Supplementary Information

Table of Contents

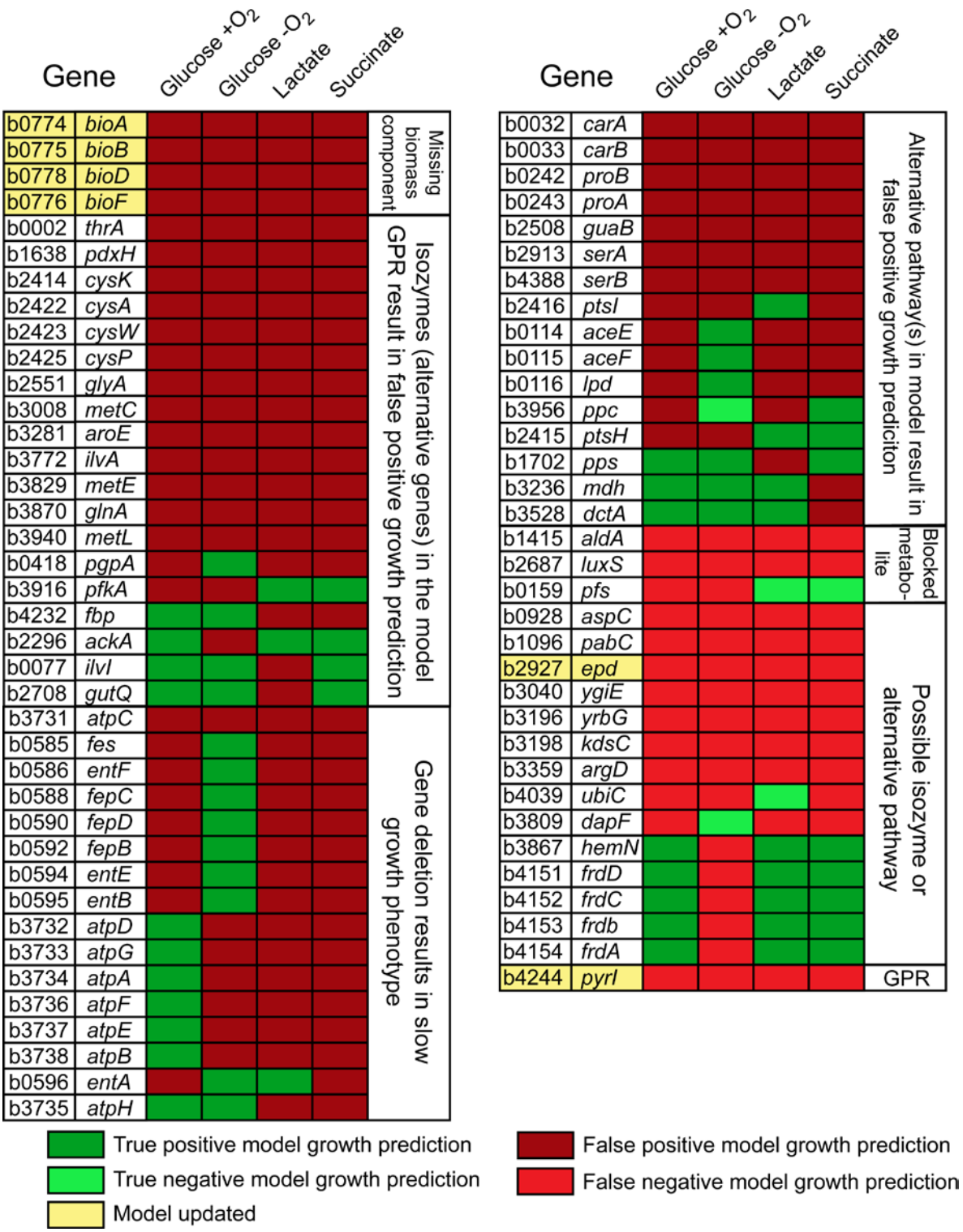
Experimental phenotypic screens

Comparisons of gene essentiality data to metabolic model predictions have been successfully used for model refinement previously (Joyce et al, 2006; Molina-Henares et al, 2010). In order to make additional comparisons to improve the *E. coli* reconstruction, we experimentally determined the conditional essentiality for a subset of *i*AF1260 genes, and compared these results to model predictions. Of 4325 ORFs in the primary *E. coli* K-12 MG1655 genome annotation (U00096.2), 3985 have corresponding knockout strains in the Keio Collection of K-12 single gene knockouts (Baba et al, 2006). Of these, only 1118 genes are present in both the Keio Collection and the *i*AF1260 *E. coli* metabolic reconstruction. Errors such as duplications of the deleted gene were recently discovered in the Keio Collection in 43 of the 1118 mutants (Yamamoto *et al*, 2009), further restricting this study to 1075 mutants. High-throughput determination of the growth phenotypes of these 1075 knockout strains was obtained using growth in liquid culture in 96-well plates. The use of defined media for the growth condition allowed precise modeling of the environment using the *i*AF1260 metabolic model. The growth phenotype (i.e., growth or no growth) of each mutant was studied in four conditions: glucose aerobic M9 minimal medium, glucose anaerobic M9 minimal medium, lactate aerobic M9 minimal medium, and succinate aerobic M9 minimal medium. Complete results of these screens can be found in Supplementary Table 1.

