The *E. coli* molecular phenotype under different growth conditions

1. Mehmet U. Caglar[1, 2, 3]*, 2. John R. Houser[3, 4, 5], 3. Craig S. Barnhart[3, 4], 4. Daniel R. Boutz[3, 4, 5], 5. Sean M. Carroll[6, 7], 6. Aurko Dasgupta[3, 4, 8], 7. Walter F. Lenoir[5], 8. Bartram L. Smith[1, 2, 3], 9. Viswanadham Sridhara[2, 3], 10. Dariya K. Sydykova[1, 2, 3], 11. Drew Vander Wood[3, 5], 12. Christopher J. Marx[9, 10], 13. Edward M. Marcotte[3, 4, 5]*, 14. Jeffrey E. Barrick[3, 4, 5]*, 15. Claus O. Wilke[3, 4, 5]*

[1]Department of Integrative Biology, The University of Texas at Austin, Austin, Texas, USA
[2]Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, Texas, USA
[3]Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, Texas, USA
[4]Center for Systems and Synthetic Biology, The University of Texas at Austin, Austin, Texas, USA
[5]Department of Molecular Biosciences, The University of Texas at Austin, Austin, Texas, USA
[6]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, USA
[7]Axcella Health Inc, Cambridge, Massachusetts, USA
[8]Center for Women's Infectious Diseases Research, Division of Infectious Diseases, Department of Internal Medicine, Washington University School of Medicine, St. Louis, Missouri, USA
[9]Department of Biological Sciences, University of Idaho, Moscow, Idaho, USA
[10]Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, Idaho, USA

*Corresponding author: umut.caglar@gmail.com (MUC); marcotte@icmb.utexas.edu (EMM); jbarrick@cm.utexas.edu (JEB); wilke@austin.utexas.edu (COW)

Abstract

Modern systems biology requires extensive, carefully curated measurements of cellular components in response to different environmental conditions. While high-throughput methods have made transcriptomics and proteomics datasets widely accessible and relatively economical to generate, systematic measurements of both mRNA and protein abundances under a wide range of different conditions are still relatively rare. Here we present a detailed, genome-wide transcriptomics and proteomics dataset of *E. coli* grown under 34 different conditions. We manipulate concentrations of sodium and magnesium in the growth media, and we consider four different carbon sources glucose, gluconate, lactate, and glycerol. Moreover, samples are taken both in exponential and stationary phase, and we include two extensive time-courses, with multiple samples taken between 3 hours and 2 weeks. We find that exponential-phase samples systematically differ from stationary-phase samples, in particular at the level of mRNA. Regulatory responses to different carbon sources or salt stresses are more moderate, but we find numerous differentially expressed genes for growth on gluconate and under salt and magnesium stress. Our data set provides a rich resource for future computational modeling of *E. coli* gene regulation, transcription, and translation.

Introduction

A goal of systems biology has been to understand how phenotype originates from genotype. The phenotype of a cell is determined by complex regulation of metabolism, gene expression, and cell signaling. Understanding the connection between phenotype and genotype is crucial to understanding disease and for engineering biology[1]. Computational models are particularly well suited to studying this problem, as they can synthesize and organize diverse and complex data in a predictive framework, but detailed experimental studies including many samples are needed to understand interactions between different types of omics data[2]. Much effort is currently being spent on understanding how to best integrate information collected about multiple cellular subsystems that is derived from different types of high-throughput measurements. For example, there are many proposed approaches for relating gene expression and protein abundances, focusing on integrative, whole-cell models[2–5].

Given the growing interest in integrative modeling approaches, there is a pressing need for high quality genome-scale data that is comparable across cellular subsystems and reflects many different external conditions. *E. coli* is an ideal organism to study genome-wide, multi-level regulatory effects of external conditions, since it is well adapted to the laboratory environment[6] and was one of the first organisms studied at the whole-genome level[7]. There have been a number of studies of the *E. coli* transcriptome and/or proteome in response to different growth conditions. For example, in cells growing at high density, expression of most amino acid biosynthesis genes is down-regulated and expression of chaperones is up-regulated, suggesting stresses that these cells experience[8]. Exposure of *E. coli* to reduced temperature leads to changes in gene-expression patterns consistent with reduced metabolism and growth[9]. Under long-term glucose starvation, mRNAs are generally down-regulated while the protein response is more varied[10]. Specifically, the copy numbers of proteins involved in energy-intensive processes decline whereas those of proteins involved in nutrient metabolism remain constant, likely to provide the cell with the ability to jump-start metabolism when nutrients become available again. A few other larger-scale studies have measured mRNA and/or protein abundances under multiple conditions[11–14].

Here, we provide a systematic analysis of *E. coli* gene expression under a wide variety of different conditions. We measure both mRNA and protein abundances, at exponential and stationary phases, for growth conditions including different carbon sources and different salt stresses. We find that mRNAs and proteins display divergent responses to the different growth conditions. Further, growth phase yields more systematic differences in gene expression than does either carbon source or salt stress, though this effect is more pronounced in mRNAs than in proteins. We expect that our data set will provide a rich resource for future modeling work.

Results

Experimental design and data collection

We grew multiple cultures of *E. coli* REL606, from the same stock, under a variety of different growth conditions. We measured RNA abundances under all conditions and matching protein abundances for approximately 2/3 of the conditions (Figure 1 and Supplementary Table S1). We also measured central metabolic fluxes for a subset of conditions using glucose as carbon source (Supplementary Table S1). Results from one of these conditions, long-term glucose starvation, have been presented previously[10]. Conditions not previously described include one

additional starvation experiment, using glycerol instead of glucose as carbon source, exponential and stationary phase cultures using either gluconate or lactate as carbon source, and conditions varying Mg^{2+} and Na^{+} concentrations.

Measurements of RNA and protein abundances were carried out as previously described[10]. All resulting data sets were checked for quality, normalized, and log-transformed. Our final data set consisted of 152 RNA samples, 105 protein samples, and 65 flux samples (Supplementary Table S1) 59 of the flux samples are associated with high Mg^{+2} and high Na^{+} experiments.

Our raw RNA-seq data covers 4279 distinct mRNAs, our protein data covers 4201 distinct proteins, and our flux data covers 13 different metabolic reactions. All raw data files are available in appropriate repositories (see Methods for details), and final processed data are available as Supplementary Tables S2, S3, and S4.

Finally, we measured growth rates in exponential phase for all experimental conditions. We found that doubling times varied between 50 and 100 minutes among the various conditions (Figure 2). Growth was the fastest when glucose was used as carbon source and the slowest when the carbon source was lactose. Growth was also reduced for high Na^{+} concentrations and very high or low Mg^{2+} concentrations. Surprisingly, we found a broad range of Mg^{2+} concentrations (0.02mM to 200mM) in which growth rate remained virtually unchanged (Figure 2).

Broad trends of gene expression differ between mRNA and proteins

To identify broad trends of gene expression among the different growth conditions, we performed hierarchical clustering on both mRNA and protein abundances (Figures 3 and 4). For mRNA, we found that differences in gene expression were primarily driven by the growth phase (exponential vs. stationary/late stationary). Nearly all exponential samples clustered together in one group, separate from the vast majority of stationary and late-stationary samples (Figure 3). Mg^{2+} levels, Na^{+} levels, and carbon source had less influence on the clustering results. We also found a similar result for protein abundances (Figure 4). The exponential-phase samples grouped together, separated from stationary and late stationary samples. Similarly, Na^{+} levels and carbon sources also seemed to be grouped together upon clustering.

To quantify the clustering patterns of mRNA and protein abundances, we defined a metric that measured how strongly clustered a given variable of the growth environment (growth phase, Mg^{2+} level, Na^{+} level, carbon source) was relative to the random expectation of no clustering. For each variable, we calculated the mean cophenetic distance between all pairs corresponding to the same condition (e.g., for growth phase, all pairs sampled at exponential phase and all pairs sampled at stationary/late stationary phase). The cophenetic distance is defined as the height of the dendrogram produced by the hierarchical clustering from the two selected leafs to the point where the two branches merge. We then converted each mean cophenetic distance into a *z*-score, by resampling mean cophenetic distances from dendograms with reshuffled leaf assignments. A *z*-score below −1.96 indicates that the mRNA or protein abundances are clustered significantly by the corresponding variable.

