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

Mr. Jeffrey D Orth, Mr. Tom M Conrad, Ms. Jessica Na, Mr. Joshua A Lerman, Hojung Nam, Adam M Feist and Bernhard Ø Palsson

Corresponding author: Bernhard Palsson, University of California San Diego

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision	24 May 2011

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the three referees who agreed to evaluate your manuscript. As you will see from the reports below, the referees generally agreed that this updated model of E. coli metabolism could provide a useful resource to the community, but they raised a series of substantial concerns, which, I am afraid to say, preclude its publication in its present form.

The reviewers main concerns feel broadly into two categories.

1. They indicated that this work currently lacked a clear and direct comparison of the performance of this new model to its predecessor (iAF1260). In fact, reviewer #1 notes that comparison with your previous work may even suggest that this model shows a small performance decrease compared to iAF1260. Similarly, the editor notes Fig. 4C as compared to Fig. 3E in Feist et al. (2007) which appears to suggest a slight decrease in the accuracy of the model predictions. In addition, all three reviewers has concerns about the appropriateness of the Schaub et al. (2008) flux data used for the flux prediction comparisons.

2. More broadly, the reviewers found the current work overly-long and indicated that many of the analyses presented could be condensed, and some could be entirely removed. Indeed, the second reviewer recommended dramatically reformatted this work using a shorter format (e.g. our Report format). We now encourage reviewers to comment on each other's reviews and during this process reviewer #1 supported this suggestion, writing:

"I think that the idea of reviewer #2 to make this work some kind of short communication/letter to the editor [is] very valuable. The model itself will

certainly be of importance to the Systems Biology community, but the scope of
the work as it stands does not warrant a full fledged article. If this work
would be reduced to the mere essence (as outlined in the reviews) it would
make a valuable contribution."

As such, we encourage you to prepare a revised manuscript that reformats this work according to
our Report format (see our Instruction for Authors for details), while also providing a more direct
comparisons of the performance of this model to iAF1260. We recognize that condensing this work
down to the Report format may be challenging, so I have provided a list of suggestions to consider
while revising your work:

-- Reports may have up to three figures; additional figures could be moved to the supplemental
material or eliminated as appropriate. I would recommend one figure summarizing the model
content (merging parts of Fig. 1 & 2), one figure with experimental validation results (from Fig. 3
and 4), and one describing the strain comparisons (Fig. 7).

-- The Knowledge index section could be removed, and the sections "Validation: Experimental
phenotypic screens" and "Prediction of gene essentiality" could be merged.

-- The lengthy comparisons to the EchoLocation database, and discussion of gaps and orphans could
be condensed and moved to the Supplementary Information.

-- The Discussion should be merged with the Results section, and the Introduction substantially
streamlined.

In addition, we ask you to address the following format and data issues when preparing your
revision:

-- Please provide any supplemental tables with more than 50 rows as separate tab-delimited text or
excel files, and remove them from the Supplementary Information pdf.

-- Please submit the iJO1362 model to a public database such as BioModels or JWS Online.

PLEASE NOTE As part of the EMBO Publications transparent editorial process initiative (see
http://www.nature.com/msb/journal/v6/n1/full/msb201072.html), Molecular Systems Biology now
publishes online a Review Process File with each accepted manuscript. Please be aware that in the
event of acceptance, your cover letter/point-by-point document will be included as part of this file,
which will be available to the scientific community. More information about this initiative is
available in our Instructions to Authors. If you have any questions about this initiative, please
contact the editorial office (msb@embo.org).

If you feel you can satisfactorily deal with these points and those listed by the referees, you may
wish to submit a revised version of your manuscript. Please attach a covering letter giving details of
the way in which you have handled each of the points raised by the referees. A revised manuscript
will be once again subject to review and you probably understand that we can give you no guarantee
at this stage that the eventual outcome will be favorable.

Yours sincerely,

Section Editor
Molecular Systems Biology

REFEREE REPORTS

Reviewer #1 (Remarks to the Author):

In their work, Orth et al. describe the reconstruction of an updated version of the genome-scale metabolic model of E. coli (iJO1362). Starting from a previous reconstruction, iAF1260, published in 2007 they 1) add new content based on new findings concerning existing metabolic pathways in E. coli 2) analyze a phenotypic screen of a large set of knockout mutants in four different conditions with respect to experimental and theoretical predictions of growth leading to additional model content 3) test the accuracy of viability predictions of the new model 4) reconstruct a large number of metabolic models for other sequenced E. coli strains and test whether these models predict viability of these strains. For a selected number of strains where inability to grow on the defined medium was predicted, existing auxotrophies were confirmed through a literature search. While the model itself will certainly be of value for the Systems Biology community I doubt that the limited scope of the expansion over the previous model (with only a small increase in predictive accuracy) and the limited scope of the analyses conducted with the expanded model are suited for the broad audience of Molecular Systems Biology. Moreover, the work includes many cases of an arbitrary choice of parameters without motivation or demonstration that this choice does not affect conclusions drawn in the work.
Major issues:

While the last reconstruction increased the number of known genes (from iJR904 to iAF1260) by 356 genes, the present reconstruction only increases this number by 102 genes. While this modest increase might be due to a limited amount of additional knowledge on E. coli metabolism accumulated through recent years, there appears to be no significant increase in the overall increase of the accuracy of the model. For instance, a large phenotypic screen of E. coli knockouts in different growth media with the iAF1260 network resulted in the modifications concerning only six genes while the remaining false positives and negatives were attributed to various causes some of which consisted of speculations (e.g. that there might exist isoenzymes for a particular knocked out enzyme). Moreover, a comparison of the predictions of the expanded model to experimental data from the Keio collection shows even a small decrease from 92% to 91% in comparison to the predecessor model (which is not mentioned by the authors).

