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

Mehmet U. Caglar1, 2, 3, John R. Houser3, 4, 5, Craig S. Barnhart3, 4, Daniel R. Boutz3, 4, 5, Sean M. Carroll6, Aurko Dasgupta3, 4, Walter F. Lenoir5, Bartram L. Smith1, 2, 3, Viswanadham Sridhara2, 3, Dariya K. Sydykova1, 2, 3, Drew Vander Wood3, 5, Christopher J. Marx7, 8, Edward M. Marcotte3, 4, 5\*, Jeffrey E. Barrick3, 4, 5\*, Claus O. Wilke3, 4, 5\*

1Department of Integrative Biology, The University of Texas at Austin

2Center for Computational Biology and Bioinformatics, The University of Texas at Austin

3Institute for Cellular and Molecular Biology, The University of Texas at Austin

4Center for Systems and Synthetic Biology, The University of Texas at Austin

5Department of Molecular Biosciences, The University of Texas at Austin

6Department of Organismic and Evolutionary Biology, Harvard University

7Department of Biological Sciences, University of Idaho, Moscow, Idaho

8Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, Idaho

\*Corresponding author: [marcotte@icmb.utexas.edu](mailto:marcotte@icmb.utexas.edu) (EMM); [jbarrick@cm.utexas.edu](mailto:jbarrick@cm.utexas.edu) (JEB); [wilke@austin.utexas.edu](mailto: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 cheap 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 synthetic engineering of biology1. 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 data2. Much effort is currently being spent on understanding how to best integrate information collected about multiple cellular subsystems and derived from different high-throughput methods. For example, there are many proposed approaches for relating gene expression and protein abundances, focusing on integrative, whole-cell models2–5.

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 under 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 environment6 and was one of the first organisms studied at the whole-genome level7. There have been a number of studies of the *E. coli* transcriptome and/or proteome in response to different growth conditions. For example, if cells are grown at high density, expression of most amino-acid biosynthesis genes is down-regulated and expression of chaperones is up-regulated, suggesting that high cell density is stressful for cells8. Exposure of *E. coli* to reduced temperature leads to changes in gene-expression patterns consistent with reduced metabolism and growth9. At the single-cell level, the *E. coli* transcriptomes and proteomes are noisy, and mRNA and protein copy numbers tend to be uncorrelated10. Under long-term glucose starvation, mRNAs are generally down-regulated while the protein response is more varied11. 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 larger-scale studies have measured mRNA and/or protein abundances under multiple different conditions12–15.

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 previously11. 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 Mg2+ and Na+ stress.

Measurements of RNA and protein abundances were carried out as previously described11. All resulting data sets were carefully checked for quality, normalized, and log-transformed. Our final data set consisted of 152 RNA samples, 105 protein samples, and 65 flux samples, 59 of which are associated with high Mg+2 and high Na+ experiments (Supplementary Table S1).

Our raw RNA-seq data covers 4279 distinct mRNAs, our protein data covers 4201 distinct proteins, and metabolic flux data covers 13 different fluxes. 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 rate 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 Mg2+ concentrations. Surprisingly, we found a broad range of Mg2+ 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). Mg2+ 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, Mg2+ 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 Mg2+ 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 DESeq216. 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 5mM Na+ and 0.8mM Mg2+. This is the baseline formulation of media used in the glucose time-course samples11. We then compared RNA and protein abundances between this reference condition and the alternative conditions (different carbon sources, elevated Na+, and elevated or reduced Mg2+). Note that a detailed comparison of reference exponential phase vs. reference stationary phase has already been published11.