We found that mRNA abundances were significantly clustered by growth phase, with a *z*-score of −23.99 (Table 1). Na^+ and Mg^{2+} levels displayed the next-largest *z*-scores by magnitude, of −1.54 and −1.46, but these were not significantly different from zero. The *z*-score for carbon source was 1.16, which implies that there is no significant clustering by carbon source in the mRNA data. Importantly, when we calculated a *z*-score for batch number, we found that batch effects also significantly influenced mRNA abundances, with $z = -2.82$. Batch number here represents cultures grown at the same time, in parallel. Thus, batch effects may represent fluctuations in incubator temperatures, slight differences in growth medium composition or water quality, or effects of reviving the initial inoculum of cells, among other possibilities.

For protein abundances, the variables Na^+ level, growth phase, and carbon source were all significantly clustered, with *z*-scores of −4.78, −4.21, and −3.15, respectively (Table 1). Batch number had a *z*-score of −23.29, which implies that there were strong batch effects present in the protein data.

In summary, the largest effect in mRNA abundances, growth phase, was similarly present in proteins. However, protein abundances clustered also by Na^+ and carbon source, effects that were not present in the mRNA data. Finally, both mRNA and protein data were influenced by batch effects, and the effect was much more pronounced for proteins than for mRNA (Table 1).

Identification of differentially expressed genes

We next asked under which conditions and to what extent RNA and protein expression were altered. To identify differentially expressed mRNAs and proteins, we used DESeq2[15]. Since our data clustered significantly by growth phase, we analyzed RNA and protein expression separately for exponential and stationary phase. For each growth phase, we defined the reference condition to be glucose as carbon source, with 5 mM Na^+ and 0.8 mM Mg^{2+}. This is the baseline formulation of media used in the glucose time-course samples[10]. We then compared RNA and protein abundances between this reference condition and the alternative conditions (different carbon sources, elevated Na^+, and elevated or reduced Mg^{2+}). Note that a detailed comparison of reference exponential phase vs. reference stationary phase has already been published[10].

We defined significantly differentially expressed genes as those whose abundance had at least a two-fold change ($\log_2$ fold change > 1) between the reference condition and a chosen experimental condition, at a false-discovery-rate (FDR) corrected *P* value < 0.05. We found that the number of significantly differentially expressed mRNAs and proteins varied substantially between exponential and stationary phase and between mRNAs and proteins (Figure 5). In general, there were fewer differentially expressed genes in stationary phase than in exponential phase. Further, protein abundances showed the most differential regulation for high Na^+ and for the carbon sources glycerol and lactate, whereas mRNA showed the most differential regulation for high Na^+ levels in stationary phase, and for low Mg^{2+} levels and for the carbon sources glycerol and lactate in exponential phase (Figure 5).

Next, we asked how much overlap there was among differentially expressed genes between the various growth conditions. To simplify this analysis, we did not distinguish between up- or down-regulated genes, and we combined low and high Mg^{2+} into one group “Mg stress” and glycerol, lactate, and gluconate into one group “carbon source”. (Note that differentially expressed genes were still identified for individual conditions, as described above, and were

combined into "Mg stress" and "carbon source" only for the final comparison.) At the mRNA level, there was some overlap (21.7%) between carbon source and Mg^{2+} stress in exponential phase. All other overlaps where minimal, ~5% or less (Figure 6). At the protein level, there was overlap between Na^{+} stress and carbon source (14.9% in exponential phase, 10.7% in stationary phase), while all other overlaps were also minimal, ~3% or less (Figure 6).

We also identified significantly altered biological pathways and molecular activities of gene products. We use the Kyoto Encyclopedia of Genes and Genomes (KEGG)[16] for biological pathways and annotations from the Gene Ontology (GO) Consortium for molecular functions[17]. Figure 7 and Supplementary Figure 1 show the top 5 significantly altered biological pathways (as defined in the KEGG database) and molecular functions (as defined by GO annotations) under different conditions, respectively, as determined by DAVID[18]. In all cases, we used a cutoff of 0.05 on false-discovery-rate (FDR)-corrected *P* values to identify significant annotations. We found numerous significantly altered KEGG pathways (Figure 7) but only two significantly altered GO annotations: "structural constituent of ribosome" and "structural molecule activity" (Figure S1).

Finally, we looked at individual, differentially expressed genes associated with specific pathways and/or functions (Supplementary Figures 2–22). As an example, the differentially expressed mRNAs associated with significantly altered KEGG pathways under high Mg^{2+} concentrations in exponential phase are shown in Figure 8A. Three pathways are significantly altered; sulfur metabolism and nitrogen metabolism are mostly up-regulated and flagellar assembly is mostly down-regulated. Changes in sulfur metabolism in this condition might reflect a linked increased in the concentration of sulfate (SO_4^{2-}), as this was the counterion in the salt that was added to increase Mg^{2+} levels. By contrast, using lactate instead of glucose as carbon source caused up-regulation of pyruvate metabolism, citrate cycle, and carbon metabolism at the protein level in exponential phase (Figure 8B).

Metabolic flux ratios under salt stress

For the high sodium and high magnesium experiments, we also determined metabolic flux through central metabolism by analyzing ^{13}C incorporation into protein-bound amino acids. We here analyzed only flux samples taken in exponential phase, since stationary-phase samples have an unclear interpretation[10]. For each condition, flux samples were analyzed in triplicate (except one, which was analyzed in duplicate only), and 13 different flux ratios were measured for each sample. The flux ratios were then averaged across replicates (Supplementary Figure 23). We saw no significant changes in flux ratios with increasing Na^{+} (linear regression, all $P > 0.05$ after FDR correction, Supplementary Table 5). Results were similar for Mg^{2+}. Due to the wide range of Mg^{2+} concentrations considered, we regressed flux ratios against log-transformed Mg^{2+} concentrations. Again, we saw no significant changes in any flux ratio with increasing Mg^{2+} (linear regression, all $P > 0.05$ after FDR correction, Supplementary Table 5).

We also asked whether the flux ratios changed with doubling time rather than with ion concentration, since doubling time is not necessarily monotonic in ion concentration (Figure 2B). For this analysis, we pooled all flux measurements and plotted flux ratios against doubling times (Figure 9). Again, we saw no significant relationship between flux ratios and doubling time after FDR correction (Supplementary Table 6). However, we note that the branches erythrose-4-phosphate from pentose-5-phosphate and pyruvate from malate (upper bound)

showed a significant relationship before correction for multiple testing (P = 0.026 and P = 0.018, respectively, Supplementary Table 6), both driven by one outlying data point for the slowest-growing condition, at 300 mM Na^+.

Discussion

We studied the regulatory response of *E. coli* under a wide variety of different growth conditions. The experimental conditions we considered include four different carbon sources, different levels of Na^+ and Mg^{2+} stress, and growth into deep stationary phase, up to two weeks post inoculation. We found that gene regulation changes the most with respect to growth phase; in general, the exponential phase under one condition is more similar to the exponential phase under another condition than to the stationary phase under the same condition. Further, we found little overlap in differentially expressed genes under different growth conditions. Finally, we found that the ratios of fluxes through alternative branches within central metabolism remained approximately constant under salt stress, despite substantial changes in doubling times.

Our data provides a comprehensive picture of *E. coli* in terms of number, range, and depth of different stresses, comparable and complementary to other recently published datasets. For example, Schmidt *et al.*[12] considered 22 unique conditions and measured abundances of >2300 proteins. mRNA abundances were not measured. Soufi *et al.*[11] considered 10 unique conditions and also measured abundances of >2300 proteins. They were interested primarily in up- and down-regulated proteins under different ethanol stresses, and they found down-regulation of genes associated with ribosomes and protein biosynthesis during ethanol stress. Such genes were similarly down-regulated in our study during stress induced by high Na^+ concentrations. Lewis *et al.*[13] considered only 3 different carbon sources but measured mRNA and protein abundances in different strains adapted to these growth conditions. Finally, Lewis *et al.*[14] compiled a database of 213 mRNA expression profiles covering 70 unique conditions, including different carbon sources, terminal electron acceptor, growth phase, and genotype. In comparison, we considered 34 unique conditions, measured 152 mRNA expression profiles, 105 protein expression profiles, and 59 flux profiles, and used the exact same *E. coli* genotype throughout.