The experimental measurements of "growth" or "no growth" agreed with *i*AF1260 predictions in 93.5% of cases. The results can be grouped into four categories: true positives (growth is both predicted and observed), true negatives (growth is neither predicted nor observed), false positives (growth was predicted but not observed), and false negatives (no growth was predicted but growth was nonetheless positively observed). Analysis of the results focused on the false positive and false negative growth predictions, since it is these failure modes that indicate missing knowledge and potential model refinements (Supplementary Figure 1). By comparing model predicted growth phenotypes to the measurements, errors in the reconstruction were found. We investigated these disagreements, and it was found that there are several possible explanations for both false negative and false positive predictions. Most false positives were found to be due to the presence of isozymes or alternative pathways in the model. Under the actual experimental conditions, these alternative genes are likely not expressed. There may also be latent pathways, which are known to exist but may require significant regulatory adjustments to be activated. False negative model predictions are most likely due to genes and pathways missing from the model, and thus represent knowledge gaps.

A total of 66 testable model predictions (1.5%) were in the FNs category. There are several reasons why FNs were observed. First, the model may miscalculate essentiality because deletion of a gene causes violation of the FBA steady state assumption (Orth et al, 2010). The *aldA* (b1415), *luxS* (b2687), or *pfs* (b0159) genes are not necessary for producing biomass constituents, but their deletion resulted in FN predictions due to the model steady-state assumption. In the model these genes are needed for the recycling or degradation of metabolic intermediates, and are thus essential. *In vivo*, it may be possible that the concentration of these intermediates can build up without inhibiting growth or that they are degraded or diluted by other mechanisms. The second cause of FNs is that a gene may not carry out the reaction specified by the model's GPRs. The *pyrI* (b4244) gene encodes the nonessential regulatory subunit of the aspartate carbamoyltransferase enzyme, but was incorrectly assigned in *i*AF1260 as being essential for catalytic activity in the GPR. This gene has been changed to a nonessential component of the aspartate carbamoyltransferase GPR in *i*JO1366. The third cause of FNs is that an isozyme or alternative pathway not present in the model may carry out the function of a deleted gene *in vivo*. For the case of *epd* (b2927), evidence for an isozyme, *gapA* (b1779), was found in the literature (Yang et al, 1998), and the GPR has been corrected in *i*JO1366. Double genetic perturbation experiments for other FN genes are expected to uncover novel isozymes and alternative pathways (Butland et al, 2008; Nakahigashi et al, 2009; Typas et al, 2008). The fourth cause of FNs is contamination of cultures. It was confirmed by PCR that FN results of four genes, *iscS* (b2530), *purA* (b4177), *purK* (b0522), and *thyA* (b2827), were due to cross-contamination in the Keio Collection strain isolates used (although no evidence of contamination was found in the original Keio Collection (Yamamoto et al, 2009)). Therefore these genes were not considered in this analysis.

We also identified 213 FP results. As with FN results, there are several reasons why FPs occurred. First, mutants may be able to grow as predicted, but grow slowly or with a poor yield and are inappropriately experimentally designated essential. The observation of FPs for ATP synthase and iron transport genes is likely due to slow growth in these mutants (Joyce *et al*, 2006). These slow growth phenotypes may be overcome through adaptive laboratory evolution (Charusanti et al, 2010; Fong & Palsson, 2004; Ibarra et al, 2002). The second cause of FPs is that isozymes and alternative pathways exist in the model but for some reason may be unable to carry sufficient flux for normal growth (i.e., the alternative genes are not expressed under the growth condition, encode inefficient enzymes, or are wrongly assigned to the GPR). The third possible cause of FPs is that genes may be needed to produce essential biomass components which are not included in the model's core biomass reaction. Because a complete biosynthetic pathway for biotin had not yet been identified in *E. coli*, biotin was not included in the biomass reaction of *i*AF1260, resulting in FP growth predictions for biotin biosynthetic genes that were included in *i*AF1260. This issue has been resolved in *i*JO1366 through addition of the complete biotin synthesis pathway.



Supplementary Figure 1. False positive and false negative model predictions from the experimental gene essentiality screen. Results are categorized by the suspected reason that the model failed to accurately predict the phenotype. The six genes highlighted in yellow have been updated in *i*JO1366 partly on the basis of these results. The 13 false negatives that were due to media contamination were omitted from this figure, as were 41 false positives (associated with 32 genes) that did not have clear explanations.