We defined significantly differentially expressed genes as those whose abundance had at least a two-fold change (log2 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 source glycerol, whereas mRNA showed the most differential regulation for altered Mg2+ levels and again for the carbon source glycerol (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 Mg2+ 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 little overlap between Na stress and either Mg stress or carbon source, and moderate overlap between Mg stress and carbon source, in particular in exponential phase (Figure 6). By contrast, at the protein level, there was more overlap between Na stress and either Mg stress in exponential phase. There is a moderate overlap with carbon source, in particular in exponential phase (Figure 6). In stationary phase Mg2+ is isolated from the rest. And there is little to no overlap between carbon source and NA+ for stationary phase

We also identified significantly altered biological pathways and molecular activities of gene products. We use the Kyoto Encyclopedia of Genes and Genomes (KEGG)17 for biological pathways and annotations from the Gene Ontology (GO) Consortium for molecular functions18. 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 DAVID19. In all cases, we used a cutoff of 0.05 on FDR-corrected *P* values to identify significant annotations. For KEGG pathways we have plenty of variation but of GO annotations associated with molecular function we find 2 altered annotations; “” and “”

Finally, we looked at individual, differentially expressed genes associated with specific pathways and/or functions (Supplementary Figures SX–SY). As an example, the differentially expressed mRNAs associated with significantly altered KEGG pathways under high Mg2+ concentrations in exponential phase are shown in Figure 8. Almost all of the most altered genes are significantly down-regulated in both situations. By contrast, using glycerol instead of glucose in exponential phase caused up-regulation of many pathways and genes (Figure 9).

**Metabolic flux under salt stress**

We collected measurements of metabolic flux through the central metabolism for the high sodium experiment and the high magnesium experiment, at both exponential and stationary phase. 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.

For exponential growth, we saw almost no change in flux ratios with increasing Na+ or Mg2+ (Supplementary Figure S2). There was possibly a decrease in pentose-5-phosphate pathway use at 300 mM Na+, and *Chris, can we say something about Mg? Also, please be aware of the paragraph on flux in the discussion. The discussion in the context of doubling times should be there, not here, I think. Here I just want to state briefly what the results were.*

For stationary phase, we found *Chris, please complete.* (Supplementary Figure S3).

## Discussion

We have studied the regulatory response of *E. coli* under a wide variety of different growth conditions. The experimental conditions we have considered include four different carbon sources, different levels of Na+ and Mg2+ stress, and growth into deep stationary phase, up to two weeks post inoculation. We have 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 have found that the smallest number of differentially expressed genes for different carbon sources and larger numbers for different salt stresses. Under salt stress, most genes tend to be down-regulated, with the exception of elevated Mg2+ in stationary phase, where we have found over twice as many up-regulated mRNAs as down-regulated mRNAs. Finally, our measurements provide a comprehensive dataset that will be useful for future modeling of *E. coli* metabolism.

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.*13 considered 22 unique conditions and measured abundances of >2300 proteins. mRNA abundances were not measured. Soufi *et al.*12 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. However, similar to our study, they found down-regulation of genes associated with arginine biosynthesis and metallic ion binding under stress. Lewis *et al.*14 considered only 3 different carbon sources but measured mRNA and protein abundances in different strains adapted to these growth conditions. Finally, Lewis *et al.*15 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 study11, we here observed clear differences in the differential expression of mRNAs and proteins. In particular, we had reported11 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 earlier11 that energy-intensive processes were down-regulated and stress-response proteins up-regulated in stationary phase. Similarly, we observed here that stress conditions also led to the down-regulation of energy-intensive processes. In particular, we saw down-regulation of translation, flagellar assembly, nucleotide related pathways such as RNA degredation, and aminoacyl-tRNA biosynthesis. These pathways seem to be broadly down-regulated under many different stress conditions8,9.

For all growth conditions, we observed patterns of altered gene expression consistent with the known function of specific differentially expressed genes. For example, under Na+ stress we saw up-regulation of *acnB* and *talB*, as previously reported20.

It is well known from large-scale, high-throughput gene-expression studies that measurements are frequently confounded by batch effects21. 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 we are aware of their existence and analyze the 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 recommended22. How to best correct for batch effects is a topic of ongoing investigation, and increasingly sophisticated methods are being developed to separate batch effects from real signal in an automated fashion23.

