# Predicting bacterial growth conditions from mRNA and protein abundances.

Mehmet U. Caglar1, Claus O. Wilke1\*

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

\*Corresponding author: [wilke@austin.utexas.edu](mailto:wilke@austin.utexas.edu)

## Questions/comments/concerns

## Do you have any negative control analyses? F1 score ranges between zero and 1 but with all the weighting, etc. is that really true? If you scrambled training data labels would you *really* get F1 scores of zero??? Not sure we need to address it now but I think this might be an important reviewer concern (at least it’s something that I’d say if I were reviewing this) I am currently running it.

## Abstract

Individual species of bacteria are capable of surviving and reproducing across a range of environmental conditions. Cells respond to changing nutrient availability and external stresses by altering the expression of individual genes. Low-cost sequencing technologies may be able to leverage these condition-specific gene expression patterns to act as sensors of environmental conditions, providing insight into the presence of various small molecules, toxins, or species-interactions. However, the ability and limitations of gene expression signatures alone to predict individual environmental growth conditions remains an open question. Here, we used machine learning to predict 16 closely-related growth conditions using 155 datasets of *E. coli* transcript and protein abundances. We show that models are able to discriminate between different environmental features with a relatively high degree of accuracy. We observed a small but significant increase in model accuracy by combining transcriptome and proteome-level data, and show that stationary phase conditions are typically more difficult to distinguish from one another compared with exponential growth. Our analysis demonstrates several clear outcomes: i) with sufficient training data, gene expression measurements from a single species are sufficient to distinguish environmental conditions that are separated by a single environmental variable ii) this ability is significantly enhanced when capturing cells during the exponential rather than stationary growth phase and iii) despite their overall quiescence, stationary phase cells nevertheless retain signatures of the external environment.

## Introduction

Generalist bacterial species must be able to sense environmental conditions and alter gene expression patterns according to nutrient availabilities. For instance, *Escherichia coli* grows inside the nutrient rich digestive tract of host organisms, but spends another portion of its life-cycle exposed to harsh environmental conditions upon being excreted and before finding another host [reference]. Maintenance and reproduction in these vastly different conditions is only accomplished by expressing a controlled set of environmentally specific gene products in addition to more ubiquitous “housekeeping” genes [reference]. Determining which genes are up- or down-regulated in response to a shift in environmental conditions can provide insight into the function of individual gene products and pathways. However, a related question is whether global gene expression patterns themselves can provide insight into the external environmental conditions.

Cells can be engineered to act as sensors for a variety of environmental conditions via rational design of synthetic genetic circuits [reference]. However, individual applications take time to develop and refine and in the end there is still a concern about releasing genetically engineered species into natural environments [reference]. Previous work has shown that the microbial species composition of an environment can serve as a rapid and low-cost “biosensor” to indicate the presence of various contaminants [reference]. In addition to species composition, the gene expression patterns of individual species may be able to provide even higher resolution into the past and current chemical composition of environments.

A current challenge is to establish what molecular level data is important, and to determine at what resolution individual environmental differences can be resolved. Combining multiple-layers of information from different ‘omics’-scale technologies is likely to provide better discriminatory capability, but integrating datasets is challenging based on the biases of individual methods1 and the inevitability of batch-level effects that occur when datasets are generated across multiple labs and platforms2,3.

Several past studies have looked into the question of predicting external conditions by using the cells internal variables4,5. Other studies have interrogated multi-omic datasets from different growth conditions to understand the function of regulatory networks, individual gene functions, and resource allocation strategies [reference]. However, the main focus of many of these studies has been to understand differences in gene expression patterns across environmental conditions so as to provide insight into cellular mechanisms and pathways or to predict cellular level phenotypes such as specific growth rates.