In the section "Prediction of metabolic phenotypes" the authors use flux balance analysis to predict growth rates using their model. There are several issues related to this analysis:
1) They use data from Schaub et al. (2008) for E. coli grown at a growth rate of 0.1 h^-1 on glucose minimal medium. I think that data of better quality should be used as it was assumed in the work of Schaub et al. that the glyoxylate bypass is not used at these growth rates. However, in a work by Fischer and Sauer (2003) it was observed that the glyoxylate bypass is indeed used by E. coli if grown on low glucose concentrations. The authors should use, for instance, the data provided by Ishi et al. (Science 2007, 316(5824):593-7) in which intracellular fluxes were measured at different growth rates (between 0.1 h^-1 and 0.7 h^-1) and the glyoxylate bypass was indeed active at low growth rates. This would also allow them to assess the accuracy of their method for different growth rates.
2) In the methods section "Constraint-based modeling" the authors write that they constrain the flux through several reactions "that are not used under typical growth states" to zero. They should provide more detail why they have chosen to set the flux through these reactions to zero. In particular they should explain what they refer to as "typical growth state". For instance, they constrain the flux through the superoxide dismutase (SPODM) to zero. However, a double mutant of the enzymes catalyzing this reaction (SodA and SodB) is not able to grow aerobically on glucose-minimal medium (Carlioz and Touati, EMBO J 1986, 5(3):623-30.). Moreover, the formate-hydrogen lyase (FHL), whose flux is also constrained to zero, is known to be active under anaerobic conditions (Sawers, Antonie Van Leeuwenhoek 1994, 66(1-3):57-88.).
3) At several points they define upper limits for the inflow of metabolites without providing supporting evidence for the validity of the corresponding values (e.g. for cobalamin on p. 18 and succinate as well as L-lactate in the phenotypic screen).
4) There is no test of the change in the accuracy of the flux predictions between iAF1260 and iJO1362. If predictions are significantly better, this could provide additional support for the importance of the new reconstruction.

The authors estimate new values for the growth- and non-growth associated maintenance (GAM and NGAM, respectively). While the value for the GAM is similar to the value in iAF1260, the NGAM dropped from 8.39 to 3.15. While the authors mention themselves that these values depend on the experimental conditions, they should explain this drastic change in the NGAM.

The overall length of the manuscript (exceeding the 60.000 character limit of MSB articles by 11.000 characters) makes it hard to follow the work. I think that the manuscript could greatly benefit from an overall shortening, especially the introduction and discussion section should be made more concise. Also the results section could be reduced since the reconstruction process follows, in large parts, a framework put forward by one of the authors (Thiele and Palsson, 2010). Additionally, the subsection "Conversion to a computational model" could be moved to the methods section and the subsection "Knowledge index of iJO1362 genes" could be made more concise or entirely removed since the only insight it provides is that metabolism is marginally better studied than other cellular systems.

While the phenotypic screen of E. coli mutants on four different media is quite interesting I think that a screen of other growth media than those used in the study would be better suited to identify knowledge gaps in the reconstruction. In particular this would support the predictions concerning different carbon, nitrogen and sulfur sources.

In the reconstruction of strain specific models, the authors use an identity of 40% for the enzymes as a cut-off. They should provide supporting evidence why this choice of parameter is valid to identify functionally identical enzymes in a related strain.

Minor issues:

Lipoate biosynthesis should be included in the model as this metabolite is an important cofactor of several central metabolic reactions.

On p. 4 the term "root no-consumption gap" is used for the first time without an explanation which might be required for those not extremely familiar with metabolic reconstructions.

On p. 6 it is mentioned that there are 66 false negative predictions and 213 false positive predictions (p. 7) of iAF1260 in comparison to the phenotypic screen. However, in Figure 3 only 55 false negative and 179 false positive predictions are shown. Thus, there appears a problem with the numbers.

On several occasions (e.g. p.12, l.1) the authors present p-values from statistical tests without mentioning the kind of tests they have used.

On p. 18, section "Growth on different substrates" the authors mention growth rates of 4^h-1 and 6^h-1 which they use for their calculations. These values are very high and thus might by typos.

In the reconstruction of several strain specific models it is nowhere mentioned which growth media is used. Even if the standard medium they define in the methods section is used, this should be mentioned.

Reviewer #2 (Remarks to the Author):

The manuscript by Palsson and co-workers presents an improved reconstruction of the E. coli metabolic network. In the past, these resources have significantly contributed to the advance of the field of microbial systems biology and also this new reconstruction will likely become a major tool in the field. Thus, I feel that this manuscript could be a valuable contribution for MSB. However, I have two concerns that I would like the authors to address.

Readability of the manuscript: The main work of the presented manuscript definitely laid in the reconstruction of the metabolic network - work that could basically be described in a few paragraphs. However, such a short description would (in most journals) not warrant a "full-fledged" article in the classical terms, incl. a M&M section, a result and a discussion section. Thus, what authors typically do is (and this can be seen with this manuscript and also was seen with previous papers presenting metabolic network models from this and other groups) to augment these manuscripts with lots of analyses that provide little useful new scientific insight, and add overly-long discussion, to generate a full-fledged article. However, in the light of more and more papers

being published every day, I would appreciate if these articles could be kept to the mere essence (e.g. a brief description of the model with the most important numbers, but not more than that). I am sure that MSB with its compelling views on future ways of publishing systems biology research would be open to any reasonable publishing formats describing such work.