Updating the biomass composition and growth requirements

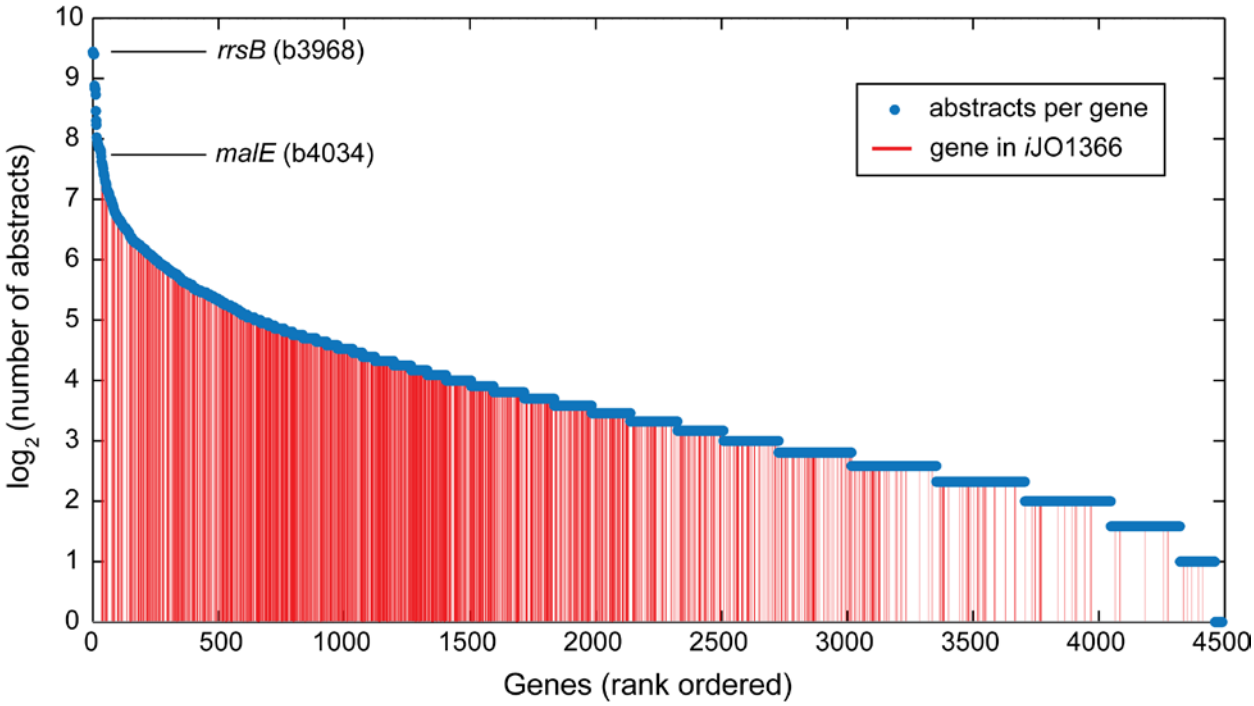
The "core" and "wild-type" biomass reactions of *i*AF1260 have been updated in *i*JO1366. These are reactions that drain biomass precursor compounds in experimentally determined ratios to simulate growth (Feist & Palsson, 2010; Varma & Palsson, 1993). Each component of a biomass reaction has the units mmol/gDW (milli-moles per gram cell dry weight), and flux through a biomass reaction has the units h^{-1}, and is equivalent to the exponential growth rate of the organism (Thiele & Palsson, 2010). The "wild-type" biomass reaction contains the precursors to all the typical wild-type cellular components of *E. coli*, while the "core" biomass reaction contains the precursors only to essential components. Now that the complete biotin synthesis pathway has been added to the reconstruction, biotin has been added to the biomass reactions along with the related cofactor lipoate. The [2Fe-2S] and [4Fe-4S] iron sulfur clusters have also been added, along with the molybdenum cofactors. Based on a recent study in which the metal content of *E. coli* was measured (Cvetkovic et al, 2010), the compositions of Cu, Mn, Zn, Ni, Mo, and Co in the biomass reactions have been adjusted.

Growth-associated maintenance (GAM) and non-growth-associated maintenance (NGAM) are the amounts of ATP consumed during cell growth and by non-growth associated processes such as maintenance of membrane gradients, respectively. GAM is a component of the biomass reaction, while NGAM is manifest as a lower bound on the separate ATP draining reaction "ATPM." These two parameters were recalculated for *i*JO1366 based on a new experimental dataset for *E. coli* K-12 MG1655 growing in a glucose minimal media chemostat (Taymaz-Nikerel *et al*, 2010). This dataset accounts for cell lysis when determining growth rate, and thus includes a slightly higher growth rate and lower apparent maintenance requirements than in the previously used datasets (Feist et al, 2007). For GAM and NGAM determination, the P/O ratio of the model was constrained to 1.375, a physiologically realistic ratio (Noguchi et al, 2004), by enforcing a flux split through the two primary NADH dehydrogenases. GAM was determined to be 53.95 mmol ATP gDW^{-1}, while NGAM was determined to be 3.15 mmol ATP gDW^{-1} h^{-1}. It should be noted that the GAM and NGAM in a strain specific biomass reaction can vary given the experimental data set from which they were calculated. As such, these values should be based on the experimental data that most closely matches the field of use for a modeling application. For the complete core and wild-type biomass reactions see Supplementary Table 6.

Knowledge index of *i*JO1366 genes

Knowledge about individual *E. coli* genes is being accumulated at a rapid pace (Figure 1D). One measure of the accumulated knowledge of an organism is its species-knowledge-index (SKI) value (Janssen et al, 2005). The SKI value is calculated as the total number of abstracts per species in Medline divided by the total number of genes in the genome, and it was shown that *E. coli* has a relatively large SKI value (over 55.1) compared to other model organisms (Reed et al, 2006a).

Here, the number of abstracts per gene of *E. coli* K-12 MG1655 in Medline was calculated to confirm that accumulated knowledge is well represented in the updated model. *i*JO1366 genes are enriched in number of abstracts (Supplementary Figure 2). The best studied *E. coli* gene is the ribosomal RNA gene *rrsB* (b3968), with 697 abstracts. The best studied gene included in *i*JO1366 is the maltose ABC transporter *malE* (b4034), with 209 abstracts. Each gene in *i*JO1366 has an average of 24.5 abstracts, while each gene in the total set of 4490 *E. coli* genes has 20.4 abstracts on average, a significant difference ($p = 2.9\times10^{-6}$, t-test). The complete list of the number of abstracts per gene can be found in Supplementary Table 7. The high knowledge index values for these genes indicate that the accumulated knowledge of *E. coli* is well represented in *i*JO1366.