Here, we are interested in determining whether gene expression patterns can discriminate between environmental conditions in the absence of prior knowledge about the role and function of individual genes. Our study leverages a large dataset of transcriptomic and proteomic measurements of *E.coli* growth under multiple distinct but closely-related conditions6. We use this data to train machine learning models and find that highly similar environmental conditions can be discriminated with a relatively high degree of accuracy. We compare the predictability of external conditions by using transcriptomic and proteomic datasets individually and in combination, and observe a small but significant increase in model accuracy when combining datasets. Interestingly, we also investigate which conditions are more- and less-challenging to discriminate, and find that stationary phase cells are highly similar to one-another. Nevertheless, we show that even stationary phase cells can retain some similarities in gene expression to their exponentially grown counterparts after up-to 2 weeks of quiescence.

## Results

### Data structure and pipeline design

We used a previously generate dataset of whole-genome *E. coli* mRNA and protein abundances, measured under 34 different conditions6,7. This dataset consists of a total of 155 samples, for which mRNA abundances are available for 152 and protein abundances for 105 (Figure 1). For 102 samples, both mRNA and protein abundances are available. The 34 different experimental conditions were generated by systematically varying four parameters, growth phase (i.e., time), carbon source, Mg2+ concentrations, and Na+ concentrations (Figure 1). We cluster some of the conditions together (i.e. 100, 200, 300 mm Na+ are labelled as high Na) and end up with 16 distinct conditions to run all further analysis. Here, we asked to what extent machine learning models would be capable of discriminating between these growth parameters given knowledge of mRNA abundances, protein abundances, or both.

We used a general cross-validation set-up by first splitting samples into training and testing datasets. We used the labeled training data to fit models capable of predict environmental conditions, and ask how accurate these models are at predicting the conditions present in the test set. We employed four different machine learning models, three based on Support Vector Machines (SVMs) with different kernels (radial, sigmoidal, and linear) and the fourth using random forest classification. We used *C-Classification*8for training classification model (move to methods). We adjusted sample weights such that each growth condition has an equal weight in order to prevent prediction bias from favoring more populated classes, and for accuracy we report the multi-conditional *F*1 score9 which weights all conditions equally and assigns equal importance to false positive and false negative predictions.

We generated a “tuning loop” to fine tune the machine learning parameters by using cross-validation within the training set. (see Materials and Methods).

Our overall pipeline is illustrated graphically in Figure 2. We start by normalizing our datasets and then dividing the data into training and testing sets. The training set is further split into training and tuning sets to optimize hyper-parameters as described above. We apply the training set model to predict the identity of the test set data to test overall prediction accuracy. Finally, we repeat this entire process 60 times so as to get a more complete picture of predictive accuracy.

### Growth conditions can be predicted accurately from both mRNA and protein abundances

We first asked whether there were major differences in how well the different machine-learning approaches performed at predicting the growth conditions. We represent “growth condition” as a four-dimensional vector of categorical variables, consisting of the dimensions growth phase (exponential, stationary, late stationary), carbon source (glucose, glycerol, gluconate, lactate), Mg2+ concentrations (base, low, high), and Na+ concentrations (base, high). In total, our dataset contains 16 distinct combinations of these four variables (Figure 1).

We trained models to predict the entire four-dimensional condition vector at once for a given sample, and we used the multi-conditional *F*1 score9 to quantify prediction accuracy. The *F*1 score is the harmonic mean of precision and recall. It approaches zero if either quantity approaches zero, and it approaches one if both quantities approach one (representing perfect predictive accuracy). We assessed model performance during the tuning stage of our pipeline, by recording which model had the best *F*1 score for each tuning run. We found that the SVM with a radial kernel clearly outcompeted the other models when fit to mRNA data, and the random forest model outcompeted the other models by a comparatively small margin when fit to protein data (Table 1).

We next compared the *F*1 scores for model predictions applied to the test set. When using mRNA abundance data alone, the distribution of *F*1 scores from our 60 independent replications werecentered around a value of 0.7 (Figure 3). The *F*1 score distributions were virtually identical for the three SVM based models, and were somewhat lower for the random forest model. Model performance on test data using only protein abundance measurements was slightly worse than those achieved with mRNA abundance data. However, it is important to note that the protein abundance data contains fewer conditions overall, which may partially explain the decreased predictive accuracy of protein only models—a point to which we return to later.