Similar to our prior study[10], we observed clear trends in the differential expression of mRNAs and proteins. In particular, we had reported previously[10] that mRNAs are widely down-regulated in stationary phase whereas only select proteins are down-regulated. Consistent with that observation, we found here that mRNAs were significantly and strongly clustered by growth phase (z = −23.21) whereas proteins were not (z = −1.26). By contrast, at the protein level we saw significant clustering by carbon source (z = −2.80), which we did not see at the mRNA level. More specifically, we had found earlier[10] that energy-intensive processes were down-regulated and stress-response proteins up-regulated in stationary phase. Similarly, we observed here that high Na^+ stress conditions also led to the down-regulation of energy-intensive processes.

A number of genes and pathways that we found to be influenced by treatment conditions are consistent with prior knowledge from the literature. For instance, we found that increasing the

concentration of Na^+ and Mg^{2+} decreased transcription of the flagellar genes during exponential growth, as seen previously[19]. We also found that high concentrations of Mg^{2+} induce an increase in mRNA expression of sulfur and nitrogen transport proteins, and an increase in the enzymes necessary to produce the siderophore enterobactin (necessary for obtaining iron from the environment). These regulatory changes could be due to the high Mg^{2+} concentrations interfering with the bacterial membrane potential, and thereby inhibiting cotransporters that are coupled to this gradient. This effect has been previously described for iron[20]. High Na^+ concentrations also significantly reduced the expression of a large number of proteins, mostly either involved in the biosynthesis of amino acids or components of the ribosome. These changes may simply reflect stress induced by the high Na^+ concentrations used in these experiments.

Altering the carbon source, as well, provided predictable changes in gene expression. For instance, providing glycerol as the sole carbon source instead of glucose increases expression of *glpX*, part of the *glp* operon, which is involved in glycerol uptake[21]. Gluconate as a carbon source increases expression of genes from the *gnt* and *idn* operons, both involved in gluconate metabolism[22,23]. Finally, using lactate as a carbon source induces the expression of *lldD (lctD)*, a gene required for lactate utilization in *E. coli*[24].

Large-scale, high-throughput gene-expression studies are frequently confounded by batch effects that can give rise to incorrect conclusions if they are not accounted for[25]. We saw such effects in our study as well. In our data, the batch number indicates bacterial samples that were grown at the same time. Not unexpectedly, our data showed significant clustering by batch number, and more so in protein data than in mRNA data (z scores of −20.54 and −2.11, respectively). Batch effects are not inherently a problem, as long as one is aware of their existence and analyzes data accordingly. Here, in our differential expression analysis, we corrected for batch effects by including batch as a distinct variable in the DESeq model (see Methods), as recommended. How to best correct for batch effects is a topic of ongoing investigations, and increasingly sophisticated methods are being developed to separate batch effects from real signal[26–29].

Given the many cellular changes observed in mRNA and protein levels, we turned to ^{13}C labeling techniques[10,30,31] to examine the extent to which these changes affected the relative flux of metabolites through different central metabolic pathway branch points during exponential growth. For this work, we concentrated upon growth on glucose during Na^+ and Mg^{2+} stresses. Across these conditions, growth rates change over nearly a two-fold range, with the doubling time changing from approximately 50 to 95 minutes. In particular, both high Na^+ and high Mg^{2+} levels reduced growth by a third. Despite this substantial effect on growth, we observed no significant changes in the relative flux through different reactions in central metabolism. The only exception was a potential decrease in pentose-5-phosphate pathway use and increase in flow through malic enzyme at 300 mM Na^+. The general picture, however, was that homeostasis in central metabolism was sufficient to ward off significant changes in relative pathway use despite large changes in overall growth rate and the pools of mRNA and proteins.

In summary, our study provides a large and comprehensive dataset for investigating the gene-regulatory response of *E. coli* under different growth conditions, both at the mRNA and the protein level. We found systematic differences in gene-expression response between

exponential and stationary phase, and between mRNAs and proteins. Our dataset provides a rich resource for future modeling of *E. coli* metabolism.

Materials and Methods

Cell growth, RNA-seq, proteomics, and metabolic flux measurements

Growth and harvesting of *E. coli* B REL606 cell pellets for the multiomic analysis was performed as previously described[10], with the following additional details. For tests of different carbon sources, the Davis Minimal (DM) medium used was supplemented with 0.5 g/L of the specified compound (glycerol, lactate, or gluconate) instead of glucose. Mg^{2+} concentrations were varied by changing the amount of $MgSO_4$ added to DM media from the concentration of 0.83 mM that is normally present. For tests of different Na^+ concentrations, NaCl was added to achieve the final concentration. The base recipe for DM already contains ~5 mM Na^+ due to the inclusion of sodium citrate, so 95 mM NaCl was added for the 100 mM Na^+ condition, for example. Exponential-phase samples were taken during growth when the OD_{600} reached 20-60% of the maximum achieved after saturating growth. Stationary phase samples were collect 20-24 hours after the corresponding exponential sample. The exact sampling times for each condition are provided in Supplementary Table S1.

After sample collection, RNA-seq, mass-spec proteomics, and metabolic flux analysis were performed exactly as described[10].

Doubling times were estimated from OD_{600} measurements. Specifically, the logarithms of all OD_{600} values in the exponential part of each growth curve, defined as when OD_{600} values were between 0.05 and 0.75 times the maximum observed OD_{600} at stationary phase, were fit to a linear model with respect to time. Doubling times were calculated as $\log_e 2$ divided by the fit slope for each biological replicate separately. Means and confidence intervals were calculated from three replicate growth curves for all conditions except for gluconate and lactate, which had measurements for only two replicates.

Normalization and quality control of RNA and protein counts

Our raw input data consisted of RNA and protein counts. Protein counts can be fractional, because some peptide spectra cannot be uniquely mapped to a single protein, so they are equally divided amongst these proteins. We rounded all protein counts to the nearest integer for subsequent analysis. We set the counts of all unobserved proteins to zero. For RNA, we only analyzed the counts of reads that overlapped annotated protein coding genes (mRNA counts). Subsequently, all mRNA and protein counts were analyzed in the same manner.

We next performed quality control, by checking replicates of the same condition for consistency. For all pairs of replicate samples, we made histograms of the log-differences of RNA or protein counts. If the two samples differ only by experimental noise, then the resulting histogram should have a mode at 0 and be approximately bell-shaped. If a sample consistently shows deviations from this expectation when compared to other samples, then there are likely systematic problems with this sample. We tested the quality of our mRNA and protein samples by looking the similarity between samples collected in similar conditions but from different

batches whenever possible, i.e., whenever we have at least 3 replicates. Out of 152 mRNA samples we found only two samples (samples MURI_091 and MURI_130, Supplementary Table 1) that seemed to deviate from their biological replicas. Among 105 protein samples we found no major deviation between biological replicas. Because of this broad consistency among all samples for the same growth conditions, we keep all samples for subsequent analysis.

After quality control, we normalized read counts using size-factors calculated via DESeq2[15]. Because we had many mRNAs and proteins with counts of zero at some condition, we added pseudo-counts of +1 to all counts before calculating size factors. We then used those size factors to normalize the original raw counts (i.e., without pseudo-counts).

Clustering

We clustered normalized mRNA and protein counts based on their Euclidian distance, using the complete linkage method implemented in the `flashclust`[32] package, which is a faster implementation of the `hclust` function in R. This method defines the cluster distance between two clusters as the maximum distance between their individual components[33]. At every stage of the clustering process, the two closest clusters are merged into the next bigger cluster. The final outcome of this process is a dendogram that measures the closeness of different samples to each other.

