Response to Reviewer Comments

We would like to thank both reviewers for their helpful and constructive comments. We have made every effort to address the reviewer comments. A detailed response to each comment is provided below. Throughout this response, reviewer comments are reproduced in blue and our response is typeset in black.

In addition to these responses, we would like to highlight one other change relative to the previous manuscript version: In the previous version, we stated that we found virtually no molecular function annotations for differentially expressed genes. It turns out that this was caused by a bug in the (then) recently released DAVID 6.8 beta. This bug has been fixed, and we have re-run all our analyses with the final DAVID 6.8 release, and we now report molecular functions for differentially expressed genes.

Reviewer #1

The study entitled “The E. coli molecular phenotype under different growth conditions” by Caglar et al. describes the production and comparison of transcriptomes and proteomes from multiple different conditions including growth phase and carbon source over a 2 week period. The dataset is huge and would be an excellent resource for the community, however the description of the comparative analysis in the current version of the manuscript is lacking and prevents a complete understanding of the broader implications of the dataset.

We would like to thank the reviewer for the overall positive evaluation of our work.

1. The authors have indicated that the data is being submitted to NCBI. This is a critical feature of this manuscript and the data should be public already according to NIH data release policies. The reviewer should be able to validate the authors findings and not have to just “trust” them that this is correct, or the analyses are completed correctly.

We agree that all data must be freely available and deposited in the appropriate repositories, and we had already provided all our processed data and analysis scripts as part of the accompanying github repository located at:
https://github.com/umutcaglar/ecoli_multiple_growth_conditions
In fact, we would like to highlight Reviewer #2’s overall evaluation of our work, which includes the following statement:
“The R code is admirably available and readable [...] and the data are readily available.”

When we submitted the present manuscript for the first time, we were still in the process of submitting raw RNA reads and protein ms data to the appropriate repositories. This step has now been completed and we provide the accession numbers in the revised manuscript, in subsection “Statistical analysis and data availability” in Methods.

2. Additionally, it is unclear how authors have published part of this dataset previously, but there are no accession numbers (page 2 last line)

We assumed that referring to the relevant publication (Houser et al., PLOS Comput. Biol. 11:e1004400, 2015) was sufficient, since the prior accession numbers are provided in that paper. However, we now also explicitly provide those accession numbers in the Methods section of the present manuscript, subsection “Statistical analysis and data availability”.

3. and how is this dataset any different from the previous one? This feels a bit like the authors are double dipping on the data.

Our present manuscript presents results from seven different experiments (Figure 1). One of these seven experiments, the glucose time course, was previously published (Houser et al., PLOS Comput. Biol. 11:e1004400, 2015), while the other six experiments are entirely new. Thus, the vast majority of the data presented is new. We only include the previously published data in our analysis because it provides a useful baseline for the new experiments. Moreover, in Houser et al. we focused on differences between exponential and stationary phase, whereas here we purposefully exclude that comparison and focus on differences between carbon sources and ion concentrations.

4. The lack of any detail on the analysis other that to indicate it is done the same way as the previous paper and dataset is lacking and needs to be expanded, since these are novel datasets.

For completeness, we have added all experimental details from the previous paper into the Methods section of this paper.

5. It must be clearly defined what are the comparators. Are the 4279 genes the total number of genes in this genome and why only 4201, why is there a 78 gene:protein discrepancy? Also there are 152 RNA samples, but 105 proteome samples – again why is there a discrepancy? Additionally, why does Figure 3 have 143/152 samples and likewise Figure 4 has only 101/105 samples? It is fine to have a difference, but there should be some explanation why these appear to be incomplete datasets?

Please see our answers to these various questions below:

- Are the 4279 genes the total number of genes in this genome and why only 4201, why is there a 78 gene:protein discrepancy?
  Thank you for catching this. Different people on our team carried out the RNA analysis and the protein analysis, and they used different reference genomes, which have minor differences in the genes that they define. To resolve this issue, we have now carried out all downstream analyses only on the 4196 mRNAs and proteins that match exactly between these two reference genomes. Thus, our final analysis is now on 4196 genes. All figures and tables have been updated accordingly.