In addition to assessing the overall predictive power using *F*1 scores, we also recorded the percentage of times specific growth conditions were accurately or erroneously predicted in the form of a confusion matrix. Here, the columns headings at the top show the predicted condition from the model on the test set and the rows show the true experimental condition (Figure 4). The large numbers/dark colorings along the diagonal highlight the percentage of true positive predictions whereas any off-diagonal element represents the percentage of incorrect predictions for a given condition. We found that the erroneous off-diagonal predictions are partially driven by the uneven sampling of different conditions in the original dataset. Even though we used sample-number-adjusted class weights in all fitted models, we observed a trend of increasing fractions of correct predictions with increasing number of samples available under training (Supplementary Figure 3).

The *F*1 score quantifies accuracy by only considering perfect predictions (i.e. when all 4 features are correctly predicted). A sample that is incorrectly classified for all four factors is treated the same as one that only differs from the true set of features by a single incorrect factor. In practice, we observed that the majority of incorrect predictions differed from their true condition vector only by a single value (Supplementary Figure 4).

### Joint consideration of mRNA and protein abundances improves model accuracy

We next asked whether predictions could be improved by simultaneously considering mRNA and protein abundances. To address this question, we limited our analysis to the subset of 102 samples for which both mRNA and protein abundances were available, and ran our analysis pipeline for mRNA abundances only, protein abundances only, and for the combined dataset containing both mRNA and protein abundances. For all four machine-learning algorithms, protein abundances yielded significantly better predictions than mRNA abundances (Figure 5, Table 2). This is in contrast to Figure 3, where we saw increased accuracy using mRNA abundance data. However, as previously noted, our dataset contains a larger number mRNA abundance samples, which results in a larger amount of training data. When compared on the same exact conditions—as depicted in Figure 5—protein abundance data appears to be more valuable for discriminating between growth conditions. The combined dataset consisting of both mRNA and protein abundance measurements yielded the best overall predictive accuracy, irrespective of machine-learning algorithm used (Figure 5, Table 2).

When considering the confusion matrices for the three scenarios (mRNA abundance, protein abundance, and combined), we found that many of the erroneous predictions arising from mRNA abundances alone were not that common when using protein abundances and vice versa (Supplementary Figures 5,6). For example, when using mRNA abundances, many conditions were erroneously predicted as being exponential phase, glycerol, base Mg2+, base Na+, or as stationary phase, glucose, base Mg2+, high Na+; these particular predictions were rare or absent when using protein abundances. By contrast, when using protein abundances, several conditions were erroneously predicted as being stationary phase, glycerol, base Mg2+, base Na+, and these predictions were virtually absent when using mRNA abundance data. For predictions made from the combined dataset, erroneous predictions unique to either mRNA or protein abundances were generally suppressed, and only those predictions that arose for both mRNA and protein abundances alone remained present in the combined dataset (Supplementary Figure 7).

### Prediction accuracy differs between environmental features

We finally wanted to better assess the sources of inaccuracy in our models. As previously noted, the majority of incorrect predictions differed by only a single factor. The environmental feature that accounted for most of these single incorrect predictions was Mg+2 concentration for protein data and carbon sources for mRNA data.

Growth phase (e.g. exponential, stationary, late-stationary) is not strictly an environmental variable and using this as a feature may partially skew our results if the goal is to predict *strictly* *external* conditions. We thus trained and tested separate models using only exponential or only stationary phase datasets and asked to what degree these samples were predictive of the remaining 3 environmental features (carbon source, [Mg2+], and [Na+]). We found that prediction accuracy was consistently better for exponential phase samples than for stationary phase samples, irrespective of the machine-learning algorithm or the data source (mRNA, protein abundances, or both) (Figure 6). We further showed that this result is not simply a consequence of the increased dataset size for exponential phase growth by down-sampling the exponential phase data to be the same size as the stationary phase datasets (Supplementary Figure XXX). Late-stationary phase data contained to few samples to run an equivalent analysis.