To assess whether the clustering process significantly grouped similar samples together, we employed a reshuffling test. For any category that we tested for significant clustering (e.g., carbon source, Na stress, or batch number), we calculated the mean cophenetic distance in the clustering dendogram between all pairs belonging to the same level of the categorical variable tested (e.g., same carbon source). We then repeatedly reshuffled the labeling within each category and recalculated the mean cophenetic distance each time. Finally, we calculated *z* scores of the original cophenetic distance relative to the distribution of reshuffled values.

Identifying differentially expressed genes

We used DESeq2[15] to identify differentially expressed mRNAs and proteins across conditions. We used two reference conditions in our comparisons, one for exponential phase and one for stationary phase. The reference conditions always had glucose as carbon source and base Na^+ and Mg^{2+} concentrations. We did not compare exponential phase to stationary phase samples, since this comparison was done in depth previously[10] for samples grown on glucose and with base Na^+ and Mg^{2+} concentrations.

We corrected for possible batch effects by including batch number as a predictor variable in the design formula of DESeq2. In general, our design formula was `~batch_number + variable_of_interest`, where `variable_of_interest` was either a categorical variable representing the carbons source or growth phase (exponential or stationary) or a quantitative variable representing Na^+ level or Mg^{2+} level.

We considered genes as differentially expressed between two conditions if their $\log_2$ fold change was > 1 and their FDR-corrected *P* value < 0.05. We subsequently annotated differentially expressed genes with DAVID[18] version 6.8 Beta released in May 2016. We considered both KEGG pathways[16] and GO annotations[17].

Statistical analysis and data availability

All statistical analyses were performed in R. The relevant R scripts and processed data are available on github: https://github.com/umutcaglar/ecoli_multiple_growth_conditions
Raw RNA reads and peptide spectra are being submitted to the appropriate repositories.

References

1. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33,** 228–237 (2003).
2. Zhang, W., Li, F. & Nie, L. Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiol. Read. Engl.* **156,** 287–301 (2010).
3. Joyce, A. R. & Palsson, B. Ø. The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell Biol.* **7,** 198–210 (2006).
4. Ideker, T. *et al.* Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. *Science* **292,** 929–934 (2001).
5. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13,** 227–232 (2012).
6. Lee, S. Y. High cell-density culture of *Escherichia coli*. *Trends Biotechnol.* **14,** 98–105 (1996).
7. Blattner, F. R. *et al.* The Complete Genome Sequence of *Escherichia coli K-12*. *Science* **277,** 1453–1462 (1997).
8. Yoon, S. H., Han, M.-J., Lee, S. Y., Jeong, K. J. & Yoo, J.-S. Combined transcriptome and proteome analysis of *Escherichia coli* during high cell density culture. *Biotechnol. Bioeng.* **81,** 753–767 (2003).

9. Gadgil, M., Kapur, V. & Hu, W.-S. Transcriptional response of *Escherichia coli* to temperature shift. *Biotechnol. Prog.* **21,** 689–699 (2005).

10. Houser, J. R. *et al.* Controlled Measurement and Comparative Analysis of Cellular Components in *E . coli* Reveals Broad Regulatory Changes in Response to Glucose Starvation. *PLOS Comput Biol* **11,** e1004400 (2015).

11. Soufi, B., Krug, K., Harst, A. & Macek, B. Characterization of the *E. coli* proteome and its modifications during growth and ethanol stress. *Front. Microbiol.* **6,** 103 (2015).

12. Schmidt, A. *et al.* The quantitative and condition-dependent *Escherichia coli* proteome. *Nat. Biotechnol.* **34,** 104–110 (2016).

13. Lewis, N. E. *et al.* Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **6,** 390 (2010).

14. Lewis, N. E., Cho, B.-K., Knight, E. M. & Palsson, B. O. Gene Expression Profiling and the Use of Genome-Scale In Silico Models of *Escherichia coli* for Analysis: Providing Context for Content. *J. Bacteriol.* **191,** 3437–3444 (2009).

15. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).

16. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

17. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).

18. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2008).

19. Shi, W., Li, C., Louise, C. J. & Adler, J. Mechanism of adverse conditions causing lack of flagella in *Escherichia coli*. *J. Bacteriol.* **175,** 2236–2240 (1993).

20. Braun, V., Hantke, K. & Köster, W. Bacterial iron transport: mechanisms, genetics, and regulation. *Met. Ions Biol. Syst.* **35,** 67–145 (1998).

21. Weissenborn, D. L., Wittekindt, N. & Larson, T. J. Structure and regulation of the glpFK operon encoding glycerol diffusion facilitator and glycerol kinase of *Escherichia coli* K-12. *J. Biol. Chem.* **267,** 6122–6131 (1992).

22. Fujita, Y., Nihashi, J. & Fujita, T. The characterization and cloning of a gluconate (gnt) operon of *Bacillus subtilis*. *J. Gen. Microbiol.* **132,** 161–169 (1986).

23. Bausch, C. *et al.* Sequence analysis of the GntII (subsidiary) system for gluconate metabolism reveals a novel pathway for L-idonic acid catabolism in *Escherichia coli*. *J. Bacteriol.* **180,** 3704–3710 (1998).

24. Dong, J. M., Taylor, J. S., Latour, D. J., Iuchi, S. & Lin, E. C. Three overlapping lct genes involved in L-lactate utilization by *Escherichia coli*. *J. Bacteriol.* **175,** 6671–6678 (1993).

25. Gilad, Y. & Mizrahi-Man, O. A reanalysis of mouse ENCODE comparative gene expression data. *F1000Research* **4,** 121 (2015).

26. Chen, C. *et al.* Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS One* **6,** e17238 (2011).

27. Lazar, C. *et al.* Batch effect removal methods for microarray gene expression data integration: a survey. *Brief. Bioinform.* **14,** 469–490 (2013).

28. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostat. Oxf. Engl.* **8,** 118–127 (2007).

29. Luo, J. *et al.* A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics J.* **10,** 278–291 (2010).

30. Zamboni, N., Fendt, S.-M., Rühl, M. & Sauer, U. 13C-based metabolic flux analysis. *Nat. Protoc.* **4,** 878–892 (2009).

31. Zamboni, N., Fischer, E. & Sauer, U. FiatFlux--a software for metabolic flux analysis from 13C-glucose experiments. *BMC Bioinformatics* **6,** 209 (2005).

32. Langfelder, P. & Horvath, S. Fast R Functions for Robust Correlations and Hierarchical Clustering. *J. Stat. Softw.* **46,** (2012).

33. Soni Madhulatha, T. An Overview on Clustering Methods. *ArXiv E-Prints* **1205,** arXiv:1205.1117 (2012).

Acknowledgments

This study was funded by Army Research Office (ARO,http://www.arl.army.mil/) grant W911NF-12-1-0390 to CJM, EMM, JEB, and COW. EMM also acknowledges support from the NIH (DP1 OD009572) and Welch Foundation (F1515). COW also acknowledges support from the NIH (R01 GM088344, R01 AI120560) and the NSF (Cooperative agreement no. DBI-0939454, BEACON Center). The Texas Advanced Computing Center (TACC) at The University of Texas at Austin provided high-performance computing resources.

Contributions

M.U.C., J.H.R., C.J.M., E.M.M., J.E.B., C.O.W. conceived the study and designed the experiments. J.R.H., C.S.B., D.R.B., S.M.C., A.D. performed the experiments. V.S., D.K.S. contributed computer code used for data analysis. M.U.C., J.H.R., W.F.L., B.L.S., V.S., D.V.W., J.E.B., C.O.W. analyzed the data. M.U.C., W.F.L., V.S., D.V.W. prepared the figures. M.U.C., B.L.S., D.V.W., C.O.W. wrote the initial paper draft. All authors reviewed and edited the final manuscript

Competing interests

The authors declare no competing financial interests.