Given the many cellular changes observed in mRNA and protein levels, we turned to 13C labeling techniques11,24,25 to examine the extent to which these changes affected the relative flux of metabolites through central metabolic pathways during exponential growth. For this work we concentrated upon growth on glucose during Na+ and Mg2+ stresses. Across these conditions, growth rates change over nearly a two-fold range, with the doubling time changing from approximately 50 to 95 minutes. High Na+ levels reduced growth by a third. Despite this substantial effect on growth, within the variation observed among flux ratios we observed no significant changes in central metabolism other than an apparent decrease in pentose-5-phosphate pathway use at 300 mM. In contrast, Mg2+ yielded optimal growth at intermediate concentrations of 50 mM to 200 mM, with approximately 10% slower growth at 8 mM and below and approximately 50% slower growth at 400 mM. Although the relative flux at several key splits in metabolism was unchanged across this range, several others changed significantly. The proportion of oxaloacetate generated from phospoenolpyruvate and pentose-5-phosphate from glucose-6-phosphate both decreased (and thus pentose-5-phosphate from glyceraldehyde-3-phosphate decreased) when growth was slow at either high or low concentrations. As these changes were not observed during NaCl stress that led to similar growth defects, they do not simply arise as an indirect consequence of slowed growth and we speculate that they may be specific to MgCl2 stress. With the exception of those changes, however, the general picture was that homeostasis in central metabolism was sufficient to ward off significant changes in relative pathway use despite large effects upon overall growth rate and the pools of mRNA and proteins.

In summary, our study provides a large and comprehensive data set investigating the gene-regulatory response of *E. coli* under different growth conditions, both at the mRNA and the protein level. We have found systematic differences in gene-expression response between exponential and stationary phase, and between mRNAs and proteins. Our data set provides a rich resource for future modeling of the *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 as previously described11 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. Mg2+ concentrations were varied by changing the amount of MgSO4 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 OD600 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 described11.

Doubling times were measured …

### Normalization and quality control of RNA and protein counts

Our raw input data consisted of mRNA and protein counts. Protein counts can be fractional, because some peptide spectra cannot be uniquely mapped to a single protein. We rounded all protein counts to the nearest neighbor for subsequent analysis. We set the counts of all unobserved proteins to zero. 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 need a sentence here stating which/how many samples where fine vs. not, and what we did (we kept all).*

After quality control, we normalized read counts using size-factors calculated via DESeq216. Because we had many RNAs 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 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 hclust function in R. This method defines the cluster distance between two clusters as the maximum distance between their individual components26. 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 DESeq216 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 NaCl and MgSO4 concentrations. We did not compare exponential phase to stationary phase samples, since this comparison was done in depth previously11 for samples grown on glucose and with base NaCl and MgSO4 concentrations.

We considered genes as differentially expressed between two conditions if their log2 fold change was >1 and their FDR-corrected *P* value <0.05. We subsequently annotated differentially expressed genes with DAVID19 version 6.8 Beta released in May 2016. We considered both KEGG pathways17 and GO annotations18.

### Statistical analysis and data availability

All statistical analyses were performed in R. The relevant R scripts and processed data are available on github: XXX Raw RNA reads and peptide spectra are also available under accession numbers XXX and YYY.

## 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. *Microbiology* **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. Taniguchi, Y. *et al.* Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science* **329,** 533–538 (2010).

11. 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).

12. 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,** (2015).

13. Schmidt, A. *et al.* The quantitative and condition-dependent Escherichia coli proteome. *Nat. Biotechnol.* **advance online publication,** (2015).

14. 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).

15. 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).

16. 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).

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

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

19. 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).

20. Weber, A., Kögl, S. A. & Jung, K. Time-Dependent Proteome Alterations under Osmotic Stress during Aerobic and Anaerobic Growth in Escherichia coli. *J. Bacteriol.* **188,** 7165–7175 (2006).

21. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11,** (2010).

22. Differential analysis of count data – the DESeq2 package. (2016). Available at: http://journals.plos.org/ploscompbiol/article/asset?id=10.1371%2Fjournal.pcbi.1004127.PDF. (Accessed: 12th April 2016)

23. Kim, M., Zorraquino, V. & Tagkopoulos, I. Microbial Forensics: Predicting Phenotypic Characteristics and Environmental Conditions from Large-Scale Gene Expression Profiles. *PLOS Comput Biol* **11,** e1004127 (2015).

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

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

26. Madhulatha, T. S. An Overview on Clustering Methods. *ArXiv12051117 Cs* (2012).