Supplementary Figure 2. Accumulated knowledge of each gene in *E. coli* K-12 MG1655 as determined by the number of abstracts in Medline containing information about each gene. The number of abstracts for each gene is plotted, and the genes are ordered by their total number of abstracts. All genes in *i*JO1366 have a red line, while other *E. coli* genes have no line. *rrsB* is the most studied *E. coli* gene, while *malE* is the most studied gene included in *i*JO1366. The genes in *i*JO1366 are among the best studied in *E. coli*.

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

Automated tools and methods for the assembly of metabolic network reconstructions are beginning to appear, and one of the most comprehensive new tools is the Model SEED (Henry *et al*, 2010). It is based on a strong annotation tool, RAST (Aziz *et al*, 2008). This framework combines the subsystem-based SEED genome annotations with gap-filling methods to create fully functional constraint-based metabolic models. In order to assess the completeness of *i*JO1366, it was compared to the Seed83333.1 V20.21 model of *E. coli* K-12 MG1655, a model that contains 1139 genes. Specifically, the set of genes contained in *i*JO1366 was compared to the genes in the SEED model. It was found that the SEED model contains 133 genes not contained in *i*JO1366, and that *i*JO1366 contains 362 not contained in the SEED model (Supplementary Table 8). The genes unique to the SEED model were investigated one at a time to determine if they have known metabolic functions and should be included in *i*JO1366. At the time of the initial comparison between these two models, four genes with characterized metabolic functions were identified in the SEED model and added to *i*JO1366: *btuE* (b1710), *yggF* (b2930), *nudF* (b3034), and *yieG* (b3714). These genes had not been found in previous model update procedures. Due to the manual literature searches performed and the large scope of the model, it is always possible that some known genes are missed, illustrating the value of quality automated tools such as Model SEED. Of the remaining 133 genes not included in the final *i*JO1366 model, 68 were determined to have non-metabolic functions. The other 65 genes in the SEED model currently have partially or completely uncharacterized functions, and their predicted functions provide hypotheses that could lead to new metabolic discoveries and could help to fill gaps in *i*JO1366.

Comparison of *i*JO1366 to the EchoLocation database

The *i*JO1366 model contains metabolites in three cellular compartments: the cytoplasm, the periplasm and the extracellular space. The set of metabolites that participate in a reaction can indicate the location of the protein that catalyzes the reaction. For example, a reaction that includes only cytoplasmic metabolites must be catalyzed by a protein in the cytoplasm or attached to the inner membrane. A periplasmic or outer membrane protein cannot catalyze this reaction. To verify the accuracy of the compartment assignments of the reactions in *i*JO1366, a comparison was made to the EchoLocation database (Horler et al, 2009). This database contains experimentally verified and computationally predicted subcellular locations for all *E. coli* K-12 proteins, sorted into 12 locations such as "cytoplasmic", "inner membrane", and "integral membrane protein." The protein locations in this database were compared to the compartments of the metabolites associated with each gene in *i*JO1366 using a set of Boolean rules. These rules are listed in Supplementary Table 9.

After testing all 1366 model genes, 170 were found to be inconsistent (Supplementary Table 9). The most common type of inconsistency was "cytoplasmic" or "periplasmic" proteins in EchoLocation that were associated with both cytoplasmic and periplasmic metabolites in *i*JO1366. There were 132 such inconsistencies. Most of these were transport reactions in *i*JO1366 with proteins that may be cytoplasmic or periplasmic subunits of multi-subunit complexes. The remaining 38 inconsistencies were investigated one at a time to determine whether *i*JO1366, EchoLocation, or both are correct. Through targeted literature searches, experimental evidence was found indicating that 12 of the locations are correct in *i*JO1366 but incorrect in EchoLocation. The remaining 26 proteins were found to be correct in both EchoLocation and *i*JO1366, and all involve multi-subunit transporters with individual proteins spanning multiple locations. After manually reconciling *i*JO1366 and EchoLocation, most genes were consistent. Interestingly, the locations that are based on experimental evidence in EchoLocation are more likely to be inconsistent with *i*JO1366 than computationally predicted locations. As most "periplasmic" proteins in EchoLocation are actually associated with cytoplasm to periplasm transport reactions in *i*JO1366, these discrepancies may simply be due to the definition of a "periplasmic" protein in EchoLocation. Still, the overall content of *i*JO1366 is consistent with EchoLocation, indicating that the compartments of the metabolites in most reactions are correct.

Gaps and orphan reactions in the *i*JO1366 reconstruction

The modified GapFind algorithm was used to identify all gaps in the final version of the *i*JO1366 reconstruction (Supplementary Table 10). 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 *i*JO1366 reconstruction contains 48 root no-production gaps, 63 root no-consumption gaps, 52 downstream gaps, and 69 upstream gaps (Supplementary Figure 3). The total number of blocked metabolites in *i*JO1366 is 208, with some metabolites occurring as more than one type of gap. In total, 11.5% of the metabolites in *i*JO1366 are blocked under all conditions due to gaps.