This observation implies that *E. coli* gene expression patterns during stationary phase are less indicative of the external environment compared to cells experiencing exponential growth. However, a fraction of incorrect stationary phase predictions were to identical conditions under exponential growth illustrating that stationary phase gene expression nevertheless retains some signature of the external environment. To further test this proposition, we took the models trained only on exponential phase data and asked to what degree they could predict conditions for stationary phase cells. We observed that XXX. Conversely, we asked whether models trained on stationary phase data could predict environmental conditions of exponential phase cells and observed XXX.

Finally, in order to better understand which conditions were the most problematic to predict, we constructed models to predict only *individual* features rather than the entire set of 4 features. When making predictions based on mRNA abundances only, models were most accurate in predicting growth phase and least accurate for carbon source, with Mg2+ and Na+ concentration falling between these two extremes. By contrast, when making predictions based on protein abundances, the most predictable feature was carbon source, the least predictable was Mg2+ concentration, and Na+ concentration and growth phase fell in-between these two extremes (Figure 7, Supplementary Figure 7). Finally, for the combined mRNA and protein abundance dataset, we found that accuracy for carbon source and Mg2+ concentration generally fell between the accuracies observed using mRNA and protein abundances individually. By contrast, accuracies for the Na+ concentration and growth phase were generally as good as—or better than—the prediction accuracies of the individual datasets (Supplementary Figure 9).

**Discussion**

Our central goal in this manuscript was to determine whether gene expression signatures from a single species of bacterium are sufficient to predict environmental growth conditions. Overall, our results indicate that machine learning models can be trained to predict closely-related growth conditions with a fairly high accuracy. We analyzed a rich dataset of 152 samples for mRNA data and 105 samples for protein data across 16 distinct laboratory conditions as a proof-of-concept to show that *E.coli* gene expression is responsive to external conditions in a measurable and consistent way. While *E. coli* is a well-characterized species, our analysis relies on no *a priori* knowledge. It is thus likely that increasing the number and diversity of training samples and conditions will produce further improvements in accuracy and discrimination between a wider array of conditions.

Interestingly, we found that consideration of mRNA and protein datasets alone are sufficient to produce accurate results, but that joint-consideration of both datasets results in superior predictive accuracy. This finding implies that post-transcriptional regulation is at least partially controlled by external conditions, which has been observed by previous studies that have investigated multi-omics datasets1,13–15. Such regulation may result from post-translational modifications16, stress coping mechanisms17, differential translation of mRNAs, or protein-specific degradation patterns.

An important finding that we explored was that cellular growth phase places important limits on the predictability of external conditions, with stationary phase cells being particularly difficult to distinguish from one-another irrespective of their external conditions. Nevertheless, discrimination of external environmental factors in stationary phase cells was still much better than random and retained some similarity to exponential phase cells grown under the same conditions. One reason for this behavior might be associated with endogenous metabolism; in which, stationary phase cells start to use the residue of other cells instead of provided carbon source. This new carbon source, which is independent of the externally provided carbon source, may suppress the differences between the cells in different external carbon source environments18,19. Another reason for this behavior might be related to strong coupling between gene expression noise and growth rate. Multiple studies concluded that lower growth rates are associated with higher gene expression noise, which might be a survival strategy in harsh enviroments20. Negative correlations between population average gene expression and noise have been shown for *E. coli* and *Saccharomyces cerevisiae,* lending support for this theory21,22.

A relevant finding to emerge from our study is that different features of the environment may be more- or less- easy to discriminate from one another and this discrimination may depend on which molecular species is being interrogated. Growth phase, for instance, can be reliably predicted from mRNA concentrations but similar predictions from protein concentrations were less accurate. A possible explanation for this is the fact that mRNAs and proteins have different life-cycles26. Given the comparably slow degradation rates of proteins, a large portion of the stationary phase proteome is likely to have been transcribed during exponential phase growth. As another example, carbon sources can be reliably predicted from protein concentrations, stronger than any other condition that we investigated, but the accuracy of carbon source predictions based on mRNA concentrations are limited. Carbon assimilation is known to be regulated by post-translational regulation28–30, which may be a possible reason for this finding (Figure 7, Supplementary Figure 8).