Figures

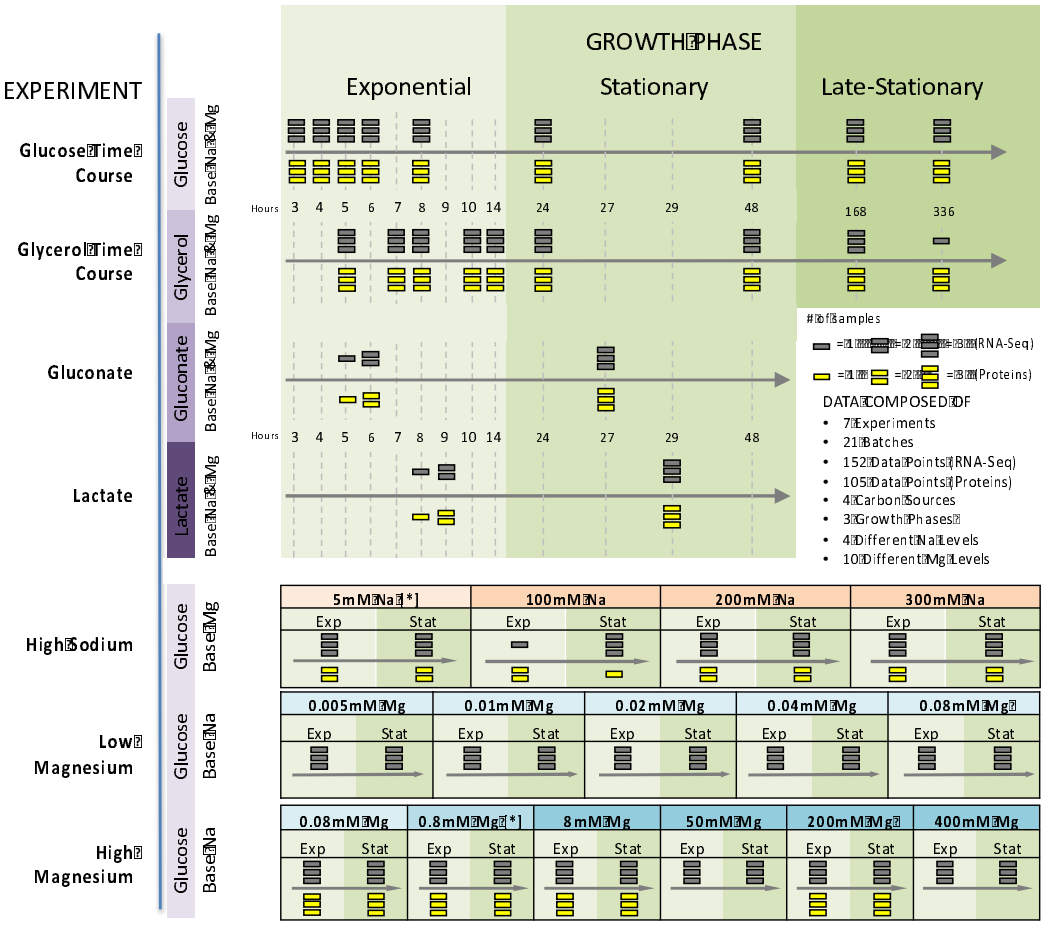

Figure 1: Experimental setup. We performed seven different experiments in which we varied the duration of growth and the temporal density of sampling, the carbon source, and ion concentrations. For each experimental condition, bacteria were grown in three biological replicates. We subsequently performed whole-transcriptome RNA-Seq for all experimental conditions and mass-spec proteomics for the majority of them. We considered four different carbon sources: glucose, glycerol, gluconate, and lactate; we also considered high sodium and both low and high magnesium levels. For the time-course and carbon-source experiments, we used base-level Na^+ (5 mM) and Mg^{2+} (0.8 mM) throughout (indicated by [*] in the sodium and magnesium experiments).

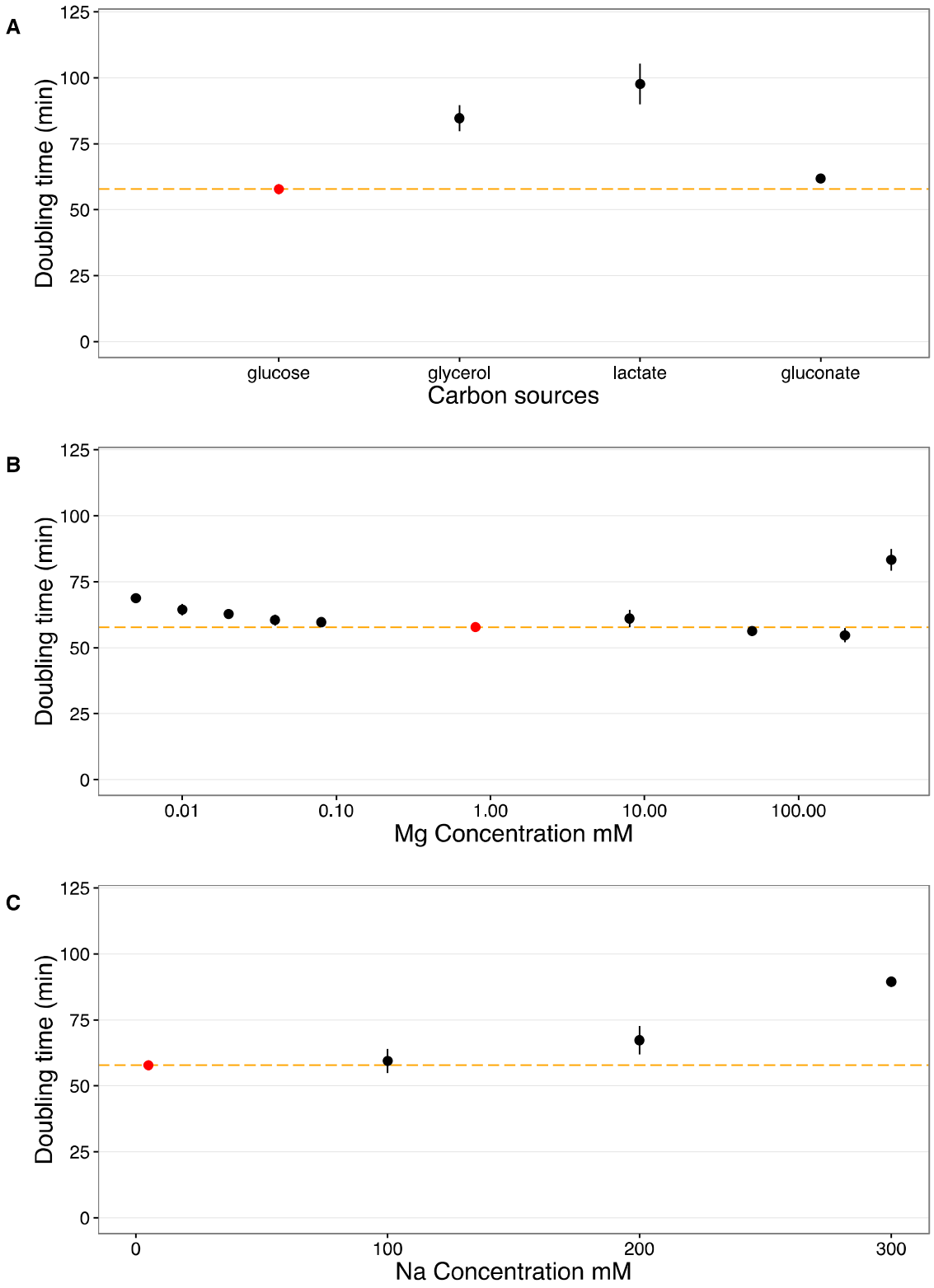

Figure 2: Doubling times under various growth conditions. We measured doubling times under exponential phase for all growth conditions. The red points and dashed orange lines represent the doubling time at the base condition (glucose, 5 mM Na^{+}, 0.8 mM Mg^{2+}). Doubling times were measured in triplicates and error bars represents 95% confidence intervals of the mean.