All 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. More than half of the total blocked metabolites in *i*JO1366 are due to scope gaps. The two main classes of scope gaps in the model are tRNA related and metal ion related. Like its

predecessor, *i*AF1260, *i*JO1366 contains charging reactions for all 20 standard amino acids as well as several non-standard amino acids such as N-formylmethionine and L-selenocysteine. These reactions are blocked because *i*JO1366 does not contain producing reactions for the uncharged tRNAs or consuming reactions for the charged tRNAs. These reactions could be used if the metabolic network is connected to a transcription and translation network (Thiele *et al*, 2009), and thus, they are included for ease of integration and completeness of the reconstruction. *i*JO1366 contains many reactions involving metal ions. Some metal ions, such as Fe^{2+}, Mg^{2+}, Ca^{2+}, and Na^{+} are included in the core and wild-type biomass reactions, providing a consuming reaction for these metabolites. Others, such as Ag^{+}, Hg^{2+}, and WO_4^{2-}, may be toxic to cells or may not serve any essential biological purpose. *E. coli* contains efflux transporters for such metals, but their exact uptake mechanisms are not known. Other scope gaps are due to metabolites that, like tRNAs, serve non-metabolic functions once they are produced. For example, *E. coli* resists osmotic stress by producing glycine betaine (Falkenberg & Strom, 1990).

Most root gaps have only one or no downstream or upstream gaps (Supplementary Figure 3B). This indicates that few long pathways in *i*JO1366 are blocked, and that most gaps have very small effects on the network as a whole. There are a few pathways blocked by gaps, however. For example, a set of nine metabolites including carnitine and carnitinyl-CoA are blocked by their downstream product γ-butyrobetainyl-CoA. This compound is not well characterized, but has been shown to be converted to crotobetainyl-CoA by *caiA* (b0039), although the mechanism and electron acceptor for this reaction are unknown (Molina-Henares et al, 2010).

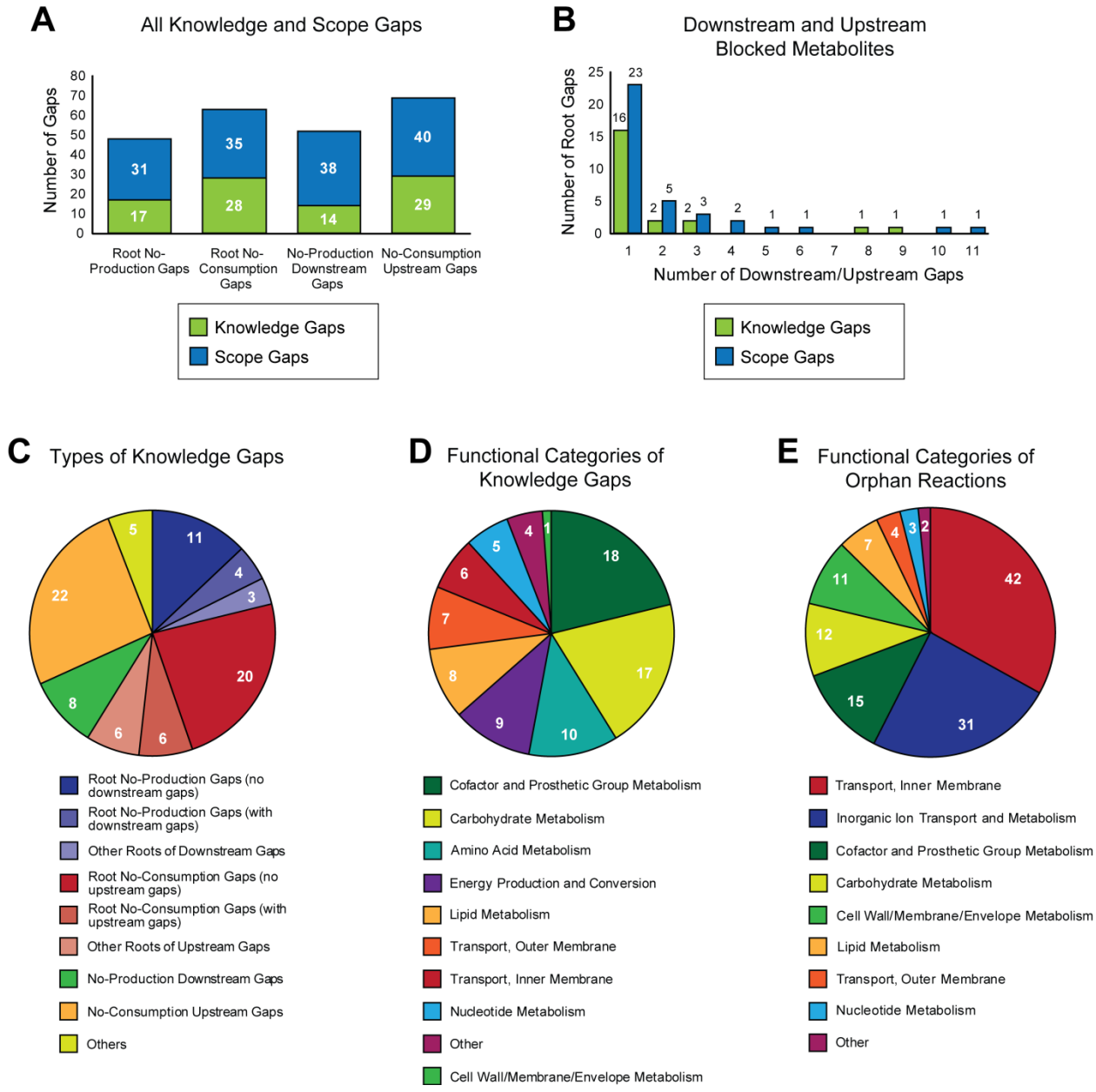
A biologically realistic gap in *E. coli* K-12 metabolism occurs in the O-antigen synthesis pathway. An IS5 insertion in the *rfb* operon has left these *E. coli* strains without a functional rhamnosyltransferase, leaving rhamnosyl-N-acetylglucosamyl-undecaprenyl diphosphate without a producing reaction (Liu & Reeves, 1994; Rubires et al, 1997). Eleven downstream metabolites are also blocked by this gap, which is listed here as a scope gap. This is a real gap in the *E. coli* metabolic network, and is not due to limited knowledge.