Anyway, this is probably still something for the future. Nevertheless, below I have a few suggestions of how the authors could shorten their manuscript and generally improve its readability:

- Please try to shorten the manuscript as much as possible.
o Several points are unnecessarily often repeated ("overly-long discussion"; see above).
o Paragraphs 2 and 3 in the introduction outline describe - in the meantime - textbook knowledge and can simply be omitted.
o The discussion is largely a repetition of earlier mentioned points. In fact, in the light of what I said above, this kind of work probably cannot be discussed in the same way as more classical work. Such, please do not bind yourself to the typically length of discussion and instead streamline and condense this section.
o Moving less important points to the supplement or completely remove them to improve the readability of the manuscript (e.g. section on Knowledge index iJO1362 genes (Page 13, Figure 8))?
- I find the first section in the result section ("Process for ..") a bit chaotic with a lot of redundancies and unnecessary information. Is it really necessary to "dump" this information on the reader?
- I find it extremely disturbing and confusing that the two sections "Validation: Experimental phenotypic screens" and "Prediction of gene essentiality" are both in the manuscript. Don't these sections basically contain the same information? Also, in one section the prediction quality is specified with 93.5% while in the other 91% is mentioned.
- The difference between the different classes of gaps in the reconstruction (i.e. root no-production, root no-consumption, no-production downstream, no-consumption upstream) is not fully clear from the manuscript. A more detailed explanation or a graph would improve the understanding of the differences and the conclusions drawn from them. (Page 4; Figure 6)

Comparison with experimental data. I have several concerns with the way how the authors have done the comparison of their model predictions with experimental results.
- Flux predictions were compared with experimental data from Schaub. My concern is that this is a somewhat "extreme" data set, for example, in the light of the PPP flux. Most other E. coli PPP fluxes are much lower (e.g. compare with Appl Environ Microbiol. 2006 February; 72(2): 1164-1172). I don't want to engage in a discussion about which experimental data is correct. Instead, I would like to see that the authors position their "model prediction" correctly! It is simply not a good thing to sort of imply good model prediction capability when comparing model predictions simply with an arbitrary data set. Asked differently: what is the point of such a comparison if there is such large experimental variability?
- Also for the other predictions of metabolic phenotypes, a short glimpse into the material and method section reveals that all different kind of rate constraints were imposed to yield an in silico flux distribution that matched with the experimental observations. If the authors would like to retain these analyses, I would ask them to honestly place the information about these constraints either directly in the main text or alternatively in the respective figure caption. Currently this information is buried in the M&M section and thus a quick reader would not stumble on this very active tuning of the model predictions. However, I feel that these analyses could also be completely omitted. After more than 10 years of E. coli FBA the whole community knows that EVERY flux distribution can be obtained if just the right rate constraints are chosen. Thus, there is simply no added value in this analysis anymore (at least they cannot used to "validate" any model prediction). Relating back to my initial comments, I do feel that these analyses provide little useful new scientific insight.
- I greatly appreciate the efforts of experimentally determining the effect of the gene deletions. However, here I would appreciate a more in depth discussion/analysis on the way a mutant computationally got classified to be "lethal". Specifically, I would also like to see the prediction quality when a strict stoichiometric criterion for lethality is assumed (no biomass production vs the currently used 5% WT-growth rate cut-off criterion).

Minor:
- At several instances, it is mentioned that some manual curations, etc. where done "one by one ...". But in no instance, it is explicitly mentioned how these manual curations particularly were done. Is it more than a PhD student using his gut feeling?

- In their last E. coli reconstruction paper (also published by MSB), the authors have added very valuable data on reaction energies. I would strongly suggest that the current manuscript is augmented for the thermodynamic information on the here newly added reactions/metabolite. Otherwise, researchers who would like to do thermodynamic analyses would always need to fall back on the more outdated iAF1260 model.
- Missing references: Certain parameters are reported without reference, i.e. default bound on uptake rates, reactions constrained to zero (Page 18).
- P/O ratio not reported: Reporting the P/O ratio that was used for the determination of the maintenance requirements, would improve the comparability of the maintenance requirements with previous reported ones (Page 5, Page 16).

Reviewer #3 (Remarks to the Author):

The manuscript describes the construction, validation and applications of a genome-scale metabolic model for E. coli. The model uses its predecessor as the template. The authors perform the update by scanning the literature for new functional annotation of genes. Based on recent literature, they fill the gaps in the model and use the model to predict metabolic phenotypes, identify growth-supporting nutrients and gene essentiality. They also compare the model with the SEED model (automated model) and validate the localization of reactions by cross-checking with EchoLocation database. Overall, they promise improved utility of the model.

1. The validation of the model was done by comparing the metabolism in E. coli at a specific growth rate of 0.1 h-1. At this conditions, the metabolism is largely respiratory and therefore, the model can predict the fluxes rather well since biomass maximization is the objective function. The authors should consider validating their model under different conditions (anaerobic, higher growth rate, etc).
2. On a related note, the authors should present the improvement in the prediction capability of the present model over its predecessor. In fact, the prediction of the fluxes (figure 4) should be compared with those from the iAF1260 model.
3. The use of the model to map other strains should be described with a greater level of caution. This is because the conservation of a gene between two strains lends the credibility to retain the reaction (that is catalyzed by the enzyme it encodes for) in the model. As the authors noted, even one SNP is sufficient to alter the phenotype. So, the even a high degree of homology between the genes does not necessarily translate into conservation of the phenotype.
4. The opening line of the discussion says that the new content added to the model demonstrates that new discoveries are made. This sentence should be rephrased to reflect that new discoveries necessitated the addition of content to the model.
5. Lastly, but definitely not the least, the format of information provided in the Supplementary material is impractical. If the reviewers are to look at any of the Supplementary tables, it is impossible to jump to that table in a file that contains over 1800 pages. This information should be fragmented into smaller files (to separate the reaction database from the analysis of the database).

1st Revision - authors' response   19 July 2011

Editor's Comments:

The reviewers main concerns feel broadly into two categories.

1. They indicated that this work currently lacked a clear and direct comparison of the performance of this new model to its predecessor (iAF1260). In fact, reviewer #1 notes that comparison with your previous work may even suggest that this model shows a small performance decrease compared to iAF1260. Similarly, the editor notes Fig. 4C as compared to Fig. 3E in Feist et al. (2007) which appears to suggest a slight decrease in the accuracy of the model predictions. In addition, all three reviewers has concerns about the appropriateness of the Schaub et al. (2008) flux data used for the flux prediction comparisons.