(A) Doubling times with respect to carbon sources. (B) Doubling times with respect to Mg^{2+} concentrations. (C) Doubling times with respect to Na^{+} concentrations.

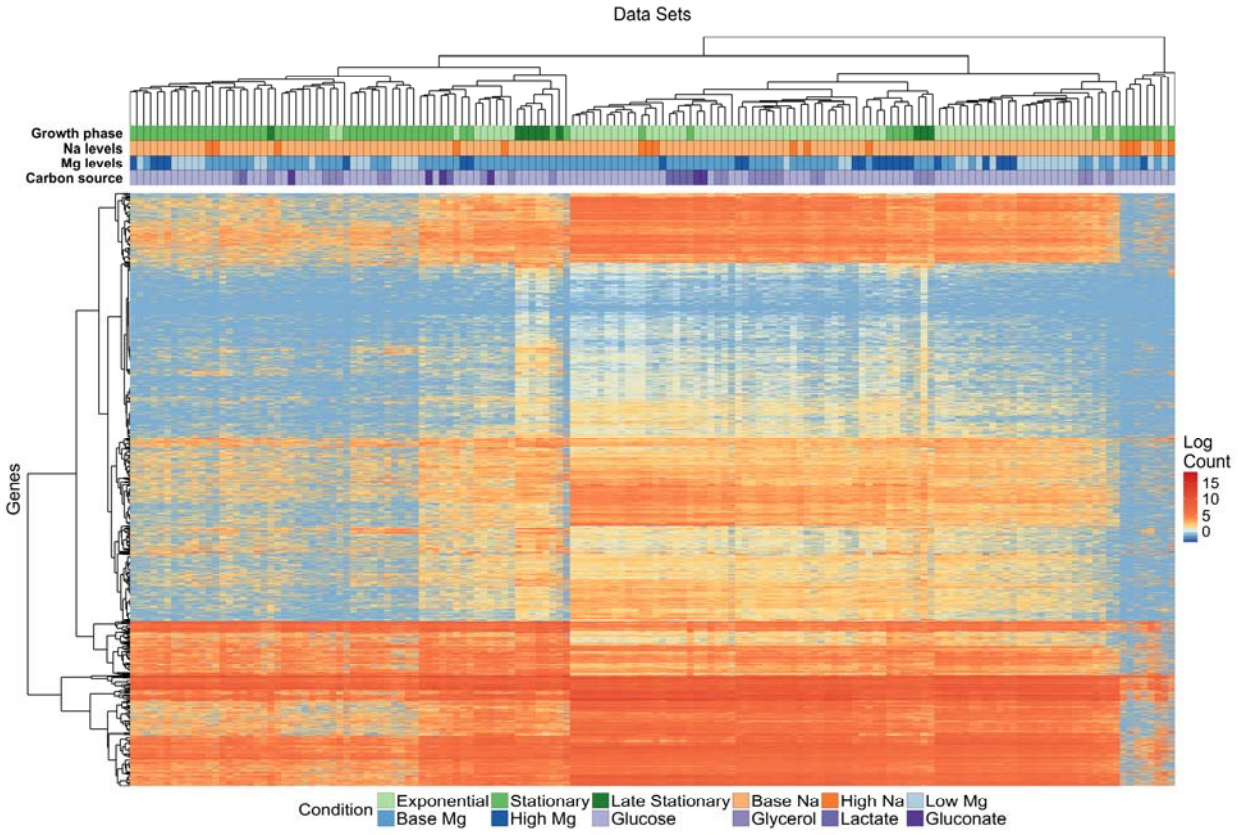

Figure 3: Clustering of mRNA abundances. The heatmap shows 4279 mRNA abundances for each of 143 samples, clustered both by similarity across genes and by similarity across samples. The growth conditions for each sample are indicated by the color coding along the top of the heatmap; the color coding is defined in the legend at the bottom.

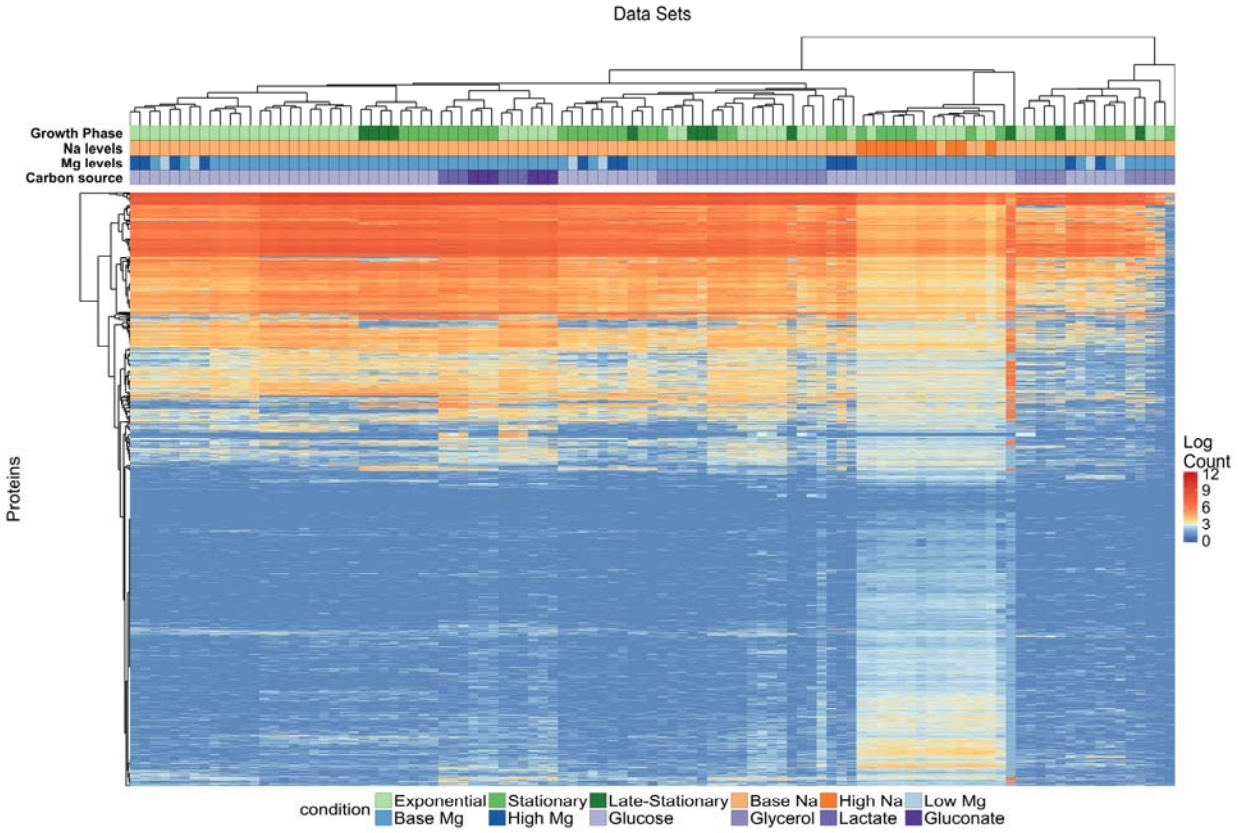

Figure 4: Clustering of protein abundances. The heatmap shows 4279 protein abundances for each of 101 samples, clustered both by similarity across genes and by similarity across samples. The growth conditions for each sample are indicated by the color coding along the top of the heatmap; the color coding is defined in the legend at the bottom.