Most downstream or upstream blocked metabolites are blocked by only one root gap (Supplementary Figure 3C). If the missing producing or consuming reaction is identified, the downstream or upstream metabolite would be unblocked. There are a few cases, however, in which a metabolite has both upstream and downstream gaps. Mercaptopyruvate, for example, has no known producing reaction in *E. coli*. This compound is consumed in a reaction catalyzed by *sseA* (b2521) that produces thiocyanate (Colnaghi et al, 2001). This product has no known consuming reactions (Adams et al, 2002). In this unusual situation, neither of these compounds can be produced until both of these gaps are filled. Knowledge gaps occur throughout the *i*JO1366 metabolic network, with the largest number occurring in cofactor and prosthetic group metabolism (Supplementary Figure 3D).

In addition to gaps, *i*JO1366 contains 128 orphan reactions. One of these, the ATP maintenance reaction, is not a real biological reaction, and is used for modeling purposes to simulate the non-growth associated maintenance requirement of *E. coli*. The other orphan reactions are due to incomplete knowledge of *E. coli* metabolism. Orphan reactions occur in all of the metabolic systems of *i*JO1366 except for 'energy production and conversion' and 'amino acid metabolism' (Supplementary Figure 3E). Most orphan reactions are inner membrane and inorganic ion transport reactions. One possible reason for this is that transport proteins tend to be more difficult to purify and assay than soluble enzymes.

Another notable feature of orphan reactions is that they are often adjacent to each other in the *i*JO1366 network. Two reactions are considered adjacent if they have a common metabolite. Orphan reactions are adjacent to an average of 3.05 other orphans, while the average for all reactions in *i*JO1366 (excluding biomass, demand, and exchange reactions) is 1.53 adjacent orphans, a significant difference (*p*

= 0.0005, t-test). This characteristic indicates that orphans are more common in certain poorly studied pathways and subsystems than in well-studied pathways.



Supplementary Figure 3. Properties of the gaps and orphan reactions in *i*JO1366. (**A**) Numbers of root no-production, root no-consumption, no-production downstream, and no-consumption upstream gaps in the network. (**B**) Histogram of the number of downstream or upstream blocked metabolites for each root gap. Most root gaps only result in one downstream gap. (**C**) The 85 knowledge gaps (no scope gaps) in *i*JO1366 by type of gap. "Others" includes special cases such as metabolites that are both root and downstream gaps. (**D**) The 85 knowledge gaps by the primary metabolic functional category (see Figure 1) of the reactions in which the blocked metabolites participate. (**E**) The 127 orphan reactions (excluding the artificial reaction ATPM) by functional category.

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

The *i*JO1366 computational model contains exchange reactions for 324 different compounds. 285 of these compounds contain at least one carbon atom, 178 contain nitrogen, 64 contain phosphorus, and 28 contain sulfur. It is therefore possible to use *i*JO1366 to predict the growth capabilities of *E. coli* on a very wide range of media conditions. As a demonstration of the prediction of growth capabilities, FBA was used to predict growth on every possible carbon, nitrogen, phosphorus, and sulfur source, one at a time, under aerobic conditions (Table II and Supplementary Table 12). For each prediction, only one of the four element source reactions was changed, and the default sources of the other three elements were used. The default carbon, nitrogen, phosphorus, and sulfur sources are glucose, ammonium, inorganic phosphate, and inorganic sulfate, respectively. If a growth rate above zero was predicted by FBA using the core biomass reaction as the objective, then a source was designated as growth supporting.

A total of 180 of the 285 possible carbon sources were found to be growth supporting. There are several reasons why a carbon containing metabolite cannot serve as a carbon source. First, not all extracellular compounds have transport reactions that allow them to enter the cell. Some may only have efflux reactions that allow them to be excreted. Second, some compounds are not connected to the central reactions of metabolism from which all essential biomass components are constructed. For example, cob(I)alamin can be converted only to vitamin B_{12}, but not to any other biomass components. Third, carbon sources must also generally serve as energy sources for *E. coli*, so a highly oxidized compound such as CO_2 cannot be growth supporting. Not all compounds can serve as nitrogen, phosphorus, and sulfur sources for similar reasons. Some compounds may serve as a source of more than one essential element, such as L-alanine, which can provide both carbon and nitrogen simultaneously. The potential growth supporting carbon, nitrogen, phosphorus, and sulfur sources were also predicted using the *i*AF1260 *E. coli* model. *i*JO1366 contains the same number of growth supporting phosphorus and sulfur sources, but has new sources for carbon and nitrogen. Thus, the scope of the environmental conditions that can be analyzed through modeling has now been increased.