6

We have removed the comparison to the Schaub flux data. Now that the manuscript has been shortened to a Report, there isn't room for direct comparisons between the predictions made by *i*JO1366 and *i*AF1260. We also believe that a direct comparison between these models would not be useful anyway, as discussed in our letter above. As mentioned by Reviewer #2, it has been known for many years that constraint-based *E. coli* metabolic models are capable of making accurate phenotypic predictions.

2. More broadly, the reviewers found the current work overly-long and indicated that many of the analyses presented could be condensed, and some could be entirely removed. Indeed, the second reviewer recommended dramatically reformatted this work using a shorter format (e.g. our Report format). We now encourage reviewers to comment on each other's reviews and during this process reviewer #1 supported this suggestion, writing:

"I think that the idea of reviewer #2 to make this work some kind of short communication/letter to the editor [is] very valuable. The model itself will certainly be of importance to the Systems Biology community, but the scope of the work as it stands does not warrant a full fledged article. If this work would be reduced to the mere essence (as outlined in the reviews) it would make a valuable contribution."

As such, we encourage you to prepare a revised manuscript that reformats this work according to our Report format (see our Instruction for Authors for details), while also providing a more direct comparisons of the performance of this model to iAF1260. We recognize that condensing this work down to the Report format may be challenging, so I have provided a list of suggestions to consider while revising your work:

-- Reports may have up to three figures; additional figures could be moved to the supplemental material or eliminated as appropriate. I would recommend one figure summarizing the model content (merging parts of Fig. 1 & 2), one figure with experimental validation results (from Fig. 3 and 4), and one describing the strain comparisons (Fig. 7).

Figures 1 and 2 have been combined. Figure 3 has been moved to the Supplementary Information, but Figure 4 was removed because as the reviewers indicated, the comparison to the Schaub flux data is unnecessary and uninformative. Figure 7 is also still included as the new Figure 2. Figures 6 and 8 have been moved to the Supplementary Information, while Figure 5 has been removed.

-- The Knowledge index section could be removed, and the sections "Validation: Experimental phenotypic screens" and "Prediction of gene essentiality" could be merged.

The knowledge index section has been moved to the Supplementary Information, while the phenotypic screen and model gene essentiality prediction sections have been combined and shortened. Details of these sections have been moved to the Supplementary Information.

-- The lengthy comparisons to the EchoLocation database, and discussion of gaps and orphans could be condensed and moved to the Supplementary Information.

These are now mentioned briefly in the main text, with the bulk of the remaining text moved to the Supplementary Information.

-- The Discussion should be merged with the Results section, and the Introduction substantially streamlined.

This has been done.

In addition, we ask you to address the following format and data issues when preparing your revision:

-- Please provide any supplemental tables with more than 50 rows as separate tab-delimited text or excel files, and remove them from the Supplementary Information pdf.

The supplementary tables are intended to be viewed and queried with Excel, and the pdf versions were a mistake resulting from the online manuscript submission process. This has been corrected.

-- *Please submit the iJO1362 model to a public database such as BioModels or JWS Online.*

The *i*JO1366 model will be submitted to BioModels upon publication. It will also be made available through the BiGG database (http://bigg.ucsd.edu/) and our lab website (http://systemsbiology.ucsd.edu/).

Reviewer #1 (Remarks to the Author):

In their work, Orth et al. describe the reconstruction of an updated version of the genome-scale metabolic model of E. coli (iJO1362). Starting from a previous reconstruction, iAF1260, published in 2007 they 1) add new content based on new findings concerning existing metabolic pathways in E. coli 2) analyze a phenotypic screen of a large set of knockout mutants in four different conditions with respect to experimental and theoretical predictions of growth leading to additional model content 3) test the accuracy of viability predictions of the new model 4) reconstruct a large number of metabolic models for other sequenced E. coli strains and test whether these models predict viability of these strains. For a selected number of strains where inability to grow on the defined medium was predicted, existing auxotrophies were confirmed through a literature search. While the model itself will certainly be of value for the Systems Biology community I doubt that the limited scope of the expansion over the previous model (with only a small increase in predictive accuracy) and the limited scope of the analyses conducted with the expanded model are suited for the broad audience of Molecular Systems Biology. Moreover, the work includes many cases of an arbitrary choice of parameters without motivation or demonstration that this choice does not affect conclusions drawn in the work.

Major issues:

While the last reconstruction increased the number of known genes (from iJR904 to iAF1260) by 356 genes, the present reconstruction only increases this number by 102 genes. While this modest increase might be due to a limited amount of additional knowledge on E. coli metabolism accumulated through recent years, there appears to be no significant increase in the overall increase of the accuracy of the model. For instance, a large phenotypic screen of E. coli knockouts in different growth media with the iAF1260 network resulted in the modifications concerning only six genes while the remaining false positives and negatives were attributed to various causes some of which consisted of speculations (e.g. that there might exist isoenzymes for a particular knocked out enzyme). Moreover, a comparison of the predictions of the expanded model to experimental data from the Keio collection shows even a small decrease from 92% to 91% in comparison to the predecessor model (which is not mentioned by the authors).

As mentioned above, a major increase in the predictive capabilities for growth on common substrates and central metabolic fluxes by *i*JO1366 is not expected since so much was already included in previous model versions. This new version contains 107 new genes, most of which have only recently been characterized and participate in outlying and less well-studied subsystems and pathways. All well-characterized central metabolic genes were already included in *i*AF1260 and its predecessors. It is thus expected that the new genes in *i*JO1366 would be slightly less accurate with respect to gene essentiality predictions, explaining the slight overall decrease in accuracy. This is now explained in the text in the "Prediction of metabolic phenotypes" section and in the Supplementary Information in the "Prediction of gene essentiality section".