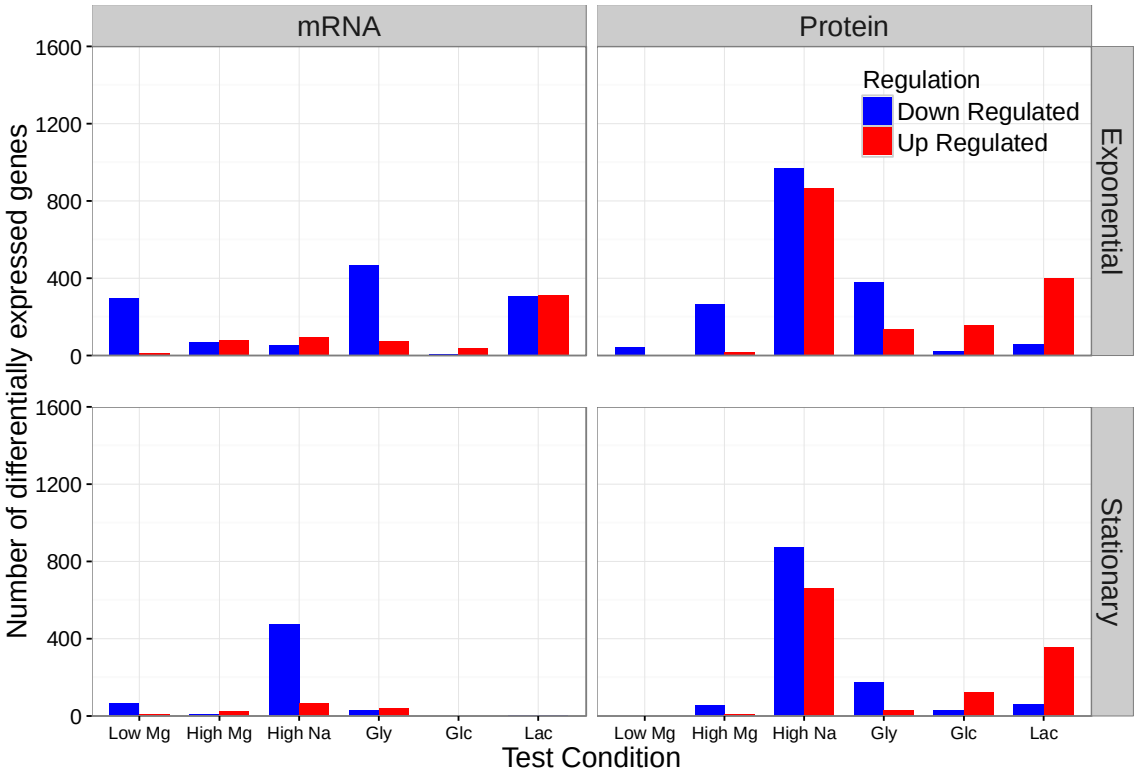

Figure 5. Number of differentially expressed genes under different conditions. We separately analyzed mRNA and protein abundances, each for both exponential and stationary growth phase. In all four cases, gene expression levels were compared to the corresponding condition with glucose as carbon source and baseline sodium and magnesium levels. Differentially expressed genes were defined has having at least a two-fold change relative to baseline and a false-discovery rate <0.05.

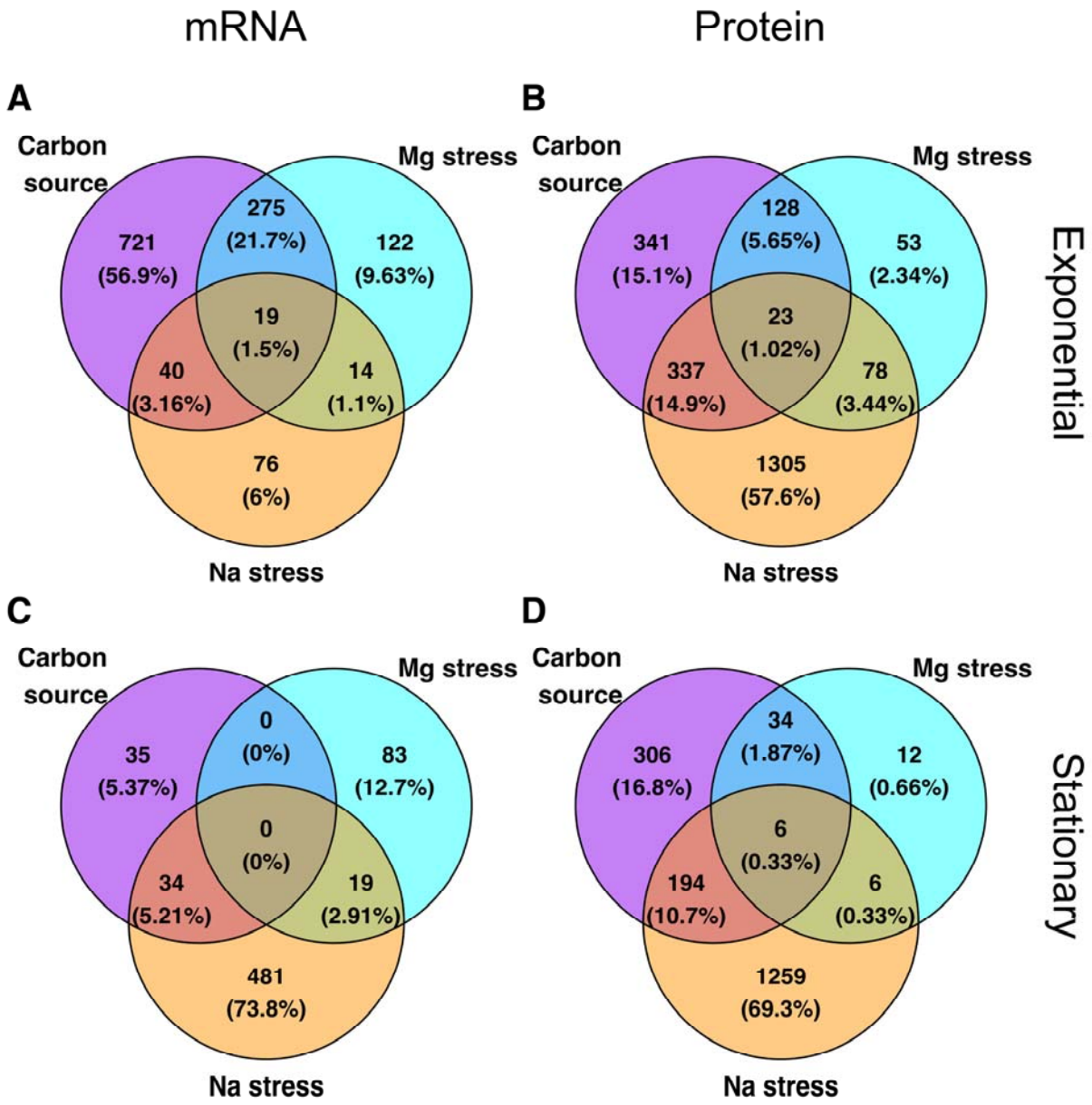

Figure 6: Overlap of differentially expressed genes among conditions. For all differentially expressed genes (identified as in Figure 5), we determined to what extent they were unique to specific conditions or appeared in multiple conditions. For simplicity, we here lumped all carbon-source experiments, all sodium experiments, and all magnesium experiments into one group each. Overall, we found relatively little overlap in the differentially expressed genes among these conditions. (A) mRNA, exponential phase. (B) protein, exponential phase. (C) mRNA, stationary phase. (D) protein, stationary phase.

Figure 7: Significantly differentially expressed KEGG pathways. For each condition, we show the top-5 differentially expressed KEGG pathways as determined by either mRNA or protein abundances. Empty boxes indicate that no differentially expressed pathways were found. The arrows next to pathway names indicate the proportion of up- and down-regulated genes among the significantly differentially expressed genes in this pathway. One up arrow indicates that 60% or more of the genes are up-regulated, two arrows correspond to 80% or more genes,

and three arrows correspond to 95% or more genes being up-regulated. Similarly, down arrows indicate the proportion of down-regulated genes. (A) Exponential phase. (B) Stationary phase.

