (http://www.genome.jp/kegg/download), and SSEARCH35 of the FASTA suite (Pearson and Lipman, 1988), an open-source implementation of the Smith–Waterman algorithm, was used to determine a PID conservation for each *i*JO1366 gene. The flags used in SSEARCH35 were '–m9 –E 1 –q –H'. Next, the deleteModelGenes COBRA Toolbox function was used to delete all genes in a strain that failed to be conserved at a particular protein sequence identity threshold. The entire cutoff range was scanned (from 0 to 100% identity in increments of 1%) and FBA was performed with an objective function of maximum growth rate using the core biomass reaction, to produce Figure 2A. Subsequently, the situation occurring at 40% identity was investigated to determine why four of the strains were unable to produce biomass at this conservation threshold. A demand reaction for each biomass component was added (excluding the components involved in the ATP hydrolysis reaction) and flux was separately maximized through each of these reactions to determine the biomass components that the strain was unable to produce. To determine the subset of genes responsible for a particular loss of function, the effects of single knockouts were computed and literature was consulted.

The preliminary network-level analysis of conservation revealed that many ORFs are not properly automatically annotated in non-model organism genomes. For this reason, this analysis was supplemented with a nucleotide-level search for conservation of each model gene in each of the 38 strains. The analysis described above was repeated, this time with the additional conservation qualification of a genomic hit displaying at least 70% identity and aligned length to the model gene. These cutoffs were determined by manual inspection of a scatter plot of all nucleotide-level results. There was a clear region of conservation in the upper right quadrant defined by these cutoffs. Genes found to be conserved only at the nucleotide level are considered to have 100% protein sequence identity (see Supplementary Table 14 for a complete list of such cases).

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Harish Nagarajan, Vasiliy Portnoy, Jennie Reed, and Ines Thiele for their helpful suggestions. This work was funded by NIH R01 GM057089.

Author contributions: JDO and BØP designed the study. JDO, JAL, and AMF performed the model updates. TMC and JN performed the phenotypic screening experiments. JAL performed the mapping to other strains. HN analyzed the knowledge index of model genes. All authors helped draft and edit the final manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Archer CT, Kim JF, Jeong H, Park JH, Vickers CE, Lee SY, Nielsen LK (2011) The genome sequence of *E. coli* W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of *E coli*. *BMC Genomics* **12:** 9

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

Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgård MJ (2007) Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat Protocols* **2:** 727–738

Covert MW, Knight EM, Reed JL, Herrgård MJ, Palsson BØ (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429:** 92–96

Cvetkovic A, Menon AL, Thorgersen MP, Scott JW, Poole II FL, Jenney Jr FE, Lancaster WA, Praissman JL, Shanmukh S, Vaccaro BJ, Trauger SA, Kalisiak E, Apon JV, Siuzdak G, Yannone SM, Tainer JA, Adams MW (2010) Microbial metalloproteomes are largely uncharacterized. *Nature* **466:** 779–782

Delli-Bovi TA, Spalding MD, Prigge ST (2010) Overexpression of biotin synthase and biotin ligase is required for efficient generation of sulfur-35 labeled biotin in *E. coli*. *BMC Biotechnol* **10:** 73

Edwards JS, Palsson BØ (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci USA* **97:** 5528–5533

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* **3:** 121

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

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

Feist AM, Zielinski DC, Orth JD, Schellenberger J, Herrgård MJ, Palsson BØ (2010) Model-driven evaluation of the production potential for growth-coupled products of *Escherichia coli*. *Metab Eng* **12:** 173–186

Fontecave M (2006) Iron-sulfur clusters: ever-expanding roles. *Nat Chem Biol* **2:** 171–174

Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC (2003) Cross-species sequence comparisons: a review of methods and available resources. *Genome Res* **13:** 1–12

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:** 977–982

Horler RS, Butcher A, Papangelopoulos N, Ashton PD, Thomas GH (2009) EchoLOCATION: an *in silico* analysis of the subcellular locations of *Escherichia coli* proteins and comparison with experimentally derived locations. *Bioinformatics (Oxford, England)* **25:** 163–166

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

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38:** D355–D360

Keseler IM, Bonavides-Martinez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, Johnson DA, Krummenacker M, Nolan LM, Paley S, Paulsen IT, Peralta-Gil M, Santos-Zavaleta A, Shearer AG, Karp PD (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res* **37:** D464–D470

Noguchi Y, Nakai Y, Shimba N, Toyosaki H, Kawahara Y, Sugimoto S, Suzuki E (2004) The energetic conversion competence of *Escherichia coli* during aerobic respiration studied by 31P NMR using a circulating fermentation system. *J Biochem (Tokyo)* **136:** 509–515

Orth JD, Fleming RM, Palsson BØ (2010a) 10.2.1 – Reconstruction and use of microbial metabolic networks: the core *Escherichia coli* metabolic model as an educational guide. In *EcoSal – Escherichia coli and Salmonella Cellular and Molecular Biology*, Karp PD (ed), 10.2.1. Washington DC: ASM Press

Orth JD, Palsson BØ (2010) Systematizing the generation of missing metabolic knowledge. *Biotechnol Bioeng* **107:** 403–412

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

Palsson BØ (2009) Metabolic systems biology. *FEBS Lett* **583:** 3900–3904

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

Reed JL, Vo TD, Schilling CH, Palsson BØ (2003) An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). *Genome Biol* **4:** R54.51–R54.12

Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, Kosuge T, Mori H, Perna NT, Plunkett III G, Rudd KE, Serres MH, Thomas GH, Thomson NR, Wishart D, Wanner BL (2006) *Escherichia coli* K-12: a cooperatively developed annotation snapshot-2005. *Nucleic Acids Res* **34:** 1–9

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

Taymaz-Nikerel H, Borujeni AE, Verheijen PJ, Heijnen JJ, van Gulik WM (2010) Genome-derived minimal metabolic models for *Escherichia coli* MG1655 with estimated *in vivo* respiratory ATP stoichiometry. *Biotechnol Bioeng* **107:** 369–381

Thiele I, Jamshidi N, Fleming RMT, Palsson BØ (2009) Genome-scale reconstruction of *Escherichia coli's* transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput Biol* **5:** e1000312

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

Vieira G, Sabarly V, Bourguignon PY, Durot M, Le Fevre F, Mornico D, Vallenet D, Bouvet O, Denamur E, Schachter V, Medigue C (2011) Core and panmetabolism in *Escherichia coli*. *J Bacteriol* **193:** 1461–1472

Yamamoto N, Nakahigashi K, Nakamichi T, Yoshino M, Takai Y, Touda Y, Furubayashi A, Kinjyo S, Dose H, Hasegawa M, Datsenko KA, Nakayashiki T, Tomita M, Wanner BL, Mori H (2009) Update on the Keio collection of *Escherichia coli* single-gene deletion mutants. *Mol Syst Biol* **5:** 335