- Also there are 152 RNA samples, but 105 proteome samples
  As shown in Figure 1, we did not perform proteomics on all samples. See also comment 4 of reviewer 2. In a nutshell, because the proteomics experiments are both laborious and expensive, we did not perform proteomics for the low-magnesium experiment.

- Additionally, why does Figure 3 have 143/152 samples and likewise Figure 4 has only 101/105 samples?
  Thank you for pointing this out. These were numbers mistakenly carried over from an earlier version of the manuscript. We have 152 mRNA and 105 protein samples, and the numbers have been updated accordingly.

6. The inclusion of the mean cophenetic distance between the pairs as a metric for comparison requires more references, or more testing on these types of data to ensure that it is a robust method. What is this usually used for? How has it been implemented in the past and why can/should it be applied here in this case?

The procedure we are following here is entirely standard in statistics: We define a statistic of interest (here the cophenetic distance) and then determine the null-distribution of this statistic by resampling. Specifically, we are carrying out a random permutation test, since we are evaluating a random sample of all possible permutations of labels in our dataset. Random permutation tests are robust, non-parametric tests that have been widely studied and applied in many different contexts.

We employ the cophenetic distance as our test statistic because it is the most commonly used quantity to assess how similar two objects have to be to be grouped in the same cluster. In fact, it is widely used in applications when one wants to compare alternative clusterings of the same data (Sokal and Rohlf 1962, Gan et al. 2007).

In the revised manuscript, we have added these explanations and the corresponding references.

R. R. Sokal and F. J. Rohlf (1962) The Comparison of Dendrograms by Objective Methods. Taxon 11:33–40.

G. Gan, C. Ma, J. Wu (2007) Data Clustering: Theory, Algorithms, and Applications. SIAM, Philadelphia, PA.

7. The suggestion that there are significant batch effects, especially in the protein data, casts shadow of doubt over the complete dataset.

Batch effects are an unfortunate fact of life in any large-scale experimental effort. The only way to prevent batch effects is to run all samples on the same machine at the same time, and this is simply not possible for a dataset comprised of a hundred or more samples.

Therefore, the next best alternative we see is to be honest and open about batch effects, and to make an effort to correct for batch effects in all downstream analyses.

8. The section on differentially expressed genes lacks any detail as to which genes are altered and there is no suggestion that the authors are going to provide this information in any useable form to the reader. Again there is the suggestion that one should just trust the authors, but this lack of detail prevents the validation or comparisons to established datasets that mimic these conditions with other isolates.

Please note that all these results were already provided as part of the github repository that contains all of our processed data and analysis scripts. However, admittedly, they were not easily discoverable.

We have made two revisions to address this issue:
1. We now provide supplementary tables that list (i) the DeSeq output for all genes (Table S8), (ii) the DeSeq output for the significantly differentially expressed genes (Table S9), and the DAVID results (Table S10).

2. We have improved the documentation in our github repository so that relevant files can be found more easily.

9. The baseline is suggested to be the glucose with Na @ 5mM and Mg @ 0.8 mM, but in which growth phase? Was each growth phase interrogated independently? This lack of detail prevents proper evaluation of the manuscript.

The reviewer is correct. This is the baseline, and we analyze exponential and stationary phase separately. We omit a comparison of exponential to stationary phase, since this was the main topic of our earlier publication (see also our response to your comment #3).

We thought we had previously explained this clearly, in the first paragraph of section “Identification of differentially expressed genes”. In the revision, we have copy-edited this paragraph for clarity.

10. It appears as though the doubling times are all based on calculations and never actually verified either in the current study or the previous study. It would be good to determine if the math reflects the biology in any way.

We are not entirely sure which aspect of our procedure the reviewer is concerned about. Doubling times were measured by experimentally determining the change in optical density at 600 nm (OD_{600}) during the exponential part of the growth curve. This is the standard approach to measure doubling times in *E. coli*.

In the revised manuscript, we have copy-edited the corresponding paragraph in the Methods to explain more clearly what was done.

11. The authors must define “experimental noise” – page 8 last paragraph.

Agreed. What we mean is any nonspecific, random measurement error that causes unbiased variation in the counts of individual RNAs or proteins. We have revised the paragraph accordingly.

12. There are many suggestions throughout the manuscript that they are managing the variability in the data, but very little or no detail is actually provided on how they are doing this.