In the section "Prediction of metabolic phenotypes" the authors use flux balance analysis to predict growth rates using their model. There are several issues related to this analysis:
1) They use data from Schaub et al. (2008) for E. coli grown at a growth rate of 0.1 h^-1 on glucose minimal medium. I think that data of better quality should be used as it was assumed in the work of Schaub et al. that the glyoxylate bypass is not used at these growth rates. However, in a work by Fischer and Sauer (2003) it was observed that the glyoxylate bypass is indeed used by E. coli if

grown on low glucose concentrations. The authors should use, for instance, the data provided by Ishi et al. (Science 2007, 316(5824):593-7) in which intracellular fluxes were measured at different growth rates (between 0.1 h^-1 and 0.7 h^-1) and the glyoxylate bypass was indeed active at low growth rates. This would also allow them to assess the accuracy of their method for different growth rates.

The comparison to the Schaub data has been removed. We have examined the data from Ishii et al in previous studies in our lab, but we concluded that this data is also flawed in several ways. Since, as pointed out by the other reviewers, it is already well established that *E. coli* metabolic reconstructions can be used to predict phenotypes, and a simple flux distribution comparison contributes no new information, we have decided simply to remove this entire section.

2) In the methods section "Constraint-based modeling" the authors write that they constrain the flux through several reactions "that are not used under typical growth states" to zero. They should provide more detail why they have chosen to set the flux through these reactions to zero. In particular they should explain what they refer to as "typical growth state". For instance, they constrain the flux through the superoxide dismutase (SPODM) to zero. However, a double mutant of the enzymes catalyzing this reaction (SodA and SodB) is not able to grow aerobically on glucose-minimal medium (Carlioz and Touati, EMBO J 1986, 5(3):623-30.). Moreover, the formate-hydrogen lyase (FHL), whose flux is also constrained to zero, is known to be active under anaerobic conditions (Sawers, Antonie Van Leeuwenhoek 1994, 66(1-3):57-88.).

The specific reasons for constraining each of these reactions to zero are now given in the Materials and Methods section. We know that FHL is expressed in *E. coli* under anaerobic conditions, but FBA wrongly predicts that it will be used under aerobic conditions as well if it is not constrained. *i*AF1260 also constrained this reaction to zero by default.

3) At several points they define upper limits for the inflow of metabolites without providing supporting evidence for the validity of the corresponding values (e.g. for cobalamin on p. 18 and succinate as well as L-lactate in the phenotypic screen).

The succinate and lactate uptake rates used in the phenotypic screen are aribitrary, but in the range of realistic uptake rates. In order to predict growth/no growth phenotypes, it is only necessary to use an uptake rate sufficiently high to assure growth. For more detailed studies of phenotypes such as predicting intracellular fluxes and byproduct secretion, the exact uptake rates can be important, but they are not important here.   Cobalamin is a precursor of vitamin B_{12}, and cannot be used for any other purpose in this model. Only a very small amount of B_{12} is required for growth, so the lower bound of 0.01 is still essentially unconstrained. This same constraint was used in the *i*AF1260 model. These details are now included in the Materials and Methods section.

4) There is no test of the change in the accuracy of the flux predictions between iAF1260 and iJO1362. If predictions are significantly better, this could provide additional support for the importance of the new reconstruction.

The predictions of common phenotypes are not significantly better. As explained above, *i*AF1260 and its predecessors were already very accurate when making such predictions, and the primary purpose of the *i*JO1366 update is not to improve the accuracy of these predictions. This is now mentioned in the text in the "Prediction of metabolic phenotypes" section.

The authors estimate new values for the growth- and non-growth associated maintenance (GAM and NGAM, respectively). While the value for the GAM is similar to the value in iAF1260, the NGAM dropped from 8.39 to 3.15. While the authors mention themselves that these values depend on the experimental conditions, they should explain this drastic change in the NGAM.

The values of *E. coli* GAM and NGAM have been calculated many times over the past 20 years for the models of Varma (4 ATP/glucose / NGAM=7.6, GAM =13), Edwards (NGAM=5.87, GAM=23), Reed (NGAM=7.6, GAM=45.73), and Feist, and they have always varied quite a bit depending on which set of experimental data was used. Given this great variability, the newly calculated NGAM value is still quite reasonable. The reason that it is lower than previous estimates is that the data from which it was calculated (Taymaz-Nikerel 2010) accounted for cell lysis rates

when determining the growth rates of *E. coli* in chemostat experiments. Thus, they determined that *E. coli* was growing at a slightly higher rate than the dilution rates of the chemostats, while in other datasets the dilution rate is assumed to be the same as the growth rate. *E. coli* was thus shown to grow slightly faster than before, indicating a lower maintenance demand. This is now explained in the Supplementary Information in the "Updating the biomass composition and growth requirements" section.

The overall length of the manuscript (exceeding the 60.000 character limit of MSB articles by 11.000 characters) makes it hard to follow the work. I think that the manuscript could greatly benefit from an overall shortening, especially the introduction and discussion section should be made more concise. Also the results section could be reduced since the reconstruction process follows, in large parts, a framework put forward by one of the authors (Thiele and Palsson, 2010). Additionally, the subsection "Conversion to a computational model" could be moved to the methods section and the subsection "Knowledge index of iJO1362 genes" could be made more concise or entirely removed since the only insight it provides is that metabolism is marginally better studied than other cellular systems.

We have shortened the manuscript to a Report, and have removed some text and moved a significant amount text to the supplement.

While the phenotypic screen of E. coli mutants on four different media is quite interesting I think that a screen of other growth media than those used in the study would be better suited to identify knowledge gaps in the reconstruction. In particular this would support the predictions concerning different carbon, nitrogen and sulfur sources.

Additional screens are in progress and are planned for a separate publication. Only the four screens presented here were used to update the current model. Completion of these additional screens would delay publication of this manuscript considerably.

In the reconstruction of strain specific models, the authors use an identity of 40% for the enzymes as a cut-off. They should provide supporting evidence why this choice of parameter is valid to identify functionally identical enzymes in a related strain.