Prediction of gene essentiality

The GPR associations of every reaction in *i*JO1366 allow this model to predict the effects of gene knockouts. We used FBA to predict the optimal growth rate of *E. coli* growing on both glucose and glycerol with all 1366 genes knocked out one at a time. These computational knockout screens were then compared to experimental screens of the entire Keio Collection (Table III and Supplementary Table 13) (Baba *et al*, 2006; Joyce *et al*, 2006; Yamamoto *et al*, 2009). Unlike the gene essentiality predictions presented above in *Experimental phenotypic screens*, the final *i*JO1366 model was used to make these predictions.

There are four possible outcomes, TP, TN, FP, and FN, when one compares computationally predicted to experimental gene essentiality data (as discussed above *Experimental phenotypic screens*). FP predictions can be made when a model contains some unrealistic capabilities, such as pathways that are normally not expressed during the particular growth conditions. Because *i*JO1366 is a metabolic network model that does not contain regulatory systems, FP predictions are possible. FN cases, on the other hand, indicate that some realistic content such as an essential transport or enzymatic reaction may be missing from the model. These predictions can be used to drive model-based biological discovery (Joyce *et al*, 2006; Reed *et al*, 2006b). When compared to the experimental gene essentiality data, most of the predictions made by *i*JO1366 are correct, confirming its overall accuracy (91%). Still, there are 80 FPs and 39 FNs among the 1366 predictions for growth on glucose minimal media. Predictions of growth on glycerol minimal media achieved similar accuracy. *i*JO1366 is slightly less accurate at predicting overall gene essentiality than *i*AF1260, when compared to the same datasets. This is because the 107 new genes added to this model

version are from less well-studied systems and pathways than the existing genes in *i*AF1260. Many of these new genes are associated with peripheral metabolic systems, while the well-studied central metabolic genes were already included in previous model versions. The overall accuracy of gene essentiality predictions for the 107 new genes is only 89%.

Mapping *i*JO1366 to closely related strains

To investigate the causes of inabilities to produce biomass, we identified the individual biomass components that each strain is unable to produce (Figure 2B). Checking the production of biomass components one at a time is a procedure commonly used during the assembly of metabolic reconstructions for microbes (Thiele & Palsson, 2010). This analysis revealed biological differences between the strains, identified auxotrophies, and brought to light many missing annotations in protein sequence databases. For example *E. coli* K-12 DH10B (KEGG organism code: ecd) cannot produce the amino acid L-leucine due to a deletion of the *leuABCD* operon (b0071-4) (Durfee et al, 2008). *Shigella flexneri* 301 (serotype 2a) (sfl) has a known L-methionine auxotrophy (Ahmed et al, 1988), which network analysis indicates is due to a lack of *metA* (b4013). This strain is also unable to produce S-adenosyl-L-methionine or biotin without L-methionine. *Shigella sonnei* (ssn) and *Shigella boydii* (sbo) were found to be unable to produce NAD^+ and $NADP^+$. Both have a niacin auxotrophy due to lack of *nadB* (b2574). This auxotrophy has been characterized experimentally in the case of *Shigella sonnei* (Pitsch & Nakamura, 1963) and is known to be a feature of many *Shigella* strains. Both strains can grow *in silico* after the free exchange of Niacin is allowed.

Model-predicted inconsistencies can also be used to prime experimental studies into metabolic differences between closely related strains. Although conservation within metabolic subsystems is strong, small differences can have large impacts on cellular physiology. For example, a single nucleotide change in *E. coli* O157:H7 EDL933 and *E. coli* O157:H7 Sakai led to altered utilization of N-acetyl-D-galactosamine (Mukherjee et al, 2008). This mutation may have been a contributing factor in the 2006 spinach outbreak, since this sugar is found in mucus glycoproteins in normal and diseased human colons (Clamp et al, 1981). Other important differences that require further investigation are components of the cell wall and outer membrane. For example, lipid A and its core oligosaccharide can take many different forms and modifications, some of which affect virulence (Raetz et al, 2007). It will be important to ensure that strain specific modifications are accounted for by manual curation. Looking to the future, constraint-based models now cover a large fraction of the core genome (40-50% depending on the particular strains and cutoffs used in the analysis) when we take into account the recent reconstruction of transcription and translation in *E. coli* K-12 MG1655 (Thiele et al, 2009) and the fact that this subsystem is strongly conserved within the species. It will be interesting and informative to see what genes remain in the core and if their functions can be brought under the umbrella of constraint-based approaches.