Despite the fact that we investigated over 150 samples spanning 16 unique conditions, a limitation of our work and conclusions is nevertheless sample size (though our study is comparable or larger than similar multi-conditional transcriptomic and/or proteomic studies)33–36. The comparison between all data and the more limited set that includes only the intersection of mRNA and protein abundance data (Figure 4 compared to Supplementary Figures 4,5) indicates that prediction accuracy decreases as the size of our training sets gets smaller for both mRNA and protein data. This trend indicates that training set sizes are still ultimately limiting the statistical power of our models. A second problem is associated with sample number bias37–39. We made corrections with weight factors40,41 and display the multi-conditional *F*1 score42 to account for the fact that some conditions contained more data samples, but the predictability of *individual* conditions nevertheless increases with the number of training samples for that (Supplementary Figure 3). This finding again highlights that increasing training data will likely result in higher prediction accuracies.

## Materials and Methods (AJH: HAVEN’T STARTED ON THIS YET)

### Data

We use the same data sets that were used in a previous study [cite]. For all single variable test, and also for multi-conditional prediction tests except for combining mRNA and proteins we use all available data. For combination tests, we use the intersection of mRNA and protein samples (102 sample).

### Prediction Methodology and Parameters

The initial preparation of the data is similar to a previous study on the same data set 7. After finding suitable subsets of the data for the tests and summing up technical replicate results for proteins, we calculate size factor normalized data with the DeSeq210 package and apply variance stabilizing transformation43 (vst) on it.

In the next step, we divide the data into two subsets; named training&tune set and test set. The division is semi-random i.e. algorithm preserves the ratios of different conditions in training&tune and test subsets. We preserve the condition labels for training&tune data but we delete the labels of the samples for the test set. We then apply frozen Surrogate Variable Analysis11 (fSVA) to get rid of the batch effects in the sample. The algorithm can correct the batch effects on both training&tune and test data without knowing the labels of the test data. After fSVA we use principal component analysis12 (PCA) to define the principle axis of training&tune set and rotate the test data set with respect to the principal axis of training&tune set. We then pick the most significant top 10 axis for both training&tune, and test data sets. The next step is to train and tune our candidate machine learning algorithms with *dimension reduced training&tune dataset* and apply those trained and tuned algorithms on *dimension reduced test dataset* to make predictions on them. In the case of combined mRNA and protein data analysis, we calculate size factors with the DeSeq2 algorithm and batch effects with the fSVA algorithm individually then combine two datasets and apply PCA on combined data. We run this process and call the training and tuning algorithm 600 times with independent divisions between training&tune, and test data sets.

We optimized the "cost" parameter for the three SVM models, and the "gamma" parameter for the SVM with radial and sigmoidal kernels. For the random forest algorithm, we optimized three parameters; "mtry", "ntrees", and "nodesize" Before the model training phase, we applied size factor normalization, normalize batch effects and apply PCA12 for dimension reduction.

The training and tuning algorithm firstly divide the data as train and tune datasets semi-randomly, trying to keep the ratios of individual conditions same in both training and tuning data. We repeat this division 10 times and generate 10 training and tuning dataset pairs for individual train&tune vs test pair which ends us up with 600 distinct training pairs. Then we generate a parameter grid for tuning process. We tune cost parameter for SVM with linear kernel, cost and gamma parameters for SVM with radial and sigmoid kernels, ntree, mtry, and nodesize parameters for random forest algorithms (Table 3). We use e1071 package44 for implementing support vector machines with linear, radial, and sigmoid kernels that based on the libSVM8 library and the randomForest package45 for the random forest algorithm.

We apply multiple sets of tests; we either try to predict an individual parameter such as growth phase, carbon source, Mg+2 concentration or Na+ concentration; or we try to predict all multiple conditions at one single run. We run our pipeline on all mRNA, all protein, and combined mRNA-protein data. We run our pipeline on three distinct phase combinations: 1.) all phases, 2.) only on exponential phase and 3.) only on stationary phase (Supplementary Table 1).

We train each four models in training set and make predictions on tune set for all variables that we are tuning and for all 10 training and tuning dataset pairs, we apply a class weight normalization for training of all four models; where class weights are inversely proportional to number of training samples and calculated independently for each training run. We average the errors calculated by score metric over training tuning dataset pairs and obtain an average error value estimate for each algorithm and for each parameter combination. Then we find the winner parameter combination for each algorithm and label this parameter combination, algorithm match as winner algorithm for each of xx independent divisions between training&tune, and test data sets.