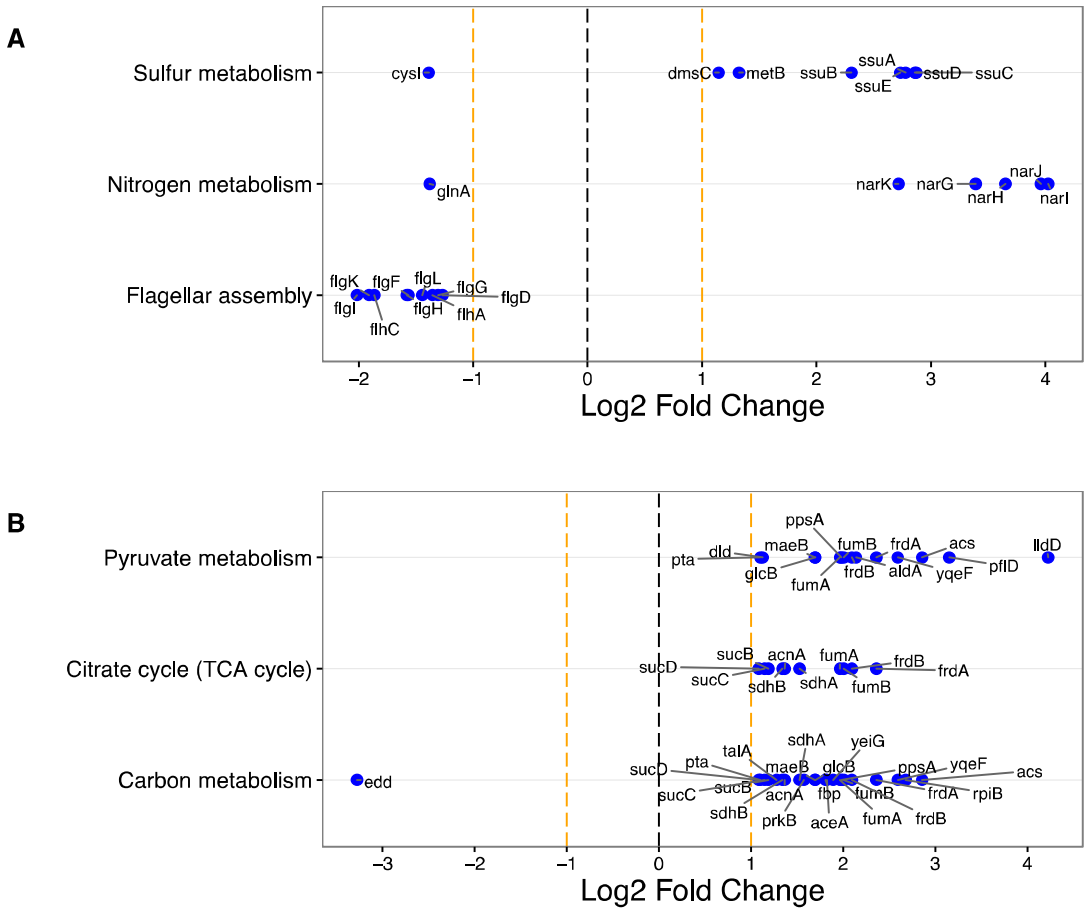

Figure 8: Examples of significantly differentially expressed KEGG pathways and associated genes. The top differentially expressed KEGG pathways are shown along the *y* axis, and the relative fold change of the corresponding genes is shown along the *x* axis. For each KEGG pathway, we show up to 10 of the most significantly changing genes. (A) Differentially expressed genes under high Mg^{2+} levels in exponential phase, as determined by mRNA abundances. (B) Differentially expressed genes under lactate as carbon source in exponential phase, as determined by protein abundances. Significant changes for all conditions are shown in Supplementary Figures 2–22.

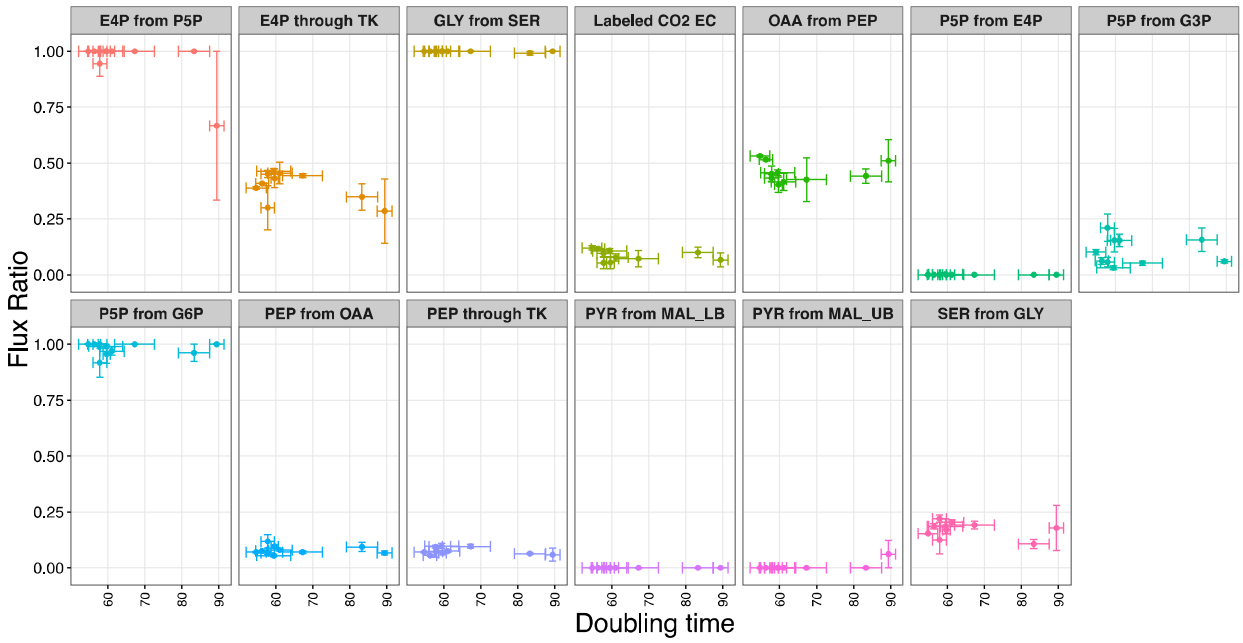

Figure 9: Flux ratios versus doubling times. 13 different flux ratios were measured for varying Na^+ and Mg^{2+} concentrations (Supplementary Figure 23). Here, these flux ratios are shown as a function of the corresponding doubling times. The specific fluxes considered and their shorthand labels as used here are defined in Ref. 31. There was no significant association between any of the flux ratios and doubling time after correction for multiple testing (Supplementary Table 6).

Tables

Table 1: Clustering of mRNA and protein abundances by different growth conditions. The *z* scores represent mean cophenetic distances between all pairs of conditions with the same label, normalized by the distribution of mean distances obtained after randomly reshuffling condition labels. The overall *z* score tests for significant clustering within a given variable, and the individual *z* score tests for significant clustering within a given condition. Significant clustering (defined as $|z|>2$) is indicated with a *.

mRNA

Variable	Overall *z* score		Condition	*z* score		# samples
Growth phase	−23.21	*	Exponential	−11.15	*	79
			Stationary	0.30		63
			Late stationary	−2.06	*	10
Carbon source	1.41		Glucose	1.53		115
			Glycerol	−1.51		25
			Lactate	−1.93		6
			Gluconate	−0.47		6
Mg Level	−1.82		Low Mg^{2+}	−0.75		36
			Base Mg^{2+}	−0.78		92
			High Mg^{2+}	−1.06		24
Na Level	−4.34	*	Base Na^{+}	−4.20	*	136
			High Na^{+}	2.85		16
Batch number	−2.11	*				

Protein

Variable	Overall *z* score		Condition	*z* score		# samples
Growth phase	−1.26		Exponential	−0.82		56
			Stationary	0.22		37
			Late stationary	−0.08		12
Carbon source	−2.80	*	Glucose	−2.34	*	66
			Glycerol	1.35		27
			Lactate	−2.73	*	6
			Gluconate	−2.63	*	6
Mg Level	−0.50		Low Mg^{2+}	0.85		6
			Base Mg^{2+}	−0.44		87
			High Mg^{2+}	−0.42		12
Na Level	−1.74		Base Na^{+}	−0.94		94
			High Na^{+}	−5.61	*	11
Batch number	−20.54	*				