This cutoff was justified through our analysis of predicted auxotrophies in the strain specific models we generated. We also investigated the effects of higher and lower cutoffs, and found that with a higher cutoff there were more auxotrophies than are known to exist according to literature. At 40%, most known auxotrophy phenotypes are accurately predicted by the models. In this analysis, FBA and the metabolic network structure in addition to sequence homology helped us to determine the most plausible metabolic gene content. We have now tried to explain this better in the main text in the "Mapping *i*JO1366 to closely related strains" section.

Minor issues:

Lipoate biosynthesis should be included in the model as this metabolite is an important cofactor of several central metabolic reactions.

We have added the lipoate biosynthesis reactions (accounting for 3 of the 4 new genes added since the initial submission), and added lipoate to the wild-type biomass function.

On p. 4 the term "root no-consumption gap" is used for the first time without an explanation which might be required for those not extremely familiar with metabolic reconstructions.

The types of gaps are now defined in the Supplementary Information in the "Gaps and orphan reactions in the *i*JO1366 reconstruction" section. There was not enough space to include this in the main text.

On p. 6 it is mentioned that there are 66 false negative predictions and 213 false positive predictions (p. 7) of iAF1260 in comparison to the phenotypic screen. However, in Figure 3 only 55 false negative and 179 false positive predictions are shown. Thus, there appears a problem with the numbers.

The numbers of false positives given in the main text are correct, but there were actually 53 FNs and 172 FPs in Supplementary Figure 1 (formerly Figure 3). The 13 missing FNs were due to contamination during the experiments, and were thus not included the figure. The 41 FPs omitted from the figure (corresponding to 32 different genes) did not have simple explanations fitting them into the defined categories. These omitted genes are now explained in the Supplementary Figure 1 caption.

On several occasions (e.g. p.12, l.1) the authors present p-values from statistical tests without mentioning the kind of tests they have used.

Standard t-tests were used. This is now mentioned with the p-values in the text (which can now be found in the Supplementary Information).

On p. 18, section "Growth on different substrates" the authors mention growth rates of 4^h-1 and 6^h-1 which they use for their calculations. These values are very high and thus might by typos.

Yes, these were just typos, but this section has now been removed entirely.

In the reconstruction of several strain specific models it is nowhere mentioned which growth media is used. Even if the standard medium they define in the methods section is used, this should be mentioned.

It was the standard glucose minimal medium. This is now mentioned in the main text.

Reviewer #2 (Remarks to the Author):

The manuscript by Palsson and co-workers presents an improved reconstruction of the E. coli metabolic network. In the past, these resources have significantly contributed to the advance of the field of microbial systems biology and also this new reconstruction will likely become a major tool in the field. Thus, I feel that this manuscript could be a valuable contribution for MSB. However, I have two concerns that I would like the authors to address.

Readability of the manuscript: The main work of the presented manuscript definitely laid in the reconstruction of the metabolic network - work that could basically be described in a few paragraphs. However, such a short description would (in most journals) not warrant a "full-fledged" article in the classical terms, incl. a M&M section, a result and a discussion section. Thus, what authors typically do is (and this can be seen with this manuscript and also was seen with previous papers presenting metabolic network models from this and other groups) to augment these manuscripts with lots of analyses that provide little useful new scientific insight, and add overly-long discussion, to generate a full-fledged article. However, in the light of more and more papers being published every day, I would appreciate if these articles could be kept to the mere essence (e.g. a brief description of the model with the most important numbers, but not more than that). I am sure that MSB with its
compelling views on future ways of publishing systems biology research would be open to any reasonable publishing formats describing such work.

Anyway, this is probably still something for the future. Nevertheless, below I have a few suggestions of how the authors could shorten their manuscript and generally improve its readability:

- Please try to shorten the manuscript as much as possible.
o Several points are unnecessarily often repeated ("overly-long discussion"; see above).

We have shortened the manuscript to a Report, trying to remove the repetitive text and moving nonessential text to the Supplementary Information.

o Paragraphs 2 and 3 in the introduction outline describe - in the meantime - textbook knowledge and can simply be omitted.

These paragraphs have been removed.

o The discussion is largely a repetition of earlier mentioned points. In fact, in the light of what I said above, this kind of work probably cannot be discussed in the same way as more classical work. Such, please do not bind yourself to the typically length of discussion and instead streamline and condense this section.

The Discussion section has been combined with the Results section, and most of its text has been removed entirely.

o Moving less important points to the supplement or completely remove them to improve the readability of the manuscript (e.g. section on Knowledge index iJO1362 genes (Page 13, Figure 8))?

The specific text on the knowledge index as well as a number of other sections has been moved to the Supplementary Information.

- I find the first section in the result section ("Process for ..") a bit chaotic with a lot of redundancies and unnecessary information. Is it really necessary to "dump" this information on the reader?

Most of this section has now been removed.

- I find it extremely disturbing and confusing that the two sections "Validation: Experimental phenotypic screens" and "Prediction of gene essentiality" are both in the manuscript. Don't these sections basically contain the same information? Also, in one section the prediction quality is specified with 93.5% while in the other 91% is mentioned.

These two sections described entirely different sets of predictions and analyses. In the first section (Validation: Experimental phenotypic screens), an experimental screen of KO strains was performed and the results were compared to the *i*AF1260 model predictions. These results were then used to help update the model. In the second section (Prediction of gene essentiality), the final *i*JO1366 model was used to predict the essentiality of all 1366 genes as a demonstration of the predictive capabilities of the model. We have moved the "Experimental Phenotypic Screens" section to the Supplementary Information, along with most of the "Prediction of gene essentiality" section. We have added text to the " Prediction of gene essentiality " section to make the distinction between these two sections clearer to readers.