At the end of all procedure we end up with 60\*4=240 winner algorithms with their tuning parameters and their average error values.

### Calculation of the score metric

The metric we use is the multi class macro *F*1 score 9,42,46 that normalized *F*1scores over individual conditions; i.e. each condition has equal weight instead of each sample. Out of two distinct macro *F*1 score calculation methods in the literature

1. ; where 42
2. 9

we picked the first one.

### Statistical analysis and data availability

All statistical analyses were performed in R. All processed data and analysis scripts are available on github: <https://github.com/umutcaglar/ecoli_multiple_growth_conditions>

Raw Illumina read data and processed files of read counts per gene and normalized expression levels per gene have been deposited in the NCBI GEO database47 (accession GSE67402 for the glucose time-course previously published7, accession GSE94117 for all other experiments). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository48 (accession PXD002140 for the glucose time-course previously published7, accession PXD005721 for all other experiments.

## References

1. Kim, M., Rai, N., Zorraquino, V. & Tagkopoulos, I. Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. *Nat. Commun.* **7,** (2016).

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

3. Scharpf, R. B. *et al.* A multilevel model to address batch effects in copy number estimation using SNP arrays. *Biostat. Oxf. Engl.* **12,** 33–50 (2011).

4. Brandes, A. *et al.* Inferring Carbon Sources from Gene Expression Profiles Using Metabolic Flux Models. *PLOS ONE* **7,** e36947 (2012).

5. Sridhara, V. *et al.* Predicting Growth Conditions from Internal Metabolic Fluxes in an In-Silico Model of E. coli. *PLOS ONE* **9,** e114608 (2014).

6. Caglar, M. U. *et al.* The E. coli molecular phenotype under different growth conditions. *Sci. Rep.* **7,** 45303 (2017).

7. 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).

8. Chang, C.-C. & Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans Intell Syst Technol* **2,** 27:1–27:27 (2011).

9. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **45,** 427–437 (2009).

10. 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).

11. Parker, H. S., Bravo, H. C. & Leek, J. T. Removing batch effects for prediction problems with frozen surrogate variable analysis. *ArXiv13013947 Stat* (2013).

12. Jolliffe, I. Principal Component Analysis. in *Wiley StatsRef: Statistics Reference Online* (John Wiley & Sons, Ltd, 2014). doi:10.1002/9781118445112.stat06472

13. Zhang, W., Li, F. & Nie, L. Integrating multiple ‘omics’ analysis for microbial biology: application and methodologies. *Microbiol. Read. Engl.* **156,** 287–301 (2010).

14. Wilmes, A. *et al.* Application of integrated transcriptomic, proteomic and metabolomic profiling for the delineation of mechanisms of drug induced cell stress. *J. Proteomics* **79,** 180–194 (2013).

15. Nie, L., Wu, G., Culley, D. E., Scholten, J. C. M. & Zhang, W. Integrative Analysis of Transcriptomic and Proteomic Data: Challenges, Solutions and Applications. *Crit. Rev. Biotechnol.* **27,** 63–75 (2007).

16. Oliveira, A. P. & Sauer, U. The importance of post-translational modifications in regulating Saccharomyces cerevisiae metabolism. *FEMS Yeast Res.* **12,** 104–117 (2012).

17. de Nadal, E., Ammerer, G. & Posas, F. Controlling gene expression in response to stress. *Nat. Rev. Genet.* **12,** 833–845 (2011).

18. R Kolter, D A Siegele & Tormo, and A. The Stationary Phase of The Bacterial Life Cycle. *Annu. Rev. Microbiol.* **47,** 855–874 (1993).

19. Maier, R. Environmental Microbiology. in **Chapter 3. Bacterial Growth,** 40 (Academic Press, 2014).

20. Keren, L. *et al.* Noise in gene expression is coupled to growth rate. *Genome Res.* gr.191635.115 (2015). doi:10.1101/gr.191635.115

21. Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nat. Genet.* **38,** 636–643 (2006).