We are not sure what the reviewer is referring to. A pointer to specific issues where details are lacking would be helpful. In any case, we have carefully re-read the entire manuscript and added clarifications where we thought they were needed. We also now provide a step-by-step explanation of how our computational pipeline is run, in file “instructions_for_pipeline.docx” in the github repository.

Reviewer #2

Caglar et al. provide an extensive dataset that includes RNA-Seq, proteomics, and growth rates under 34 different conditions, and also central metabolic fluxes under a subset of conditions including carbon sources, salt stresses, and multiple timepoints. The authors identified strong batch effects in their proteomics data and also to a lesser extent in RNA-Seq. To correct for batch effects, the authors included batch number as a predictor variable in DESeq2, which was used for normalization and differential expression analysis in both RNA-Seq and proteomics. The R code is admirably available and readable. Other than some further clarification on removal of batch effects in their analysis and other minor comments, the article is well written and the data are readily available.

We would like to thank the reviewer for the overall positive evaluation of our work.

1. Page 7: “we corrected for batch effects by including batch as a distinct variable in the DESeq model (see Methods), as recommended.” This approach seems reasonable. However, please do include the citation that “recommended” this approach. Is it the DESeq2 manual that recommends this approach?

Yes, indeed, the DESeq2 manual recommends this approach, and we now state this explicitly.

2. Global expression changes are known to be associated with growth rate (Klumpp et al., 2009, doi:10.1016/j.cell.2009.12.001). Therefore, for the measurements taken under exponential growth, would the differential expression calls change if growth rate is also a separate variable in DESeq, thus removing the global growth-associated differential expression?

This is a good question. We have run this additional analysis and have added the results to the manuscript.

We have found that most of the time the change is minimal; i.e. most of the significantly changed genes are the same whether or not we include a term for doubling time in the analysis. The main exception is protein data for different carbon sources. Many new genes show up as significantly differentially expressed when controlling for doubling time in those comparisons (see new Supplementary Figure 34). The main pathways associated with these genes are the same for the various conditions and are related to biosynthesis for both exponential and stationary phases (see new Supplementary Tables S11 and S12)

3. Please cite Kim et al. (doi:10.1038/ncomms13090). In particular, Kim et al. developed a normalization pipeline that removes batch effects. Does this normalization pipeline apply to the

authors' own data?

Thank you for pointing us to this paper. We now cite it in the discussion as an alternative approach to analyze datasets of the magnitude and complexity we are dealing with here.

4. Figure 1: The low and high magnesium .08mM Mg conditions appear to be the same, but the bottom box shows 3 proteomics measurements for both exponential and stationary phases and the top box shows none. Is the bottom box missing the proteomics symbols or are these conditions different?

As the reviewer notes the conditions are the same but the experiments are different. The high-magnesium experiment was performed first, and it was originally thought of just as an experiment of varying magnesium conditions, including concentrations above and below the base level. The low-magnesium experiment was subsequently performed to probe particularly low magnesium concentrations, and it included a repeat of the 0.08mM Mg condition.

Because the proteomics experiments are both laborious and expensive, we ultimately decided to not perform proteomics on the low-magnesium experiment.

We now point out in the caption to Figure 1 that conditions that are listed multiple times represent independent replicates of those conditions. We have also switched the high and low magnesium experiments in the figure, so that the one with protein data is listed first, and we explicitly state that proteomics was not done for the low-magnesium experiment.

5. Figure 5: Please define the abbreviations (is "glc" glucose or gluconate?)

We now define these abbreviations in the caption to Figure 5.

6. Figure 8: What is the baseline dataset for these differential expression results?

The baseline datasets are explained in the first paragraph of the subsection "Identification of differentially expressed genes" in Results. See also our response to comment 9 of Reviewer #1. In the revised manuscript, we have changed the caption of Fig. 8 as well to clearly mention the control datasets.

7. In the abstract, it is worth stating that growth rates and metabolic fluxes (for a subset of conditions) were also measured.

We have made this change.

8. Following from the comment above, the sentence in the abstract: “systematic measurements of both mRNA and protein abundances under a wide range of different conditions are still relatively rare” could be revised to emphasize that RNA-Seq, proteomics, growth rates and metabolic fluxes were measured in different conditions.

We have made this change.