Supplementary References

Adams H, Teertstra W, Koster M, Tommassen J (2002) PspE (phage-shock protein E) of *Escherichia coli* is a rhodanese. *FEBS letters* **518:** 173-176

Ahmed ZU, Sarker MR, Sack DA (1988) Nutritional requirements of *Shigellae* for growth in a minimal medium. *Infect Immun* **56:** 1007-1009

Aziz RK, Bartels D, Best AA, Dejongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD et al (2008) The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9:** 75

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. *Molecular systems biology* **2:** 2006.0008

Butland G, Babu M, Diaz-Mejia JJ, Bohdana F, Phanse S, Gold B, Yang W, Li J, Gagarinova AG, Pogoutse O, Mori H, Wanner BL, Lo H, Wasniewski J, Christopolous C, Ali M, Venn P, Safavi-Naini A, Sourour N, Caron S et al (2008) eSGA: *E. coli* synthetic genetic array analysis. *Nat Methods* **5:** 789-795

Charusanti P, Conrad TM, Knight EM, Venkataraman K, Fong NL, Xie B, Gao Y, Palsson BØ (2010) Genetic basis of growth adaptation of *Escherichia coli* after deletion of *pgi*, a major metabolic gene. *PLoS Genet* **6:** e1001186

Clamp JR, Fraser G, Read AE (1981) Study of the carbohydrate content of mucus glycoproteins from normal and diseased colons. *Clin Sci (Lond)* **61:** 229-234

Colnaghi R, Cassinelli G, Drummond M, Forlani F, Pagani S (2001) Properties of the *Escherichia coli* rhodanese-like protein SseA: contribution of the active-site residue Ser240 to sulfur donor recognition. *FEBS letters* **500:** 153-156

Cvetkovic A, Menon AL, Thorgersen MP, Scott JW, Poole FL, 2nd, Jenney FE, Jr., 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

Durfee T, Nelson R, Baldwin S, Plunkett G, 3rd, Burland V, Mau B, Petrosino JF, Qin X, Muzny DM, Ayele M, Gibbs RA, Csorgo B, Posfai G, Weinstock GM, Blattner FR (2008) The complete genome sequence of *Escherichia coli* DH10B: insights into the biology of a laboratory workhorse. *Journal of bacteriology* **190:** 2597-2606

Falkenberg P, Strom AR (1990) Purification and characterization of osmoregulatory betaine aldehyde dehydrogenase of *Escherichia coli*. *Biochimica et biophysica acta* **1034:** 253-259

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. *Molecular systems biology* **3**

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

Fong SS, Palsson BØ (2004) Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nature genetics* **36:** 1056-1058

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology* **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

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

Janssen P, Goldovsky L, Kunin V, Darzentas N, Ouzounis CA (2005) Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications. *EMBO Rep* **6:** 397-399

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

Liu D, Reeves PR (1994) *Escherichia coli* K12 regains its O antigen. *Microbiology (Reading, England)* **140 (Pt 1):** 49-57

Molina-Henares MA, de la Torre J, Garcia-Salamanca A, Molina-Henares AJ, Herrera MC, Ramos JL, Duque E (2010) Identification of conditionally essential genes for growth of *Pseudomonas putida* KT2440 on minimal medium through the screening of a genome-wide mutant library. *Environ Microbiol* **12:** 1468-1485

Mukherjee A, Mammel MK, LeClerc JE, Cebula TA (2008) Altered utilization of N-acetyl-D-galactosamine by *Escherichia coli* O157:H7 from the 2006 spinach outbreak. *Journal of bacteriology* **190:** 1710-1717

Nakahigashi K, Toya Y, Ishii N, Soga T, Hasegawa M, Watanabe H, Takai Y, Honma M, Mori H, Tomita M (2009) Systematic phenome analysis of *Escherichia coli* multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Molecular systems biology* **5:** 306

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, Thiele I, Palsson BØ (2010) What is flux balance analysis? *Nature biotechnology* **28:** 245-248

Pitsch BL, Nakamura M (1963) Replacement of Nicotinic Acid Requirement of *Shigella Sonnei* Pyridine-3-Sulfonic Acid. *Journal of bacteriology* **86:** 159-160

Raetz CR, Reynolds CM, Trent MS, Bishop RE (2007) Lipid A modification systems in gram-negative bacteria. *Annu Rev Biochem* **76:** 295-329

Reed JL, Famili I, Thiele I, Palsson BØ (2006a) Towards multidimensional genome annotation. *Nat Rev Genet* **7:** 130-141

Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BØ (2006b) Systems approach to refining genome annotation. *Proceedings of the National Academy of Sciences of the United States of America* **103:** 17480-17484

Rubires X, Saigi F, Pique N, Climent N, Merino S, Alberti S, Tomas JM, Regue M (1997) A gene (wbbL) from *Serratia marcescens* N28b (O4) complements the rfb-50 mutation of *Escherichia coli* K-12 derivatives. *Journal of bacteriology* **179:** 7581-7586

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. *Biotechnology and bioengineering* **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 computational biology* **5:** e1000312

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

Typas A, Nichols RJ, Siegele DA, Shales M, Collins SR, Lim B, Braberg H, Yamamoto N, Takeuchi R, Wanner BL, Mori H, Weissman JS, Krogan NJ, Gross CA (2008) High-throughput, quantitative analyses of genetic interactions in *E. coli*. *Nat Methods* **5:** 781-787

Varma A, Palsson BØ (1993) Metabolic capabilities of *Escherichia coli*: II. Optimal growth patterns. *Journal of theoretical biology* **165:** 503-522

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. *Molecular systems biology* **5:** 335

Yang Y, Zhao G, Man TK, Winkler ME (1998) Involvement of the *gapA*- and *epd* (*gapB*)-encoded dehydrogenases in pyridoxal 5'-phosphate coenzyme biosynthesis in *Escherichia coli* K-12. *Journal of bacteriology* **180:** 4294-4299