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

23. Klumpp, S., Zhang, Z. & Hwa, T. Growth Rate-Dependent Global Effects on Gene Expression in Bacteria. *Cell* **139,** 1366–1375 (2009).

24. Keren, L. *et al.* Promoters maintain their relative activity levels under different growth conditions. *Mol. Syst. Biol.* **9,** 701 (2013).

25. Ingraham, J. L., Maaløe, O. & Neidhardt, F. C. *Growth of the bacterial cell*. (Sinauer Associates, 1983).

26. Milo, R., Jorgensen, P., Moran, U., Weber, G. & Springer, M. how fast do rnas and proteins degrade? in *BioNumbers—the database of key numbers in molecular and cell biology* **38,** (2010).

27. Kendrew, J. *Encylopaedia of Molecular Biology*. (John Wiley & Sons, 2009).

28. Martínez-Gómez, K. *et al.* New insights into Escherichia coli metabolism: carbon scavenging, acetate metabolism and carbon recycling responses during growth on glycerol. *Microb. Cell Factories* **11,** 46 (2012).

29. Perrenoud, A. & Sauer, U. Impact of Global Transcriptional Regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on Glucose Catabolism in Escherichia coli. *J. Bacteriol.* **187,** 3171–3179 (2005).

30. Kumar, R. & Shimizu, K. Transcriptional regulation of main metabolic pathways of cyoA, cydB, fnr, and fur gene knockout Escherichia coli in C-limited and N-limited aerobic continuous cultures. *Microb. Cell Factories* **10,** 3 (2011).

31. Sokal, R. R. & Rohlf, F. J. The Comparison of Dendrograms by Objective Methods. *Taxon* **11,** 33–40 (1962).

32. Gan, G., Ma, C. & Wu, J. *Data Clustering: Theory, Algorithms, and Applications*. (SIAM, 2007).

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

34. 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).

35. 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).

36. Yoon, S. H. *et al.* Comparative multi-omics systems analysis of Escherichia coli strains B and K-12. *Genome Biol.* **13,** R37 (2012).

37. Batista, G. E. A. P. A., Prati, R. C. & Monard, M. C. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explor Newsl* **6,** 20–29 (2004).

38. Chawla, N. V. Data Mining for Imbalanced Datasets: An Overview. in *Data Mining and Knowledge Discovery Handbook* (eds. Maimon, O. & Rokach, L.) 853–867 (Springer US, 2005). doi:10.1007/0-387-25465-X\_40

39. He, H. & Garcia, E. A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **21,** 1263–1284 (2009).

40. Huang, Y.-M. & Du, S.-X. Weighted support vector machine for classification with uneven training class sizes. in *2005 International Conference on Machine Learning and Cybernetics* **7,** 4365–4369 Vol. 7 (2005).

41. Support Vector Machines. Available at: http://www.di.fc.ul.pt/~jpn/r/svm/svm.html. (Accessed: 24th April 2017)

42. Yang, Y. An Evaluation of Statistical Approaches to Text Categorization. *Inf. Retr.* **1,** 69–90 (1999).

43. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

44. Meyer, D. & Wien, T. U. *Support Vector Machines. The Interface to libsvm in package e1071. Online-Documentation of the package e1071 for "R*. (2001).

45. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2,** 18–22 (2002).

46. Ghamrawi, N. & McCallum, A. Collective Multi-label Classification. in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* 195–200 (ACM, 2005). doi:10.1145/1099554.1099591

47. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* **41,** D991-995 (2013).

48. Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32,** 223–226 (2014).

## Acknowledgments

This study was funded by Army Research Office (ARO) grant W911NF-12-1-0390 to COW. The Texas Advanced Computing Center (TACC) at The University of Texas at Austin provided high-performance computing resources.

## Contributions

M.U.C and C.O.W. conceived of the study.

M.U.C. designed the analysis pipeline, analyzed the data, and prepared the figures.

M.U.C. and C.O.W. wrote the manuscript.

## Competing interests

The authors declare no competing financial interests.