- The difference between the different classes of gaps in the reconstruction (i.e. root no-production, root no-consumption, no-production downstream, no-consumption upstream) is not fully clear from the manuscript. A more detailed explanation or a graph would improve the understanding of the differences and the conclusions drawn from them. (Page 4; Figure 6)

The types of gaps are now defined in the Supplementary Information in the "Gaps and orphan reactions in the *i*JO1366 reconstruction" section. The discussion of remaining gaps in the model has also been moved to the Supplementary Information. With the reformatting of the manuscript to a Report, there is not enough space to include these definitions in the main text.

Comparison with experimental data. I have several concerns with the way how the authors have done the comparison of their model predictions with experimental results.
- Flux predictions were compared with experimental data from Schaub. My concern is that this is a somewhat "extreme" data set, for example, in the light of the PPP flux. Most other E. coli PPP fluxes are much lower (e.g. compare with Appl Environ Microbiol. 2006 February; 72(2): 1164-1172). I don't want to engage in a discussion about which experimental data is correct. Instead, I would like to see that the authors position their "model prediction" correctly! It is simply not a good thing to sort of imply good model prediction capability when comparing model predictions simply with an arbitrary data set. Asked differently: what is the point of such a comparison if there is such large experimental variability?

The comparison to the Schaub flux data has been removed.

- Also for the other predictions of metabolic phenotypes, a short glimpse into the material and method section reveals that all different kind of rate constraints were imposed to yield an in silico flux distribution that matched with the experimental observations. If the authors would like to retain these analyses, I would ask them to honestly place the information about these constraints either directly in the main text or alternatively in the respective figure caption. Currently this information is buried in the M&M section and thus a quick reader would not stumble on this very active tuning of the model predictions. However, I feel that these analyses could also be completely omitted. After more than 10 years of E. coli FBA the whole community knows that EVERY flux distribution can be obtained if just the right rate constraints are chosen. Thus, there is simply no added value in this analysis anymore (at least they cannot used to "validate" any model prediction). Relating back to my initial comments, I do feel that these analyses provide little useful new scientific insight.

This section has been removed since it is not useful, and as explained above, the *i*JO1366 model does not make more accurate phenotypic predictions than its predecessors for conditions involving primarily central metabolic reactions, such as for growth on glucose, succinate, or acetate.

- I greatly appreciate the efforts of experimentally determining the effect of the gene deletions. However, here I would appreciate a more in depth discussion/analysis on the way a mutant computationally got classified to be "lethal". Specifically, I would also like to see the prediction quality when a strict stoichiometric criterion for lethality is assumed (no biomass production vs the currently used 5% WT-growth rate cut-off criterion).

We reran the computational screen with a strict threshold of zero growth instead of 5% WT, and found that it only made a difference in one case out of the entire screen. Only the *lldD* (b3605) knockout grown on lactate minimal medium has a computationally predicted growth rate between 0 and 5% of WT. The experimental screen classified this gene as essential under these conditions, so an additional false positive case is created when the zero threshold is used. Still, the fact that the "lethal" threshold only affected one case out of 4300 indicates that exact value of this threshold does not significantly affect the results. This is now mentioned in the Materials and Methods section.

Minor:
- At several instances, it is mentioned that some manual curations, etc. where done "one by one ...". But in no instance, it is explicitly mentioned how these manual curations particularly were done. Is it more than a PhD student using his gut feeling?

The published 96-step protocol of Thiele and Palsson was used for manual curation, and this is now mentioned in the main text in the "Process for updating the reconstruction and its content" section.

- In their last E. coli reconstruction paper (also published by MSB), the authors have added very valuable data on reaction energies. I would strongly suggest that the current manuscript is augmented for the thermodynamic information on the here newly added reactions/metabolite. Otherwise, researchers who would like to do thermodynamic analyses would always need to fall back on the more outdated iAF1260 model.

We agree that a thermodynamics update could be useful to the community. Unfortunately, it would be very difficult to add the thermodynamic information to the updated reconstruction at this time. This analysis was performed by Chris Henry (now at Argonne National Laboratory), who is not currently collaborating with us on this project.

- Missing references: Certain parameters are reported without reference, i.e. default bound on uptake rates, reactions constrained to zero (Page 18).

The reactions constrained to zero and the bounds on various uptake rates are now explained in the Materials and Methods section. In general the bounds used are arbitrary but realistic and do not affect the qualitative results (i.e. growth/no growth phenotypes).

- P/O ratio not reported: Reporting the P/O ratio that was used for the determination of the maintenance requirements, would improve the comparability of the maintenance requirements with

previous reported ones (Page 5, Page 16).

The P/O ratio used for determination of maintenance requirements was 1.375. This is now explicitly stated in the "Updating the biomass composition and growth requirements" in the Supplementary Information.

Reviewer #3 (Remarks to the Author):

The manuscript describes the construction, validation and applications of a genome-scale metabolic model for E. coli. The model uses its predecessor as the template. The authors perform the update by scanning the literature for new functional annotation of genes. Based on recent literature, they fill the gaps in the model and use the model to predict metabolic phenotypes, identify growth-supporting nutrients and gene essentiality. They also compare the model with the SEED model (automated model) and validate the localization of reactions by cross-checking with EchoLocation database. Overall, they promise improved utility of the model.

1. The validation of the model was done by comparing the metabolism in E. coli at a specific growth rate of 0.1 h-1. At this conditions, the metabolism is largely respiratory and therefore, the model can predict the fluxes rather well since biomass maximization is the objective function. The authors should consider validating their model under different conditions (anaerobic, higher growth rate, etc).

The comparison to the Schaub flux data has been removed. It is already known that the genome-scale *E. coli* models can make accurate predictions of phenotypes.

2. On a related note, the authors should present the improvement in the prediction capability of the present model over its predecessor. In fact, the prediction of the fluxes (figure 4) should be compared with those from the iAF1260 model.

As mentioned above several times, the predictions of central fluxes and growth on central metabolic substrates are not, nor should they be expected to be, better than those of *i*AF1260. The section making phenotypic predictions has been removed.

3. The use of the model to map other strains should be described with a greater level of caution. This is because the conservation of a gene between two strains lends the credibility to retain the reaction (that is catalyzed by the enzyme it encodes for) in the model. As the authors noted, even one SNP is sufficient to alter the phenotype. So, the even a high degree of homology between the genes does not necessarily translate into conservation of the phenotype.

It is certainly true that homology alone does not guarantee identical function. This is why we consider the models of the other strains to be draft reconstructions, and suggest that more curation is necessary to confirm their completeness. Still, the use of homology for gene annotation is extremely common and in general quite effective. We have added more text addressing these facts to the "Mapping *i*JO1366 to closely related strains" section.

4. The opening line of the discussion says that the new content added to the model demonstrates that new discoveries are made. This sentence should be rephrased to reflect that new discoveries necessitated the addition of content to the model.

A very good point. This is now mentioned in the last paragraph of the Introduction.

5. Lastly, but definitely not the least, the format of information provided in the Supplementary material is impractical. If the reviewers are to look at any of the Supplementary tables, it is impossible to jump to that table in a file that contains over 1800 pages. This information should be fragmented into smaller files (to separate the reaction database from the analysis of the database).

We intended these tables to be viewed in Excel, and they should be available as Excel files upon resubmission.

Thank you again for submitting your revised work to Molecular Systems Biology. We have now heard back from the three referees evaluating the study. As you will see, the referees felt that the revisions had significantly improved this work, and they are now largely supportive. They do make, however, some suggestions for clarifications and modifications, which we would like to ask you to carefully address in a final minor revision of the present work.

In addition, while preparing this revised work please submit the iJO1366 model to BioModels and incorporate the resulting accession number into the Methods section of this work.

Please resubmit your revised manuscript online, with a covering letter listing amendments and responses to each point raised by the referees. Please resubmit the paper **within one month** and ideally as soon as possible. If we do not receive the revised manuscript within this time period, the file might be closed and any subsequent resubmission would be treated as a new manuscript. Please use the Manuscript Number (above) in all correspondence.

Thank you for submitting this paper to Molecular Systems Biology.

Sincerely,

Editor - Molecular Systems Biology
msb@embo.org

REFEREE REPORTS

Reviewer #1 (Remarks to the Author):

I feel that the shortening has considerably improved the manuscript of Orth et al. and I think that it is suitable for publication now in particular due to the expansion of the metabolic model. I also appreciate the expansion of the section "Constraint-based modelling" which gives much more detail about the side assumptions of the model, even though I find explanations such as "Certain reactions are by default constrained to carry zero flux to avoid unrealistic behaviors" pretty dubious. Besides that I have two points the authors should consider addressing.

First, on p. 3 the authors state that for E. coli, „there is little variability with respect to metabolic gene content within the species (Vieira et al, 2011)". However, in the referenced work it was found that on average (only) 59% of metabolic reactions were present in all metabolic networks of a large number of E. coli and Shigella strains. I wouldn't refer to this number as "little variability". Moreover, these results appear to stand in contrast to a finding the authors report a little later in that paragraph where they state: "The average genome in this analysis contains approximately 97% of the genes in iJO1366 (Figure 2C)". While Vieira et al. based their numbers on reactions and in comparison to the "core metabolism" (present in all species) the number of on average 97% overlap with iJO1366 found by Orth et al. appears to be pretty high. The authors should comment on that.

A second (minor) issue is that the authors still use "root no-production gap" and related terms in the main manuscript without giving an explanation what they refer to (also noted by reviewer #2 in the previous review). Even though it might consume a few lines to explain these terms, the reader would much better understand what is referred to. Otherwise, the authors could also just state that there is a particular number of metabolites that cannot be produced and a particular number of metabolites that cannot be consumed at steady state. The detailed numbers are already given in the Supplementary material along with a detailed explanation of the terms.

Reviewer #2 (Remarks to the Author):

After having addressed most of my earlier concerns, I now recommend this manuscript for publication.

Reviewer #3 (Remarks to the Author):

In the revised version of the manuscript, the authors have addressed many of the issues raised by the reviewers. This version is concise and better highlights the need to have a updated genome-scale metabolic model for E. coli.

An important aspect of the model is the parameter values. While the values for maintenance have roots in solid experimental data, assumption that the two NADH dehydrogenases have equal importance could have a big impact on proton translocation and ATP generation. Therefore, it is important that there should be some justification given to the assumption.

2nd Revision - authors' response   17 August 2011

Thank you for the prompt review of our resubmitted manuscript, "A comprehensive genome-scale reconstruction of *Escherichia coli* – 2011," Manuscript Number: MSB-11-2860R. We have now made the minor revisions requested by the reviewers. The *i*JO1366 model in SBML format has been submitted to the BioModels database (accession: MODEL1108160000). In response to Reviewer #1's concerns, we changed "little variability" to a "moderate level of variability" in paragraph 1 of the "Mapping *i*JO1366 to closely related strains" subsection of the Results section (pg 4). We also added a short explanation of why our "core metabolism" appears to be relatively larger than the "core metabolism" of Vieira et al to the end of the last paragraph of this section. We added definitions of orphans and scope and knowledge gaps to paragraph 1 of the "Process for updating the reconstruction and its content" subsection of the Results Section (pg 2), and added the definitions of root and upstream/downstream gaps to the third paragraph of this subsection (pg 3). In response to Reviewer #2's suggestion, we have added a brief explanation of the NADH dehydrogenase flux split to the last paragraph of the "Metabolic network reconstruction procedure" subsection of the Methods Section (pg 7). The purpose of this flux split is to achieve the realistic P/O ratio of 1.375, and this was already explained in the Supplementary Information in the "Updating the biomass composition and growth requirements" section, and may have been missed by the reviewer. The Supplementary Information and all tables and figures remain unchanged from our previous submission.